Journal of Software

ISSN 1796-217X

Volume 9, Number 11, November 2014

Contents

REGULAR PAPERS

An Approach for Identifying Detecting Objects of Null Dereference

Yukun Dong
College of Computer and Communication Engineering, China University of Petroleum, Qingdao, China
Email: dongyk@upc.edu.cn

Abstract—**On account of the complexity of programs, it is difficult to identify all detecting objects of null dereference, which is one of the preconditions of null dereference detection. This paper introduces an approach for identifying all detecting objects of null dereference of C programs. First, based on the relationship of dereference expressions with nodes of abstract syntax tree (AST), we identify referenced pointers; then based on the abstract storage described by region-based three value logic (RSTVL) and function summary, we identify detecting objects of null dereference. In order to validate the adequacy of our approach, five real-world projects are utilized for experimental analysis, and the results show that our approach could identify all detecting objects of null dereference.**

Index Terms—**null pointer dereference, defect detection, addressable expression, function summary**

I. Introduction

With increasing of software scale and complexity, software security becomes increasingly apparent. Particularly, null dereference has become one of the main causes of software security vulnerabilities, and it is one of the most common and difficult defects to eliminate.

At present, null pointer testing methods can be divided into dynamic methods [1, 2] and static methods [3-7]. Static methods check pointers dereference on the precondition of without running programs, which can be divided two categories: null dereference detection [3-5] and dereference validation [6, 7]. Generally, null dereference detection will first implement dataflow analysis or points-to analysis, then check if the pointer being referenced is null based on the analysis result; dereference validation is demand-driven, identify the pointer being referenced first, then analyses along the control flow backwards from the program point of a pointer dereference, checks if the pointer being referenced may be null.

Both static null pointer testing methods need to identify pointers being referenced, and identify detecting objects of null dereference based on associations between expressions. It is difficult to identify all detecting objects of null dereference, because pointer, struct and array exist in C programs, which cause alias, hierarchical, logic relationships exist among variables, and pointer parameter, especially complex type parameter.

If some detecting objects of null dereference unidentified, will lead to false positive of null dereference defects. The difficulties of identification lie in two aspects: First, some pointer expressions have complex grammatical structure; second, complex relationships among expressions, including alias, hierarchy, parameters with arguments, etc.

To solve these problems, we first establish mapping relationship between addressable expressions [8] with nodes of AST, and then apply RSTVL [9] to describe memory state of any memory object and all kinds of associations. Based on the analysis result, we identify detecting objects of null dereference by the following two steps. At the first step, we identify pointer expressions from AST based on the mapping between addressable expressions and nodes of AST, so we identify referenced pointers; at the second step, we identify detecting objects of null dereference for each pointer being referenced based on the result of data flow analysis and function summary [10].

This paper makes the following contributions:

- We introduce an approach for identifying pointer expressions from AST based on the relationship of addressable and nodes of AST.
- We show how to identify various detecting objects of null dereference based on RSTVL and function summary.

The remainder of this paper is organized as follows. Section II presents background on defect detection and motivation examples. Section III introduces addressable expression and RSTVL. Section IV and Section V introduce identifying detecting objects of null dereference. We present experimental results in Section VI, related work in Section VII and conclusion is in Section VIII.

II. Background And Motivation

During a program execution, the temporal safety property indicates a series of operations that must be executed in a specified manner.

Definition 1: *Defect pattern*. Syntax or semantics feature presented by defect that occurred frequently in programs.

Defect pattern describe a kinds of property of program, satisfy it will lead defects. For example, null dereference as a defect pattern appears to be a null pointer is referenced.

Definition 2: *Defect feature*. For a defect pattern, it can detect whether some properties violate syntax or

doi:10.4304/jsw.9.11.2771-2777

semantics rules of program, these properties related variables are defect features of defect pattern.

Defect feature can be understood whether defect detection related to some addressable expressions, and these addressable expressions are called detecting objects.

Definition 3: *Detecting object of null dereference*. A pointer with definite points-to attribute and related to null dereference defect detection.

We have implemented a defect detecting tool DTSGCC, which is a defect testing system for C written in Java. DTSGCC analyzes programs in five stages as shows in Figure 1. The last step of analysis stage is defect detecting, all detecting objects of null dereference can be identified in this step.

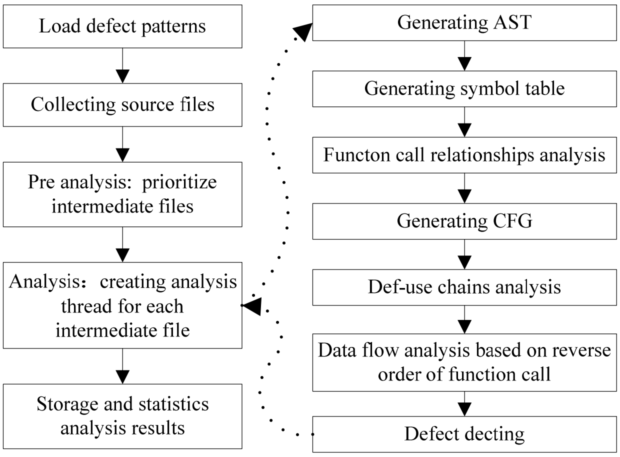

Load defect patterns
Collecting source files
Pre analysis: prioritize intermediate files
Analysis：creating analysis thread for each intermediate file
Storage and statistics analysis results
Generating AST
Generating symbol table
Functon call relationships analysis
Generating CFG
Def-use chains analysis
Data flow analysis based on reverse order of function call
Defect decting

Figure 1. Analysis stages of DTSGCC

We use two examples in Figure 2 to illustrate some obstacles of identifying defecting objects of dereference. For the example of Figure 2(a), pointer expression *pst[i]->m at line 7 actually implies three pointers being referenced: pst, pst[i], pst[i]->m, all of which need to be identified, or may lead to false negative null dereference defect. But pst[i] and pst[i]->m are not top-level variable, they can not be identified by analysing variable declaration, it's not easy to identify them.

For the example of Figure 2(b), *p* is referenced at line 4, there is an alias between *p* and *ps->a*, since *ps* is a formal parameter, we can't deduce the real point information of *ps->a* in function *f2* and wheather *ps->a* is null pointer. It can only be determined based on the calling context at call site. Because *ps->a* is not top-level parameter, the mapping between of parameters with arguments is unknown because the hierarchy and alias relationship between expressions. In fact, *f3* calls *f2* at line 9, and *s.a* maps *ps->a*, if we check whether *p* is referenced safely at line 4, we should check whether *s.a* is a safe pointer at line 9, and treat *s.a* as a detecting object of null dereference.

```
typedef struct{
   int *m;
}st;
void f1(st **pst){
   int i = 0;
   for(; i <9; i++){
     int j = *pst[i]->m;
   }
}
```

(a)

```
typedef struct{ int *a; }st;
int f2(st *ps){
   int *p = ps->a;
   *p = 2;
}
void f3(){
   st s;
        s.a = NULL;
   f2(&s);
}
```

(b)

Figure 2. Motivating examples

III. Region-Based Symbolic Three-Valued Logic

A. Ddressable Expression

Definition 4: *Memory Object*. The expression that corresponds to allocated memory when running programs, which can be top-level variable *v*, a member of a complex memory object, a dynamically allocated memory.

For all types of expressions defined by C99, we describe a C memory object expression *var* by the following grammar:

var::=v | *var.f* | *var*[*n*] | *malloc*(*exp*). Where *v* is top-level variable, *exp* is parameter.

Definition 5: *Addressable Expression*. The expression which has l-value and can be assigned.

For all types of expressions defined by C99, we describe a C addressable expression *aexp* by the following grammar:

aexp::= var | aexp.f | aexp->f | aexp[exp] | (aexp) | *aexp | id(exp)

*aexp can be defined as: *aexp::=*aexp' | *(++aexp') | *(--aexp') | *(aexp'++) | *(aexp' --) | *(aexp' op exp'), the type of aexp' is pointer, op= + | -, the type of exp' is integer.

For *id*(*exp*), where *id* is a method and return type is pointer, *exp* means parameters.

A memory object is an addressable expression, and there exist three relationships among addressable expressions as relationships between l-value and r-value.

- **Hierarchy**, relationship among l-values. It exists in addressable expression of compound type with its members.
- **Points-to relationship**, relationship of l-value and r-value. It exists in a pointer with the target that the pointer point to.
- **Linear and logical relationship**, relationship among r-values. The r-ralue of a memory unit has linear or logical relationship with the r-value of another memory unit.

Based on hierarchy and points-to relationships, we give the concept of parent addressable expression.

Definition 6: *Parent Addressable Expression*. Complex addressable expression is the parent of its members; Pointer is the parent of the addressable expressions that it points to.

For seven kinds of addressable expressions, *aexp i*s the parent of *aexp.f*, *aexp->f*, *aexp*[*exp*], **aexp*; *aexp->f* is equivalent to (**aexp*).*f*, whose parent is **aexp*.

When a function is called in different calling contexts, the points-to information of its pointer arguments maybe different. In order to map the points-to information at call site to the called function, we introduce extended variables to represent the points-to information of pointer parameters and global variables.

Let *e* as the variable that need to be extended, the extended rules are as follows:

- if *e* is a pointer, and the maximum level of dereference from *e* is *n*, then we create *n* extend variables, include **e*, ***e*,…and so on;
- if *e* is a variable with compound type and has *n* member, then we create *n* extend variables, include $e.f_1$, $e.f_2$,…and so on.

For example, parameter *ps* of function *f2* in Figure 2(b), the maximum level of dereference from *ps* is 1, so extended variable **ps* is introduced. **ps* is an extended variable and its type is struct and has a child *a*, so we generate an extended variable (**ps*).*a*. And (**ps*).*a* is pointer and the maximum level of dereference from it is 1, so we generate extended variable *(**ps*).*a*.

B. Rstvl

Definition 7: *Region-based Symbolic Three-Valued Logic*. RSTVL is a model of quadruple <*Var*, *Region*, S_{Exp}, *Domain*>, where *Var* is memory object, *Region* is abstract memory, S_{Exp} is symbolic expression, and *Domain* is the domain of value.

Quadruple RSTVL describes scalar memory object, and complex memory object can be decomposed into combination of scalar elements. Complex type memory object can be described by triple <*Var*, *Region*, *x*>, where *x* is determined by the type of *Var*, if the type of *Var* is array, *x* is {<*i*, *Region*>}, *i*∈N, *i* is the index of array *Var*; if the type of *Var* is struct, *x* is {< *f*, *Region*>}, *f* is the member of struct *Var*.

For different types of memory objects, different types of regions are applied. *PrimitiveRegion* describes primitive type memory object, *PointerRegion* describes pointer, *ArrayRegion* describes array, and *StructRegion* describes struct.

Each region has the only number, the numbering form of *PrimitiveRegion* is *bm_i* (*i*∈N), the numbering form of *PointerRegion* is *pm_i*, the numbering form of *ArrayRegion* is *am_i*, and the numbering form of *StructRegion* is *sm_i*. For the region dynamically allocated memory, its number is *mxm_i_n*(*x* means the type of the region, the value is '*b*', '*p*', '*a*' or '*s*'), *n* is bytes of memory size. The number of null address is "*null*", and the number of wild address is "*wild*". If the initial letter of the number of a region is '*u*' or '*g*', this region describes a parameter or global variable.

We call the region that maps *v*, *var.f*, *var*[*n*] is safe region, dynamically allocated region is dynamic region, the region that maps parameter or global variable is unknown region, these three kinds regions collectively call operable region; the region identified null or wild is an inoperable region. Dynamic region and unknown region will become safe region after not null judgement, dynamic region and unknown region will become inoperable region after is null judgement.

We divide domain [11] into two types: numeric and pointer, and apply *PointTos* to describe points-set in pointer domain *PointerDomain*, the elements of *PointTos* is the number of a region.

Domains of RSTVL and operators to them constitute complete lattice $< L, \le, \sqcup, \sqcap, \bot, \top >$. $\bot$ is empty set; $\top$ of numeric domain is $[-\infty, +\infty]$; $\top$ of pointer domain is the union of null, wild and all numbers of operable region; $\sqcup$ is merge operation of sets; $\sqcap$ is intersection operation of sets. Static data flow analysis based on RSTVL can be transferred to operation on lattice.

RSTVL describes all three associations among addressable expressions; and is suitable for flow-sensitive, field-sensitive, context-sensitive and path-insensitive static analysis. Given a program point, a region abstraction based on RSTVL consists of the following:

- At each program point *l*, a set of regions R^l that models the locations that may access at *l*, a set S^l expresses symbols that may be used at *l*.
- At each program point *l*, exists an abstract store: $\rho^l = (\rho_v^l, \rho_r^l, \rho_f^l)$, where $\rho_v^l : V \to R^l$ maps memory objects to their regions; $\rho_r^l : R^l \to R^l$ expresses the points-to relationship among regions; $\rho_f^l : (R^l \times F) \to R^l$ maps members of a complex addressable expression to their regions.

To analyse an addressable expression, we need to get potentially associated regions first, and an addressable expression may associates several regions. At a program point *l*, if the abstract store is ρ , we use $R^l[\![e]\!]$ to express region set that addressable expression *e* associated. Then strategies can be given for achieving region set that all kinds of addressable expressions associated.

- $R^l[\![v]\!] = \rho_v^l(v)$;
- $R^l[\![e.f]\!] = \bigcup\limits_{r \in R^l[e]} \rho_f^l(r, f)$;
- $R^l[\![e[i]]\!] = \bigcup\limits_{r \in R^l[e]} \rho_f^l(r, i)$;
- $R^l[\![*e]\!] = \bigcup\limits_{r \in R^l[e]} \rho_r^l(r)$;
- $R^l[\![*e]\!] = \bigcup\limits_{r \in R^l[e]} \rho_r^l(r)$;
- $R^l[\![(e)]\!] = R^l[\![e]\!]$;
- $R^l[\![e->f]\!] = \bigcup\limits_{r \in R^l[e]} \left\{ \bigcup\limits_{r' \in \rho_r^l(r)} \rho_f^l(r', f) \right\}$.

We have applied RSTVL to data flow analysis in DTSGCC [9]; the analysis is flow-sensitive, field-sensitive, and context-sensitive based on symbolic function summary, it can analyzes the over-approximation of every memory objects in every program point.

IV. Identifying Detecting Objects of Intraprocedural Null Dereference

A. Identifying Referenced Pointers

Based on the grammar defined by BNF, we generate AST for the C file under test. At the generating scope table stage, we identify addressable expressions from AST, and bind each addressable expression to the related node of AST [8].

There are three kinds of nodes that are closely related to addressable expressions, which are UnaryExpression, PostfixExpression and PrimaryExpression; and their grammars are described by BNF as follows:

UnaryExpression::= PostfixExpression | "++" UnaryExpression | "--" UnaryExpression | <SIZEOF> (UnaryExpression | "("TypeName") ") | UnaryOperator CastExpression, UnaryOperator::= "&" | "*" | "+" | "-" | "~" | "! ");

PostfixExpression ::= PrimaryExpression (". " <IDENTIFIER> | "["Expression"] " | "("(ArgumentExpressionList)? ") " | "->" <IDENTIFIER> | "++" | "--")*;

PrimaryExpression ::= <IDENTIFIER> |"("Expression")"| Constant.

According to grammatical features, pointer expression e_p as a kind of addressable expression can be divided in to three types: $*e_p$, e_p->f and $e_p[exp]$, so we can identify all dereference expressions from searing AST, and identified all referenced pointers. We apply XPath to search AST, and the query statement of $*e_p$ is:

.//AssignmentExpression//UnaryExpression[/UnaryOperator[@Operators='*']]/UnaryExpression.

The query statement of e_p->f and $e_p[exp]$ is:

.//AssignmentExpression//UnaryExpression/PostfixExpression[./PrimaryExpression][contains(@Operators,'[')or contains(@Operators, '->')].

For the example of Figure2(a) , *pst[i]->m is a $*e_p$ type pointer expression, so we can identify it from the related UnaryExpression node of AST, and deduce the pointer being referenced is *pst*[*i*]->*m*; *pst*[*i*]->*m* is a e_p->f type pointer expression, so we can identify it from the related PostfixExpression node of AST, and deduce the pointer being referenced is *pst*[*i*]; *pst*[*i*] is a $e_p[exp]$ type pointer expression, so we can identify it from the related PostfixExpression node of AST, and deduce the pointer being referenced is *pst*. Above all, we identify three referenced pointers from *pst[i]->m: *pst*[*i*]->*m*, *pst*[*i*], *pst*.

B. Points-to Attribute

We can decide whether a pointer being referenced is null dereference or not based on its points-to attribute. Points-to attribute is described as a lattice: AL_{PTR} = (V_{PTR}, F_{join}, F_{meet}), and its Hesse table is shown in Figure 3. V_{PTR} depicts the value set of points-to attribute, which can describe security of a pointer being referenced effectively, and can be conveniently applied to null dereference detection. EMPTY expresses initial value of attribute lattice, NULL expresses a pointer points to null address, NOTNULL expressed a pointer points to a safe memory address, NON (NULL_OR_NOTNULL) expresses a pointer may be points to null address. When a pointer is referenced, null dereference will inevitably occur if points-to attribute of the pointer is NULL, may occur if points-to attribute of the pointer is NON.

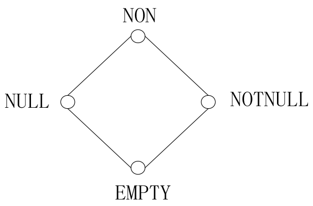


Figure 3. Hasse table of AL_{PTR}

F_{join}: $V_{PTR}\times V_{PTR}\rightarrow V_{PTR}$ is the greatest lower bound function of AL_{PTR}.

F_{meet}: $V_{PTR}\times V_{PTR}\rightarrow V_{PTR}$ is the least upper bound function of AL_{PTR}.

In order to comprehensive express points-to information, UNKNOWN is introduced to express uncertainty of points-to of a pointer; it is applied to initialize points-to attribute of pointer parameters and global variables. Operations about UNKNOWN with other attribute value X as follows:

$$F_{join}\text{(X, UNKNOWN) = X}$$

$$F_{meet}\text{(NOTNULL, UNKNOWN) = UNKNOWN}$$

$$F_{meet}\text{(NULL, UNKNOWN) = NON}$$

$$F_{meet}\text{(NON, UNKNOWN) = NON}$$

$$F_{meet}\text{(EMPTY, UNKNOWN) = UNKNOWN}$$

$$F_{meet}\text{(UNKNOWN, UNKNOWN) = UNKNOWN}$$

At program point l, let $T^{l}[\![r_{name}]\!]$ express the type of the region numbered r_{name}, pd express the domain of pointer e_p, the abstraction function α_{ρ}^{l} of points-to attribute is defined as follows:

$$\alpha_{\rho}^{l}(pd)=\begin{cases}\text{EMPTY} & pd=\varnothing \\ \text{NULL} & pd=\{\text{null}\} \\ \text{NOTNULL} & \forall pt\in pd, T^{l}[\![pt]\!]\text{ is safe} \\ \text{UNKNOWN} & (\forall pt\in pd, T^{l}[\![pt]\!]\text{ is safe or} \\ & \quad\text{unknown) and }(\exists pt\in pd, \\ & T^{l}[\![pt]\!]\text{ is unknown)} \\ \text{NON} & \textit{others}\end{cases}$$

V. Identifying Detecting Objects Of Interprocedural Null Dereference

A. Function Summary

Each function call might affect its concrete call site context in four aspects:

- the callee function might cause side effects to actual-parameters and global variables;
- the caller's dataflow and control flow might be transformed by callee's return value;

- potential interrupt instructions, such as exit, assert, exception, etc;
- pre-condition, the call site context must obey the callee's invocation constraints to avoid defects.

In this paper, our function summary only focuses on pre-condition of null dereference NPDPreSummary. For the example in Figure 2(b), pointer parameter *ps* is referenced at line 3, and the points-to attribute of *ps* is UNKNOWN, so we must add *ps* can not be null as a pre-condition of function *f4*.

B. Generating Pre-condition of Null Dereference

If $\alpha_\rho(V^l[\![e_p]\!]) ==$ UNKNOWN and the *PointerDomain* of e_p is *pd*; then for each region number r_{Name} in *pd* and the region named r_{Name} is an unknown region, if the momory object mapping to the region is *exp*, then we set the parent addressable expression of *exp* can not be null as pre-condition. Let $R_n^l[\![r_{name}]\!]$ express the region numbered r_{name} at program point *l*; Let $E_r[\![r]\!]$ express the memory object that related to *r*. The generating pre-condition of null dereference is detailed in algorithm 1.

Algorithm 1 Generating pre-condition of null dereference

Input: pointer being referenced e_p, NPDPreSummary

Output: NPDPreSummary

Declare: *getParent*(*para*): get father addressable expression of *para*.

```
for each $pt \in V^l[\![e_p]\!] \&\& T^l[\![pt]\!]$ is unknown region
    let $var = E_r[\![R_n^l[\![pt]\!]]\!]$;
    let $fvar = getParent(var)$;
    add $fvar$ to NPDPreSummary;
end for
return NPDPreSummary;
```

C. Instantiating Pre-condition of Null Dereference

For each function call, we get its function summary first, and instantiate the function summary based on the calling context at the call site.

Function call expression is a kinds of addressable expression, the grammar of it is *id*(*exp*). *id*(*exp*) maps to PrimaryExpression, it can be identified by searching AST, the query statement is:

.//PrimaryExpression[@Method='true']

If the called function has function summary, and the pre-condition constraint some pointers can not be null, then we instantiate it.

To instantiate the pre-condition of null dereference, the key is for each constrained pointer e_{cp} in pre-condition, get related addressable expression set e_pList at the call site; and based on the abstract store state at the call site described by RSTVL, get the points-to attribute for each pointer of e_pList. If the points-to attribute of a pointer in e_pList is UNKNOWN, then we add this pointer into pre-condition of null dereference applying algorithm 1, otherwise , the pointer is a detecting object of null dereference.

In all of above steps, the key is getting the addressable expression set for each constrained pointer in pre-condition of null dereference, which is a problem that maps a parameter to arguments; the details is shown in algorithm 2.

Algorithm 2 Mapping a Parameter to Arguments.

Input: *para*, R^n

Output: *VarsList*<Variable>

Declare:

getParents(*para*): get parent addressable expressions of *para* sorted according to parent-child relationship.

getArgument(*var*, *n*): get the corresponding arguments of top-level parameter *var* at the calling point *n*.

getParent(*var*): get parent addressable expression of *var*.

getType(*e*): get the addressable expression type of *e*,where 0: *v*, 1: *e.f*, 2: *e->f*, 3: *e*[*exp*], 4: (*e*), 5: **e*, 6: *m*(*exp*).

getMemName(*s*, *var*): for addressable expression *s* and the type of *s* is struct, get the member name of its child addressable expression *var*.

```
let $args$<Variable> = $\varnothing$;
let $parents$<Variable> = $getParents(para)$;
get first variable $v_0$ in $parents$;
$args$ = { $getArgument(v_0, n)$ };
for each $p \in parents \&\& p \neq v_0$
    let $v_p = getParent(p)$;
    let $vars$<Variable> = $\varnothing$;
    for each $v \in args$
        for each $r \in R^{\bullet n}[\![v]\!]$
            if $getExpType(var) == 1$ then
                let $m$ = getMemName($v_p$, $p$);
                $vars \cup= \{ R_f^l[\![r, m]\!] \}$;
            else if $getExpType(var) == 5$ then
                $vars \cup= \{ V_r^{\bullet n}[\![r]\!] \}$;
            end if
        end for
    end for
    $args = vars$;
end for
return $args$;
```

For the function *f2* in Figure2 (b), parameter *ps* is referenced at line 3; dereference *p* at line 4 is actually access the region that pointed by (**ps*).*a*. the points-to attribute of *ps* and (**ps*).*a* are UNKNOWN, so the pre-condition of null dereference of *f2* is: {*ps*[NOTNULL], (**ps*).*a* [NOTNULL]}.

Function *f3* calls *f2* at line 9, *ps* is a top-level parameter and constrained can not be null in pre-condition of null dereference, based on the numerical order, we can deduce *ps* maps to *&s* at call site, it's a safe dereference. (**ps*).*a* is also constrained can not be null in pre-condition, its parent addressable expression set is {*ps*, **ps*, (**ps*).*a*}. Based on the abstract store state at the call site described by RSTVL at line 9, we can deduce that *ps* maps {*&s*}, **ps* maps {*s*}, (*ps).a maps {s.a}. So detecting objects of null dereference is &s, s.a, and the points-to attribute of s.a is NULL, so a null dereference defect will be reported at line 9.

VI. Experimental Analysis

We choose five C projects to validate the effectiveness of our approach.

A. Identifying Defecting Objects of Null Dereference

Pointers being referenced of five C projects identified by our approach are shown in TABLE I. Pointers being referenced can be divided as: local pointer and points-to attribute is known (LKP), local pointer and points-to attribute is unknown (LUP), external pointer and points-to attribute is known (EKP), external pointer and points-to attribute is unknown (EUP), function pointer (FP).

Table I
Statistics Of Referenced Pointers

Benchmark	KLOC	Referenced pointer					
		LKP	*LUP*	*EKP*	*EUP*	*FP*	*Total*
antiword-0.37	24.2	770	142	50	1716	73	2751
uucp-1.07	52.6	1849	1112	401	2397	379	6138
sphinxbase-0.3	22.5	691	203	602	2011	82	3589
optipng-0.6	27	415	239	178	1258	53	2143
barcode-0.98	3.4	176	52	249	378	18	873
Total	130	3901	1748	1480	7760	605	15494

It is shown in TABLE I that pointer dereference occur frequently in C functions, about 120/KLOC, and more than 60% pointers being referenced can not be determined their points-to attribute in respective function, if function pointers are considered, more than 65% pointers being referenced need interprocedural identified.

For lib functions, we construct their function summary artificially. And detecting objects of null dereference identified by function summary can be divided two kinds: identified based on custom function (CFP), identified based on lib function precondition pointer (LFP). For pointers being referenced that can not be determined points-to attribute in TABLE I, we identify their relative detecting objects of null dereference by interprocedural identifying approach, the result is shown in TABLE II.

Table II
Statistics Of Detecting Objects Of Null Dereference

Benchmark	Detecting objects of null dereference					
	LKP	*EKP*	*FP*	*CFP*	*LFP*	*Total*
antiword-0.37	770	50	73	372	105	1370
uucp-1.07	1849	401	379	1382	1314	5325
sphinxbase-0.3	691	602	82	170	222	1767
optipng-0.6	415	178	53	579	358	1583
barcode-0.98	176	249	18	56	259	758
Total	3901	1480	605	2559	2258	10803

For the example in Figure 1(b), *ps* is a EUP type referenced pointer, *p* is a LUP type referenced pointer, *s.a* is a CFP type detecting object of null dereference.

Applying our approach, more detecting objects of null dereference can be identified, and more null dereference defects can be detected. There is a null dereference in Figure 4, *Barcode_128_make_array* calls lib function *strlen* at line 321, the pre-condition of null dereference in function summary of *strlen* constrain its parameter can be null, and the points-to attribute of *bc->ascii* is UNKNOWN at line 321, so we add (**bc*).*ascii*(equals to *bc->ascii*) can not be null pointer into pre-condition of null dereference of *Barcode_128_make_array*. *Barcode_128_encode* calls *Barcode_128_make_array* at line 439, we can deduce that parameter (**bc*).*ascii* maps argument *bc->ascii*, and the points-to attribute of *bc->ascii* at line 439 is NON, so we make as *bc->ascii* a detecting object of null dereference, and report *bc->ascii* is a null dereference defect at line 439. Klocwork9 [12] and Saturn [13] can not detect these defects.

```
File:barcode\code128.c
The called functions  at line 314:
static int *Barcode_128_make_array(struct Barcode_Item *bc, *)
321: len = 2 * strlen(bc->ascii) + 5;

The calle function at line 414:
int Barcode_128_encode(struct Barcode_Item *bc)
433: text = bc->ascii;
434: if (!text) {
……;
}
439: codes = Barcode_128_make_array(bc, &len);  //NPD
```

Figure 4. A detecting object of null dereference identified by our approach

VII. Related Work

There is some research in the area of expression recognition that related to our work. Maksim O et al. [14] present core expression as canonical representation, they also identify expression from AST, but can not guarantee to identify all addressable expressions; so their method can not guarantee identify all referenced pointers. PenAnalysis [15] applies expression tree to representation expression, which is more complex than our method. S. Blazy et al.

In order to analysis expressions comprehensively, relations among expressions must be considered, otherwise the result will be inaccurate. Alias set and points-to set only focus on alias relationship, can not express hierarchy of compound variables; applying them can't analyze complex pointers effectively. As a region model, RSTVL is similar to Brian Hackett's memory model [16], and appropriated for shape analysis.

Specific to null dereference testing, it is an important work to identify detecting objects of null dereference. Although null dereference testing has been extensively studied, only few of past researches mention how to fully identify detecting objects. PSE [3] defines a simple pointer language, and regulated source code patterns for the null dereference property; but it can't guarantee identifying multilevel pointers effectively, especially interprocedural multilevel pointers. B. Cheng, etc apply access path for interprocedural pointer analysis [17], their access path is similar to our parent addressable expression, and they also use function summary.

VIII. Conclusions

In this paper, we introduce an approach for identifying detecting objects of null dereference. RSTVL describes abstract storage of each program point and relationships of addressable expressions, uses region number set to express pointer point to. Based on the correspondence between addressable expressions and nodes of AST, we identify all pointers being referenced from AST based on the grammar of pointer expression. If the points-to attribute of a pointer being referenced is can be determined, then we add the related pointer can not be null into the pre-condition of null dereference in function summary, and identified related detecting objects of null dereference at call site based on the abstract storage described by RSTVL.

References

[1] Michael D, Graham Z, Samuel Z, "Breadcrumbs: efficient context sensitivity for dynamic bug detection analyses, " In *Proceedings of the 2010 ACM SIGPLAN conference on Programming language design and implementation*, 2010, pp. 13-24.

[2] J. Huang, M. Bond. "Efficient context sensitivity for dynamic analyses via calling context uptrees and customized memory management," In *Proceedings of the 2013 ACM SIGPLAN international conference on Object oriented programming systems languages & applications*. 2013, pp. 53-72.

[3] R. Manevich, M. Sridharan, S. Adams, "PSE: Explainint program failure via psotmortem static analysis, " In *Proceedings of the 12th ACM SIGSOFT twelfth International Symposium on Foundations of Software Engineering*, 2004, pp. 63-72.

[4] X. Ma, J. Wang, D. Wang. "Computing must and may alias to detect null pointer dereference, " *Leveraging Applications of Formal Methods, Verification and Validation*, 17(17): 252-261, 2008.

[5] M. Buss. Summary-based pointer analysis framework for modular bug finding [D]. Columbia: Columbia University, 2008

[6] Y. Xie, A. Aiken. "Saturn: A scalable framework for error detection using Boolean satisfiability," *ACM Transactions on Programming Languages and Systems*, 29(3): 1-43, 2007.

[7] M. Ravichandhran, K. Raghavan. Null dereference verification via over-approximated weakest pre-conditions analysis. In *Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications*, 1033-1052. ACM, 2011.

[8] Y. Dong, Y. Xing, D. Jin, Y. Gong. "An approach to fully recognizing addressable expression," In *The 13th International Conference on Quality Software*, 2013, pp. 149-152.

[9] Y. Dong, D. Jin, Y. Gong, Y. Xing. "Static analysis of C programs via region-based memory model," *Journal of Software*, 25(2): 357-372, 2014 (in Chinese with English abstract).

[10] Y. Dong, D. Jin, Y. Gong. "Symbolic procedure summary using region-based symbolic three-valued logic," *Journal of Computers*, 9(3): 774-780, 2014.

[11] Y. Wang, Y. Gong, Q. Xiao, Z. Yang. "A Method of Variable Range Analysis Based on Abstract Interpretation and Its Applications," *Acta Electronica Sinica*, 39(2): 296-303, 2011 (in Chinese with English abstract).

[12] M. Webster. "Leveraging static analysis for a multidimensional view of software quality and security: Klocwork's solution," White paper, IDC. 2005.

[13] I. Dillig, T. Dillig, A. Aiken. "Sound, complete and scalable path-sensitive analysis," *ACM SIGPLAN Notices*, 43(6): 270-280, 2008.

[14] M. Orlovich and R. Rugina. "Core expressions: An intermediate representation for expressions in C," *In Submitted to Compiler Construction'06*. Available at http://www.cs.cornell.edu/~rugina.

[15] M. Strout, J. Mellor-Crummey, P. Hovland. "Representation-independent program analysis," In *Proceedings of the The sixth ACM SIGPLAN-SIGSOFT Workshop on Program Analysis for Software Tools and Engineering*, 2005, pp. 64-74.

[16] B. Hackett, R. Rugina. "Region-Based Shape Analysis with Tracked Locations," In *Proceedings of the 32nd ACM SIGPLAN- SIGACT symposium on Principles of programming languages*, 2005, pp. 310-323.

[17] B. Cheng, W. Hwu. "Modular interprocedural pointer analysis using access paths: design, implementation, and evaluation," *Acm Sigplan Notices*, 35(5): 57-69, 2000.

Yukun Dong, received his PhD in computer science from School of Beijing University of Posts and Telecommunications, Beijing, China, in 2014. He currently serves as a lecturer in College of Computer and Communication Engineering, China University of Petroleum, Qingdao, China. His research interests include software testing and program static analysis.

A Defense Model of Reactive Worms Based on Dynamic Time

Haokun Tang
College of Economics and Management, SouthWest University, Chongqing, P.R.China
HanHua Financial Holding, P.R.China
Email: tanghaokun@hanhua.com

[1]Shitong Zhu, [2]Jun Huang and [3]Hong Liu
[1,2]School of communication and information engineering, Chongqing University of Posts and Telecommunications, Chongqing, P.R.China
[3]School of Software, Chongqing University of Posts and Telecommunications, Chongqing, P.R.China
Email: [1]zstchina1993@gmail.com; [2]xiaoniuadmin@gmail.com; [3]liuhong1@cqupt.edu.cn

***Abstract*—The popularity of reactive worms, whose attacking behavior inherits characteristics from both active worms and passive worms, has brought great threat to P2P networks in recent years. Most existing defense models only focus on the effects of P2P churn on reactive worm's propagation, but neglect the impact of user behaviors on the spread of worms. This paper proposes a defense model of reactive worms based on dynamic time with full consideration of various dynamic factors that restrict the propagation of reactive worms in real networks; then compares major distinctions of several key parameters in worms' propagation between models based on mean-field theory and the presented dynamic-time-based one; and deduces the crucial periods of time within a particular 24-hour day for defending against reactive worms' attack. Eventually simulation experiment shows this defense model is feasible and effective.**

***Index Terms*—dynamic time, probability theory, reactive worm, defense model, mean-field theory**

I. Introduction

Based on distributed system and computer network, Peer-to-Peer (P2P) is currently the most popular networking technology for data sharing, instant messaging, enterprise collaboration, etc. However, current P2P networks are facing serious security threats since they show some facilities towards worm attack and propagation. With the emergence of P2P worms, they bring harm to P2P networks and even pose an underlying threat to Internet.

P2P worms can be generally categorized into three groups: passive worms, reactive worms and active worms. Unlike passive worms that hide themselves in malicious files and trick users into downloading and executing them for propagation, active worms that automatically connect to potential targets by using topological information for propagation, reactive worms propagate themselves with legitimate network activities using security vulnerabilities on particular P2P nodes. This type of normal network connections can be initiated by a user or one reactive worm personating a legitimate user. If an infected node finds exploitable security vulnerabilities in a connection, it will pass a worm's copy to the uninfected node. Then once the worm copy is executed by the uninfected node, it will be injected and the newly-infected node will continue infecting its neighborhood nodes. Since reactive worms can propagate through normal connections, they are relatively tough against detection or common firewalls. Characteristics above make the propagation of reactive worms in P2P networks speedy and conceal.

The transmission mode of reactive worms can be generally divided into three categories due to different infection directions: source infection, target infection and mixed infection. Among these three, mixed infection reactive worms infect not only the source host (download port) but also the target host (upload port) in one connection, making it the most harmful one to P2P networks. This paper mainly lays emphasis on this kind of reactive worms.

Most research into reactive worm defense at present is mostly based on the epidemic model. Early in 1926, McKendrick et al. first modeled the spread of biological viruses by the means of mathematics [1], then proposed the epidemic mathematical model. From then on, the epidemic model was widely introduced into the process of modeling on computer viruses and reached some truly remarkable achievements. Some representative ones of them are listed as follows: Kephart et al. first introduced the epidemic model into computer virus modeling process in the early 90's [2]. Yu et al. developed a P2P worm propagation model on the basis of simple epidemic model, and discussed the prevention strategy of P2P worms [3]; Sun et al. studied the propagation process of active P2P worms by a dynamic model [4], took both the entire

Supported by China Postdoctoral Science Foundation funded project (Grant No. 2014M552308), NCET, NSFC (Grant No. 61272400,Grant No. 61309032), Chongqing Innovative Team Fund for College Development Project (Grant No. KJTD201310), Foundation of CEC (Grant No. KJ130525, KJ130523), and CQUPT Research Fund for Young Scholars (Grant No. A2012-79).

doi:10.4304/jsw.9.11.2778-2788

network's status and each single node's action into consideration; Li et al. presented a stochastic model of worm propagation on the basis of the epidemic model, analyzed the process of state transition of nodes by using the state space of a Markov chain [5]; Yang et al. proposed a new method to integrate the delivery predictability of ProPHET-Routing and verified their proposed OOPProPHET-Routing method was better than Epidemic-Routing method by NS2 network simulator [6]; Xie et al. built a new multi-agents risk assessment model based on attack graph(MRAMBAG) and had shown that the MRAMBAG was a more feasible and effective way for evaluate the network security risk [7]; Xing et al. analyzed the security threat to the virtual network and brought forward a security guarantee embedding algorithm for virtual network [8], the simulation results showed that the algorithm was effective; Wang et al. gave a simulation analysis to reactive worms [9], and simulated the defense process of reactive worms under Internet environment and P2P environment. Also, the keys to defense in these environment had been proposed. Qin et al. and Feng et al. respectively modeled the propagation process and immune process of reactive worms on the basis of Kermack-Mckendrick model [10] [11], and also the prevalence condition of reactive worms was presented in these papers, pointing out the direction for defending against reactive worms in P2P environment. Yang et al. designed a dynamic quarantine protocol to defend active worms in P2P networks by quarantining the suspicious hosts, and he developed a mathematical model of PWPQ to prove the effectiveness of this defense method [12]. Ouyang et al. analyzed the trust mechanism and application model, set up a new kind of trust model based on P2P network [13], the simulation experiment showed this trust model helps improve the success rate of transaction.

The referred works above described the propagation process of worms to some extent and provided some valuable references for establishing the corresponding defense system of worm in P2P networks. But we notice quite few works focus on the defense model of reactive worms especially in modeling reactive worms' propagation with considerations of the dynamic environment such as user behavior, network size, and network bandwidth, etc. In some degree they are basically untouched. This paper attempts to address this issue, and mainly makes the following four contributions.

(1) We present a propagation strategy of reactive worms in dynamic environment, and provide the dynamic process of state transition of nodes when reactive worms spread in accordance with the strategy.

(2) On the basis of analyzing pros and cons of existing defense models of reactive worms, we develop an improved defense model based on mean-field theory and deduce a number of key parameters affecting propagation speed of reactive worms in dynamic environment.

(3) We analyze the shortages of the foregoing model, put forward some improvement methods. That is, we analyze factors including network size and user behavior at different time periods and simulate network size using probability theory, finally propose a defense model of reactive worms based on dynamic time.

(4) We conduct mathematical analysis to study the improved defense model, compare the difference of key parameters that affect reactive worm defense between the defense model that is based on mean-field theory and the one based on dynamic time, and deduce the most crucial period within a day for defending against reactive worm attack.

The rest of this paper is organized as follows. Section 2 presents a propagation strategy of reactive worms in dynamic environment, and elaborates the dynamic process of state transition of nodes when reactive worms spread in accordance with the strategy, section 3 develops an improved defense model of reactive worms in P2P networks based on mean-field theory, section 4 analyzes shortages of the foregoing model and puts forward some improvement methods, section 5 compares the difference of key parameters that affects reactive worm defense between the defense model based on mean-field theory and the one based on dynamic time, deduces the most crucial period of a particular day for defending against reactive worm attack, section 6 proposes the conclusion and some future work directions, the acknowledgment is put forward in section 7.

II. A Propagation Strategy for Reactive Worms in Dynamic Environment

A. State Transition of Nodes When Reactive Worms Spread

A P2P nodes has least six states in different stages of reactive worms' propagation, The summary of these states are listed as follows:

(1) Susceptible infected state (S state): This is the state when an online node is vulnerable to worm attack for its secure vulnerability. Yet it hasn't downloaded the worm file.

(2) Latent state (L state): This is the state when a node in S state has downloaded a worm file from another online worm node but the worm file hasn't been executed. At this stage, the node cannot be invaded by the same type of reactive worm infection. It is not contagious either.

(3) Infected state (I state): Once a worm file is executed by a node in L state, the state of this node changes from latent state to infected state. At this stage, the node is contagious and has already become a worm node.

(4) Quarantined state (Q state): Once a node in I state is detected by monitoring software for transmitting reactive worms, it will be quarantined and its state will be converted into quarantined state. At this stage, the node is no longer contagious.

(5) Immune state (R state): This is the state when an online node has been patched by security software. At this stage, the node cannot be infected by reactive worms anymore, nor is contagious.

(6) Offline state (O state): This is the state when the node has left P2P networks.

State transition of nodes is shown in "Fig. 1," the description is as follows:

When a benign node containing security vulnerabilities joins P2P networks, it is in susceptible infected state(S); if it has been patched, it is in immune state(R); when an infectious malignant node just joins P2P networks, it is in infected state(I); when a node in I state connects to a node in S state, the infected node would inject a worm file into the uninfected one with a probability φ, when the node in S state has downloaded the worm without execution, the state of this node would be converted into latent state(L); and a node in L state would execute worm files with a probability η, then its state will be converted into infected state(I); when a node in I state is detected by monitoring software with the probability χ for transmitting reactive worms, it will be quarantined and its state would be converted into quarantined state (Q); if an online node in S state, L state, I state, or Q state is found with security vulnerabilities by security software in periodic inspection, it would be patched and its state would be converted to immune state(R) at a probability r_1 , r_2 , r_3 and r_4 respectively; all online nodes would choose to leave P2P networks with a probability α , and if so, their states would then be converted into offline state (O); meanwhile, all offline nodes would choose to join P2P networks with a probability β , and their states would then return to original states before being offline. And users of some offline nodes will reinstall their operation systems with a probability δ , thus their states would be converted into susceptible infected state (S) when being back online.

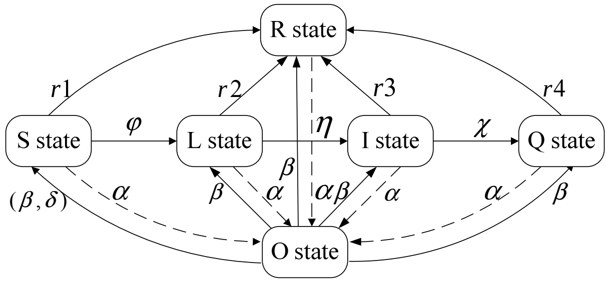


Figure 1. State transitions of nodes

III. A Defense Model Of Reactive Worms Based On Mean-Field Theory

A. Assumptions and Parameters in the Defense Model Based on Mean-field Theory

In order to simplify the modeling process of reactive worms based on mean-field theory, the following assumptions could be made:

(1) The number of nodes in P2P networks keeps constant.

(2) Each node has the same online rate and offline rate in a unit of time whichever its state is. Each offline node will be back to its original state if its operation system hasn't been reinstalled.

(3) A node in L state can finish downloading all worm fragmentations from other infected nodes in a unit of time.

(4) Only infected node can spread worm fragmentations, which also possibly make it quarantined.

(5) Although all the worm fragmentations in those nodes in Q state have been cleaned up, there are still some vulnerabilities in them and some of them may go back to S state before they are patched.

(6) Reactive worms will launch the attack based on the mixed infection strategy.

Table I lists all variables in the model.

TABLE I.
VARIABLES IN THE DEFENSE MODEL OF REACTIVE WORMS BASED ON MEAN-FIELD THEORY

Variable	Description	Initial value
N	The total number of nodes in P2P network	$N = 100000$
λ_d	Downloading rate of a node (The number of files that any P2P node can download in a unit of time)	$\lambda_d = 20$
λ_e	Execution rate of a node (The number of files that any P2P node can execute in a unit of time)	$\lambda_e = 8$
φ_d	Downloading infection rate (The probability of a node in S state that gets infected by downloading a file)	$\varphi_d = 0.3$
φ_u	Uploading infection rate (The probability of a node in S state that gets infected by uploading a file)	$\varphi_u = 0.2$
α	Offline rate of an online node	$\alpha = 0.01$
β	Online rate of an offline node	$\beta = 0.9$
δ	The probability for an offline node that will be back online after reinstalling OS	$\delta = 0.05$
η	Execution infection rate (The probability of a node in L state that is infected by executing a download file and will be converted into infected state)	$\eta = 0.05$
χ	Detection rate (The probability of a node in I state is that is detected by monitoring software for transmitting reactive worms. Then it will be converted into quarantined state)	$\chi = 0.03$
λ	The probability of a node in Q state that will go back to S state after clearing up worm fragmentation	$\lambda = 0.3$
r_1	The probability of a node in S state found to contain security vulnerabilities by security software. Then it will be patched and its state will be converted into immune state	$r_1 = 0.01$
r_2	The probability of a node in L state found to contain security vulnerabilities by security software. Then it will be patched and its state will be converted into immune state	$r_2 = 0.05$
r_3	The probability of a node in I state found to contain security vulnerabilities by security software. Then it will be patched and its state will be converted into immune state	$r_3 = 0.08$
r_4	The probability for a node in Q state is found to contain security vulnerabilities by security software, it will be patched and its state will be converted into immune state	$r_4 = 0.1$
$S_N(t)$	The number of online nodes in susceptible infected state at the time where $S_N(0)$ indicates the total number of nodes in susceptible infected state in P2P networks initially	$S_N(t) = 9900$
$S_O(t)$	The number of offline nodes in susceptible infected state at time t	$S_O(t) = 0$

$L_N(t)$	The number of online nodes in latent state at time t	$L_N(0)=0$
$L_O(t)$	The number of offline nodes in latent state at time t	$L_O(0)=0$
$I_N(t)$	The number of online nodes in infected state at time t, where $\mathrm{I}_N(0)$ indicates the total number of nodes in infected state in P2P networks initially	$I_N(0)=1000$
$I_O(t)$	The number of offline nodes in infected state at time t	$I_O(t)=0$
$Q_N(t)$	The number of online nodes in quarantined state at time t	$Q_N(0)=0$
$Q_O(t)$	The number of offline nodes in quarantined state at time t	$Q_O(0)=0$
$R_N(t)$	The number of online nodes in immune state at time t	$R_N(0)=0$
$R_O(t)$	The number of offline nodes in immune state at time t	$R_O(0)=0$
$A(t)$	The number of additional online nodes whose states have converted from susceptible infected state to latent state at time t	$A(0)=0$
$O(t)$	The number of nodes in offline state at time t	$O(0)=0$

B. A Defense Model of Reactive Worms Based on Mean-Field Theory

The defense model of reactive worms based on mean-field theory should meet the following theorems:

Theorem 1:

$$S_o(t)=\alpha\sum_{i=0}^{t-1}S_N(i)(1-\beta)^{t-i}$$

$$E_o(t)=\alpha\sum_{i=0}^{t-1}E_N(i)(1-\beta)^{t-i}$$

$$I_o(t)=\alpha\sum_{i=0}^{t-1}I_N(i)(1-\beta)^{t-i}$$

$$Q_o(t)=\alpha\sum_{i=0}^{t-1}Q_N(i)(1-\beta)^{t-i}$$

$$R_o(t)=\alpha\sum_{i=0}^{t-1}R_N(i)(1-\beta)^{t-i}$$

Theorem 2:

$$\begin{aligned}A(t)={}&S_N(t)\{\varphi_d\lambda_d I_N(t)/(N-O(t))\\&+\varphi_u[1-(1-1/(N-O(t)))^{\lambda_d I_N(t)}]\}\end{aligned}$$

Theorem 3:

$$\begin{aligned}dS_N(t)/dt={}&(1-\delta)\beta S_o(t)+\delta\beta O(t)+\lambda Q(t)\\&-A(t)-(\alpha+r_1)S_N(t)\end{aligned}$$

Theorem 4:

$$\begin{aligned}dL_N(t)/dt={}&\beta(1-\delta)L_o(t)+A(t)-(\alpha+r_2)\\&\bullet L_N(t)-L_N(t)[1-(1-\eta)^{\lambda_e}]\end{aligned}$$

Theorem 5:

$$\begin{aligned}dI_N(t)/dt={}&\beta(1-\delta)I_o(t)+L_N(t)[1-(1-\eta)^{\lambda_e}]\\&-(\alpha+r_3+\chi)I_N(t)\end{aligned}$$

Theorem 6:

$$\begin{aligned}dQ_N(t)/dt={}&\beta(1-\delta)Q_o(t)+\chi I_N(t)\\&-(\alpha+r_4+\lambda)Q_N(t)\end{aligned}$$

Theorem 7:

$$\begin{aligned}dR_N(t)/dt={}&\beta(1-\delta)R_o(t)+r_1S_N(t)+r_2L_N(t)\\&+r_3I_N(t)+r_4Q_N(t)-\alpha R_N(t)\end{aligned}$$

Theorem 8:

$$\begin{aligned}dO(t)/dt={}&\alpha[S_N(t)+L_N(t)+I_N(t)\\&+Q_N(t)+R_N(t)]-\beta O(t)\end{aligned}$$

Due to space limitation, we leave their proof omitted. And we advise interested readers pay attention to the author's follow-up papers for it.

C. Numerical Simulation and Analysis of Defense Model Based on Mean-field

The defense model of reactive worms based on mean-field theory should meet the following theorems:

Having developed the defense model of reactive worms, the next stage is to conduct simulation experiments by MATLAB. Some important experimental results are listed as follows.

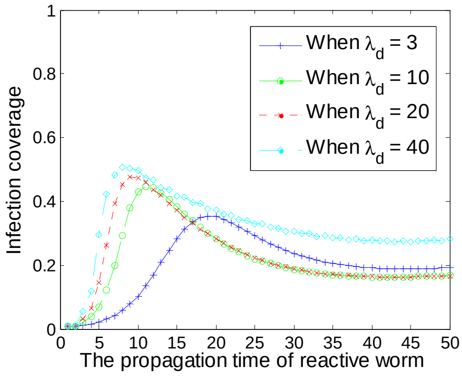


Figure 2. How the downloading rate of a node affects reactive worm propagation.

"Fig. 2" shows the influence of downloading rate of a node on the propagation speed of reactive worms. It can be seen from the figure that the higher the downloading rate of a node is, the faster reactive worms will spread. When the downloading rate of a node exceeds 10, the propagation speed of reactive worms will not be significantly increased.

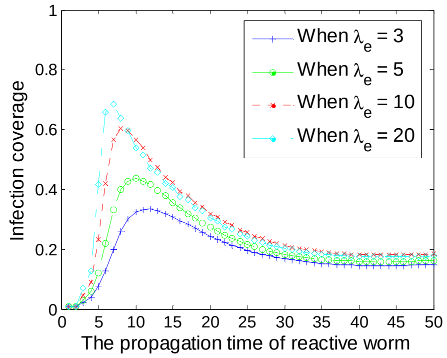


Figure 3. How the execution rate of a node affects reactive worm propagation

"Fig. 3" shows the influence of execution rate of a node on the propagation speed of reactive worms. It can be seen from the figure that the higher the execution rate of a node is, the faster reactive worms will spread. When the execution rate of a node is limited to be less than 3, the propagation speed of reactive worms can be effectively delayed.

"Fig. 4" shows the influence of offline rate of an online node on the propagation speed of reactive worms. It can be seen from the figure that the higher the offline rate of

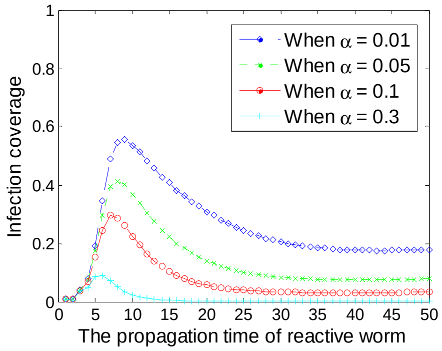


Figure 4. How the offline rate of a node affects reactive worm propagation

an online node is, the slower reactive worms will spread. When the offline rate of a node is greater than 0.1, the propagation speed of reactive worms will be effectively delayed.

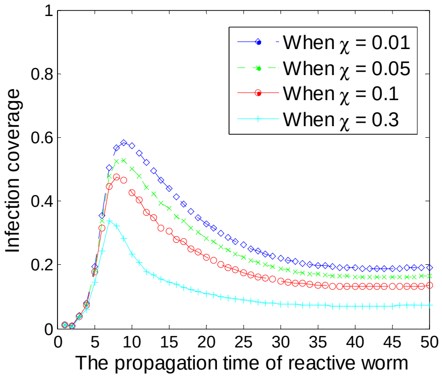


Figure 5. How the detection rate of a node affects reactive worms' propagation

"Fig. 5" shows the influence of detection rate of monitoring software on the propagation speed of reactive worms. It can be seen from the figure that the higher the detection rate of monitoring software is, the slower reactive worms will spread. When the detection rate exceeds 0.3, the propagation speed of reactive worms can be obviously delayed. This suggests that the propagation speed of reactive worms can be effectively delayed by improving detection density of monitoring software.

IV. Researches And Modification To Defense Model Of Reactive Worms Based On Mean-Field Theory

Although defense models of reactive worms based on mean-field and epidemiologic theories can roughly predict the infection ratio, the spread trend and the key points to defense reactive worms, they do not accurately match the defense process of reactive worms in dynamic environment. Because lots of parameters that influence the accuracy of these defense models are estimated under ideal conditions, these estimations are not adequate in reality. To address this issue, this paper presents a defense model based on dynamic time, then makes some improved methods to estimate the key parameters for ensuring the reliability and validity of our model. This section first analyses the deficiency of the foregoing defense model.

A. Deficiency of Defense Models Based on Mean-Field Theory

Such foregoing defense models based on mean-field theory studied the effect of P2P churn on defense effect of reactive worms under the hypothesis that the number of nodes in P2P networks remains basically unchanged within 24 hours, a day. This assumption is obviously unsuited to users' online habits.

Such foregoing defense models based on mean-field theory assume that all the nodes in L state can finish downloading each worm fragmentation within a unit of time. This assumption fails to consider the impact of the fragmentation size, network bandwidth, security awareness of user nodes and the number of seed nodes that probably provide worm fragmentations for user nodes to download during propagation.

Such foregoing defense models based on mean-field theory define the execution infection rate η as a constant. This is obviously inaccurate. Similarly, parameters including λ_e , χ , r_1 , r_2 , r_3 and r_4 should not be defined as constants either.

B. Improvements to Defense Models Based on Mean-Field Theory

Realistically, the size of online nodes is considerably different within a particular day. Most worms will take long time to reach the maximal infection peak from the beginning of attacks, the propagation of reactive worms is much more influenced by the network scale change, which heavily depends on users' habits during this period. Therefore the change rate of network scale has also been taken into consideration in our improved defense model.

In reality, the bigger and the smaller the worm fragmentation size and the network bandwidth are respectively, the lower the security awareness of user nodes is; the fewer the number of seed nodes providing worm fragmentation is, the longer a node in L state will take to download all the worm fragmentation and the larger the probability that a node in L state is detected by monitoring software is. Hence four parameters *WormSize* (represents the average size of worm fragmentation), *Bandwidth* (represents the average network bandwidth), *SecAw* (represents the security awareness of user nodes) and *SeedNum* (represents the number of seed nodes) are added in our improved defense model. Meanwhile, the probability θ that a node in L state is converted into S state has also been added.

In our improved defense model, those dynamic parameters such as η λ_e χ r_1 r_2 r_3 and r_4 are defined as mathematical functions associated with *SecAw* and ρ (represents infection coverage of reactive worms).

Two given parameters δ and λ have limited effect on the propagation of reactive worms, they are ignored in our improved defense model. In the same way, offline state will not be considered in our improved model.

V. A Defense Model Of Reactive Worms Based On Dynamic Time

A. Analysis on the Network Size with Consideraton of Users' Online Habits

Time consumed on Internet of the public shows certain regularity in real world. "Fig. 6" shows the distribution within 24 hours a day according to the CNNIC's statistics in the twentieth statistical report of Internet development in China [14].

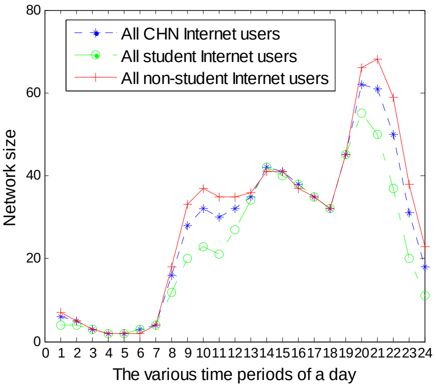


Figure 6. Online user distribution within 24 hours a day

As we see from "Fig. 6", the number of online users stays lowest from 1 a.m. to 7 a.m., and this number will gradually increase after 8 a.m. The trend will continue ascending until it reaches its first local maximum at 10 a.m. with roughly 30 percent of user nodes online. Then the number of online users will slightly drop down at around 11 a.m., while the figure keeps rising from 12 a.m. to 3 p.m. and reach the second local peak of a day with around 40 percent of user nodes online. Then the percentage falls slowly again after 3 p.m. There is a sharp rise in the number of online users around 6 p.m., and the figure will reach its third peak, the global maximum of a day at 9 p.m. with about 60 percent of user nodes online. Later on, the percentage falls rapidly and this trend will continue until 5 a.m. with only 2 percent of user nodes online still, also the minimum of a day. Moreover, the statistical report notes only 13 percent of user nodes have no fixed Internet time, while the remaining 87 percent does and follows the pattern mentioned.

On the basis of the above analytic results, we make the following assumptions about the dynamic changing regularity of the network size within 24 hours in a day.

Since most ordinary users are sleeping from 4 a.m. to 6 a.m., we define these nodes who stay online during these hours as "forever online nodes", and the number of this kind of user nodes remains constant of a day.

User nodes in our improved model are classified into two categories: working nodes and leisure nodes. The time consumed on Internet of working nodes is mainly during the working hours between 9 a.m. to 5 p.m. and the online time of leisure nodes is mainly spent during leisure hours which we assume as 7 p.m. to 12 p.m.

To simplify this model, we assume that all the working nodes or leisure nodes are online at their peak of a day.

As addressed before, there are only 8 working hours in a day and most users surf the Internet using leisure nodes at home. We set the number of leisure nodes equal to be 1.5 times than the amount of working nodes.

Both the number of working online nodes and leisure online nodes of a day are subject to the distributing rules mentioned above.

B. Assumptions and Parameters in the Defense Model Based on Dynamic Time

In order to simplify the modeling process of reactive worms, following assumptions are made:

(1) The number of online nodes in P2P networks is a dynamic variable that varies with time.

(2) P2P nodes are divided into two categories: one is forever online nodes that account for 4% of the total; the other is temporarily online nodes accounting for 96% of the total. As noted above, the temporarily online nodes can be further classified as working online nodes and leisure online nodes. The former accounts for 38.4% of the total, while the latter occupies 57.6%. Besides, the number of online nodes in various states is subject to this proportion.

(3) In various periods of a day, the number of temporarily online nodes is subject to a range of Poisson distribution with different parameters.

(4) The probability θ and r_2 are directly proportional to parameters $WormSize$ and $SecAw$, and inversely proportional to parameters $Bandwidth$ and $SeedNum$. While the probability η follows the opposite law to θ or r_2 with parameters.

Table II lists all variables in our improved defense model.

TABLE II.
Variables In The Defense Model Of Reactive Worms Based On Dynamic Time

Variable	Description	Initial values
$N(t)$	The total number of nodes in P2P networks at time t	$N(1) = 16417$
N_d	The total number of forever online nodes in P2P networks	$N_d = 10417$
$N_t(t)$	The total number of temporarily online nodes in P2P networks at time t	$N_t(1) = 6000$
$N_{tw}(t)$	The total number of working online nodes at time t	$N_{tw}(1) = 0$
$N_{tr}(t)$	The total number of leisure online nodes at time t	$N_{tr}(1) = 6000$
P_{tw}	The maximum peak of working online nodes of a day	$P_{tw} = 100000$
P_{tr}	The maximum peak of leisure online nodes of a day	$P_{tr} = 150000$
$WormSize$	The average size of worm fragmentation	$WormSize = 6MB$
$SeedNu$ $-m(t)$	The number of seed nodes at time t	$SeedNum(1) = 8.2$
$\rho(t)$	The infection coverage of reactive worms at time t, and $\rho(t) = I(t) / N(t)$	$\rho(1) = 0.01$
$SecAw(t)$	The security awareness function of user nodes at time t, and $SecAw(t) = 0.3 + 1.1 \times \rho(t)$	$SecAw(1) = 0.311$

$\theta(t)$	The probability for a node in L state to be converted into S state at time t	$\theta(1)=0.0724$
$\eta(t)$	The probability for a node in L state infected by executing a download file at time t	$\eta(1)=0.051$
λ_d	Downloading rate of a node (The number of files that any P2P node can download in a unit of time)	$\lambda_d=20$
$\lambda_e(t)$	Execution rate of a node at time t (The number of files that this P2P node can execute in t unit of time)	$\lambda_e(1)=20$
φ_d	Downloading infection rate (The probability of a node in S state infected by downloading a file)	$\varphi_d=0.3$
φ_u	Uploading infection rate (The probability of a node in S state infected by uploading a file)	$\varphi_u=0.2$
$\chi(t)$	The probability of a node in I state that is converted into Q state at time t, where $\chi(t)=0.1\times SecAw(t)$	$\chi(1)=0.0311$
$r_1(t)$	The probability of a node in S state found to contain security vulnerabilities by security software, then its state will be converted into immune state at time t. $r_1(t)=0.08\times SecAw(t)$	$r_1(1)=0.0249$
$r_2(t)$	The probability of a node in L state found to contain security vulnerabilities by security software, then its state will be converted into immune state at time t. $r_2(t)=0.08\times SecAw(t)\lg[SeedNum(t)$ $+Bandwidth/WormSize]$	$r_2(1)=0.0241$
$r_3(t)$	The probability of a node in I state found to contain security vulnerabilities by security software, then its state will be converted into immune state at time t. $r_3(t)=0.15\times SecAw(t)$	$r_3(1)=0.0466$
$r_4(t)$	The probability for a node in Q state found to contain security vulnerabilities by security software, then it will be patched and its state will be converted into immune state at time t. $r_4(t)=0.2\times SecAw(t)$	$r_4(1)=0.0622$
$S(t)$	The number of online nodes in S state at time t, where $S(1)$ indicates the total number of nodes in S state in P2P networks initially	$S(1)=16253$
$L(t)$	The number of online nodes in L state at time t	$L(1)=0$
$I(t)$	The number of online nodes in I state at time t, where $\mathrm{I}(1)$ indicates the total number of nodes in I state in P2P networks initially	$I(1)=164$
$Q(t)$	The number of online nodes in Q state at time t	$Q(1)=0$
$R(t)$	The number of online nodes in R state at time t	$R(1)=0$

C. A Defense Model of Reactive Worms Based on Dynamic Time

Having proposed model assumptions and parameter elucidations, our next stage is to develop the improved defense model based on dynamic time. The modeling methodology based on dynamic time is similar to the one based on mean-field theory. The key of the improved modeling methodology is how to estimate the number of online nodes in various states at different times within a day. This section will focus on resolving this problem.

In this model, variable t represents different hours of a day, and the number of online nodes at different times of a day can be discussed the way as follows:

(1) The discussion on the number of working online nodes at different times of a day.

When $t=1-6$, sleeping time for ordinary users, the number of working online nodes at this period is set as 0. That is $N_{tw}(t)=0\ (t\in[1,2,..6])$ (1)

When $t=7-15$, the number of working online nodes gradually increases because most users begin working and major stock exchanges throughout the world open one after another. The number of working online nodes keeps growing within this period, and this figure will reach the peak at 3 p.m. The changing trend of working online nodes during this period obeys the Poisson distribution of parameter 2.6. That is

$$N_{tw}(t)=P_{tw}\bullet \sum_{k=0}^{t-7}(\lambda^k/k!)\bullet e^{-\lambda}\ (\lambda=2.6,t\in[7,8,...,15]) \quad (2)$$

When $t=16-24$, almost the end of daily working hours, the number of working online nodes continue to retreat from its peak at 3 p.m. until it finally reduces to be zero. Given individual users might do extra work at night, the descending trend of the number of working online nodes will continue to 12 p.m. The changing trend of working online nodes during this period also conforms to the Poisson distribution of parameter 3.3. That is

$$N_{tw}(t)=P_{tw}(1-\sum_{k=0}^{t-16}(\lambda^k/k!)\bullet e^{-\lambda}) \quad (\lambda=3.3,t\in[16,17,...,24]) \quad (3)$$

When $t>15$, $P_{tw}\sum_{k=0}^{t-7}(\lambda^k/k!)\bullet e^{-\lambda}=1$. Combining the three Equations (1) (2) (3), the change of working online nodes within 24 hours a day can be calculated as follows:

$$N_{tw}(t)=P_{tw}(\sum_{k=0}^{t-7}(2.6^k/k!)\bullet e^{-2.6}-\sum_{k=0}^{t-7}(3.3^k/k!)\bullet e^{-3.3}),$$

$$\text{And } 1\le t\le 24 \quad (4)$$

(2) The discussion on the number of leisure online nodes at different times of a day.

When $t=1-3$, most ordinary users are sleeping, while only a few users are playing online games or watching online videos. Therefore the number of this part of users is on the decline, the change of leisure online nodes during this period conforms to the Poisson distribution of parameter 1.3. That is

$$N_{tr}(t)=P_{tr}(1-\sum_{k=0}^{t+2}(\lambda^k/k!)\bullet e^{-\lambda})\ (\lambda=1.3,t\in[1,2,3]) \quad (5)$$

When $t=4-18$,most users are either resting or working, few of leisure online nodes are being used at this period, so the number of leisure online nodes at this period is set as 0. That is $N_{tr}(t)=0\ (t\in[4,5,...,18])$ (6)

When $t=19-22$, the leisure online nodes are being used in large amounts, so the number of leisure online nodes will rapidly climb to the maximum peak of a day. The change of leisure online nodes during this period also conforms to the Poisson distribution of parameter 0.5. That is

$$N_{tr}(t) = P_{tr}\sum_{k=0}^{t-19}(\lambda^k / k!)\bullet e^{-\lambda}, \quad \lambda = 0.5, \quad t \in [19, 20, 21, 22] \qquad (7)$$

When $t = 23-24$, it's again time for bed, causing the number of leisure online nodes rapidly decreases. The change of leisure online nodes during this period obeys the Poisson distribution of parameter 1.3. That is

$$N_{tr}(t) = P_{tr}(1-\sum_{k=0}^{t-23}(\lambda^k / k!)\bullet e^{-\lambda}) \; (\lambda = 1.3, t \in [23, 24]) \qquad (8)$$

When $t > 3$, the equation $1-\sum_{k=0}^{t+2}(1.3^k / k!)\bullet e^{-1.3} = 0$ is valid, no matter what value the variable t is. And when $t > 22$, the equation $\sum_{k=0}^{t-19}(0.5^k / k!)\bullet e^{-0.5} = 1$ is valid, no matter what value the variable t is. Combining the four Equations (5) (6) (7) (8), the change of leisure online nodes within 24 hours a day can be calculated as following equation (9).

$$N_{tr}(t) = P_{tr}(1-\sum_{k=0}^{t+2}\frac{1.3^k}{k!}e^{-1.3} + \sum_{k=0}^{t-19}\frac{0.5^k}{k!}e^{-0.5} - \sum_{k=0}^{t-23}\frac{1.3^k}{k!}e^{-1.3})$$

And $1 \le t \le 24$. (9)

(3) In conclusion, the change of all the online nodes within 24 hours a day can be calculated as following Equation (10).

$$\begin{aligned} N(t) &= N_d + N_t(t) = N_d + N_{tw}(t) + N_{tr}(t) \\ &= N_d + P_{tw}(\sum_{k=0}^{t-7}\frac{2.6^k}{k!}e^{-2.6} - \sum_{k=0}^{t-7}\frac{3.3^k}{k!}e^{-3.3}) + P_{tr}(1 \\ &\quad -\sum_{k=0}^{t+2}\frac{1.3^k}{k!}e^{-1.3} + \sum_{k=0}^{t-19}\frac{0.5^k}{k!}e^{-0.5} - \sum_{k=0}^{t-23}\frac{1.3^k}{k!}e^{-1.3}) \end{aligned}$$

And $1 \le t \le 24$ (10)

In this improved defense model, the change of nodes in S state in a unit of time is composed of four parts. The first part will be converted into L state for downloading worm fragmentation from infected nodes; the second part will be converted into L state for uploading some resource to infected nodes; the third part is converted from nodes in L state because these latent nodes are found to contain worm fragmentation by security software before their states are converted into S with worm fragmentation being removed; the fourth part will be converted into R state because these nodes in S state are found to contain security vulnerabilities by security software, they will be patched and their states will be converted into immune state. Given the above, the change rate of nodes in S state satisfies the following Equation (11).

$$\begin{aligned} dS(t)/dt = \theta(t)L(t) - [&\varphi_d\lambda_d I(t)/N(t) + \\ &\varphi_u\{1-[1-1/N(t)]^{\lambda_d I(t)}\} + r_1(t)]S(t) \end{aligned} \qquad (11)$$

The same theory proves that the change rate of nodes in L state satisfies the following Equation (12).

$$\begin{aligned} dL(t)/dt = [&\varphi_d\lambda_d I(t)/N(t) + \varphi_u\{1-[1-1/N(t)]^{\lambda_d I(t)}\}] \\ &S(t) - \{\theta(t) + r_2(t) + \{1-[1-\eta(t)]^{\lambda_e(t)}\}\}L(t) \end{aligned} \qquad (12)$$

The change rate of nodes in I state satisfies the following Equation (13).

$$dI(t)/dt = \{1-[1-\eta(t)]^{\lambda_e(t)}\}L(t) - [r_3(t) + \chi(t)]I(t) \qquad (13)$$

The change rate of nodes in Q state satisfies the following Equation (14).

$$dQ(t)/dt = \chi(t)I(t) - r_4(t)Q(t) \qquad (14)$$

And the change rate of nodes in R state satisfies the following Equation (15).

$$dR(t)/dt = r_1(t)S(t) + r_2(t)L(t) + r_3(t)I(t) + r_4(t)Q(t) \qquad (15)$$

For the sake of brevity, we leave their proof omitted.

D. Numerical Simulation and Analysis of Defense Model Based on Dynamic Time

There are three steps to count the number of infected nodes:

The first step is to initialize the number of online nodes in all states during a first time period, the second step is to calculate the change in numbers of online nodes in all states during the same period of time according to the formula (11-15), the third step is to reckon the actual number of online nodes during a second time period according to the formula (10), the fourth step is to initialize the number of online nodes in all states during a second time period according to the proportion of different states of online nodes that has been calculated in second step and the actual number of online nodes that has been reckon in third step, then the change in numbers of online nodes in all states during the second period of time can be calculated in the fifth step, The rest can be done in the same manner, until the number change of online nodes in all states for 24 hours within a day has been calculated.

Key parameters affecting reactive worm defense in real environment can be deduced by adjusting the parameters in our improved defense model.

"Fig. 7" shows the influence of downloading rate of a node on the defense of reactive worms in this improved model. It can be seen from the figure that the higher the downloading rate of a node, the larger the number of online nodes infected by reactive worms will be. This

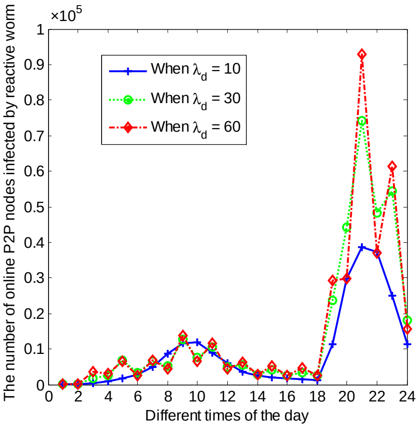


Figure 7. How the downloading rate of a node affects reactive worm propagation in improved model

parameter has great effects on the defense of reactive worms, if large amount of reactive worms have been found in P2P networks, the propagation of reactive worms can be effectively controlled by restricting the downloading rate of each node.

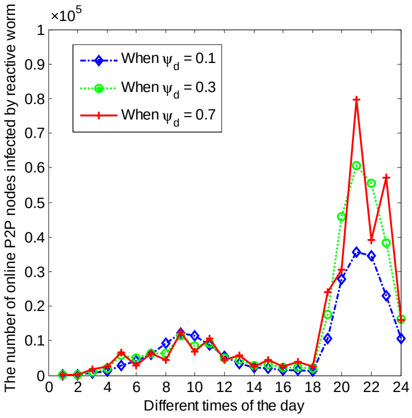


Figure 8. How the downloading infection rate of a node affects reactive worm propagation in improved model

“Fig. 8” shows the influence of downloading infection rate of a node on the defense of reactive worms in this improved model. It can be seen from the figure that the higher the downloading infection rate of a node stays, the larger the number of online nodes infected by reactive worms will be.

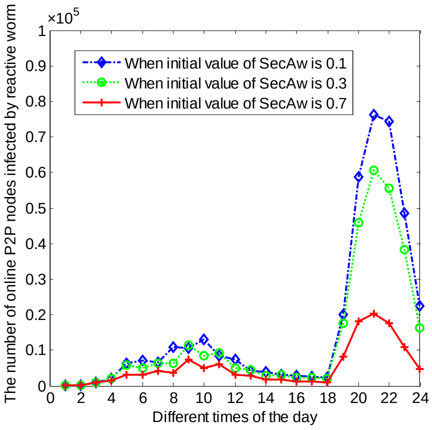


Figure 9. How the initial value of SecAw affects reactive worm propagation in improved model

“Fig. 9” shows the influence of the initial value of the security awareness function of user nodes on the defense of reactive worms in this improved model. In this improved model, those parameters such as $\theta(t)$, $\lambda_e(t)$, $\chi(t)$, $r_1(t)$, $r_2(t)$, $r_3(t)$, $r_4(t)$ are all related to the security awareness function of user nodes. It can be seen from the figure that the higher the security awareness of a user node is, the fewer the number of online nodes infected by reactive worms will be and also the better the defense effect of reactive worm can be obtained. Therefore cyber-safety education should be expanded to all ranges of Internet users for raising their knowledge level and safe consciousness, which effectively help defense against reactive worms in P2P networks.

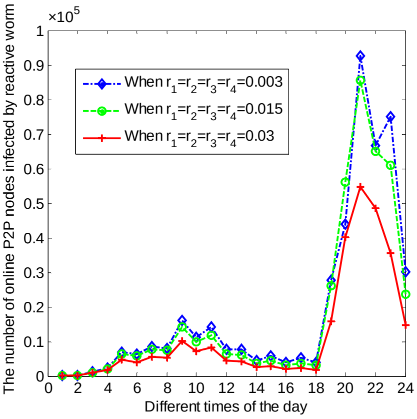


Figure 10. How the initial value of immunity system affects reactive worm propagation in improved model

“Fig. 10” shows the influence of the initial value of immunity system on the defense of reactive worms in this improved model. It can be seen from the figure that the higher the initial value of immunity system is, the fewer the number of online nodes infected by reactive worm will be and the better the defense effect of reactive worm can be.

In conclusion, user's online habits give it a rise to the most significant impact on the worm attack according to these simulation results. As you can see from these figures, only few of online P2P nodes are infected by reactive worms between 1 a.m. and 4 a.m. because most users are sleeping; the number of online P2P nodes infected by reactive worms begin ascending between 4 a.m. and 11 a.m. because most users go to work during this period; however the number of online infected nodes remains very limited even during peak hours. The major reason for this is that most leisure nodes that occupy the mainstream of P2P networks are not at working hours during this period, meanwhile only few working nodes join P2P networks, seriously restricting the developing speed of reactive worms; the number of online nodes in P2P networks further reduces form 11 a.m. to 6 p.m. because major stock exchanges close and most users go home from working; and the number of P2P nodes infected by reactive worms will fall from the previous peak to relatively lower level; the peak of infection of a day occurs during 6 p.m. and 12 p.m.; the maximum of the number of infected nodes appears about 8 p.m. because most leisure nodes have joined to P2P networks by that time; the peak hour of surfing on the internet also appears at 8 p.m., offering an opportunity for reactive worms’ sudden spread, meanwhile the number of online P2P nodes infected by reactive worms is booming; the number of infected P2P nodes will decrease rapidly within bed time after 10 p.m. Obviously the most crucial time of defending reactive worms is from 6 p.m. to 10 p.m. In order to effectively guard against reactive worms’ attack in P2P networks and ensure the availability of normal operations of P2P networks, we should increase strength on key nodes’ supervision, speed up the frequency of scanning vulnerability during particular periods within a day.

VI. Conclusion And Future Work

In the paper, firstly we proposed a propagation strategy of reactive worms in dynamic environment and provided the process of nodes' state transition when reactive worms spread in accordance with the strategy proposed; second we developed a defense model based on mean-field theory; third, we analyzed shortages of the foregoing model and proposed an improved defense model of reactive worms based on dynamic time; finally we compared the difference among key parameters that affect reactive worm defense between the model based on mean-field theory and the one based on dynamic time and deduced the most important period of a day for defending against reactive worm attack.

Future work will involve how to improve the detection rate of monitoring software according to the characteristics of reactive worms, how to improve the accuracy of the defense model of reactive worms by considering the trust relationship and social nature between P2P nodes upon the propagation of reactive worms and how to build an efficient defense system to prevent reactive worms based on these works.

Acknowledgment

This work is supported by China Postdoctoral Science Foundation funded project (Grant No. 2014M552308), NCET, NSFC (Grant No. 61272400,Grant No. 61309032), Chongqing Innovative Team Fund for College Development Project (Grant No. KJTD201310), Foundation of CEC (Grant No. KJ130525, KJ130523), and CQUPT Research Fund for Young Scholars (Grant No. A2012-79).

References

[1] A. G. McKendrick. "Applications of mathematics to medical problems," Proc. the 44th Edinburgh Mathematica Society. Edinburgh, pp. 98-130, 1926.

[2] J. O. Kephart, S. R. White. "Directed-graph epidemiological models of computer viruses". Proc. the IEEE Symp on Security and Privacy. Piscataway, NJ,pp. 343-359, 1991.

[3] W. Yu, "Analyze the Worm-Based Attack in Large Scale P2P Networks", Proc. the 8th IEEE International Symposium on High Assurance Systems Engineering (HASE'04), pp.308-309, 2004.

[4] Q. D. Sun, Q. Wang, J. Ren, "Modeling and Analysis of the Proactive Worm in Unstructured Peer-to-Peer Network", Journal of Convergence Information Technology, vol. 5, No. 5, pp.111-117, 2010.

[5] Q. R. Li, W. Han，"An Analysis for Stochastic Model of Worm Propagation"，Journal of IJACT, Korea, vol. 4, No. 3, pp. 156-164, 2012.

[6] S. Y. Yang, J. T. Jiang, P. Z. Chen, "OOPProPHET: A New Routing Method to Integrate the Delivery Predictability of ProPHETRouting with OOP-Routing in Delay Tolerant Networks, " Journal of Computers, vol.8, no.7, pp. 1656-1663, July 2013.

[7] L. X. Xie, X. Zhang, J. Y. Zhang, "Network Security Risk Assessment Based on Attack Graph, " Journal of Computers, vol.8, no.9, pp. 2339-2347, September 2013.

[8] C. Q. Xing, J. L. Lan, Y. X. Hu, "Virtual Network with Security Guarantee Embedding Algorithms," Journal of Computers, vol.8, no.11,pp. 2782-2788, November 2013.

[9] Y. W. Wang, J. W. Jing, J. Xiang, "Contagion worm propagation simulation and analysis", Journal of Computer Research and Development, China, vol. 45, pp.207-216, 2008.

[10] Z. G. Qin, C. S. Feng, F. L. Zhang et al. "Modeling propagation of reactive worm in P2P networks". Proc. Communications, Circuits and Systems 2009(ICCCAS 2009), Milpitas, CA, 2009, 335-340.

[11] C. S. Feng, Z. G. Qin, L. Cuthbert.. "Reactive Worms Propagation Modeling and Analysis in Peer-to-Peer Networks", Journal of Computer Research and Development, China, vol. 47, pp.500-507, 2010.

[12] W. Yang, M. G. R. Chang, Y. Yao, and X. M. Shen, "Stability Analysis of P2P Worm Propagation Model with Dynamic Quarantine Defense," Journal of Networks. Finland, vol. 6, pp. 153–162, January 2011.

[13] G. Ouyang, X. Chen. "Trust Model Based on P2P Network," Journal of Networks. Finland, vol. 8, pp. 2013–2020, September 2013.

[14] CNNIC, The twentieth statistical report on Internet development in China [EB/OL]. http://www.cnnic.cn/gywm/xwzx/rdxw/2007nrd/201207/t20120710_31532.htm, July 18, 2007.

Haokun Tang received B.S. degree in computer application from Southwestern Normal University, China, in 1999. M.S. degree in computer application from Southwestern Normal University, China, in 2003. And Ph.D. in Computer Systems Organization from the University of Electronic Science and Technology of China. China, in 2013. Now he is a researcher at postdoctoral workstation, HanHua Financial Holding. He has published more than 16 refereed journal/conference papers. His current research interests are in network security, P2P applications, cloud computing.

Shitong Zhu is currently a full-time sophomore student at School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, China. His current research interests include Device-to-Device communications, LTE-A etc.

Jun Huang received B.S. degree in computer science from Hubei University of Automotive Technology, China, in 2005. M.S. degree (with honor) in computer science from Chongqing University of Posts and Telecommunications, China, in 2009. And Ph.D. degree (with honor) from Institute of Network Technology, Beijing University of Posts and Telecommunications, China, in 2012. Now he is an Associate Professor at School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications. He was a visiting researcher at Global Information and Telecommunication Institute, Waseda University, Tokyo, from Sept. 2010 to Sept.

2011. He is a member of IEEE and IEICE. A best paper award winner of AsiaFI 2011. He has published more than 20 refereed journal/conference papers. His current research interests include network optimization, future Internet, and Cloud computing etc.

Hong Liu currently studies for her Ph.D. in Computer Engineering from the University of Chongqing. She is an associate professor in Chongqing University of Posts and Telecommunications in the Department of Software Engineering, China. She published several papers in the area of wireless networks and image security. Her research interests include sensor networks, image processing and image security.

An Improved Algorithm of Quantum Particle Swarm Optimization

Yan-xia JIN
North University of China/Department of Computer Science and Technology, Taiyuan, China
Email:jinyanxia_730128@163.com

Jing XUE and Zhi-bin SHI
North University of China/Department of Computer Science and Technology, Taiyuan, China
Email: 150301757@qq.com & shizb@nuc.edu.cn

***Abstract*—Based on the classical particle optimization algorithm and the quantum behavioral theory, this paper proposes an improved QPSO algorithm---GLQPSO to perfect the global and local convergence speed ability and speed of classical particle swarm. To achieve this purpose, the author introduces an improved Logistic chaotic mapping theory [1] to conduct chaotic search for the initial particle and chaotic remolding of the locally optimized particle swarm. The test of the classical function has proved the success of this effort.**

***Index Terms*—quantum behavior, PSO algorithm, chaotic mapping thoughts, local search**

I. Introduction

In 1995, Kennedy and Eberhart put forward the Particle Swarm Optimization, PSO algorithm. PSO is an evolutionary computation technique, developed for optimization of continuous nonlinear, constrained and unconstrained, non differentiable multimodal functions [2]. It is a random search algorithm of group cooperation and developed by imitating the foraging behavior of Bery. Kennedy and Suganthan [3] analyze the performance of this algorithm in neighborhood operator and its difference from the standard GA. PSO has better computational efficiency, i.e. it requires less memory space and less speed of CPU, and it has less number of parameters to adjust [4]. PSO has gained popularity lately and has been widely applied in different fields [5]. The development of this algorithm benefits from observations of social behaviors of animals, such as bird flocking and fish schooling. PSO was first used to optimize the nonlinear continuous function and train the neural network. Then it is applied to solve the constraint optimizing issue, the multi-objective issue and the dynamic optimizing issue. Now, it gradually becomes a good tool in data classification, clustering, mode recognition, telecommunications Qos management, biological system modeling, flow layout, signal process, robot control [6],vector machines [7], Micro-grid [8], decision support, simulation and system discrimination [9,10].

Despite the great efforts made by many researchers in improving the performance of this algorithm and certain success achieved in this process, the large part of the simplicity and convenience of the algorithm has been sacrificed and its calculated quantity has been increased, which is apparently against the original intention of presenting this optimization algorithm. Therefore, to find a better optimization way while without increasing the calculated quantity becomes pressing.

To solve this issue, the author of this paper introduces an improved logistic chaotic mapping to describe the initial population based on the Quantum PSO, QPSO. The reason is that in spite of its better global convergence ability, the global search capacity of QPSO will be relatively weakened and its local search capacity will be strengthened with the continual increase of iterations, which will result in the local optimal point. When parts of particles reach the local extreme points, the logistic mapping is again introduced to locally initializing these points. It can not only improve the quality of the initial population but also the local optimization ability of the QPSO, and further enhance its computational accuracy.

Ⅱ.BRIFE INTRODUCTION OF THE BASIC PSO

Human being have their own previous experience, set beliefs and set rules of doing some work, based on which they take their actions also human follow the path set by society or group. This path is supposed to be the best according to the whole global best position [11]. Similar to other population-based algorithms, such as evolutionary algorithms, PSO can solve a variety of difficult optimization problems, and has shown a faster convergence rate than other evolutionary algorithms on some problems [12,13]. PSO is an evolutionary computation technique motivated by the simulation of social behavior. Namely, each individual (agent) utilizes two important kinds of information in decision process [14]. PSO is a method for performing numerical optimization without explicit knowledge of gradient of the problem to be optimized. It was originally developed for nonlinear optimization problems with continuous variables so that it can easily be expanded to treat a problem with discrete variables [15].

The basic principle of PSO algorithm is described as the following: in the D-dimensional space, an aggregate of n particles is flying at certain speed. In the search space, each

doi:10.4304/jsw.9.11.2789-2795

particle's movement will be depicted through three parameters: x_i, v_i, p_i [15].

An individual particle i is composed of these vectors: its position in the n-dimensional search space $x_i=(x_{i1}, x_{i2}, \ldots, x_{in})$, the best position that it has individually found $p_i=(p_{i1},p_{i2},\ldots,p_{in})$, and its velocity $v_i=(v_{i1},v_{i2},\ldots,v_{in})$. Particles were originally initialized in a uniform random manner throughout the search space; velocity is also randomly initialized. Each particle adjusts its trajectory toward its own previous best position P_{best} and the previous best position G_{best} attained by the whole swarm [16].

Formula (1) and (2) [17,18]are introduced to get the particle's position and velocity, among which, c_1 and c_2 are called study gene, normally equal to 2; Parameter r and R are two false random numbers evenly distributing in the range [0,1]; x_i and v_i are repeatedly limited within the maximum translocation and maximum velocity.

The modified velocity and position of each particle can be manipulated according to the following equations: where i=1,2,…,n; w is a weight factor which controls the velocity's magnitude; c_1 and c_2 are two positive constants, known as acceleration coefficients; r_1 and r_2 are random numbers within the range[0,1][19,20]. The x_i and v_i are limited to the maximum displacement and the corresponding speed.

$$v_i^{k+1} = wv_i^k + c_1 r_1 (p_i - x_i^k) + c_2 r_2 (p_g - x_i^k) \qquad (1)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \qquad (2)$$

The PSO process is given as the followings:

- Initial positions and velocities within the initialization are generated. pi represents the current position of each particle. pg represents the pi of the best value.
- Calculating the fitness value of each particle;
- The fitness value of each particle is compared to the current best position. It is current as the optimal value of the individual particles. The current location update the personal best position;
- The pi and pg of each particle is compared, if the new pg value is better than the previous pg value, the previous pg value is replaced by the new pg value;
- According to Formula (1) and (2), velocity and position of the particles are updated;
- Go to Step 2 until termination condition is reached (Termination condition is generally set to a sufficiently good fitness value or reach a preset maximum number of iterations).

III. Quantum Particle Swarm Optimization Algorithm

In 2004, Sun and other researchers present a new PSO algorithm model from the perspective of quantum mechanics. This model is based on the DELTA trap that the particle has a quantum behavior. Compared with the original PSO algorithm, this new QPSO has only radius vector in its evolution equation [21], which greatly simplifies the equation, reduces its parameters and thereby makes the equation more controllable [22].

In contrast with the classical PSO algorithm, the convergence speed of QPSO is fast and QPSO can not easily fall into the local optimum. This will be helpful for finding the optimal parameters [23]. In the QPSO algorithm, the quantum state of a microscopic particle is described with wave function $\varphi(r, t)$. Once $\varphi(r, t)$ is set, the average value and probability measure of any dynamical variable become certain. That is because, in the quantum world, the moving track of a particle is unlimited. It only changes with the change of time. Its movement can be described with time-dependent Schrodinger equation, which means Particle x_i and v_i cannot be determined simultaneously.

Based on Reference 9, the motion formula (3) of particle is introduced, in which, p、a、$Mbest_i$ and $Mbest^i$ are presented in Formula (4), (5), (6) and (7) respectively[24]. U is a random number in interval [0,1]. Generally, a_1 and a_2 are respectively the beginning and ending value of variable shrinkage factor t. MAXITER is the maximum interactions. Since a_1=2.5，a_2=0.5，in most cases, the value of a is between 0.5 and 2 [25]. $Mbest_i$ represents the average value of each particle when it is in the best position in the local searching. p_{id} reveals the best position of the particle in such a searching, while $Mbest_i$ and p_{id} represent the corresponding number in an overall searching.

$$x_i(t+1) = p \pm a \times |Mbest_i - x_i(t)| \times \ln(1/u) \qquad (3)$$

$$p = (c_1 p_{id} + c_2 p_{gd}^i)/(c_1 + c_2)\ (c_1+c_2=1) \qquad (4)$$

$$a=(a_1-a_2)*((\text{MAXIER}-t)/\text{MAXIER}+a_2) \qquad (5)$$

$$Mbest_i=\frac{1}{M}\sum_{d=1}^{M} p_{id} \qquad (6)$$

$$Mbest^i=\frac{1}{M}\sum_{d=1}^{M} p_{gd}^i \qquad (7)$$

The procedure of a QPSO is shown as the following:

- The position and speed of each particle in a randomly initialized particle group in the D-dimensional space;
- The overall and local best position of each particle determined based on Formula (6) and (7);
- The present best value of each particle by comparing its overall position with the local one;
- The best value of variable shrinkage factor got by changing t in Formula (5) ;
- Updating the present speed and position of each particle by substituting the result of Step (3) to Formula (1) and (3)[26];
- If the above result fails to exceed the maximum speed and displacement value or the expected best state, please return to Step 2 to repeat the whole process.

Ⅳ. THE IDENTICAL PARTICLE SYSTEM

In the quantum mechanics, particles of the same type are called identical particles. The exchanging symmetry of such a particle system will set a strong limit on the wave function. In general, wave-function $\psi(q_1, q_2, \ldots, q_n)$ of this system is not necessarily the inherent property of certain p_{ij}. All of the p_{ij} should be equally important. A detailed analysis has proved that the common inherent property of all the p_{ij} exists.

Introduction of the identical particle system, the corresponding wave function must conform to the situation set in Formula (8).

$$p_{ij}^2\varphi = cp_{ij}\varphi = c^2\varphi \quad (P_{ij}^2=1,\ C^2=1,\ i\neq j,\ i=1,2,3,\ldots,\ n, j=1,2,3,\ldots.) \tag{8}$$

Just as what mentioned above, the motion formula of the particle group is changed into Formula (9), which is used to get a new position for the particle in the present position. And the identical particles are introduced to limit the displacement equation of each particle so as to update its displacement and speed.

$$x_i(t+1) = C^2(p \pm \alpha\times|Mbest_i - x_i(t)|\times\ln(1/u)\ (u\in[0,1]) \tag{9}$$

Ⅴ. An improved Logistic Chaotic Mapping

A. Unidimensional nonlinear logistic chaotic mapping

Chaotic is a widely existing nonlinear phenomenon in the nature. Commonly, people call the random motion state got in the deterministic equation as chaotic. Logistic mapping is a typical chaotic system, its iterative formula is shown as Formula (10): in which, μ is control parameter; when μ=4，$0\leq z_0\leq 1$, Logistic is in a complete chaotic state. Due to the randomness, traversal, and sensitivity to the initial situation of the chaotic motion, the search technology based on chaotic will be more effective that other random search techniques.

$$x_{n+1} = \mu x_n(1-x_n) \quad n=1, 2, \ldots \quad \mu\in(2, 4] \tag{10}$$

B. An improved chaotic mapping

Because of the pseudo randomness of the chaotic, probability and statistics can be applied to conduct quantitative research on the property of the chaotic sequence. Based on reference [1], the probability distribution density function of such sequence produced through Schuster H.G Formula (11) is:

$$\rho(x)=\begin{cases}\dfrac{1}{\pi\sqrt{x(1-x)}} & 0<x<1\\ 0 & x\leq 0, x\geq 1\end{cases} \tag{11}$$

The related experiment proves the inconsistent distribution of Logistic mapping. To get a random system with a consistent distribution, Formula (10) can be changed as the following:

$$y^n = \frac{2}{\pi}sin^{-1}(\sqrt{x_n}), \qquad \text{n=1,2,3,…} \tag{12}$$

The time average of x, that is, the average value of the chaotic sequence tracing point is:

$$\bar{x} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}x_i = \int_0^0 x\rho(x)dx = 0 \tag{13}$$

The distribution function of variable y is:

$$F\{y\leq Y\}=F\{x\leq sin(\frac{\pi Y}{2})\}=$$

$$\int_0^{sin\left(\frac{\pi Y}{2}\right)}\rho(x)dx = \int_0^{sin\left(\frac{\pi Y}{2}\right)}\frac{1}{\pi\sqrt{x(1-x)}}dx = Y \tag{14}$$

Formula (12) conforms to the consistent distribution within the (0,1) interval, having better random distribution than Formula (10).

Thus, the probability density function of the variable y is:

$$\rho(Y)=\frac{dF}{dY}\{y\leq Y\}=1 \tag{15}$$

Ⅵ. An QPSO algorith based on the improved Logistic chaotic mapping

In recent years, researchers both home and abroad have raised a lot of versions of PSO algorithms, with many of which use normal, Cauchy, uniform and exponential distribution to produce random sequence to update the speed of this algorithm. In this paper, a QPSO algorithm based on the improved Logistic chaotic mapping is proposed. The foundation of chaotic theory is to use the traversal of the chaotic motion to produce a large amount of particle population and select the optimal ones from them. It can further prevent the particles from beginning to conduct local search too early and help them to find their optimal position.

In QPSO algorithm, the chaotic sequence replaces the random sequence to achieve the diversity of QPSO population and the improved performance of this algorithm, which is very useful in restrain the minimization of local convergence. The improved Logistic chaotic mapping mentioned in Reference[1] means transforming Formula (10) into Formula (12), which can result in more even variable distribution and better randomness. By applying the above improved algorithm to QPSO, the author can not only ensure the even distribution of the particle in their initial situation, but also further initialize part of the local extreme points in the later local convergence process. The detailed procedure is shown as the following:

- Based on the improved one-dimensional Logistic chaotic mapping system presented in Reference[1], a large number of initial population are produced and the best of them are selected;
- Using the formula (17) to determine the position of each particle in the identical particle system [27];
- Using the formula (16) and (17) to determine the global and local optimal position of each particle respectively, in which, pgd is their global optimal position and pid is the local one; Then comparing the two positions of each particle to select the best value as the present optimal value of this particle.

$$Mbest=\frac{1}{M}\sum_{i=1}^{N} p_{gd}^{i} \qquad (16)$$

$$Mbest_i=\frac{1}{M}\sum_{i=1}^{N} p_{id} \qquad (17)$$

- Subtracting the local and global optimal value of each particle; if the result is smaller than certain order of magnitude or is a minus, it means this particle has fallen into the local minimum sector and the improved Logistic system can be used to locally optimize this part of particles, thereby to prevent them from beginning local minimization too soon and from failing to find their optimal positions. That is why this new search technology is more effective than other techniques of the same type;
- Substituting the result of Step (3) into Formula (9) to update the present position of the particle;
- If the above result does not exceed the maximum value of the speed and displacement or fails to achieve the expected state, return to the Step 2 to repeat the procedure.

Ⅶ.PERFORMANCE ANALYSIS OF THE IMPROVED QPSO ALGORITHM

In the following, the property of GLQPSO is tested by comparing with that of PSO, CPSO (Chaos particle swarm optimization) and QPSO through optimizing the Ackley, Rosenbrock and Rastrigin function when they are in their ten-dimensional conditions. According to Reference [28], the globally optimal value of Rastrigin function is 0. It is a nonlinear unction of multiple peak values, having many local optimal points and very difficult of find its globally optimal value and therefore also very difficult to be optimized by the optimization algorithm. The globally optimal values of Griewank, Ackley functions are also 0. They are invariably nonlinear functions of multiple peak values. Their local optimal points are distributed regularly, whose quantity gradually setting of the parameters in the algorithm: $c_1=c_2=2$, the value of w will drop from 0.9 linearity to 0.4 with the change of iteration step, its maximum iteration is 1000; that of a will drop from 1.0 to 0.5.

TABLE Ⅰ.
THE INITIAL VALUES OF RASTRIGIN , ACKLEY AND GRIEWANK TRIAL FUNCTION

test function	search interval	initialization interval	space dimension
Ackley	(-32.768,32.768)	(-32.768,32.768)	10,3
Rastrigin	[-10,10]	[-10,10]	10,3
Rosenbrock	[-5,10]	(-5,10)	10,3

$$\text{Rastrigin} \quad \sum_{i=1}^{n}(x_i^2-10cos(2\pi x_i))+10 \qquad (18)$$

$$\text{Rosenbrock} \quad \sum_{i=1}^{n-1}[100(x_i^2-x_{i+1})^2+(x_i-1)^2] \qquad (19)$$

$$\text{Ackley} \quad 20+e-20exp(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2})-exp(\frac{1}{n}\sum_{i=1}^{n}cos(2\pi x_i)) \qquad (20)$$

TABLE Ⅱ.
THE COMPARED OF PSO、CPSO、QPSO 、GLQPSO RESULTS

		global optimum value（max）	global optimum value（min）	Mean global optimum value	Global optimum value variance
Ackley	PSO	2.3172	0.0001201	1.1898	0.64965
	CPSO	3.0271	0.000472	1.0604	0.82805
	QPSO	0.02339	5.46E-02	4.0053	1.23
	GLQPSO	0.011339	9.46E-05	0.002382	1.14E-05
Rosenbrock	PSO	9.5865	0.24229	6.3559	4.4032
	CPSO	9.7047	0.25196	6.7629	5.1954
	QPSO	4.6148	2.97E-07	0.5778	0.60297
	GLQPSO	4.1003	5.94E-10	0.16944	0.63E-04
Rastrigin	PSO	38.803	1.99	17.003	67.836
	CPSO	32.841	4.0241	15.274	27.937
	QPSO	23.0337	7.26E-06	5.4415	35.472
	GLQPSO	19.9	2.74E-07	5.3436	24.568

Table 1 lists the value range of variable initialization of the three functions. Table 2 lists the optimization results of the four algorithms with the number of iterations being 1000. It reveals that arranging from high to low, the order of their convergence values should be that of GLQPSO, of QPSO, of CPSO and of PSO, which means GLQPSO has the best convergent tendency of the four. As for their speed to reach the global optimum the order is quite the same, with GLQPSO being the fastest. It proves that the employment of the probability density function helps to improve the chaotic randomness, restrain the local convergence, increase the iterations and facilitate the group to reach the global optimum quicker. Mean square error refers to the distance square of each datum drifting from the average, disclosing the dispersion degree of a Datasets. Among the four algorithms, GLQPSO's mean square error is the lowest because this algorithm can not only help to improve the quality of the initial population through the identical particle, but also use the bettered chaotic system to prevent the particle from beginning the local minimum search, improve its local search ability, save the search time and thus make it find the optimal position faster. And this, in turn, can improve the local optimization ability and convergence. And all this has also been proven through the related experiment. In addition, the convergence of the four algorithms is also tested in this essay by optimizing the three functions when they are in their three-dimensional conditions. The result finds that among the four algorithms, QPSO has the highest speed to reach the global optimum.

Figure 1~2 show the simulation results of the one experiment. The two figures show the change in the average fitness curve and global optimal cure of the particle population when the number of particle changes while the iteration time remains the same. Due to the fewness of the particle number and their sparse distribution, the fluctuation of the curve in Figure1 is larger than

Figure2. In these two figures, there will appear an upward fluctuation when the iteration times reach 50~200 interval. That is because GLQPSO has increased the diversity of the population and enhanced the trial period of the particle. In addition, by using the chaotic sequence to initialize the population and the chaotic disturbance to improve the optimization ability, too early local optimization process can be avoided and the search efficiency can be greatly raised.

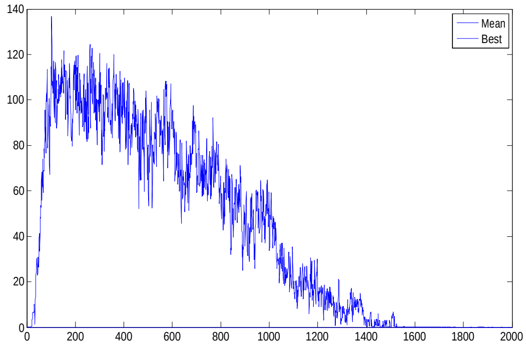


Figure1. the simulation result of the GLQPSO algorithm in the two-dimensional space with its iteration times being 2000 and the number of particle being 30 (Rastrigin Function)

Figure 3~8 show the simulation results of 50 experiments. Fig3, 4 and 5 shows the convergence curves of the three functions in their ten- dimensional conditions. They reveal that GLQPSO has the fastest convergence speed. Fig 6~8 show the same type of curves when these functions are in their three-dimensional conditions. The weakness of GLQPSO can be found clearly in these curves, which means that despite its great convergence in the high dimension, this ability will be weakened in the low dimension.

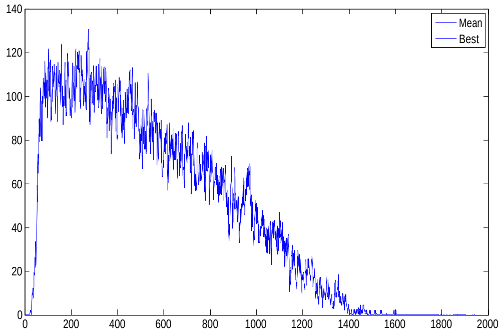


Figure2. the simulation result of the GLQPSO algorithm in the two-dimensional space with its iteration times being 2000 and the number of particle being 45 (Rastrigin Function)

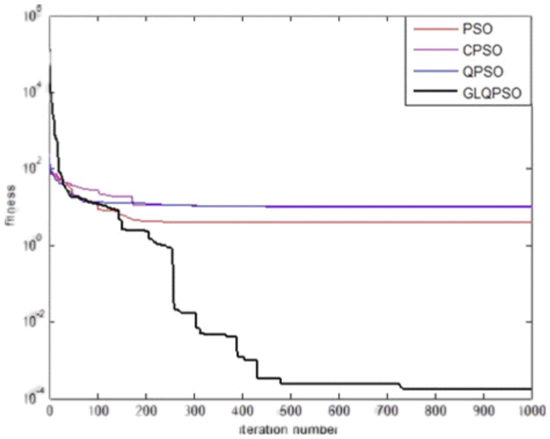


Figure3. the simulation results of the four algorithms in the ten-dimensional space with its iteration times being 1000 and the number of particle being 30 (Rosenbrock Function)

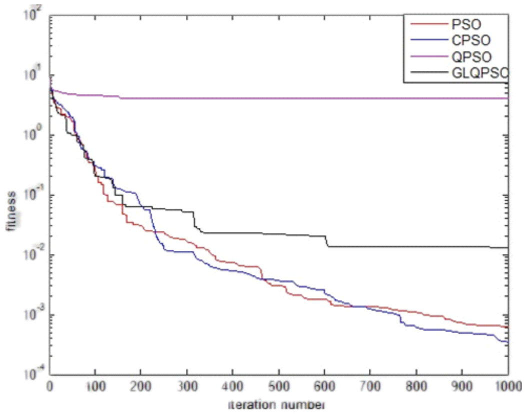


Figure4. the simulation results of the four algorithms in the ten-dimensional space with its iteration times being 1000 and the number of particle being 30 (Ackley Function)

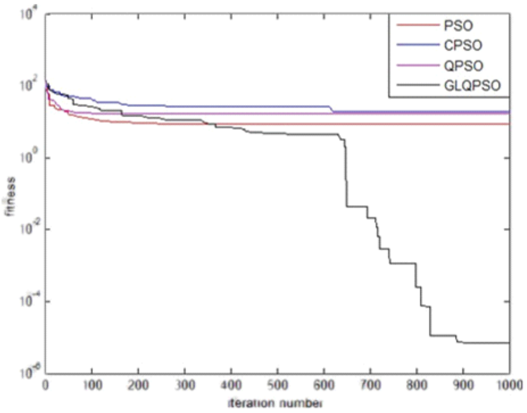


Figure5. the simulation results of the four algorithms in the ten-dimensional space with its iteration times being 1000 and the number of particle being 30 (Rastrigin Function)

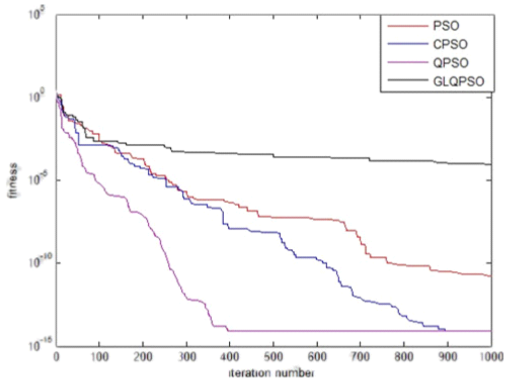


Figure6. the simulation results of the four algorithms in the three-dimensional space with its iteration times being 1000 and the number of particle being 30 (Ackley Function)

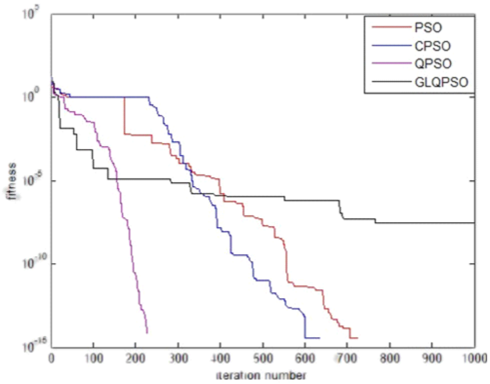


Figure7. the simulation results of the four algorithms in the three-dimensional space with its iteration times being 1000 and the number of particle being 30 (Rastrigin Function)

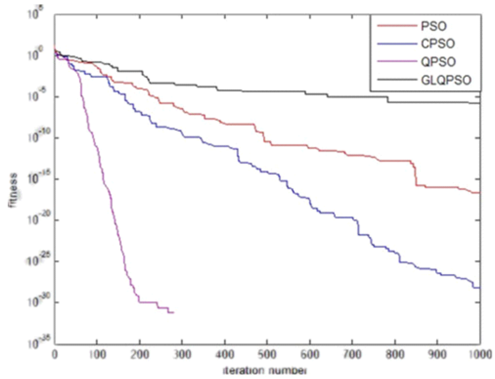


Figure8. the simulation results of the four algorithms in the three-dimensional space with its iteration times being 1000 and the number of particle being 30 (Rosenbrock Function)

Ⅷ. Conclusion

This paper proposes a new PSO algorithm based on the improved quantum behavior---GLQPSO algorithm by using identical particle system to update the particle position and the chaotic theory to conduct chaotic disturbance and initialization on each particle. The test shows that compared with the classical PSO, CPSO and QPSO, the new algorithm greatly improves the local search capacity and convergence speed of the particle swarm. The author hopes that this algorithm can be further perfected in the future application.

ACKNOWLEDGEMENTS

This work was financially supported by the two Shanxi Province Natural Science Foundations (2012011011-3, 2013011017-7) and the Shanxi University Innovation Project.

References

[1] HAN Feng-ying. Image Encryption Algorithm Based on Improved Logistic Chaotic System. Journal of Central South University of Forestry & Technology. Vol.28, No.1, 153-157, 2008.

[2] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of the 1995 IEEE International Conference on Neural Networks. vol. 4, pp.1942–1948, 1995.

[3] P. N. Suganthan. Particle swarm optimiser with neighbourhood operator. Proceedings of the IEEE Congress of Evolutionary Computation. Vol. 3, pp. 1958 -1961,1999.

[4] Anula Khare, Saroj Rangnekar. A review of particle swarm optimization and its applications in Solar Photovoltaic system. Applied Soft Computing. Vol.13, No. 5, pp. 2997-3006, May 2013.

[5] M.M. Ali, P. Kealo. Improved particle swarm algorithms for global optimization. Applied Mathematics and Computation. Vol. 196, Issue 2, pp. 578-593, 2008.

[6] Dun-wei Gong, Jian-hua Zhang, Yong Zhang. ulti-objective Particle Swarm Optimization for Robot Path Planning in Environment with Danger Sources. JOURNAL OF COMPUTERS. Vol.6, No. 8, pp.1554 -1561, 2011.

[7] Shifei Ding, Junzhao Yu, Huajuan Huang, Han Zhao. Twin Support Vector Machines Based on Particle Swarm Optimization. JOURNAL OF COMPUTERS. Vol.8, No. 9, pp. 2296-2303, 2013.

[8] Ning Lu, Ying Liu. On Predicted Research Methods of Supply Capacity of Micro-grid Based on Improved Particle Swarm Optimization. JOURNAL OF COMPUTERS. Vol.8, No. 10, pp.2706-2710, 2013.

[9] Wei-Bing Liu, Xian-Jia Wang. An evolutionary game based particle swarm optimization algorithm. Journal of Computational and Applied Mathematics. Vol. 214, Issue 1, pp. 30-35, 2008.

[10] GAO Shang, YANG Jing-yu. Swarm Intelligence Algorithm and Application. BEIJING: China Water Power Press.pp.2-8, 2006.

[11] Anula Khare, Saroj Rangnekar. A review of particle swarm optimization and its applications in Solar Photovoltaic system. Applied Soft Computing. Vol. 13, Issue 5, pp. 2997-3006, May 2013.

[12]J.Kennedy, R.C.Eberhart. Swarm Intelligence. Morgan Kaufmann Publishers. 2001.

[13] Zhongping Wan, Guangmin Wang, Bin Sun. A hybrid intelligent algorithm by combining particle swarm optimization with chaos searching technique for solving nonlinear bilevel programming problems. Swarm and Evolutionary Computation. Vol. 8, pp. 26- 32,2013.

[14] A.A.Mousa, M.A.El-Shorbagy, W.F.Abd-El-Wahed. Local search based hybrid particle swarm optimization algorithm for multiobjective optimization. Swarm and Evolutionary Compution. Vol. 3, pp. 1-14, 2012.

[15] Weiguo Zhao. BP Neural Network based on PSO Algorithm for Temperature Characteristics of Gas Nanosensor. Journal of Computers, Vol.7, No.9, 2318-2323,2012.

[16] Yuxin Zhao, Wei Zu, Haitao Zeng, “Amodified particle swarm optimization via particle visual modeling analysis”. Computers & Mathematics with Applications, Vol.57, pp.2022 - 2029, 2009.

[17] Yu Wang, Bin Li, Thomas Weise, Jianyu Wang, Bo Yuan, Qiongjie Tian. Self-adaptive learning based particle swarm optimization. Information Sciences, Vol. 181, Issue 20, pp.4515-4538, October 2011.

[18] Y. Shi, R.C. Eberhart. A modified particle swarm optimizer. in: Proceeding in IEEE Congress on Evolutionary Computation (CEC), pp.69 – 73, 1998.

[19] R. Poli, J. Kennedy, T. Blackwell. Particle swarm optimization: an overview, Swarm Intelligence.Vol.1, No.1,pp. 33–57,2007.

[20] J.F. Schutte, A.A. Groenwold, A study of global optimization using particle swarms. Journal of Global Optimization. Vol. 31, pp. 93–108,2005.

[21] Sun J, Feng B, Xu WB. Particle swarm optimization with particles having quantum behavior. Proceedings of Congress on Evolutionary Computation. Piscataway: IEEE Press, pp.325-331, 2004

[22] SHEN Jia-ning, XU Wen-bo, SUN Jun. Analyzing Convergent Ability of QPSO Algorithm. MICROCOMPUTER INFORMATION. Vol 25, No.2 -3, pp:218-219,2009.

[23] Shifei Ding, Fulin Wu, Ru Nie, Junzhao Yu, Huajuan Huang. Twin Support Vector Machines Based on Quantum Particle Swarm Optimization. JOURNAL OF SOFTWARE. VOL.8, NO.7,pp. 1743-1750, JULY 2013.

[24] Leandro dos Santos Coelho. A Quantum Particle Swarm Optimizer with Chaotic Mutation Operator. Chaos, Solitons and Fractals. Vol.37, pp.1409-1418, 2008.

[25] KANG Yan, SUN Jun, XU Wenbo. Parameter selection of quantum-behaved particle swarm optimization. Computer Engineering and Applications. Vol. 43, No.23, pp. 40-42,2007

[26] Zeng jinyan. Introduction to Quantum Mechanics (Second Edition). Beijing: Peking University Press. pp.135-162,2006.

[27] Jin Yan-xia, HAN Xie, ZHOU Han-chang. Improved particle swarm optimization algorithm. Computer Engineering and Design.Vol.30, No.17,pp.4074-4076, 2009

[28] SUN Jun. Quantum-behaved particle swarm optimization algorithm. Doctoral Dissertation of Jiangnan University. pp.48 – 49, 2009

My name is Yan-xia JIN. I am 40 years old, born in shanxi province in china. I am Associate professor. I got my bachelor degree from NORTH UNIVERSITY OF CHINA in 1997.I am master instructor. My major is computer. In 2004, i got master degree from NORTH UNIVERSITY OF CHINA in shanxi province. My major is measuring and testing technology and instruments. In 2010, i got doctor from NORTH UNIVERSITY OF CHINA. My major is measuring and testing technology and instruments too. My major field of study should be virtual reality, image processing and optimization theory. (E-mail:jinyanxia_730128@163.com)

The past five years, i have main results: 10 papers, including one by SCI, four by EI, two textbooks, a book, a national invention patent.

The second author is Jing XUE. He is a graduate.

The third author is Zhi-bin SHI. She is Associate professor. She has main results: 10 papers, including three by EI, one textbook.

Haptic Data Compression Based on a Linear Prediction Model and Quadratic Curve Reconstruction

Fenghua Guo
School of Computer Science and Technology, Shandong University, Jinan, China
Email: tuxiang201304@163.com

Caiming Zhang
School of Computer Science and Technology, Shandong University, Jinan, China
School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China
Email: czhang@sdu.edu.cn

Yan He
School of Computer Science and Technology, Shandong University, Jinan, China
Email: hy_1986503@163.com

***Abstract*—In this paper, a new haptic data compression algorithm is presented. The algorithm partitions haptic data samples into subsets based on knowledge from human haptic perception. To reduce the number of data subsets, a prediction model based on a tangential direction concept is derived that adapts to local geometric changes of haptic signals. Furthermore, to improve signal approximation precision, each haptic data subset is fitted by a quadratic curve. Accordingly, only the coefficients of the quadratic curves are encoded. Experiments are performed on datasets acquired using a six-degrees-of-freedom haptic-enabled telepresence system. The experimental results demonstrate that the proposed haptic data compression algorithm can potentially outperform existed methods in the literature.**

***Index Terms*—haptics, linear prediction, compression, curve reconstruction**

I. Introduction

Recently, haptic technology has been recognized as being compelling to further augment human-to-human and human-to-machine interaction [1]. Many haptic applications involve the transmission, storage, and management of haptic data, which depict trajectory, cutaneous, and kinesthetic information (i.e., force, torque, position, orientation, velocity, etc). The multidimensional, high-frequency, and data intensive nature of haptic media motivates the research and development of both online and offline haptic data compression algorithms. In particular, online compression (e.g., haptic enabled telepresence [2–4]) is restricted by strict delay constraints in order to guarantee control loop stability. Moreover, in online compression, every set of sample data is transmitted in individual packets to limit packetization and transmission delays. In contrast, offline compression algorithms can process blocks of haptic samples as stability (due to delay) is not a concern. Moreover, offline compression techniques primarily address file size reduction. Examples of applications in which haptic data compression is of great importance include: (1) In telehaptic environments (i.e., online compression), it is highly desirable to use compression techniques to reduce haptic data traffic and improve system performance while maintaining a high-quality telehaptic user experience (e.g., real-time remote haptic collaboration [31]); (2) The compression of voluminous haptic data files typically produced during a haptic session (i.e., offline compression). For example, a haptic training session where the user learns handwriting can be stored to be later analyzed, or even replayed [32].

Haptic data compression is a fairly new research area that has attracted much interest in recent years. The techniques currently available in the literature can be classified into three distinct categories: (1) lossless compression; (2) lossy compression without a model of human haptic perception; and (3) perceptual compression (lossy compression with a model of human haptic perception). Methods that fall in the third category exploit the limitations of human haptic perception to efficiently and transparently compress haptic data in online and offline conditions (e.g., in haptic playback or telepresence systems). All three categories of methods in the literature have been exploited for both online and offline haptic data compression [2,3,5–14].The related research work discussed here is focused primarily on offline haptic data compression methods as this is the primary application of the proposed algorithm. In [5,6], low-delay compression methods were introduced that exploit Differential Pulse

Manuscript received October 22, 2013; revised May 12, 2014; accepted May 24, 2014.

doi:10.4304/jsw.9.11.2796-2803

Code Modulation (DPCM) and quantization methods, together with Huffman entropy coding. Similarly, in [7] a method is proposed that combines adaptive sampling and adaptive DPCM for the reduction of haptic data samples. In [8], a data reduction technique based on lossy uniform and nonuniform quantization of first-order differences is used. This is performed to explore the potential data reduction rate that may be achieved before compression-induced artifacts are haptically perceived. In more recent work[9],Sakr *et al.* presented a linear predictor that relies on an autoregressive model to predict haptic data; the method's application in an offline haptic data compression scheme is discussed in [10]. In [15], an offline coding technique for haptic data is introduced. It is intended primarily for the compression of haptic data files used in haptic playback applications. Similar to [10], it is a compression algorithm that relies on the concept of the Just Noticeable Differences as well as predictive and entropy coding modules. In particular, the method encodes strictly perceptually significant haptic data samples using a minimum number of bits. For a detailed review of recent literature on haptic data compression refer to [1,28].

In this paper, a new haptic data reduction algorithm is proposed. The algorithm relies on a basic concept that exploits tangential directions at different haptic data points to construct a high-accuracy linear predictor. The linear predictor is used to partition haptic data samples into subsets, while relying on knowledge from human haptic perception. Moreover, in order to further improve approximation precision, each haptic data subset is fitted by a quadratic curve. Accordingly, only the coefficients of the quadratic curves are encoded rather than the original haptic data samples.

The rest of the paper is organized as follows. In Section II the proposed haptic data compression technique is presented. Section III discusses the experimental settings. Section IV presents the experimental results. Finally, conclusive remarks are outlined in Section V.

II. Haptic Data Compression

In this section, the proposed haptic data compression method is presented. The objective is to enable a high data reduction performance and approximation precision, while preserving a high-quality haptic experience during playback of the compressed haptic data streams. The suggested data compression strategy relies on linear prediction combined with quadratic curve reconstruction. Knowledge from human haptic perception is incorporated into the architecture to assess the perceptual quality of the compressed haptic signals.

A. Haptic Perceptibility

The suggested data compression method relies on the limitations of human haptic perception. In particular, human haptic perception is analyzed using Weber's law of Just Noticeable Differences (JND). The JND consists of the minimum amount of change in stimulus intensity which results in a noticeable variation in sensory experience. This relation can be expressed as

$$\Delta I / I = k \, , \tag{1}$$

where I is the stimulus intensity, ΔI is the so-called difference threshold or the JND and k is a constant called the Weber fraction. Generally, the JND for human haptic perception ranges from 5% to 15% [16–18]. This suggests that, if a change in haptic force (or movement) magnitude is less than the JND, the user would not perceive a force-feedback (or movement variation). Weber's law defines a very simple mathematical model to characterise human haptic perception. It essentially provides an approximate model that allows the detection of perceptible changes in haptic signals. In the haptic data compression literature, a haptic perception threshold is often referred to as a *deadband* [2]. Generally, the deadband principle states that haptic signal changes (e.g., due to prediction) do not need to be stored or transmitted, unless they exceed a certain perceptual threshold.

The proposed algorithm relies on a general formulation of the deadband principle intended for multiple multidimensional haptic data types (e.g., force, torque) [1,2,4]. This is due to the fact that the algorithm in Section IV will be evaluated using 6-DoF haptic datasets which consist of force feedback (force/torque) information. It should be emphasized that with minor or no modifications, the algorithm can be easily applied to haptic datasets acquired from devices with fewer or more degrees-of-freedom.

To determine which haptic force-feedback data samples should be encoded, human haptic perceptual limitations with respect to the exerted force and torque must be considered together as follows:

If $(D_{\mathrm{F}}(\boldsymbol{F}_{\mathrm{pred}}, \boldsymbol{F}_{\mathrm{real}}) > f(\boldsymbol{F}_{\mathrm{real}})$ or $D_{\mathrm{T}}(\boldsymbol{T}_{\mathrm{pred}}, \boldsymbol{T}_{\mathrm{real}}) > f(\boldsymbol{T}_{\mathrm{real}})$
Then $\boldsymbol{F}_{\mathrm{real}}, \boldsymbol{T}_{\mathrm{real}}$ must be encoded, can not be predicted
Else $\boldsymbol{F}_{\mathrm{real}}, \boldsymbol{T}_{\mathrm{real}}$ can be predicted.

where $\boldsymbol{F}_{\mathrm{pred}}$ and $\boldsymbol{F}_{\mathrm{real}}$ denote the predicted and actual force vectors, whereas $\boldsymbol{T}_{\mathrm{pred}}$ and $\boldsymbol{T}_{\mathrm{real}}$ correspond to the predicted and actual torque vectors, $D_{\mathrm{F}}(\cdot)$ and $D_{\mathrm{T}}(\cdot)$ consist of distance metrics used to measure the proximity between force and torque vectors respectively, whereas the functions $f(\boldsymbol{F}_{\mathrm{real}})$ and $f(\boldsymbol{T}_{\mathrm{real}})$ define the human perceptual thresholds for different force and torque values. More specifically, the method relies on a force deadband d_{F} and torque deadband d_{T}. Also, $f(\boldsymbol{h}) = d_{\mathrm{h}} \cdot |\boldsymbol{h}|$, and $\boldsymbol{h} \in [\boldsymbol{F}, \boldsymbol{T}]$. It should be observed that $f(\boldsymbol{h})$ is defined using Weber's law.

B. Distance Measurement

In order to measure the proximity of two haptic vectors $\boldsymbol{V}_{\mathrm{pred}}$ and $\boldsymbol{V}_{\mathrm{real}}$ ($\boldsymbol{V} \in [\boldsymbol{F}, \boldsymbol{T}]$), different distance measurements are considered. The Euclidean distance is very commonly used in the haptic compression literature [1,2,4,19]. In the proposed haptic data compression method, the orthogonal distance is exploited when evaluating prediction errors. The orthogonal distance consists of the smallest Euclidean distance among all distances between a haptic vector $\boldsymbol{V}_{\text{real}}$ ($\boldsymbol{V} \in [\boldsymbol{F}, \boldsymbol{T}]$) and the line which represents the output of the linear predictor. Fig. 1 provides a visual depiction of the difference between the Euclidean distance and the orthogonal distance. Specifically, point $\boldsymbol{V}_{\text{real}}$ denotes the original haptic data sample, line L is the output of the linear predictor, whereas point $\boldsymbol{V}_{\text{predict}}$ (equivalent to $\boldsymbol{V}_{\text{pred}}$) is its predicted point (that falls on line L). Generally, the nonlinear nature of a haptic signal makes it difficult to predict (with high accuracy) using a linear prediction model. Moreover, the Euclidean distance $|\boldsymbol{V}_{\text{pred}} - \boldsymbol{V}_{\text{real}}|$ does not represent the true minimal distance between $\boldsymbol{V}_{\text{real}}$ and the potential predictor output. In fact, the minimal distance corresponds to the orthogonal distance from $\boldsymbol{V}_{\text{real}}$ to the prediction line L (point $\boldsymbol{V}_{\text{min}}$). Consequently, using the orthogonal distance approach, haptic data reduction performance improvement should be expected.

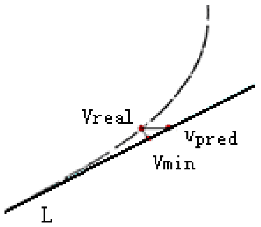
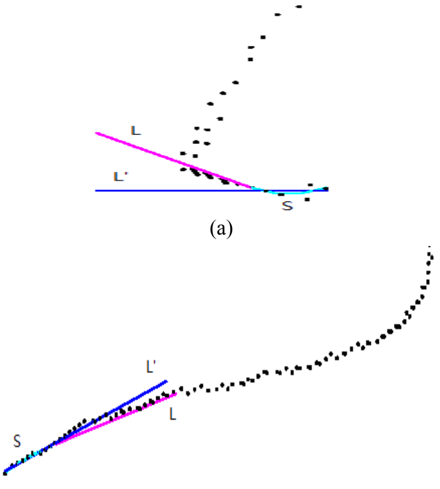


Figure 1. A visual illustration of the difference between the Euclidean and Orthogonal distances (adapted from [14]).

C. Tangential Direction-based Linear Prediction

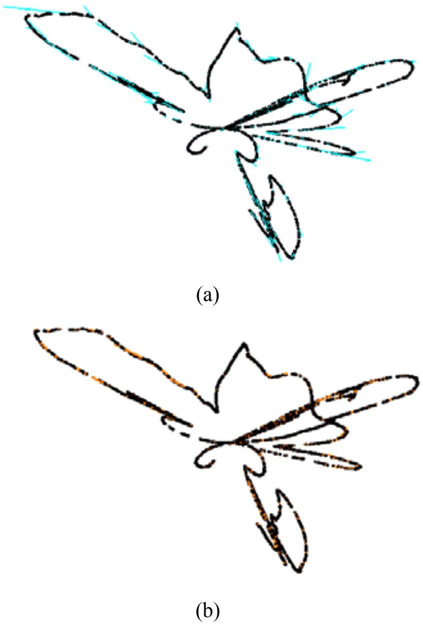
Several haptic prediction techniques have already been exploited in the haptic data compression literature. In [2,15,20], prediction is performed using a simple linear extrapolation procedure, which solely relies on two previously received sample values. In [21], Clarke *et al.* presented a haptic data prediction method based on a double exponential smoothing approach which essentially models a time series using a basic linear regression equation. In [9], Sakr *et al.* presented a linear predictor that relies on an autoregressive model to predict haptic data; a small number of the initial data is normally required in order to initiate the prediction process. In more recent work [3,4,11,19], a haptic prediction model is introduced that relies on the least-squares estimation method (referred to in this paper as LSE-LP). The authors evaluated their algorithm in both, offline and networked haptic applications. In this paper, a linear prediction model based on curve reconstruction and a tangential direction concept is presented. The details of the algorithm are as follows.

Curve reconstruction has been widely studied in computer graphics and geometric modeling in the past decade and it has various applications in CAD/CAM, computer vision, and many other disciplines [22,23,33, 34]. Quadratic curves and surfaces own a lot of elegant properties which make them a powerful tool for shape modeling [24–26]. In this paper, quadratic curves are used to derive the haptic linear predictor and to improve approximation precision. Specifically, based on M data samples $\boldsymbol{V}_0, \boldsymbol{V}_1, ..., \boldsymbol{V}_{M\text{-}1}$ ($\boldsymbol{V} \in [\boldsymbol{F}, \boldsymbol{T}]$), a short smooth quadratic curve segment



(b)

Figure 2. Two examples illustrating the differences between the least square estimation-based linear prediction (LSE-LP) approach and linear prediction based on the tangential directions concept (adapted from [14]).



(b)

Figure 3. Plots showing the difference in prediction performance between the Least Square Estimation-based linear predictor (LSE-LP) method (a), and the tangential directions-based prediction model (b), when force data are considered.

$$\boldsymbol{S}(u) = \boldsymbol{d}_0 + \boldsymbol{d}_1 u + \boldsymbol{d}_2 u^2 \tag{2}$$

is initially fitted on the data, where $\boldsymbol{d}_0, \boldsymbol{d}_1, \boldsymbol{d}_2$ are the coefficients and $u \in [0,1]$ is the parameter. The linear predictor is defined as follows

$$\boldsymbol{Q}(t) = \boldsymbol{b}_0 + \boldsymbol{b}_1 t \tag{3}$$

where $\boldsymbol{b}_0 = \boldsymbol{V}_{M\text{-}1}$, $\boldsymbol{b}_1 = \boldsymbol{S}'(u=1)$. As illustrated in Fig. 2, haptic data samples are first fitted with a curve segment S . Regardless of whether S has a significant curvature (Fig. 2(a)) or a relatively small curvature (Fig. 2(b)), the linear predictor (L) based on this tangential direction concept will follow the local geometry of the point cloud [27]. Compared with the LSE-LP method ($L^{'}$, which is directly fitted with the same M data samples $\boldsymbol{V}_0, \boldsymbol{V}_1, ..., \boldsymbol{V}_{M\text{-}1}$), the linear predictor based on tangential directions can be expected to predict more accurately. For example, a visual comparison between the linear predictor based on the tangential direction concept and the Least Square Estimation based linear predictor (LSE-LP, as aforementioned this is a popular prediction method used in numerous recent papers in the offline and online haptic data compression literature [3,4,11,19]) is shown in Fig.3. Both algorithms are evaluated in off-line settings. It can be clearly seen that with the same force perception threshold (tolerable signal distortion threshold), the proposed linear predictor (based on the tangential direction concept, the orthogonal distance and quadratic curve reconstruction) outperforms the LSE-LP/Euclidean distance approach. A more detailed analysis of the proposed data reduction algorithm will be provided in Section IV.

D. The Proposed Algorithm

A detailed description of the proposed algorithm is as follows. First, the algorithm constructs a linear predictor. The predictor is then used to divide a haptic signal into subsets, i.e., haptic data samples in the same subset are those that can be predicted within a tolerable perceptual error. Subsequently, samples in each subset are fitted by a quadratic curve. The procedure of the proposed method is divided into the following steps. Given the predefined perception thresholds (deadbands for force and torque) and a haptic dataset, repeat the following three steps until there are no more haptic data samples left (these steps are repeated for each subset).

1- Given the ordered haptic data samples set, the algorithm uses the techniques presented in Section C to construct a quadratic curve based on the initial M data points $\boldsymbol{V}_0, \boldsymbol{V}_1, ..., \boldsymbol{V}_{M\text{-}1}$ ($\boldsymbol{V} \in [\boldsymbol{F}, \boldsymbol{T}]$) , and subsequently compute the linear predictors L_V based on the tangential directions of the corresponding curve segments. Each predictor L_V is represented by a parametric line $\boldsymbol{W}(u) = \boldsymbol{c}_0 + \boldsymbol{c}_1 u$, where $\boldsymbol{c}_0 {}_{\cdot} \boldsymbol{c}_1$ are the coefficients of the line and u is a parameter that denotes the index of each sample in the sequence.

2- For the successive sample $\boldsymbol{F}_{\text{real}}$ and $\boldsymbol{T}_{\text{real}}$ in the haptic dataset do the following:

If (the distance from $\boldsymbol{F}_{\text{real}}$ to line $L_{\mathbf{F}}$ [linear predictor] is less than $d_{\mathbf{F}} \cdot |\boldsymbol{F}_{\text{real}}|$ AND the distance from $\boldsymbol{T}_{\text{real}}$ to line $L_{\mathbf{T}}$ is less than $d_{\mathbf{T}} \cdot |\boldsymbol{T}_{\text{real}}|$;

Then $\boldsymbol{F}_{\text{real}}$ and $\boldsymbol{T}_{\text{real}}$ can be approximated or predicted by the linear equations $L_{\mathbf{F}}$ and $L_{\mathbf{T}}$ respectively, go to step 2;

Else $\boldsymbol{F}_{\text{real}}$ and $\boldsymbol{T}_{\text{real}}$ cannot be approximated [or predicted] by the linear equations $L_{\mathbf{F}}$ and $L_{\mathbf{T}}$, $\boldsymbol{F}_{\text{real}}$ and $\boldsymbol{T}_{\text{real}}$ will be the first points of the next point subset.

3- All the data samples in the subset which can be approximated by a linear equation (i.e., the predictor) $L_{\mathbf{V}}$ (where $\boldsymbol{V} \in [\boldsymbol{F}, \boldsymbol{T}]$) are fitted by a new parametric quadratic curve $\boldsymbol{S}(u) = \boldsymbol{a}_0 + \boldsymbol{a}_1 u + ... + \boldsymbol{a}_n u^n$ where $\boldsymbol{a}_0, \boldsymbol{a}_1, ...\boldsymbol{a}_n$ are the coefficients of the curve ($n = 2$), u is a parameter that denotes the index of each sample in the subset sequence. Then store (or transmit) only $\boldsymbol{a}_0, \boldsymbol{a}_1, ...\boldsymbol{a}_n$ and the number of data samples in the subset. Accordingly, the haptic subset can be reconstructed by simply using the curve coefficients $\boldsymbol{a}_0, \boldsymbol{a}_1, ...\boldsymbol{a}_n$ and the number of data samples in the subset.

In order to obtain a precise quadratic parametric curve, the least-squares method is used. A quadratic parametric curve is computed in the suggested formulation and it can be expressed as follows:

$$\boldsymbol{S}(u) = \sum\nolimits_{i=0}^{2} \boldsymbol{a}_i u^i \ , \tag{4}$$

where $\boldsymbol{a}_i$, $i \in [0,2]$ are the coefficients. The quadratic parametric curve that would best fit the original haptic data samples $\boldsymbol{S} = [\boldsymbol{S}_0, \boldsymbol{S}_1, ..., \boldsymbol{S}_{l-1}]^T$ where $\boldsymbol{S} \in [\boldsymbol{F}, \boldsymbol{T}]$ in the least square sense is derived. The equation is then formulated as

$$\boldsymbol{\phi A} = \boldsymbol{S} \ , \tag{5}$$

where,

$$\boldsymbol{\phi} = \begin{bmatrix} 1 & u_0 & \cdots & u_0^n \\ 1 & u_1 & \cdots & u_1^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & u_{l-1} & \cdots & u_{l-1}^n \end{bmatrix}, \ \boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_0 \\ \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}, \ \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_0 \\ \boldsymbol{S}_1 \\ \vdots \\ \boldsymbol{S}_{l-1} \end{bmatrix} \ ,$$

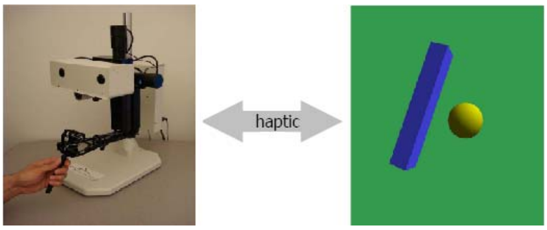
$\boldsymbol{A}$ is the unknown to be solved, u_0 , u_1 ,…, u_{l-1} are the parameter corresponding to haptic data $\boldsymbol{S} = [\boldsymbol{S}_0, \boldsymbol{S}_1, ...,$. $\boldsymbol{S}_{l-1}]^T$. The solution to (5) in the least-squares sense is

$$\boldsymbol{A} = (\boldsymbol{\phi}^T \boldsymbol{\phi})^{-1} \boldsymbol{\phi}^T \boldsymbol{S} \ . \tag{6}$$

It should be emphasized that only the coefficients of the curve segments and the number of data samples in a subset are encoded. Entropy coding (e.g., Huffman, Arithmetic coding, etc.) can be applied as a subsequent step (to the coefficients of the curve segment and the number of data samples in different subsets) to further improve the compression.

III. Experimental Settings

For our evaluation, an experimental telemanipulation is used. On the operator side, an MPB high fidelity haptic device tracks the hand movements which control the teleoperator. The haptic device is a six-axis force-feedback hand controller that can generate both translational force and rotational torque (twist force). The



(a) (b)

Figure 4. Experimental setup which consists of (a) an MPB Freedom 6S haptic device used by the operator, to interact with (b) a remote virtual environment (adapted from [4]).

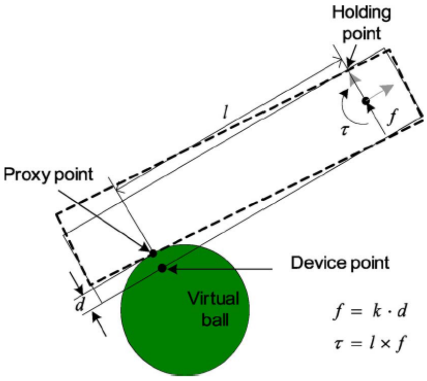


Figure 4. Force/torque feedback in 6-DoF haptic rendering (adapted from [30]).

remote environment is a virtual environment consisting of a stick-on-ball application. Fig. 4 shows a snapshot of the telemanipulation setup. As soon as the virtual 3D stick is in contact with the virtual 3D ball, corresponding force feedback is generated and displayed to the human operator. Furthermore, the force/torque computation to enable 6-DoF haptic rendering is performed while using the direct rendering method [29]. This technique is illustrated in Fig. 5, where a virtual 3D stick mimics the motion of the device stylus (device tool) in free space, but is constrained to the surface of the virtual 3D ball when a collision between the two objects is detected. Moreover, when the 3D stick is in contact with the virtual 3D ball, the deepest penetration point (device point) and its corresponding contact point located at the surface of the ball (proxy point) are first acquired. A feedback force $\boldsymbol{F(F_x, F_y, F_z)} \in R^3$ is then calculated using the Hooke's (linear) law $\boldsymbol{F} = k \bullet \boldsymbol{d}$ where $\boldsymbol{d} \in \boldsymbol{R}^3$ is the penetration vector between the two points and $k \in R$ is the stiffness constant. The force $\boldsymbol{F}$ is then used to generate a torque $\boldsymbol{T} = l \times \boldsymbol{F}$ around a predefined holding point on the virtual tool where $\boldsymbol{T} \in \boldsymbol{R}^3$, l is the distance between the predefined user holding point and the collision point between the 3D stick and the 3D ball. The resulting contact force/torque pair is then directly displayed (haptically) back to the user.

Before the conducted evaluation, we recorded 10 tele-manipulation sessions of the described telepresence setup. In order to ensure stable and realistic haptic interaction, exerted force (force/torque) data acquisition is performed at 1 kHz. Specifically, 10 haptic datasets were recorded, each consists of 6000 instances. Moreover, each instance of haptic force feedback signal encompasses 6 data samples, i.e., 3D force $\boldsymbol{(F_x, F_y, F_z)}$ and torque $\boldsymbol{(T_x, T_y, T_z)}$ data. Therefore, each dataset encompasses 6×6000= 36000 force/torque samples.

Haptic compression experiments using the proposed algorithm are performed on all 10 datasets. Force-feedback (force/torque) samples in each dataset are compressed using different deadband values. The purpose is to achieve a high compression ratio while ensuring that perceptual haptic distortions introduced by the algorithm remain relatively imperceptible (if a user chooses to decode the haptic data and play back the signals using a 6-DoF haptic device, e.g., the MPB Freedom 6S device). Accordingly, the haptic compression method was evaluated using eight different force/torque (%/%) deadband values: 0.5/0.5, 1.0/1.0, 1.5/1.5, 3.0/3.0, 5.0/5.0, 7.0/7.0, 10.0/9.0 and 15.0/12.0.

The compression algorithm runs on a Pentium(R) Dual-Core 2.20 GHz PC, with 2.00 GB of RAM and a 32-bit Operating System (Windows 7 Home Basic). The software used for the implementation is Microsoft Visual Studio C++ 2008 and OpenGL.

IV. Results

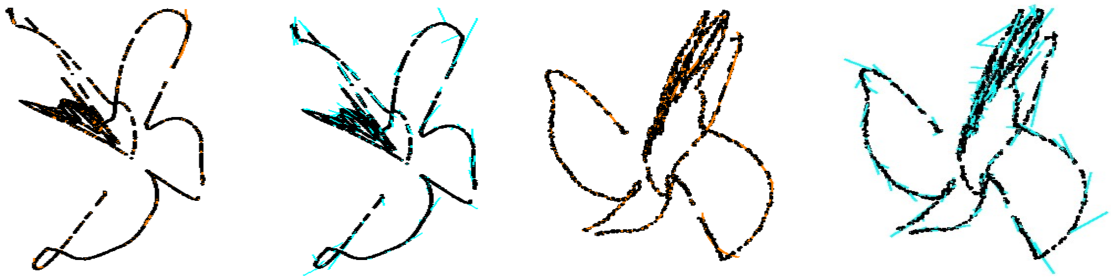
We are basing the comparison of the proposed algorithm to that of the popular Least Square Estimation-based Linear Prediction (LSE-LP) method which was proofed to deliver perceptually accepted results based on subjective evaluations. Consequently in this paper we objectively compare our results to those of LSE-LP.

Fig. 6 provides a visual performance comparison between the proposed prediction model and the LSE-LP method, using one of the 10 recorded experimental 6-DoF haptic datasets which encompass force/torque data.

Furthermore, Table I compares the haptic data reduction performance of the proposed method with the LSE-LP approach when force feedback (force/torque) data are considered. Specifically, Tables I shows the data reduction rates and the corresponding Mean Square Errors between the original and compressed signals, i.e.,

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=0}^{N-1}\left(\left\|\boldsymbol{V}(i) - \hat{\boldsymbol{V}}(i)\right\|\right)^2 \ , \ \boldsymbol{V} \in [\boldsymbol{F}, \boldsymbol{T}]$$

using the proposed method and the LSE-LP approach for different force/torque deadband values, respectively. It should be emphasized that Tables I presents the average haptic data compression results obtained upon evaluating the 10 aforementioned 6-DoF datasets. Additionally, data samples encoded (stored) using the proposed compression method consist of 3D coefficients associated with computed/fitted quadratic curves of force, torque data, and the number of data samples in the subsets. Conversely, data samples encoded using the LSE-LP method (as typically performed in the offline and online compression methods that use this approach [3,4,11,19]) consist of the original 3D force, torque data, and the



(a) Force – Proposed (b) Force – LSE-LP (c) Torque – Proposed (d) Torque – LSE-LP

Figure 5. Plots comparing the performance of the proposed prediction model with the LSE-LP method, using force/torque data contained in one of the experimental datasets (force/ torque deadbands = 15%/12%).

TABLE
COMPARISON OF THE PROPOSED COMPRESSION METHOD WITH THE LSE-LP APPROACH WHEN FORCE FEEDBACK (FORCE/TORQUE) DATA ARE CONSIDERED.

Force/Torque Deadband [%/%]	Reduction Ratio Proposed/LSE-LP	Force MSE Proposed/LSE-LP	Torque MSE Proposed/LSE-LP
0.5/0.5	41.41% / 13.91%	$5.3 \bullet 10^{-7}/6.9 \bullet 10^{-7}$	$3.9 \bullet 10^{-9}/5.2 \bullet 10^{-9}$
1.0/1.0	53.57% / 35.12%	$1.7 \bullet 10^{-6}/3.9 \bullet 10^{-6}$	$2.0 \bullet 10^{-8}/3.5 \bullet 10^{-8}$
1.5/1.5	62.01% / 48.06%	$3.2 \bullet 10^{-6}/1.0 \bullet 10^{-5}$	$4.4 \bullet 10^{-8}/1.0 \bullet 10^{-7}$
3.0/3.0	75.70% / 68.87%	$7.9 \bullet 10^{-6}/4.6 \bullet 10^{-5}$	$1.3 \bullet 10^{-7}/4.9 \bullet 10^{-7}$
5.0/5.0	83.67% / 80.15%	$1.9 \bullet 10^{-5}/1.3 \bullet 10^{-4}$	$2.7 \bullet 10^{-7}/1.4 \bullet 10^{-6}$
7.0/7.0	88.41% / 85.41%	$4.6 \bullet 10^{-5}/2.9 \bullet 10^{-4}$	$6.1 \bullet 10^{-7}/2.7 \bullet 10^{-6}$
10.0/9.0	89.74% / 88.33%	$9.4 \bullet 10^{-5}/5.5 \bullet 10^{-4}$	$9.9 \bullet 10^{-6}/4.7 \bullet 10^{-6}$
15.0/12.0	91.12% / 90.74%	$1.6 \bullet 10^{-4}/9.6 \bullet 10^{-4}$	$1.5 \bullet 10^{-6}/8.2 \bullet 10^{-6}$

lengths of predicted line segments.

From Table I and Fig. 6 (a – d), it can be observed that for different force and torque deadbands, the proposed haptic data compression strategy outperforms the LSE-LP method. For all considered deadbands, the number of encoded data samples is substantially less when the proposed data reduction algorithm is used. For example, for a relative force/torque deadband pair=3.0/3.0[1], the average data reduction rate using the proposed method is 75.70%. Conversely, the average data reduction rate using the LSE-LP method is 68.87%. Moreover, for a relative force/torque deadband=0.5/0.5, the average data reduction rate using the proposed method is 41.41%. For the same relative deadband value pair, the average data reduction rate using the LSE-LP method is 13.91%.

Furthermore, from Table I it can be seen that MSE values obtained when compression is performed using the proposed algorithm are better than those obtained when the LSE-LP method is considered.

V. CONCLUSION

In this paper, an offline haptic data reduction algorithm is proposed. The algorithm is based on a prediction model that exploits tangential directions at different haptic data points and quadratic curve reconstruction to improve haptic data reduction performance and signal approximation precision. Furthermore, the limitations of human haptic perception are considered in the method to ensure that compression artifacts are imperceptible to the user in haptic playback systems. The experimental results demonstrate that the proposed haptic data reduction strategy can potentially outperform other related methods in the literature which typically rely on a general linear prediction method.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their invaluable comments. This work was supported in part by the Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 61020106001), the National Natural Science Foundation of China (Grant No. 61202148, 61373078), Doctoral Fund of Ministry of Education of China (Grant No.20110131130004) and Natural Science Foundation of Shandong Province, China (Grant No. ZR2011FM031).

REFERENCES

[1] E. Steinbach, S. Hirche, J. Kammerl, I. Vittorias, and R. Chaudhari. Haptic data compression and communication for telepresence and teleaction. *IEEE Signal Processing Magazine,* 2011, 28(1):87–96.

[2] P. Hinterseer, S. Hirche, S. Chaudhuri, E. Steinbach, and M. Buss. Perception-based data reduction and transmission of haptic data in telepresence and teleaction systems. *IEEE Transactions on Signal Processing*, 2008,56(2):588–597.

[3] N. Sakr, J. Zhou, N.D. Georganas, and J. Zhao. Prediction based haptic data reduction and transmission in telementor-

[1] In [4], it was determined that in haptic data compression of 6-DOF data, force/torque deadbands of 3.0%/3.0% will result in little or no influence on the quality of haptic- enabled interaction.

ing systems. *IEEE Transactions on Instrumentation and Measurement*, 2009,58(5):1727–1736.

[4] N. Sakr, N. D. Georganas, and J. Zhao. Human perception-based data reduction for haptic communication in six-dof telepresence systems. *IEEE Transactions on Instrumentation and Measurement*, 2011, 60(11): 3534–3546.

[5] A. Kron, G. Schmidt, B. Petzold, M.F. Zah, P. Hinterseer, and E. Steinbach. Disposal of explosive ordnances by use of a bimanual haptic telepresence system. In *Proc. of IEEE International Conference Robotics and Automation,* 2004: 1968–1973.

[6] M.L. McLaughlin, J.P. Hespanha, and G.S. Skhatme. Lossy compression of haptic data, in *Touch in virtual environment: haptics and the design of interactive systems*. M. McLaughlin, Ed., IMSC Press, Prentice Hall, 2002.

[7] C. Shahabi, A. Ortega, and M.R. Kolahdouzan. A comparison of different haptic compression techniques. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2002: 657–660.

[8] C. W. Borst. Predictive coding for efficient host-device communication in a pneumatic force-feedback display. In *Proc. of Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, 2005:596–599.

[9] N. Sakr, N. D. Georganas, J. Zhao, and X. Shen. Motion and force prediction in haptic media. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2007: 2242–2245.

[10] N. Sakr, N. D. Georganas, J. Zhao, and X. Shen. Towards an architecture for the compression of haptic media. In *Proc. of IEEE Int. Conf. on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, 2007: 13–18.

[11] N. Sakr, J. Zhou, N. D. Georganas, J. Zhao, and X. Shen. Prediction-based haptic data reduction and compression in telementoring systems. In *Proc. of IEEE Instrumentation and Measurement Technology Conference*, 2008:1828–1832.

[12] N. Sakr, J. Zhou, N.D. Georganas, J. Zhao, and E.M Petriu. Robust perception-based data reduction and transmission in telehaptic systems. In *Proc. of World Haptics*, 2009:214–219.

[13] F. Guo, Y. He, N. Sakr, J. Zhao, and A. El Saddik. Haptic data compression based on curve reconstruction. In *Proc. of International Conference on Autonomous and intelligent systems (Springer-Verlag)*, 2011:343–354.

[14] F. Guo, Y. He, N. Sakr, J. Zhao, and A. El Saddik. Haptic data compression based on quadratic curve reconstruction and prediction. In *Proc. of ACM International Conference on Internet Multimedia Computing and Service*, 2011:193–196.

[15] J. Kammerl and E. Steinbach. Deadband-based offline-coding of haptic media. In *Proc. of ACM international conference on Multimedia*, 2008:549–558.

[16] M. Zadeh, D.Wang, and E. Kubica. Perception-based lossy haptic compression considerations for velocity-based interactions. *Multimedia Systems*, 2008,13(4):275–282.

[17] V. Nitsch, J. Kammerl, B. Faerber, and E.Steinbach. On the impact of haptic data reduction and feedback modality on quality and task performance in a telepresence and teleaction system. In *Haptics: Generating and Perceiving Tangible Sensations, Lecture Notes in Computer Science*,2010,volume 6191, pages 169–176.

[18] H. Pongrac, B. Farber, P. Hinterseer, J. Kammerl, and E. Steinbach. Limitations of human 3D force discrimination. In *Proc. Of Human-Centered Robotics Systems*, 2006:109–114.

[19] N. Sakr, N. D. Georganas, and J. Zhao. Exploring human perception-based data reduction for haptic communication in 6-dof telepresence systems. In *Proc. of IEEE International Symposium on Haptic Audio-Visual Environments and Games*, 2010:1–6.

[20] P. Hinterseer, E. Steinbach, and S. Chaudhuri. Model based data compression for 3d virtual haptic teleinteraction. In *Proc. of the IEEE International Conference on Consumer Electronics*,2006: 23–24.

[21] S. Clarke, G. Schillhuber, M. F. Zaeh, and H. Ulbrich. Telepresence across delayed networks: a combined prediction and compression approach. In *Proc. of IEEE International Workshop on Haptic Audio Visual Environments and their Applications*, 2006:171–175.

[22] A. Ardeshir Goshtasby. Grouping and parameterizing irregularly spaced points for curve fitting. *ACM Transactions on Graphics*, 2000,19(3):185–203.

[23] Z. Yang, J. Deng, and F. Chen. Fitting unorganized point clouds with active implicit b-spline curves. *The Visual Computer*, 2005,21(8-10):831–839.

[24] Y. J. Ahn. Conic approximation of planar curves. *Computer-Aided Design*, 2001,33(12):867–872.

[25] L. Piegl. Techniques for smoothing scattered data with conics sections. *Computer Industry*, 1987, 9(3):223–237.

[26] V. Pratt. Techniques for conic splines. *SIGGRAPH Computer Graphics*, 1985,19(3):151–160.

[27] Y. Liu, H. Yang, and W. Wang. Reconstructing B-spline curves from point clouds - a tangential flow approach using least squares minimization. In *Proc. of International Conference on Shape Modeling and Applications*, 2005:4–12.

[28] E. Steinbach, S. Hirche, M. Ernst, F. Brandi, R. Chaudhari, J. Kammerl, et.al. Haptic communications. In *Proc. of the IEEE, frontiers of audiovisual communications: convergences of broadband, computing and rich media*, 2012,100(4):937–956.

[29] M. A. Otaduy and M. C. Lin. *High Fidelity Haptic Rendering.* San Rafael, CA: Morgan & Claypool Publishers, 2006.

[30] J. Zhou, F. Malric, E. M. Petriu, and N. D. Georganas. Uniform hardness perception in 6-DOF haptic rendering, *IEEE Transactions on Instrumentation and Measurement*, 2011, 60(1): 214–225.

[31] Z. Jiang, Z. Gao, X. Chen and W. Sun. Remote haptic collaboration for virtual training of lumbar puncture. *Journal of Computers*,2013,8(12):3103-3110.

[32] Min Xiong, Isabelle Milleville-pennel, Cedric Dumas, and Richard Palluel-Germain. Comparing haptic and visual training method of learning Chinese handwriting with a haptic guidance. *Journal of Computers,* 2013, 8(7):1815-1820.

[33] X. Hao, L. Hao, H. Zhou, S. Jiang and Y. Li. Application of curve fitting extrapolation in measuring transient surface temperature. *Journal of Computers,* 2013, 8(8):2003-2010.

[34] L. Li, J. Liu and Y. Yang. Research and development of intelligent motor test system. *Journal of Computers,* 2012, 7(9):2192-2199.

Fenghua Guo received her BS degree and Master degree in Computer Science from Shandong University, China, in 1992 and 1995, respectively. She received her Ph.D. degree in Computer Software and Theory from Shandong University in 2007. She is an associate professor of School of Computer Science and Technology at Shandong University. From August 2010 to August 2011, she had been a visiting professor in the

Multimedia Communications Research Laboratory in the School of Information Technology and Engineering at the University of Ottawa, Canada. Her research interests include computer graphics, virtual reality and haptic communications.

Caiming Zhang is a professor and doctoral supervisor of the school of computer science and technology at the Shandong University. He is now also the dean and professor of the school of computer science and technology at the Shandong University of Finance and Economics. He received a BS and an ME in computer science from the Shandong University in 1982 and 1984, respectively, and a Dr. Eng. degree in computer science from the Tokyo Institute of Technology, Japan, in 1994. From 1997 to 2000, Dr. Zhang has held visiting position at the University of Kentucky, USA. His research interests include CAGD, CG, information visualization and medical image processing.

Yan He is a graduate student of the school of computer science and technology, Shandong University, China. She is interested in haptic communications and computer graphics.

Business Intelligence Fusion Based on Multi-agent and Complex Network

Mingliang Chen
College of management, Zhejiang University, Hangzhou, China
Email: Chenml@zju.edu.cn

Sainan Liu
College of management, Zhejiang University,
College of digital media and artistic design, Hangzhou Dianzi University, Hangzhou, China
Email: Liusn@hdu.edu.cn

***Abstract*—A fusion method of heterogeneous business intelligence (BI) technologies is put forward, named agent-network. This method treats BI system as a complex network composed of agents as its nodes. One agent is a unit of intelligence resource (IR) representing a computing model or an algorithm. A BI technology is a group of agents. Three basic mechanisms are discussed in detail. The IR aggregating and optimizing mechanisms can improve BI software to be a dynamical and flexible system. New technologies can be added into BI software continuously and less value existing technologies can be deleted from it. The IR using mechanism can always let the BI system to select an or a group of optimal technology (technologies) to respond to every specific user request by using some marketing mechanism such as negotiation, bidding and auction. The IR optimizing mechanism can keep the best agent and delete inferior agent by the performance of agents. The BI architecture is proposed based on our method. Different enterprises can customize their own BI service at a lower cost by using our method. In the future, we will develop a prototype software system based on our agent-network method to improve the decision level of enterprise.**

***Index Terms*—business intelligence, agent-network, multi-agent, complex network**

I. Introduction

At present, more and more enterprises need BI software to support their statistics, analysis, forecast and decision-making. Some big companies, such as Oracle, SAS, BO, Cognos, MS, SAS and SPSS, have developed their own BI software. But those BI softwares are developed for some special application and only used in a limited range. Currently, the existing BI software is deficient in three points. Firstly, the current BI software can only provide solution for specified situation. If the situation is beyond its range, it can’t recognize it and respond to it. Secondly, the current BI software can’t deal with the dynamical requirement of enterprise. If the requirement is changed with time, the responding capability of the BI software is weakened. Lastly, the update speed of current BI software is slow. The source code of BI software must be always rewritten when new requirement is added. So the updating cost is high. In the current market, there is a need for an universal BI software that can meet the dynamical and various requirement of heterogeneous enterprise and can fuse all kinds of heterogeneous intelligence technologies together on one BI software.

In order to develop such a BI software, we propose an fusion method for intelligence resource named agent-network based on multi-agent and complex network. This method treats BI as a complex network composed of agents as its nodes. One agent is a unit of intelligence resource which represents computing model or algorithm. Each BI technology is composed of a group of agents. The massive accumulation ability of complex network can help aggregating all the useful and new agents continuously into the system to make the system update more seamlessly and inexpensively. The optimal reorganization feature of multi-agent can help selecting the best agents to deal with the dynamical requirement of user’s service at any time. Therefore, not only all the useful intelligence resource (IR) can be aggregated in the BI system, but also the optimal agents which represent the most suitable BI technology can be selected to respond to the service request of user at any time. With our method, the BI software can be updated without modifying the source code, only adjusting the construction of agent-network by adding or deleting some agents. So our agent-network method is meaningful for providing a new solution to deal with the adaptability and compatibility of BI.

The rest of the paper is organized as follows. Section 2 discusses related research. Section 3 discusses our proposed fusion method of BI and the three mechanisms of agent-network for BI fusion. Section 4 discusses the fusion levels of our proposed method. Section 5 analyzes the feasibility of our method. Section 6 summarizes our research work and discusses directions for future work.

II. Related Work

A. *Multi-agent Oriented BI System*

Multi-agent systems as a standard communication

doi:10.4304/jsw.9.11.2804-2812

platform can interchange data and tasks [1]. The multi-agent systems have characteristics such as autonomy, reasoning, reactivity, social abilities, pro-activity, usability and adaptability [2-3].

Recently, multi-agent technology is used more and more in BI system for modeling and studying. Using multi-agent technology can make the BI system more adaptable, renewable, flexible, and extensible. The goal of intelligence fusion can be realized by multi-agent. I. Perko et al. [4] proposed a solution for multiple prediction models management and a uniform result representation by using multi-agent system and knowledge reasoning. The proposed system is adaptive, allowing the modifications and upgrading more easily and inexpensively. J. Bajo et al. [5] proposed a multi-agent system aimed at providing advanced capacities for risk management in small and medium enterprises. The agents in their system are characterized by their capacities for learning and adaptation in dynamic environments. F. Borrajo et al. [6] introduced a new business simulator named SIMBA based on web-based platform for business education and business intelligence. In SIMBA, the simulated market can be more complicated by using intelligent agents that is to assume the role of competitors. The proposed system has several key advantages in learning objectives, the development of work skills and the teaching function. K.I.K.Wang et al [7] proposed a novel ambient intelligence platform to facilitate fast integration of different control algorithms, device networks and user interfaces. The intelligence platform consists of four layers, including ubiquitous environment, middleware, multi-agent system and application layer. The multi-agent system can incorporate multiple control algorithms as agents for managing different tasks. For this, the offline control errors can be reduced greatly in comparison with single process control algorithms. The system seems to be more flexible development and future improvement. M. Janssen [8] developed a semi-cooperative architecture based on multi-agent in which human-beings or other agents can substitute agents without affecting other parts. The initial system can start with a few agents having relatively simple behavior and then be extended into a more comprehensive system. And some researches use multi-agent and other technologies to make the BI system more efficient, intelligent and automatically. A.L.Symeonidis et al. [9] proposed a method that can dynamically extract knowledge to improve agent intelligence by using data-mining and multi-agent technology together. The concept of training and retraining are described in detail. By this way, the system can be more efficient and intelligent. H.Pham [10] proposed an agent-based hypothetic agent-based model for carrying out business automation in large, distributed, and real-time business system. In their model, the agent-based components of a business organization can be created and integrated automatically into the system. They focused on controlling the agent interactions to achieve system reliability and regulate the agent visibility. Zhisong Hou et al.[11] designed a distributed intrusion detection system based on mobile agent. The dynamic adaption of the system could be implemented while false alarm rate and false negative rate would be reduced. Weidong Zhao et al.[12] designed a multi-agent middleware for mobile supply chain management, aiming to solve integration problems and achieve mobile supply chain dynamic integration. Walaa H.E. et al.[13] studied a cooperative search of autonomous agents that represent agents' coalition formation to enjoy a price discount for each of its requested service to achieve a goal. Besides the above research, multi-agent technology is used widely in dealing with all kinds of business tasks, such as trades[14], negotiation[15], bidding[16], auctions[17], supply chain[18] and warehouse management[19] for decision-making and data analysis. It seems that it's feasible to use multi-agent technology in BI system. But currently, all the existing researches are put forward to dealing with specified business tasks for one or such a kind of enterprise. And there is no such an BI software that can meet the requirement of all kinds of heterogeneous enterprises. So in this paper, we propose a method that tries to solve this problem.

B. Multi-agent and Complex Network

Multi-agent system is a distributed system based on network [20]. And especially, the internet as an important environment of multi-agent is a typical complex network[21]. The relationship between agents is a kind of complex network. [22-24] studied the statistic feature of entity in large distributed system by graphical analysis method. The result showed that the relationship of agents which denotes the entity has features of complex network, such as small-world and scale-free. J.Delgado[21] also pointed out that the topology of agents treated as complex network is more suitable than ruled-network. And at present, some existing researches develop some models and software based on multi-agent and complex network. N.Celik [25] developed an optimal workforce assignment module based on multi-agent to resolve the problem of short-term and long term tasks of alliance-based multiple organizations which forms a complex social network. The behavior of the complex social network can be predicted by using agent-based simulation. Each agent represented an individual in the organization network and had its own characteristic. M.Tran[26] developed an agent-based model to investigating the role of individual behavior and studying the complex network influence on energy innovation diffusion. M.B.Hu [27] studied the wealth distribution in different social networks. In their proposed model, they used agents to play as nodes of the complex social network and studied the agents' personal wealth to find the law of wealth distribution. All the above researches show that the agents can be treated as nodes of complex network. The agent as node has its independent functions and interacts with each other. From the complex network perspective, the relationship of agents can be described more clearly.

III. Agent-Network Method For BI

On the basis of existing researches, a new intelligence fusion method named agent-network based on multi-agent and complex network is proposed. In this method, all kinds of useful BI resources are formed to be a series of agents which are aggregated organically as nodes of complex network. The complex network acts as a container of agents. Then the optimal agents are selected according to their performance. The intelligence fusion mechanism of agent-network is in figure 1. There are three fusion mechanisms, including IR optimizing mechanism, IR using mechanism and IR aggregating mechanism. In the following, we will discuss the three mechanisms in detail.

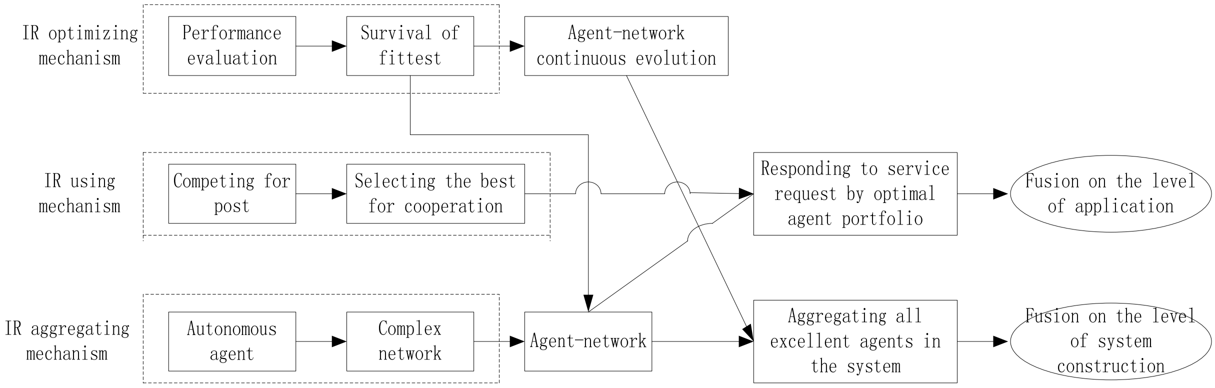

IR optimizing mechanism
Performance evaluation
Survival of fittest
Agent-network continuous evolution
IR using mechanism
Competing for post
Selecting the best for cooperation
Responding to service request by optimal agent portfolio
Fusion on the level of application
IR aggregating mechanism
Autonomous agent
Complex network
Agent-network
Aggregating all excellent agents in the system
Fusion on the level of system construction

Figure 1. Intelligence fusion of agent-network

A. IR Aggregating Mechanism Based on Autonomous Agent and Complex Network

An agent is one unit of packaged IR. The agents represent the corresponding IR. The representation of agents brings great convenience for reorganization and reuse of IR. There are various BI technologies from different fields, such as artificial intelligence, mathematics and so on. Each technology has a series of computing models or algorithms for constructing IR. Computing model is the basic unit of IR. Each model has a certain computing ability which can complete one or several complex computing tasks. Therefore, each computing model can be packaged into an agent.

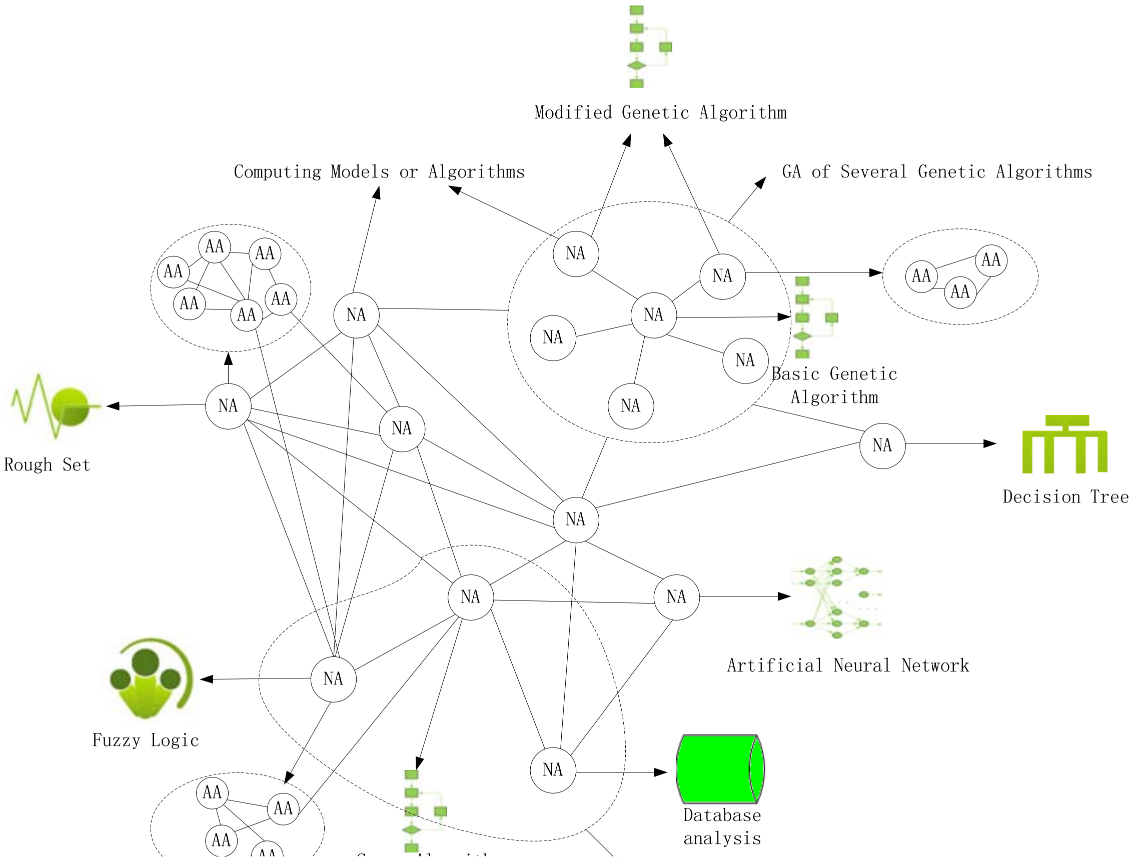
The agents can be classified into three levels: atomic agent (AA), natural agent (NA) and group agent (GA). Their relationship is in figure 2. A NA denotes a natural computing model or algorithm. A BI technology is composed of a series of mutually cooperative NA which construct a GA in agent-network. In order to reorganize and reuse agents, a natural agent can be split into some atomic agents which are the indivisible agent unit. In figure 2, we can see there are two modes of GA construction. One is that the NAs similar to each other can construct a GA. For example, some NAs representing algorithms or models based on genetic algorithm can construct a GA of genetic algorithm. In this GA, a basic genetic algorithm is the base of other modified genetic algorithms. And other NAs can cooperate with this GA, not only one agent of this GA. Another is that some cooperative NAs from different field can also work together to construct a GA for responding to the same user's service.

IR aggregating mechanism based on complex network is to aggregate the agents in the complex network. Therefore, each agent is treated as one node of the complex network. The cooperation between agents can be judged by the edge of complex network. As in figure 3, the weighted edge denotes the degree of close relationship between agents. So, a complex network of weighted agent is formed. For interactive relevance and infinite expandability of complex network, not only the existing BI resource can be aggregated, but also the new resource can be added in the system at any time. The agent at one node can be NA or AA, but not GA which is an agent sub-working net in actually. In figure 3, the initial edge weight is supposed to be 1. With the development of this agent-network after completing several tasks, the edge weight is changeable. If the edge weight is far more than 1, it denotes that the agent is active and valuable. If the edge weight is minus, it denotes that the agent is bad and can't cooperate with other agents. If the edge weight is equal to 1, it seems that the agent has no cooperative experience record with other agent. The edge weight can be between NA and AA or NA or GA.

B. IR Using Mechanism Based on "Competing for Post" and "Select the Best for Cooperation"

"Competing for post" is a service mechanism for user's dynamic requirement. A group of agents are selected by competing against other groups for the post of one service request. And one agent is a special case. How to select a group of agents to cooperate with others becomes a result of competition. The agents are not pre-designated by the system. In traditional multi-agent system, the cooperative agents are selected by the system through auto-matching or pre-designation. The cooperative relationship is rigid and lack of competition. "Competing for post" mechanism breaks up the rigid cooperation. The selection of competitive agents is the key point to realize this mechanism. So, besides remaining the two traditional modes, typically competitive mechanism in realistic society, such as negotiation, bidding and auction, are introduced into the system. The auto-matching or pre-designation can be treated as special "competing for post" mechanism when there is no other method can be chosen for agents selection or the edge weight of agents reaches a high point.

"Selecting the best for cooperation" is to select a group of optimal agents to respond to current user's service request by using “Competing for post”mechanism. Face to each service request of user, the system can flexibly choose one or several cooperative mechanisms according to current service property, service scale and status of respondent agent resource. Then, the optimal cooperation agents can be selected according to the rules of selecting the best for cooperation” mechanism. Of course, the rules and process of each mechanism are different for different situation in different complex degree. Therefore, the system must design a set of rules to guide the optimal agents’ selection at the current situation when responding to each user’s request. As in figure 4, we can use decision tree to build up the optimal selection rules to choose the optimal group of agents in different situation. The rule database can be divided into enterprise type, service type, service requirement and corresponding groups of optimal agents. Face to a service requirement, if the service requirement is not new, we can search for the group of optimal agents in the decision tree. If the service requirement is new, we can add it to our decision tree for next service selection.


Modified Genetic Algorithm
Computing Models or Algorithms
GA of Several Genetic Algorithms
AA
NA
Basic Genetic Algorithm
Rough Set
Decision Tree
Artificial Neural Network
Fuzzy Logic
Database analysis

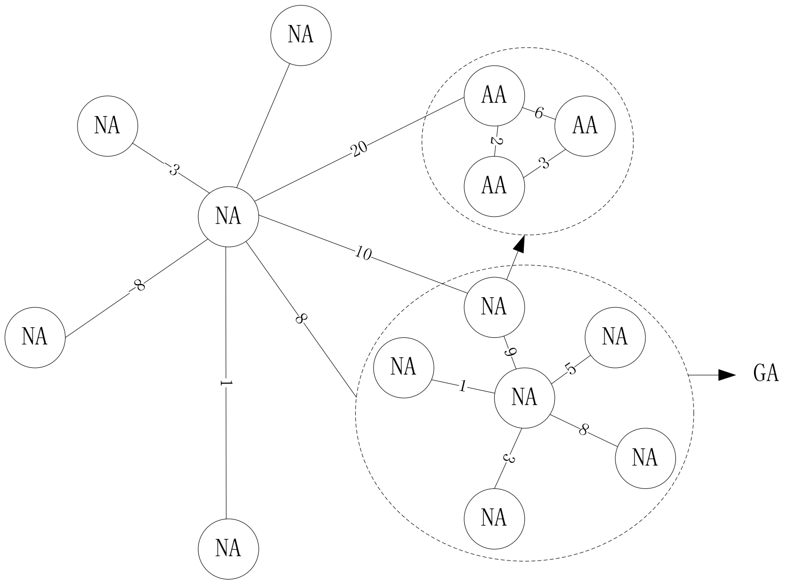


Figure 3.Weighted edge of agents

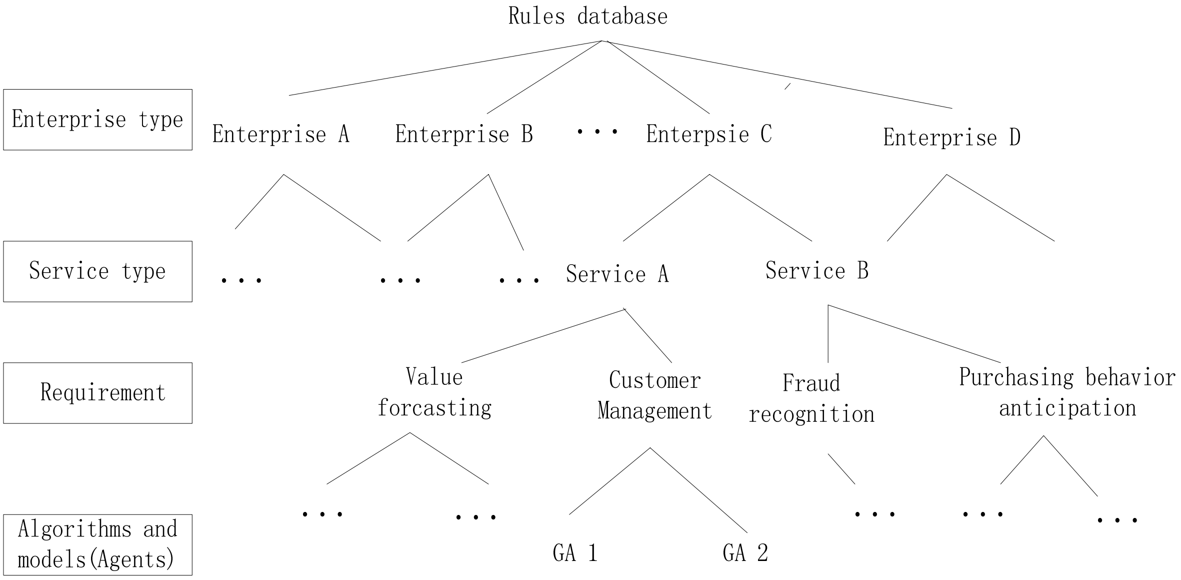


Figure 4.Decision tree for optimal group agents' selection

C. IR Optimizing Mechanism Based on Performance Evaluation and the Rule of "Survival of the Fittest"

In the agent-network, the distributed agents as nodes of complex network have independent functions and can be interactive with each other. The cooperative agents can provide intelligence services, including profitable customer relationship management, market forecasting, deal deception identification, users' monetary contribution perception and so on. Each service is a respondent for users' requirement. Performance evaluation is a mechanism to measure the service performance of agents, including real-time evaluation and periodic evaluation. Real-time evaluation is to evaluate the performance of all the agents participating in the current cooperation after the service completion. The evaluation result is the basis for updating edge weight which reflects the cooperative relationship between agents. The edge weight is bigger when the number of successful cooperation is more and the performance result is better. The cooperation opportunity is determined by edge weight which is an important judgment for periodic evaluation. Periodic evaluation is to evaluate the performance of all the agents in a long time. The change of edge weight is very important for periodic evaluation. As in figure 3, if the edge weight of one agent is not changed after a period of working time, it seems that there is no cooperation record between the agent and other agents. So the edge weight is equal to the initial edge weight which is always 1 in our proposed agent-network. It shows that the agent has no chance in cooperation with other agents. The agent is called inert agent which is similar to the worthless goods in the warehouse of enterprise. If the edge weight of one agent is minus, it shows that the agent has several times of failing cooperative experience. The agent is called inferior agent which is similar to inferior-quality goods. If the edge weight of one agent is far bigger than the initial edge weight, it shows that the agent has several times of successfully cooperation experience. The agent denotes the superior IR. The basic task of periodic evaluation is to identify inert agents and inferior agents.

"Survival of the fittest" is a mechanism like the juggle-law of nature. The inert agents and inferior agents can be eliminated from agent-network according to a certain rules. And the excellent agents can be remained in the agent-network. So the aggregated IRs in agent-network has the continuous evolution characteristic which is like ecological features of juggle-law.

IV. Agent-Network Fusion Level

As shown in figure 1, IRs can be fused organically on the level of system structure and system application by use of three agent-network mechanisms.

A. Intelligence Fusion on the Level of System Structure

Intelligence fusion on system structure means that all kinds of superior BI resource persistently can be aggregated in BI system. IR aggregating mechanism based on autonomous agent and complex network ensures that the existing and new IRs can enter in system in form of conveniently used agent and form a dynamical agent-network whose capacity is infinite. IR using mechanism based on "competing for post" and "selecting the best for cooperation" ensures eliminating inferior IRs and remaining superior IRs. Therefore, the fusion on the level of system structure -ensures that the system is the best at any time. The fusion on the level of system structure is the base of intelligence fusion on the level of system application.

B. Intelligence Fusion on the Level of System Application

Intelligence fusion on the level of system application means that the system can select the most appropriate BI resource to provide best service for users in a flexible and variable way. The fundamental purpose of intelligence fusion is to respond to user’s service request better. If there is no intelligence fusion on the level of system application, the fusion on the level of system structure is meaningless. IR using mechanism based on ”competing for post” and ”selecting the best for cooperation” ensures selecting the best agents dynamically to respond to current user’s request. In this way, the fusion on the level of system application is realized. The granularity of agent (*basic cooperation unit of agent*) in system application can be classified into three levels: GA, NA and AA. The cooperative agent portfolio is more plentiful with smaller granularity of agents. And the adaptability of agents is more widely. The three levels of agent granularity can be used cooperatively at the same time when responding to user’s service request.

V. Feasibility Analysis

In order to realize above fusion mechanisms, we must have two key technologies, including agent representation of heterogeneous BI resource and software technology of ”selecting the best for cooperation”. In the following, the feasibility of the two technologies is discussed in detail.

A. Feasibility of Agent Representation

Agent representation of BI resource is to package BI resource into a series of agents which have their characteristic. Each agent denotes a computing unit. Each BI technology becomes a group of agents after representation. Agent representation is the base of realizing intelligence fusion. There are two meanings: firstly, multidisciplinary BI resource agent can be homogenous after representation. The agent is similar to component which can be reused and used repeatedly. Secondly, the characteristic of agent, including subjectivity, intelligence, adaptability and society [28](Shi C.Y. et al.,2007), provides necessary premise for agent to take part in competition like individual and enterprise in realistic society.

Can the multidisciplinary BI resource be represented in the form of agent? The answer is yes. Many researchers propose various BI methods based on agent. For example, genetic algorithm based on multi-agent[29], production orders resolution by employing an expert system and a neural network based on multi-agent[30], a fuzzy logic controller based on multi-agent[31], data-mining for extract knowledge based on multi-agent[12]. The existing researches show that the combination of BI and agent becomes a research trend and its realization is feasible.

B. Feasibility of Competition Mechanism

Competition mechanism, such as negotiation, bidding and auction, is the key point for realizing ”competing for post” and ”selecting the best for cooperation”. The existing researches show that the Competition mechanism can be used to effectively allocate task and resource in multi-agent system[28]. For example, some researchers proposed several agent negotiation methods based on reasoning-case, consultation theory, confliction theory and sort facility[32]. A multi-agent system based on auction negotiation is used to resolve distributed multi-project scheduling[17]. A software architecture called”market-like” becomes one of the three typical architecture of multi-agent system [28]. In this architecture, negotiation, bidding and auction are realized all. And currently, in e-commence or e-market, multi-agent technologies are also used in dealing with negotiation [15] and bidding [16] to improve the efficiency of system. All the above researches show that it’s feasible in technology to build up competitive cooperation of agents by introducing negotiation, bidding and auction.

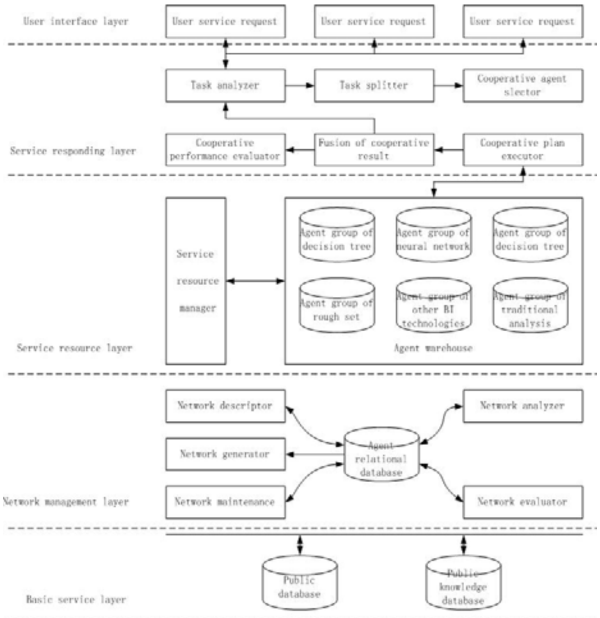


Figure 5. Software architecture of BI system based on agent-network

VI. Architecture Of Bi System Based On This Method

In this paper, we propose an architecture of BI system composed of user interface layer, service responding layer, service resource layer, network management layer and basic service layer, as figure 5 showing. Several technologies can be fused by use of this architecture.

A. User Interface Layer

There are two functions in user interface layer: request acceptation and service result return. Request acceptation provides registration, reservation and logout for users. At the same time, the service task is generated and then sent to service responding layer. Service result return provides return result display and satisfaction enquiry for users. Then the information of satisfaction is sent to service responding layer for evaluating performance of current agent combination.

B. Service Responding Layer

Each responding of user's request is a new cooperation of a group of agents in agent-network. The dispatching mechanism is to determine which agent can participate in cooperation with others according to the dynamical property of task. The service responding layer is an intelligent dispatching mechanism and a control center. All kinds of BI services are fused on system application after task completion in service responding layer. The service responding layer is composed of task analyzer, task splitter, cooperative agent selector, cooperative plan executor, fusion of cooperative result and cooperative performance evaluator. Task analyzer is to recognize the structured degree, complex degree, work load and decomposability and provide basis for selecting task decomposing strategy and task responding strategy. Task splitter is composed of task decomposing strategy set, suitable strategy condition set and selective strategy rules set. The task decomposing plan is determined by task splitter according to the suitable decomposing strategy and the decomposing granularity. Cooperative agent selector is composed of cooperative strategy set, suitable condition set, selective strategy rules set, operating process and algorithms. The optimal cooperation that can realize fusion of all kinds of BI on system application is ensured by selecting the most suitable cooperative strategy of current task and completing the final agent selection according to the corresponding process and algorithms. Cooperative plan executor is to send command to the selected agents and receive task execution result according to cooperative plan. At the same time, it's responsible for communicating and coordinating between agents. Fusion of cooperative result is to gain the final result by dealing with the return result of cooperative agents. For complex cooperative task, especially non-constructive computing task belonging to soft computing, some information fusion technologies, such as Bayesian inferences, D-S evidence theory, fuzzy set theory, expert system, artificial neural network, are used to deal with the cooperative result comprehensively. Cooperative performance evaluator is to organize related subjects users and related agent to evaluate the performance of cooperative agents, store performance result in public database, trigger network management layer to update the edge weight, and inform service resource layer to record the cooperative performance.

C. Service Resource Layer

Service resource layer is composed of agent warehouse and service resource manager. The agent warehouse is the physical place for computing resource fusion on system construction in the form of agent. Each computing resource of BI technology is packaged into a series of subjectivation agent. The agents of the same method constitute a group of close relationship agents. These agents are the entity of executing service task and the base of BI system. Each agent can be participate in cooperation as an individual or part of a agent group. The agent warehouse is dynamic open, not only including various existing BI technologies, such as rough set, fuzzy logic, decision tree, group decision, swarm algorithm, data mining, genetic algorithm, artificial neural network and other traditional statistics and analytical technologies, but also adding the new technologies into the agent warehouse at any time.

The service managers is to load, upload, import, export, and maintain agents' record which is including list maintenance, warehousing registration, cooperation registration and performance records. The service manager is also to assist network management layer to maintain agent list in the warehouse consistenting with node list of network. The agent list in the warehouse must contain node list of network, otherwise the no-existing agent would be assigned to respond to service request.

D. Network Management Layer

The network management layer is composed of network describer, network generator, network maintenance device, performance evaluator and network analyzer. Network describer provides an effective operation and management tool for users to construct the initial agent cooperative network. Users can easily describe the network composed of agents as nodes and their mutual relations, including node list of network, relational path and strength between nodes. Furthermore, the network relationship attribute database is formed for editing, adding, deleting, modifying and storing of complete network information. Network generator is to generate agent cooperative network according to data of relationship attribute. The generated network is a weighted cooperation network which has cluster structure. The weight reflects the relationship between agents. Cluster is a group of close relationship agents which denote one BI technology. Network maintenance device is to add new nodes, new edges and assign weight to new edge according to addition commands and algorithms. Network maintenance device is to delete the corresponding nodes according to deletion command and algorithms, update edge weight according to weight updating command and algorithms. Performance evaluator is to evaluate all the agents of the whole network in a long period, recognize dull agent and inferior agent according to a certain rules, inform network maintenance device to delete dull agent and inferior agent from network and inform service resource manager to clear away the dull agent and inferior agent. Network analyzer is to calculate and analyze the network statistical characteristics, such as all nodes, all edges, average degree, average path length and cluster coefficient.

E. Basic Service Layer

Basic service layer is composed of public database and public knowledge base. The data and knowledge of other layers can be stored and managed in this layer.

All kinds of saving data and information is stored in public database, including history information of network construction data of cooperative agent, history information of network evolution and performance result of cooperative agents. All kinds of knowledge which is needed by network generation and task response is stored in public knowledge base, including node evolution rules, weight evolution rules, selective rules of task decomposition strategy, selective rules of task cooperative strategy, selective rules of cooperative objects and selective rules of cooperative result fusion method.

VII. Conclusions

In this paper, we proposed a new method named agent-network for BI resource fusion. The heterogeneous technologies can be fused and aggregated together in an effective way by using our method. The three fusion mechanisms are described in detail to explain how to add new and best technologies in the agent-network to meet the dynamical requirement of enterprise service. They are IR aggregating mechanism, IR using mechanism and IR optimizing mechanism. In this way, the intelligence of BI system can be evolution and suitable for the development of enterprise. The software system architecture based on our agent-network method is discussed. The agent entering in the network can be combined optimally to respond to users' service request. The excellent agents tested in practical can be remained in the network. So the current system reflects the highest level of group intelligence. The intelligence of system is constantly evolved for the endless creativity of group.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China under grant 70971116 and Zhejiang Province Natural Science Foundation of China under grant Y6090332.

References

[1] F.Bellifemine, A.Poggi, G.Rimassa, "Developing multi-agent systems with a FIPA-compliant agent framework," Software-Practice & Experience, vol.31, pp.103–128, 2001.

[2] D.I.Tapia, J.A. Fraile, R. Sara, R.S. Alonso, et al, "Integrating hardware agents into a enhanced multi-agent architecture for ambient intelligence systems," Information Sciences, vol.222, pp. 47-65, 2013.

[3] Y.Q.Duan, V.K.Ong, M.Xu, et al, "Supporting decision making process with "ideal" software agents-What do business executives want?," Expert Systems with Applications, vol.39, pp.5534-5547, 2012.

[4] I.Perko, M.Gradisar, S.Bobek, "Evaluating probability of default: Intelligent agents in managing a multi-model system," Expert System with Applications, vol.38, pp.5336-5345, 2011.

[5] J.Bajo, M.L.Borrajo, J.F.D.Paz, et al, "A multi-agent system for web-based risk management in small and medium business," Expert System with Applications, vol.39, pp.6821-6931, 2012.

[6] F.Borrajo, Y.Bueno, I.D.Pablo, et al, "SIMBA: A simulator for business education and research," Decision Support System, vol.48, pp. 498-506, 2010.

[7] K.I.K.Wang, W.H.Abdulla, Z.Salcic, "Ambient intelligence platform using multi-agent system and mobile ubiquitous hardware," Pervasive and Mobile Computing, vol.5, pp.558-573, 2009.

[8] M.Janssen, "The architecture and business value of a semi-cooperative, agent-based supply chain management system," Electronic Commerce Research and Applications, vol.4, pp.315-328, 2005.

[9] A.L.Symeonidis, I.N.Athanasidis, P.A.Mitkas, "A retraining methodology for enhancing agent intelligence," Knowledge-Based Systems, vol.20, pp.388-396, 2007.

[10] H.Phan, Y.Ye, "An agent-based business automated system with self-adjusting visibility for reliability," Electronic Commerce Research and Application, vol.2, pp.97-113, 2003.

[11] Weidong Zhao, Haifeng Wu, Weihui Dai, et al, "Multi-agent middleware for the integration of mobile supply chain," Journal of Computers, vol.6, no.7, pp.1469-1476, 2011.

[12] Zhisong, Hou, Zhou Yu, Wei Zheng, et al, "Research on distributed intrusion detection system based on mobile agent," Journal of Computers, Journal of Computers, vol.8, no.10, pp.1919-1926,2012.

[13] Walla H. El-Ashmawi, Hu Jun, Li Renfa, "Proposed discount group formation model based on cooperative search in agent graph," vol.8, no.10, pp.2497-2506, 2013.

[14] M.Y.Cha, J.W.Lee, D.S.Lee, et al, "Wealth dynamics in world trade," Computer Physics Communications, vol.182, pp.216-218, 2011.

[15] C.B.Cheng, C.C.H.Chan, K.C.Lin, "Intelligent agents for e-marketplace: negotiation with issue trade-offs by fuzzy inference system," Decision Support System, vol.42, pp.626-638, 2006.

[16] M.Mahvi, M.M.Ardehali, "Optimal bidding strategy in a competitive electricity market based on agent-based approach and numerical sensitivity analysis," Energy, vol.36, pp.6367-6374, 2011.

[17] S.Adhau, M.L.Mittal, A.Mittal, "A multi-agent system for distributed multi-project scheduling: An auction-based negotiation approach," Engineering Applications of Artificial Intelligence, vol.25, pp.1738-1751, 2012.

[18] M.Giannakis, M.Louis, "A multi-agent based framework for supply chain risk management," Journal of Purchasing and Supply Management, vol.17, pp.23-31, 2011.

[19] J.I.U.Rubrico, T.Higashi, H.Tamura, et al, "Online rescheduling of multiple picking agents for warehouse management," Robotics and Computer-Integrated Manufacturing, vol.27, pp.62-71, 2011.

[20] Y.C.Jiang, J.C.Jiang, "A multi-agent coordination model for the variation of underlying network topology," Expert System With Application, vol.29, pp.372-382, 2005.

[21] J.Delgado, "Emergence of social conventions in complex networks," Artificial Intelligence, vol.141, pp.171-185,

2002.

[22] S.Jenkins, S.R.Kirk, "Software architecture graphs as complex networks: a novel partitioning scheme to measure stability and evolution," Information Scieneces, vol.177, pp.2587-2601, 20

[23] J.Sudeikat, W.Renz, "On complex networks in software: How agent-orientation effects software structure," CEEMAS, pp.215-224, 2007.

[24] X.L.Zhang, D.Zeng, H.Q.Li, et al, "Analyzing open-source software systems as complex networks," Physica A, vol.387, pp.6190-6200, 2008.

[25] N.Celik, S.Lee,E.Mazhari, et al, "Simulation-based workforce assignment in a multi-organizational social network for alliance-based software development," Simulation Modeling Practice and Theory, vol.19, pp.2169-2188, 2011.

[26] M.Tran, "Agent-behaviour and network influence on energy innovation diffusion," Common Nonlinear Sci Number Simulate, vol.17, pp.3682-3695, 2012.

[27] M.B.Hu, R.Jiang, Y.H.Wu, et al, "Properties of wealth distribution in multi-agent systems of a complex network," Physica A, vol.387, pp.5862-5867, 2008.

[28] Shi C.Y., Zhang W., "Calculation based on agent", Beijing: University of Tsinghua Press, 2007.

[29] L.Asadzadeh, K.Zamanifar, "An agent-based parallel approach for the job shop scheduling problem with genetic algorithms," Mathematical and Computer Modeling, vol.52, pp.1957-1965, 2010.

[30] O.López-Ortega, I.Villar-Medina, "A multi-agent system to construct production orders by employing an expert system and a neural network," Expert Systems with Applications, vol.36, pp.2937-2946, 2009.

[31] E.A.Olajubu, O.A. Ajayi, G.A.Aderounmu, "A fuzzy logic based multi-agents controller," Expert Systems with Applications, vol.38, pp.4860-4865, 2011.

[32] Malama T., Ioanna R., Lambros P, "An intelligent agent negotiation strategy in the electronic marketplace environment," European Journal of Operational Research, vol.187, pp.1327-1345,2008.

Mingliang Chen was born in Jiangsu Province, China in 1963.

He received his PH.D from Xi'an Jiaotong University in 2001. He is currently a professor in college of management of Zhejiang University. His research interests include business intelligent and customer relationship management.

Sainan Liu was born in Hubei Province, China in 1977. She received his PH.D from Zhejiang University in 2007. She is currently a postdoctoral in college of management of Zhejiang University and a lecturer in Hangzhou Dianzi University. Her research interests include business intelligent, public-opinion management, logistics and e-commerce.

A Graph Based Approach to Trace Models Composition

Youness Laghouaouta[a], Adil Anwar[b], Mahmoud Nassar[a], Bernard Coulette[c]
[a] IMS-SIME ENSIAS, Mohamed Vth Soussi University, Rabat, Morocco
Email: y.laghouaouta@um5s.net.ma, nassar@ensias.ma
[b] Siweb, EMI, Mohamed Vth Agdal Universtity, Rabat, Morocco
Email: anwar@emi.ac.ma
[c] IRIT-UTM, University of Toulouse II, Toulouse, France
Email: coulette@univ-tlse2.fr

Abstract— **A model driven engineering process involves different and heterogeneous models that represent various perspectives of the system under development. The model composition operation allows combining those sub-models into an integrated view, but remains a tedious activity. For that, traceability information must be maintained to comprehend the composition effects and better manage the operation itself. Against this context, the current paper describes a framework for model composition traceability. We consider the traces generation concern as a crosscutting concern where the weaving mechanism is performed using graph transformations. A composition specification case study is presented to illustrate our contribution.**

Index Terms— **traceability, model composition, model transformation, aspect oriented modeling, graph transformation.**

I. Introduction

One of the main Model Driven Engineering (MDE) principles is to reduce system complexity by raising the abstraction level. In MDE, the primary focus is on models rather than computing concepts. Models represent all artifacts handled by the software development process and can be used as first class entities in dedicated model management operations. Therefore, the gap between the requirements definition and the solution is reduced by metamodeling and transformation tools [1].

Usually, complex and large systems are built based on different models; each one representing a view of the system according to a different perspective, a different set of concerns, and a different group of components [2]. The main purpose is to separate concerns in order to represent the software system as a set of less complex sub-models. Hence, the complexity of the analysis/design activities is reduced in the earlier phase of the software development process.

However, several issues are raised, among them the need to synchronize contributing models. This task can be handled through the generation of views that cross different perspectives in order to propagate changes occurring in sub-models. Combining those models can be performed using a model composition approach. Nevertheless, even if model-oriented decomposition is interesting; model composition remains a laborious activity.

Traceability is a necessary system characteristic [3] that reveals the software process maturity. Model composition, as all other model management operations, requires a traceability mechanism for manifold uses: model validation, co-evolution of models and model composition optimization. Indeed, traceability management provides support to better manage the composition operation. It specifies how source artifacts participate in the production of the composed model. Those links detail the flow of execution and are useful to analyze the impact of changing sub-models during the evolution of the system and help to optimize composition chains.

This paper deals with the tracing of the composition of heterogeneous models. Our approach is based on a generic and extensible metamodel accounting for structuring trace links. Essentially, we aim at minimizing the trace links management effort and expressing highly configurable trace models. This paper extends our initial work presented in [4]. It focuses on the generation of traces by using aspect oriented modeling (AOM) principles [5] and graph transformations [6]. In fact, the weaving of the traceability aspect is specified by a set of graph transformation rules.

Graph transformations theory provides a formal support for defining some activities related to model management such as: model transformation, model refactoring and model integration. We intend to populate our traceability metamodel regardless of the composition language. To that end, we believe that graph transformation is a powerful technology for specifying and applying the weaving mechanism of the traces generation code in a composition specification in a more abstract manner.

The rest of the paper is organized as follows: in Section II we review related approaches concerning model transformation traceability; Section III represents an overview of our approach, while Section IV details the generation of the trace model. Thereafter, in Section V we present a concrete working example, followed by a discussion of our contribution in Section VI. Finally Section VII summarizes this paper and presents future works.

doi:10.4304/jsw.9.11.2813-2822

II. Background And Related Work

A. Traceability management in MDE

Traceability is recognized as an essential issue in software engineering, and model driven engineering is no exception. In the literature there are several definitions of traceability, which, differ depending on the artifacts abstraction level and the traceability intensions. The IEEE Standard Glossary of Software Engineering Terminology [7] defines traceability as: *the degree to which relationship can be established between two or more products of the development process, especially products having a predecessor-successor or master-subordinate relationship to one other; for example, the degree to which the requirements and design of a given software component match.*

More definitions concerning model transformation traceability have been proposed, notably:

- Dirvalos et al. [3] consider traceability as: *Any relationship that exists between artifacts involved in the software engineering life cycle*.
- Grammel and Voigt [8] define traceability as: *The runtime footprint of model transformation .Essentially, trace links provide this kind of information by associating source and target model element with respect of the execution of a certain model transformation*.

Traceability refers to the ability to capture and reuse links between a set of artifacts handled by a model driven development operation. This information represents the changes that have occurred in these elements and reveals the complexity of logical relations [9] existing among them. In MDE, traceability is a matter of three concerns [10]:

- What: Decide which concepts described in the models will be traced.
- How: Determine how to generate, represent and manage trace links.
- Why: Identify the intentions of capturing trace links.

B. Related work

Several researches address model transformation traceability issue. In this section we briefly outline the main approaches.

Jouault [11] presents an approach to trace transformations written in the ATL language. It addresses the problem of implicit traceability persistence. This approach is the basis for several future researches addressing traceability management. The author considers traces as a model generated in the same way as other target models. The traces generation code can be automatically inserted into any existing ATL program, through the application of a higher order transformation called *traceAdder* [11]. Since the author uses a simple metamodel to represent the trace model structure, traces are not configurable. Nevertheless, the use of the *traceAdder* transformation enhances scalability and allows reusability of the trace model stored externally.

Falleri et al. [12] suggest a framework for traceability of imperative model transformations written in the Kermeta language. The authors consider the trace model as a bipartite graph where the nodes are of two types: source nodes and target nodes. Trace links are stored in a separate model conforms to a generic traceability metamodel and can be reused. Besides, the manual adding of the traces generation code allows the user to select elements to trace, but this reduces scalability.

Amar et al. [13] propose a traceability framework for imperative transformations. The authors present a generic traceability metamodel based on the *"composite"* design pattern, while the trace generation is based on aspect-oriented programming using AspectJ. Thus, it builds traces without modifying the transformation code and supports scalability and reusability of aspects too. The framework defines categories of traceable operations and their respective poincuts. Since it does not take into account all the operations to trace, the programmer can define new custom categories or restrict the predefined ones. Furthermore, the application of the *"composite"* design pattern as well as the link type concept, allow defining configurable trace models.

Grammel and Kastenholz [14] have defined a generic traceability framework for model transformation approaches. It is based on a generic metamodel extensible through facets to simplify hierarchical structure. The approach offers two mechanisms for traces generation: transformation of the implicit trace model to another model conforms to the suggested metamodel and generation of traceability data based on aspect oriented programming. These two mechanisms make the approach scalable and enhance reusability. As for the trace model configuration, it entails the choice of artifacts to trace and the granularity level through the use of facets.

The confusion between model transformation and model composition is a debate topic. Some researches perceive model composition as a transformation with two input models and one output model, when others discern model transformation as a specific model composition which computes the source model with an empty model to produce the target one. Therefore, the presented approaches can be used to trace the model composition operation; however, we judge that the proposal for a model composition traceability approach proves advantageous. Actually, model composition has specific intensions (model synchronization, model integration...) and a particular process (matching step, merging step...) that have to drive the traceability approach.

C. Traceability requirements

In [4], we have detailed an evaluation of the presented approaches based on three comparison criteria: configuration, portability, and scalability. These criteria are inspired from the traceability challenges stated by the Center of Excellence for Software Traceability [15]. According to the results of this analysis, we derived four traceability requirements that have driven our approach.

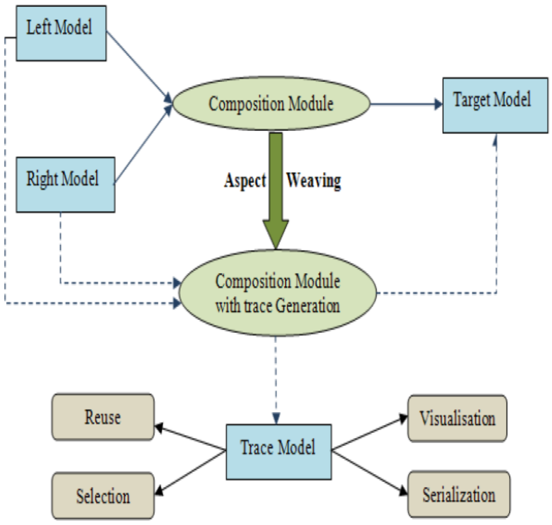


Figure 1. The trace generation process

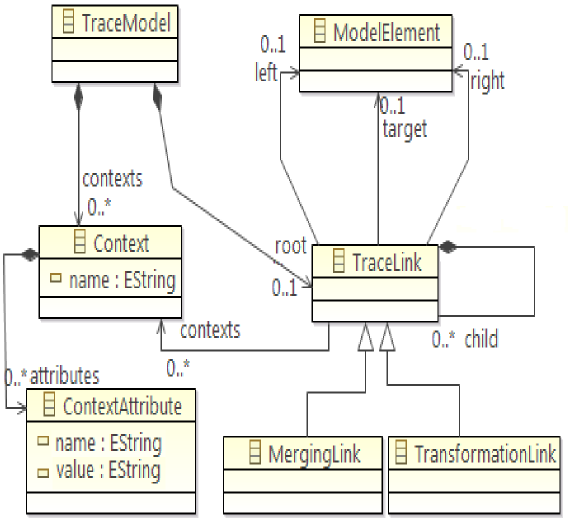


Figure 2. Composition traceability metamodel

- In order to address the scalability challenge, the trace model generation must be automatic; so as to reduce the effort required to achieve traceability. Furthermore, human intervention is useful to configure model elements to trace.
- The code necessary to generate traces must not be intrusive in the primary transformation in order to allow its reuse.
- Traceability data has to be stored in a separate model which conforms to a generic metamodel to reduce trace links management effort. Thus, it supports reusability of the trace model.
- The traceability metamodel has to be expanded with an extensibility mechanism. Essentially, this mechanism allows expressing configurable trace links depending on the traceability scenario and the models specifications.

In our approach, we propose to achieve the two first requirements by using graph transformation rules to generate the trace model. The other points are achieved by using generic composition traceability metamodel to represent the traceability data structure.

III. Overview Of The Approach

A. Trace generation process

We consider the trace model as an additional target model of the composition operation (Fig. 1). To generate it, we propose to use a weaving mechanism of the code responsible of creating traceability elements in the composition specification. We describe in section IV the aspect weaving process in more details. The trace model can be visualized as a graph, or invoked by a selection request. Furthermore, it can be used to validate the composition by checking the consistency and the completeness of the composed model. The co-evolution of models [16] can be supported by analyzing the impact of changing source elements through their corresponding trace links. Finally, we aim to optimize model composition chains; indeed, some trace links are valuable for following steps.

B. The composition traceability metamodel

Several approaches address the model composition operation: AMW [17], EML [18], Kompose [19]. We take into account the typical composition process which involves two major steps: matching and merging. During the first step, similarities between left and right model elements are calculated. Matching elements are merged while other elements are eventually transformed to target model elements or temporary modified to be merged.

We propose a traceability metamodel (see Fig. 2), which defines the different kinds of relationships between model elements independently from any given application domain. We have extended the core traceability metamodel proposed in [4] to support composition traceability requirements and complement it with well-formedness rules. Hereafter, we present the key elements of this metamodel.

A *MergingLink* element connects the left and right elements to the composed element, while a *transformationLink* element represents a transition from each left or right element to the target one. A transformation link may have no source or target element to allow tracing a deleted element or a newly created one. Moreover, we represent multi-scaled trace model that show the imbrications of the rule calls, through a parent-child relation among trace links. The nesting of traces allows the final user to configure the granularity degree he desires.

In order to enhance the trace model semantic richness, we use the *Context* concept to assign additional information to trace links depending on the traceability point of view and the models to compose specifications. Indeed, this extensibility mechanism is based on the definition of the relevant context attributes that capture the further expressiveness data to be assigned to a sub-set of traces, such as: the composition rule name, the traceability intention…

We present thereafter some well-formedness rules specifying the static semantics of the traceability metamodel.

1) A merging link must have two source elements and one target element.

```
context TraceLink
inv : self.oclIsTypeOf(MergingLink)
implies self.left->notEmpty() and self
    .right->notEmpty() and self.target
    ->notEmpty()
```

2) A transformation link has at most one source element.

```
context TraceLink
inv : self.oclIsTypeOf(
    TransformationLink)
implies  self.left.oclIsUndefined() or
     self.left.oclIsUndefined()
```

3) There is one and only one root trace link of type *MergingLink*.

```
context TraceModel
inv : let root:TraceLink = self.root
    in
if root.oclIsUndefined()
then true
else
root.oclIsTypeOf(MergingLink) and root
    .parent.ocIsUndefined()
endif
```

We illustrate a simple trace model in Fig. 3. The purpose of the composition to trace is to merge two simple class diagrams (*Left* model and *Right* model) each one containing one class *A*. The trace model contains one root element of type *MergingLink* that links the source class diagrams with the target one. Childs of this element represent the merging of the classes *A* and the types *int* in the source models. Finally the copy of the class attributes to the target model is represented by two nested transformation links.

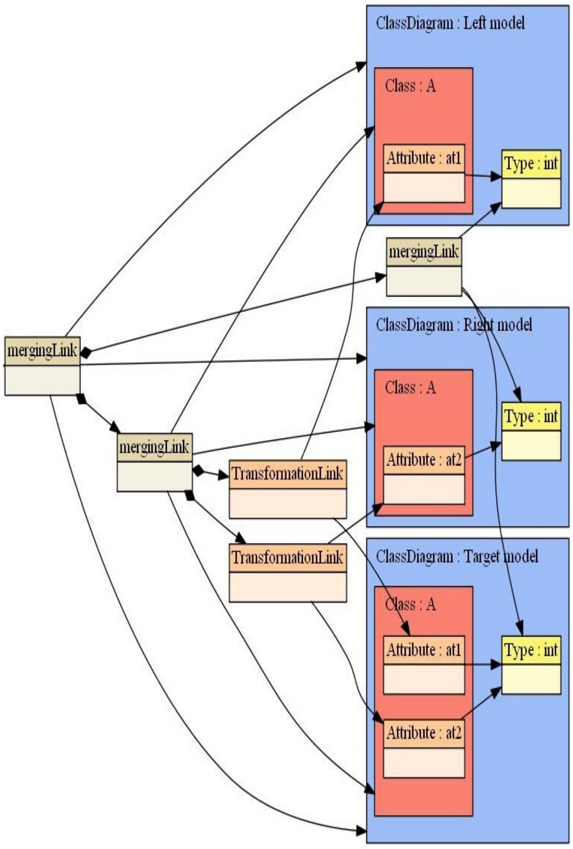


Figure 3. A simple trace model

IV. Trace Model Generation

In this section, we describe how we can use aspect oriented modeling with graph transformations to trace the model composition operation. Our objective is to address our traceability requirements in order to automatically build the trace model without modifying the code of the composition specification by hand. Indeed, we consider the insertion of the trace generation code as a weaving of the base model (which represents the composition specification) with the aspect model (describes the traceability concern). This weaving scenario is specified by graph transformation rules.

A. AOM and graph transformations concepts

Aspect oriented modeling applies aspect oriented programming [20] in the context of MDE, and focuses on modularizing and composing crosscutting concerns during the design phase of a software system. Indeed, the aspect that encapsulates the crosscutting structure and the base model it crosscuts are both models. An aspect is defined principally by:

- A pointcut: it is a predicate over a model used to determine the places where the aspect should be applied (joinpoints).
- An advice: It is the new structure that replaces the relevant jointpoints.

A graph rewriting rule consists of two parts, a left-hand side (LHS) and a right-hand side (RHS). A rule is applied by substituting the objects of the left-hand side with the objects of the right-hand side, only if the pattern of the left-hand side can be matched to a given graph [21].

A formal definition of a graph transformation rule is given in [22]: A graph transformation is a rule $r : L \rightarrow R$ from a left-hand side (LHS) graph L to a right-hand side (RHS) graph R. The process of applying r to a graph G involves finding a graph morphism, h, from L to G and replacing $h(L)$ in G with $h(R)$. To avoid dangling edges i.e., edges with a missing source or target node $h(R)$ must be pasted into G in such a way that all edges connected to a removed node in $h(L)$ are reconnected to a replacement node in $h(R)$.

We establish the following correspondences to simulate aspect weaving operation with graph transformation rules: A set of rules correspond to an aspect, the LHS part defines the points where the aspect should be applied (the pointcut), and the RHS part defines the crosscutting structure that should be inserted at those points (the advice). Note that we have chosen the Henshin project [23] to implement the weaving process.

Henshin is a transformation language and tool environment based on graph transformation concepts and operating on EMF models [23]. It provides features needed to express complex transformation such as: negative application conditions (NACs) which specify the non-existence of model patterns in certain contexts and transformation units to control the rules application sequence.

B. The weaving operation

We have chosen the Epsilon Merging Language EML [18] as an example of dedicated composition language, which is used to express a model merging specification. EML belongs to the Epsilon platform, which is a model driven framework for developing integrated languages for model management tasks such as comparison, transformation, validation, etc. This language proposes to merge models through three categories of rules: match rules, merge rules and transformation rules.

An EML specification can be represented as a graph, since the abstract syntax of EML can be considered as a graph. Hence, the transformation of an EML module to another EML module, which contains traceability generation code, can be considered as a graph transformation. Fig. 4 depicts an excerpt of the EML abstract syntax [24]. Note that the definition of some model elements has been modified to simplify the specification of graph transformations that deal with the generation task.

1) Trace link declaration for merge rules: The rule presented in Fig. 5 allows declaring the traceability element that captures the correspondence between the two source elements matched by the application of an EML merge rule and the merged one. This rule searches for a *MergeRule* node with its connected parameters corresponding to the left, right and target parameters. Thereafter, it adds a new *ParameterDeclaration* node stereotyped with *create*, referencing the merging link to be generated. Besides, the added *AssignStatement* nodes attribute the reference of the corresponding element to the appropriate trace link property (*left*, *right*, and *target*).

2) Trace link declaration for transformation rules: The graph transformation rule presented in Fig. 6 aims to add the trace link declaration to EML transformation rules. As with merge rules, it searches for a *TransformationRule* node and appends to it a new parameter of type *TransformationLink*. This newly added parameter allows generating a trace link that captures the transition from the source element to the target one. Furthermore, the added assign statements attribute the references of the matched *ParameterDeclaration* nodes stereotyped with *preserve* to the generated trace link.

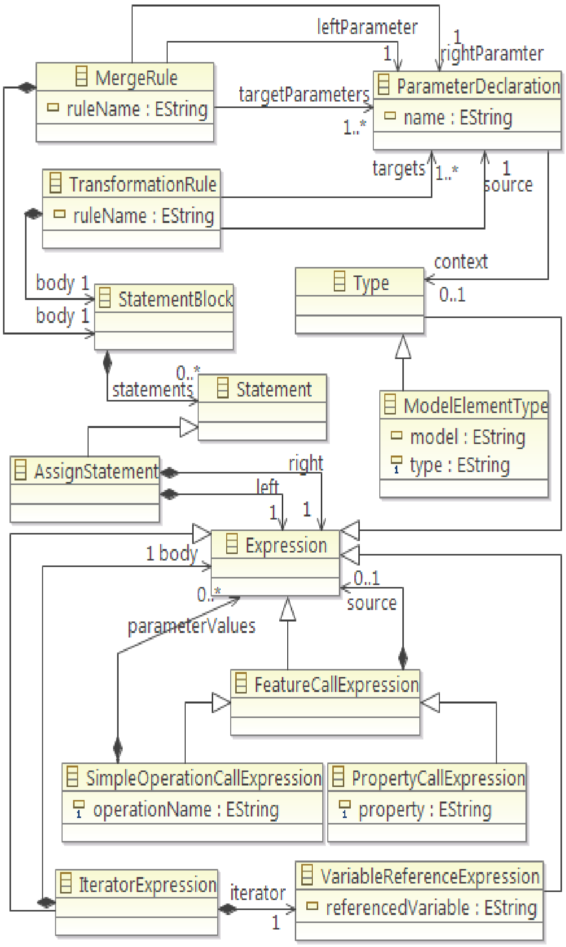


Figure 4. Excerpt of the EML abstract syntax

Note that in the EML abstract syntax, no distinction is made between the left and the right elements (the transformation rule connects the source element to the target one). Consequently, we can't automatically resolve the origin of the element (left or right model) without user's assistance.

3) Trace links nesting rule: Within EML, the rule call is implicitly performed using the *equivalent* operation that automatically resolves source elements to their transformed counterparts in the target models [24]. This target equivalent is produced by an anterior application of a given rule. We propose to structure traces conforming to the rule invocation sequence. Indeed, the application of the two previous rules allows generating extra-outputs corresponding to trace links that are resolved as potential target equivalents. Hence, we trace a rule call by assigning the trace link generated by the called rule as a child of the link generated by the calling rule.

Accordingly, the rule depicted in Fig. 7 searches for a call of the *equivalent* operation. Thereafter, it copies the reference of the element to resolve (which corresponds to the source of the *SimpleOperationCallExpression* node stereotyped with *delete*) to the variable named *element*. Then, the target equivalents are divided on two subsets: those corresponding to the traceability data that are used to bind the traceability element to its parent and the other element used to copy the original call of the *equivalent* operation. This filtering mechanism is made by applying the *select* operation.

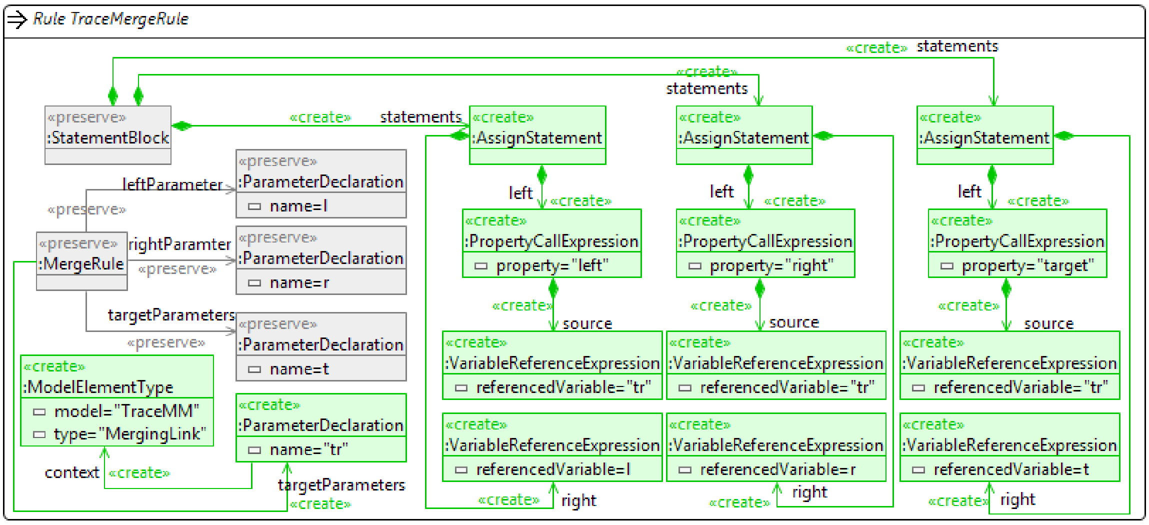

Rule TraceMergeRule
«preserve» :StatementBlock
«create» :AssignStatement
«preserve» :ParameterDeclaration name=l
«preserve» :ParameterDeclaration name=r
«preserve» :ParameterDeclaration name=t
«create» :ParameterDeclaration name="tr"
«preserve» :MergeRule
leftParameter
rightParamter
targetParameters
«create» :ModelElementType model="TraceMM" type="MergingLink"
context
«create» :PropertyCallExpression property="left"
«create» :PropertyCallExpression property="right"
«create» :PropertyCallExpression property="target"
«create» :VariableReferenceExpression referencedVariable="tr"
«create» :VariableReferenceExpression referencedVariable=l
«create» :VariableReferenceExpression referencedVariable=r
«create» :VariableReferenceExpression referencedVariable=t
statements
left
source
right

Figure 5. Trace link declaration for merge rules

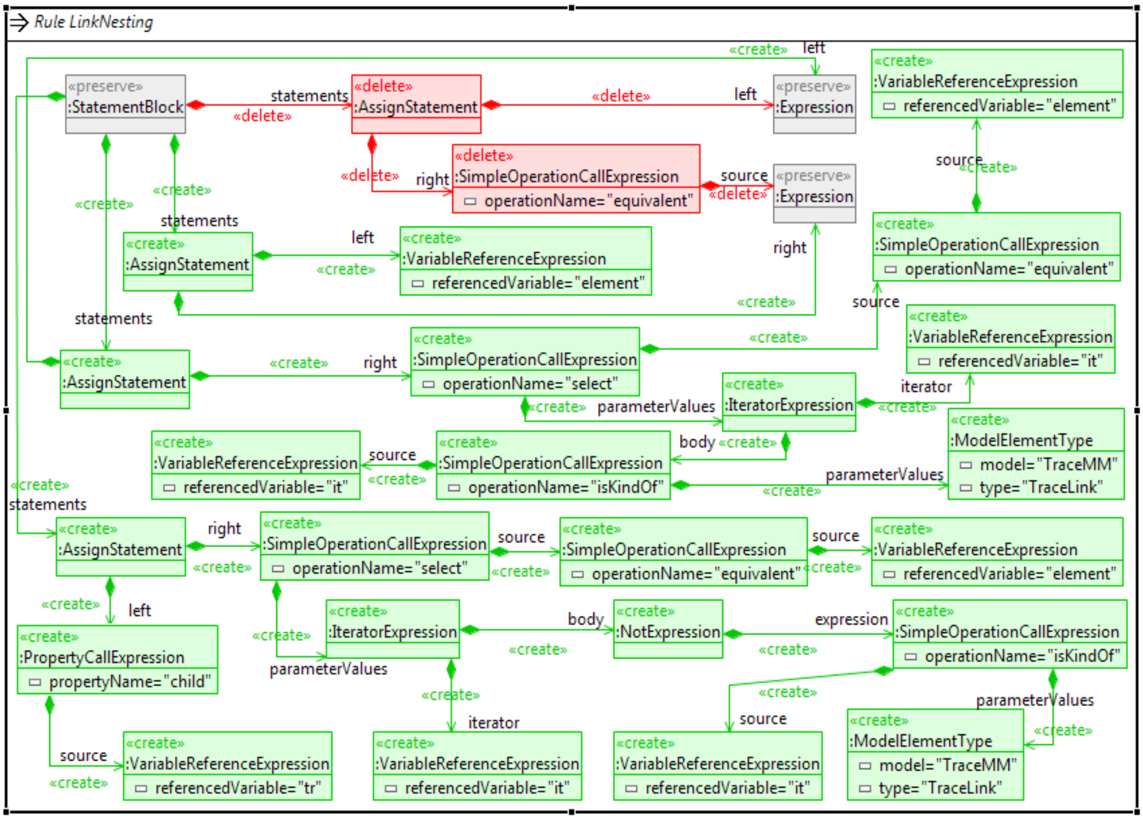

Rule LinkNesting
«preserve» :StatementBlock
«delete» :AssignStatement
«delete» :SimpleOperationCallExpression operationName="equivalent"
«preserve» :Expression
«create» :VariableReferenceExpression referencedVariable="element"
«create» :SimpleOperationCallExpression operationName="equivalent"
«create» :AssignStatement
«create» :VariableReferenceExpression referencedVariable="it"
«create» :SimpleOperationCallExpression operationName="select"
«create» :IteratorExpression
«create» :SimpleOperationCallExpression operationName="isKindOf"
«create» :ModelElementType model="TraceMM" type="TraceLink"
«create» :NotExpression
«create» :PropertyCallExpression propertyName="child"
«create» :VariableReferenceExpression referencedVariable="tr"
statements
left
right
source
iterator
body
parameterValues
expression

Figure 7. Trace links nesting rule

C. The tool architecture

Fig. 8 depicts a high level view of the tool architecture. Basically, it contains two major layers: the composition and traceability layer and the serialization and visualization layer. The first layer constitutes the core of our architecture while the second one offers facilities to perform the traceability management.

The serialization service is implemented using the EMFText project [25]; it involves a text to model parser and a model to text printer for the EML language. Essentially, this allows transforming the textual EML specification to the corresponding model conforms to the EML abstract syntax. Thereafter, a specific graph transformation unit (which is specified using the Henshin project cf. Section IV-B) weaves the traces generation patterns in the corresponding model. Finally, we reproduce the concrete specification by using the model to text printer.

The execution of the resulting specification generates exta-outputs corresponding to the traceability elements

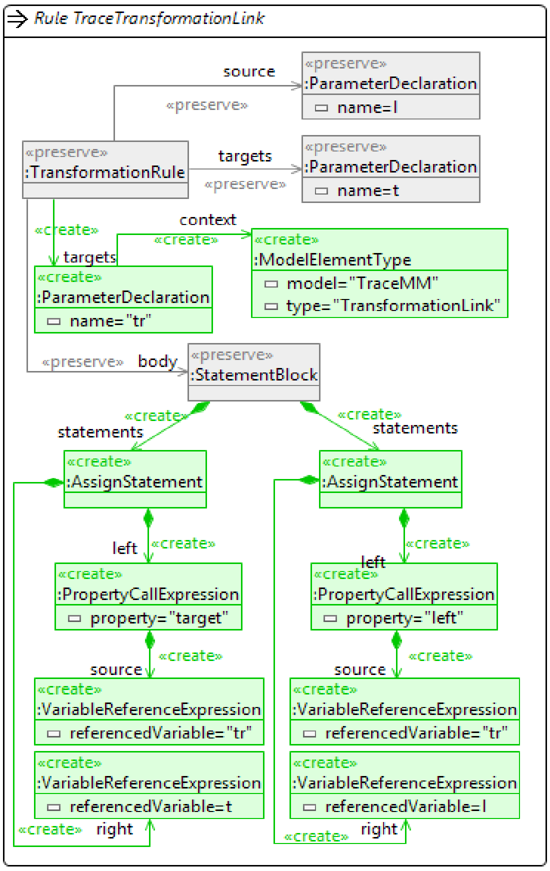

Rule TraceTransformationLink
source
«preserve» :ParameterDeclaration
name=l
«preserve»
«preserve» :TransformationRule
targets
«preserve» :ParameterDeclaration
name=t
«preserve»
context
«create»
«create»
targets
«create» :ParameterDeclaration
name="tr"
«create» :ModelElementType
model="TraceMM"
type="TransformationLink"
«preserve» body
«preserve» :StatementBlock
«create» statements
«create» statements
«create» :AssignStatement
«create» :AssignStatement
left «create»
left «create»
«create» :PropertyCallExpression
property="target"
«create» :PropertyCallExpression
property="left"
source «create»
source «create»
«create» :VariableReferenceExpression
referencedVariable="tr"
«create» :VariableReferenceExpression
referencedVariable="tr"
«create» :VariableReferenceExpression
referencedVariable=t
«create» :VariableReferenceExpression
referencedVariable=l
«create» right
«create» right

Figure 6. Trace link declaration for transformation rules

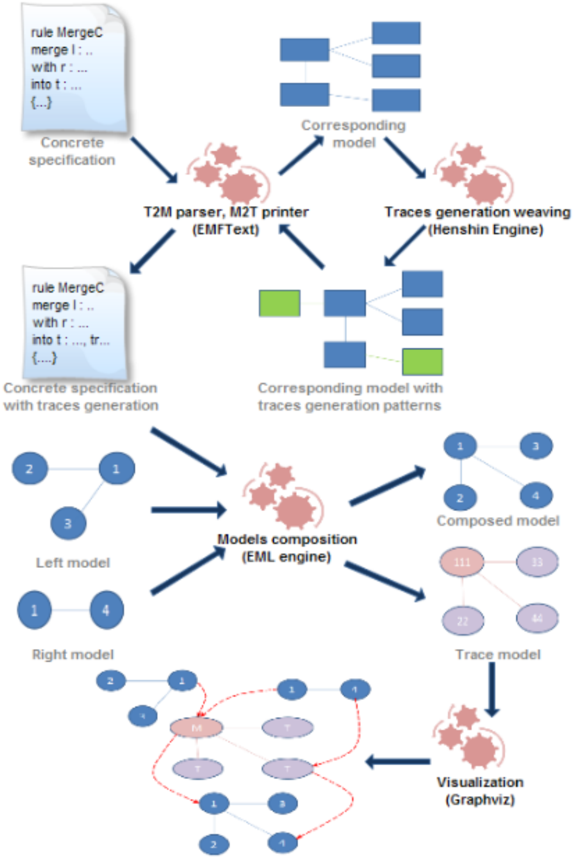


Figure 8. The tool architecture

while producing the composed model. Besides, the visualization service provides support to transform the generated trace model in a human friendly representation.

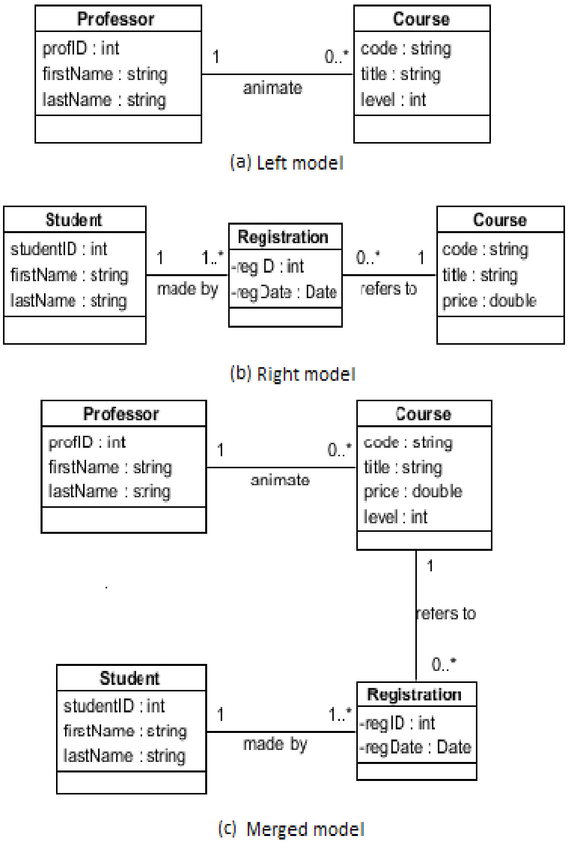
V. Case Study

In this section, we provide an example to illustrate the application of our approach. The merging scenario we have chosen is the merging of two UML models represented by class diagrams into a target model. The source models as well as the merged model are displayed in Fig. 9. Listing 1 represents the EML rule that merges two source classes, while Listing 2 depicts the resulting modifications over this rule.

```
rule MergeClassWithClass
merge l : left!Class
with r : right!Class
into t : target!Class
{
t.name = l.name;
t.ownedAttribute = l.ownedAttribute.includingAll
    (r.ownedAttribute).equivalent();
}
```

Listing 1. Merge two classes rule

Depending on the rule type (*merge* or *transformation*), the two first rules of the traces generation weaving unit, declare the traceability parameter as another target parameter, and assign the traceability information to it (Listing



(a) Left model

(b) Right model

(c) Merged model

Figure 9. Illustrative example

2: lines 8,11-13). Besides, the call of the *equivalent* operation (Listing 1: line 7) has been captured and replaced with the fragment that divides its return to trace model elements and default target elements (Listing 2: lines 1-4,14-16). The first sub-set is used to copy the original call of the *equivalent* operation (Listing 2: line 15), while the traceability element is assigned as a child of current trace link (Listing 2: line 16).

```
 1 pre
 2 {
 3 var element : new Any ;
 4 }
 5 rule MergeClassWithClass
 6 merge l : left!Class
 7 with r : right!Class
 8 into t : target!Class , tr:trace!MergingLink
 9 {
10 t.name = l.name;
11 tr.left=l;
12 tr.right=r;
13 tr.target=t;
14 element = l.ownedAttribute.includingAll(r.
       ownedAttribute);
15 t.ownedAttribute = element.equivalent().select(
       it | not it.isKindOf(trace!TraceLink));
16 tr.child = element.equivalent().select(it | it.
       isKindOf(trace!TraceLink));
17 }
```

Listing 2. Merge two classes with traces generation

Fig. 10 depicts an excerpt of the generated trace model. This model conforms to our composition traceability metamodel and contains two types of trace links (merging links and transformation links) that are generated with respect to the composition relationships kinds. Those links are nested with respect to the rules invocation sequence. Essentially, the multi-scaled character of trace links allows the user to navigate over the trace model, from rough to precise. Note that we have used the Emf2gv project[1] to visualize the trace model.

VI. Discussion

As presented in section II, traceability involves three concerns: what to trace, how to manage the traceability information, and why we require it. Our approach allows the user to identify a subset of elements to trace through the selection of the relevant aspects (graph transformation rules) to apply. On the other hand, the use of aspect oriented modeling and graph transformation rules automatically insert the code responsible for generating the traceability information. In order to reduce effort to achieve traceability and support reusability of aspects, the trace model conforms to a generic metamodel. Besides, our metamodel is extensible to express traces regarding traceability scenarios. Finally, we identified three major intensions of capturing traces: validation in model composition, co-evolution of models and optimization of composition chains. These intensions will guide our future work. We consider that the challenge is to make our approach aware of the ”why”, in order to automatically select the elements to trace and configure the trace management process depending on the user’s intension.

The use of graph transformation proves to be advantageous for augmenting composition tools with a traceability support, since, existing graph based tools as Henshin project can perform this operation. It provides features needed to express complex transformation such as: application conditions and the control flow of graph transformation rules. Furthermore, the plurality of the composition languages and their characteristics (textual, model-based, and graph-based) make traceability difficult to manage; however, we believe that the exploration of graph transformation options provides ways to overcome this problem.

We aim to abstract as much as possible the composition specification to the corresponding graph. Thereby, our approach can be used to trace model composition regardless its nature: textual specifications written in EML, model-based specification in ATL, and graph based composition [2]. However, our contribution is currently a language dependent approach, since the definition of the graph transformation rules takes into account the composition language. As a solution, we are considering a pivot language.

VII. Conclusion And Future Work

In this paper, we presented an approach dedicated to manage the traceability concern in a model composition operation. Our solution fits a set of traceability requirements we have deduced from the analysis of the main model transformation traceability approaches. Indeed, we consider traceability as a cross-cutting concern and we generate the trace model automatically based on the aspect oriented paradigm. The aspect weaving is implemented using graph transformation rules. Moreover, we use a generic and extensible traceability metamodel that deals with the configuration challenge.

Several perspectives to our work are under consideration. We are completing the visualization of the generated trace model in a human-friendly representation using the graphviz tool [26]. Besides, we intend to work on a pre-configuration tool support to generate the trace model according to the user’s requirements. Finally, we believe that traceability data is useful for optimizing the establishment of correspondences between contributing models, by automatically refining the matching model.

References

[1] S. Kent, “Model driven engineering,” in *Integrated formal methods*. Springer, 2002, pp. 286–298.

[2] A. Anwar, A. Benelallam, M. Nassar, and B. Coulette, “A graphical specification of model composition with triple graph grammars,” in *Model-Based Methodologies for Pervasive and Embedded Software*. Springer, 2013, pp. 1–18.

[3] N. Drivalos, R. F. Paige, K. J. Fernandes, and D. S. Kolovos, “Towards rigorously defined model-to-model traceability,” in *ECMDA Traceability Workshop (ECMDA-TW’08)*, 2008, pp. 17–26.

[1] See `http://sourceforge.net/projects/emf2gv`.

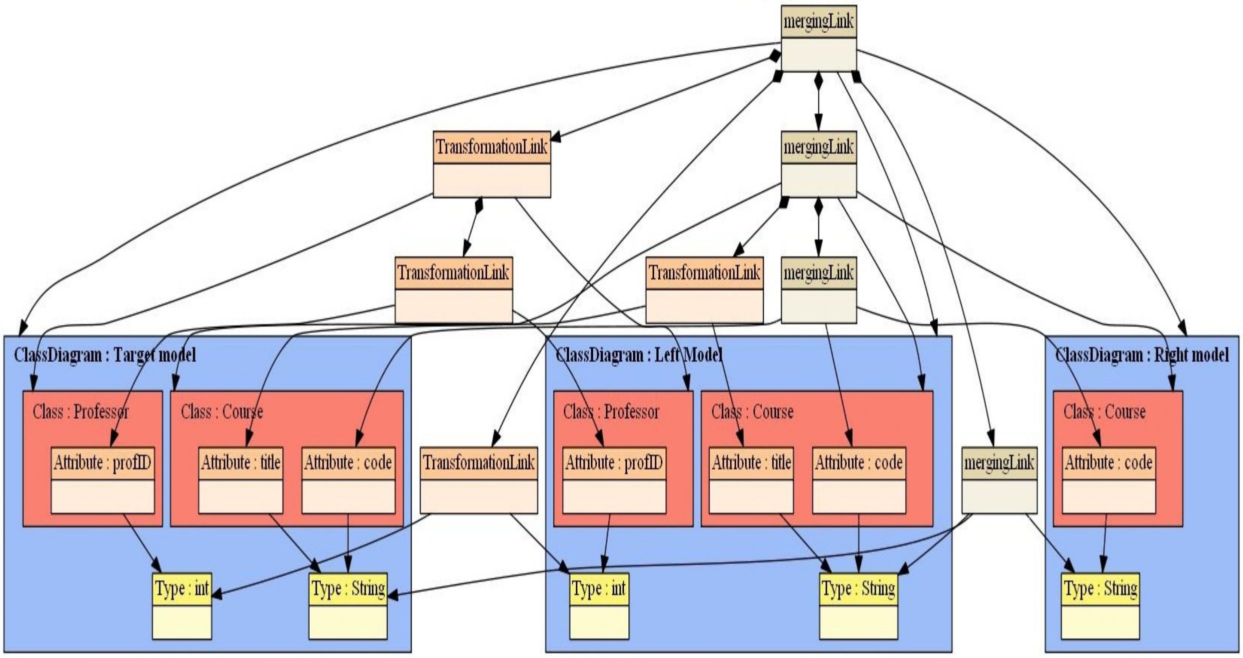


Figure 10. Excerpt of generated trace model

[4] Y. Laghouaouta, A. Anwar, and M. Nassar, "A traceability approach for model composition," in *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2013, pp. 1–4.

[5] R. France, I. Ray, G. Georg, and S. Ghosh, "Aspect-oriented approach to early design modelling," *IEE Proceedings-Software*, vol. 151, no. 4, pp. 173–185, 2004.

[6] G. Rozenberg and H. Ehrig, *Handbook of graph grammars and computing by graph transformation*. World Scientific Singapore, 1997, vol. 1.

[7] J. Radatz, A. Geraci, and F. Katki, "Ieee standard glossary of software engineering terminology," *IEEE Std*, vol. 610121990, p. 121990, 1990.

[8] B. Grammel and K. Voigt, "Foundations for a generic traceability framework in model-driven software engineering," in *ECMDA Traceability Workshop (ECMDA-TW'09)*, 2009.

[9] N. Anquetil, B. Grammel, I. Galvão, J. Noppen, S. S. Khan, H. Arboleda, A. Rashid, *et al.*, "Traceability for model driven, software product line engineering," in *ECMDA Traceability Workshop (ECMDA-TW'08)*, 2008, pp. 77–86.

[10] G. Spanoudakis and A. Zisman, "Software traceability: a roadmap," *Handbook of Software Engineering and Knowledge Engineering*, vol. 3, pp. 395–428, 2005.

[11] F. Jouault, "Loosely coupled traceability for atl," in *ECMDA Traceability Workshop (ECMDA-TW'05)*, vol. 91. Citeseer, 2005.

[12] J.-R. Falleri, M. Huchard, C. Nebut, *et al.*, "Towards a traceability framework for model transformations in kermeta," in *ECMDA Traceability Workshop (ECMDA-TW'08*, 2006, pp. 31–40.

[13] B. Amar, H. Leblanc, and B. Coulette, "A traceability engine dedicated to model transformation for software engineering," in *ECMDA Traceability Workshop (ECMDA-TW'08)*, 2008, pp. 7–16.

[14] B. Grammel and S. Kastenholz, "A generic traceability framework for facet-based traceability data extraction in model-driven software development," in *ECMDA Traceability Workshop (ECMDA-TW'10)*, 2010, pp. 7–14.

[15] O. Gotel, J. Cleland-Huang, J. H. Hayes, A. Zisman, A. Egyed, P. Grünbacher, A. Dekhtyar, G. Antoniol, and J. Maletic, "The grand challenge of traceability (v1. 0)," in *Software and Systems Traceability*. Springer, 2012, pp. 343–409.

[16] B. Amar, H. Le Blanc, P. Dhaussy, B. Coulette, *et al.*, "Trace transformation reuse to guide co-evolution of models," in *5th Int. Conference on Software and Data Technologies (ICSOFT'10)*, 2010.

[17] M. D. Del Fabro, J. Bézivin, F. Jouault, E. Breton, G. Gueltas, *et al.*, "Amw: a generic model weaver," *Procs. of IDM05*, 2005.

[18] D. S. Kolovos, R. F. Paige, and F. A. Polack, "Merging models with the epsilon merging language (eml)," in *Model Driven Engineering Languages and Systems*. Springer, 2006, pp. 215–229.

[19] F. Fleurey, B. Baudry, R. France, and S. Ghosh, "A generic approach for automatic model composition," in *Models in Software Engineering: Workshops and Symposia at MODELS 2007*, vol. 5002. Springer, 2008, p. 7.

[20] G. Kiczales, J. Lamping, A. Mendhekar, C. Maeda, C. V. Lopes, J.-M. Loingtier, and J. Irwin, "Aspect-oriented programming," in *ECOOP*, 1997, pp. 220–242.

[21] L. Lambers, H. Ehrig, and F. Orejas, "Conflict detection for graph transformation with negative application conditions," in *Graph Transformations*. Springer, 2006, pp. 61–76.

[22] J. Whittle, J. Araújo, and A. Moreira, "Composing aspect models with graph transformations," in *Proceedings of the 2006 international workshop on Early aspects at ICSE*. ACM, 2006, pp. 59–65.

[23] T. Arendt, E. Biermann, S. Jurack, C. Krause, and G. Taentzer, "Henshin: advanced concepts and tools for in-place emf model transformations," in *Model Driven Engineering Languages and Systems*. Springer, 2010, pp. 121–135.

[24] D. Kolovos, L. Rose, R. Paige, and A. Garcıa-Domınguez, "The epsilon book," *Structure*, vol. 178, 2010.

[25] F. Heidenreich, J. Johannes, S. Karol, M. Seifert, and C. Wende, "Derivation and refinement of textual syntax for models," in *Model Driven Architecture-Foundations and*

Applications. Springer, 2009, pp. 114–129.

[26] E. R. Gansner, "Drawing graphs with graphviz," Technical report, AT&T Bell Laboratories, Murray, Tech. Rep., 2009.

Youness Laghouaouta received the Engineer of state degree in Software Engineering from National High School of Computer Science and Systems Analysis (ENSIAS) in 2009. He is currently a PhD student in the IMS (Models and Systems Engineering) Team of SIME Laboratory at ENSIAS. His research interests are model traceability, model composition, Aspect Oriented Engineering, and Model-Driven Engineering.

Adil Anwar works as an assistant professor in computer science at the university of Mohammed-V Rabat, and as a member of the Siweb research team of Mohammadia school of engineers. In 2009, he received a Ph.D degree in Computer Science at the University of Toulouse. He is interested in software engineering, including model driven software engineering, mainly by heterogeneous software language modelling, traceability management in MDE, combining formal and semi-formals methods in software development.

Mahmoud Nassar is Professor and Head of the Software Engineering Department at National Higher School for Computer Science and Systems Analysis (ENSIAS), Rabat, Morocco. He is also Head of IMS (Models and Systems Engineering) Team of SIME Laboratory. He received his PhD in Computer Science from the INPT Institute of Toulouse, France. His research interests are integration of viewpoints in Object-Oriented Analysis/Design (VUML profile), Model-Driven Engineering, and Context-Aware Service-Oriented Computing.

Bernard Coulette works as a full professor at the University of Toulouse, and as a member of the MACAO team of IRIT laboratory. His research fields of interest are mainly integration of viewpoints in Object-Oriented Analysis/Design (VUML profile), modeling and enactment of Model Driven Processes. He has directed several PHD thesis in the context of international collaborations (Vietnam, Morocco).

Automatic Indexing for Research Papers Using References

Wei Liu
Institute of Scientific and Technical Information of China
China, 100038
Email: liuw@istic.ac.cn

***Abstract*—An effective way to reveal the contents of research papers is assigning a group of terms against a controlled vocabulary. To the best of our knowledge, a variety of automatic indexing techniques have been studied to enhance the effectiveness and the efficiency. However, the current approaches depended on the content of a research paper, such as title, abstract, etc., which suffering from limitations on the automatic indexing performance. In this paper, we propose a new approach of automatic indexing for single research paper with its references based on Genetic Algorithm (GA). The extensive experiments on four subjects show the effectiveness of the proposed approach.**

***Index Terms*—information retrieval, automatic indexing, research paper, genetic algorithm**

I. Introduction

Research paper indexing refers to the process of assigning a group of meaningful terms against a controlled vocabulary for one research paper, which the terms can be representative of the research paper. The indexing terms can be regarded as the knowledge summary for one research paper. Indexing terms are facilitated for users to grasp the highlights without going through the whole research paper, and they can also be used as low cost measures of text similarity for research paper search system. Many applications can benefit from it, e.g., research paper retrieval, automatic classification & clustering, content summarization, topic-oriented information visualization. Obviously, manual assignment of high-quality indexing terms is expensive and time-consuming, which is almost an impossible task to handle a large number of papers annually for a library. Therefore, automatic indexing approaches are in great demand, which try to make process of assigning index terms with minimum human intervention.

To the best of our knowledge, a variety of automatic indexing approaches have been proposed for general documents[1,2,4,5,7,14,18] or special types of documents (e.g., news articles, blogs)[3,6,7,8,9,10,11]. Research papers, as a special type of documents, require more precise terms as the highly condensed summary. Most present automatic indexing approaches for research papers only make use of the main content of research papers with statistics, linguistics, machine learning and other techniques. However, they suffer from two limitations in practice. First, many papers are in form of scanned images instead of texts in the database due to some reasons. Thus the content-based approaches could not deal with such papers. Though OCR(Optical Character Recognition) techniques could be used to recognize the texts in scanned images, the accuracy is disturbing for many factors, such as image resolution and scan angle. And usually watermarking techniques[12] are used to prevent illegal copying, which also increases the difficulties of image recognition. Second, word segmentation is one of the prerequisites for analyzing the contents of research papers. It is widely known that this problem has not been well addressed for research papers Chinese). Further, one important fact ignored by the precious approaches is that research papers are the heritage and improvement of academic knowledge, which is manifested as referenced papers. As a result, the indexing terms of the references can be considered as important clues to generate the indexing terms of the paper.

Motivated by this idea, we tries to find a practical approach which can accomplish the automatic indexing task only using the papers' references. This is very important for many applications. For instance, it is always a heavy burden for a library to index a quantity of purchased research papers every year. In this way, the newly published papers can be automatically indexed if all previous papers have been assigned appropriate index terms already. The prophase survey based on the journal paper database, dissertation database and conference paper database of National Science and Technology Digital Library[13] shows the feasibility of our approach: on average, 94.2% index terms of one paper appear in the indexing terms of its references, and 4.3% indexing terms have semantic similar terms(synonym, hypernym or hyponym) in the indexing terms of its references, and only 1.5% indexing terms are completely new for the indexing terms of its references. Without the ambiguity

As a direct way, selecting high-frequency ones or low-frequency ones from the indexing terms of the references is not a good idea because high-frequency ones are too common and low-frequency ones are too rare to represent the topic of the paper. Hence, the key is how to select appropriate ones from the indexing terms of the references? To this end, we regard the indexing task as the knowledge inheritance, and the genetic algorithm is

doi:10.4304/jsw.9.11.2823-2829

adopted help us fulfill the filter process of indexing terms. We believe this is the first work which handles the automatic indexing problem from the knowledge inheritance point of view. Based on the proposed approach, a prototype system AIRPUR(Automatic Indexing for Research Paper Using References), has been implemented and is on trial. The latest experimental results will be reported in Section 5.

II. Related Works

The keywords assigned by the authors in most papers are usually are not professional for literature indexing. As a result, the indexing quality could not be stable, and the papers have to be assigned indexing terms manually by pro. Obviously, this is a labour-intensive and error-prone task. To this end, automatic indexing is being a hot topic in the literature retrieval field. Various techniques have been adopted for solutions, such as statistics, machine learning, linguistics, and so on.

The statistics-based solutions are simple for implementation, and no complicated training is necessary. Cohen et al[14] proposed a fast statistics method for generating indexing terms. This method utilizes N-gram to compute term' weight, and only the word split technique is required. Barker and Cornacchia [15] propose a system that takes into account not only the frequency of a "noun phrase" but also the head noun. In addition, other statistics-based solutions, such as words co-occurrence method[16] and PAT tree based method[17]. Yih et al[19] uses a number of features, such as term frequency of each potential indexing term and inverse document frequency to extract new indexing terms from previously unseen pages. The linguistics-based solutions generate indexing terms through lexical analysis, syntactic analysis and discourse analysis. Ercan et al[18] construct lexical chains and extract and qualify effective features from texts. The machine-learning-based solutions obtain the statistical parameters, and then extract indexing terms from samples. Zhang et al[19] use Conditional Random Fields model for training. Thuy Dung Nguyen et al[20] deduce the weights of candidate terms based on the layout structure of scientific and technical papers. [5] presents a term extraction approach which uses the features based on lexical chains in the selection of keywords for a document. Wan et al[2,4] proposes using a small number of nearest neighbor documents to improve document summarization and term extraction for the specified document, under the assumption that the neighbor documents could provide additional knowledge and more clues.

A conclusion is easy to be drawn that all the existing methods only make use of the internal features of literatures, which put forward high demand for natural language analysis techniques. As mentioned above, they heavily limited by the word splitting algorithm and OCR(Optical Character Recognition) techniques. In the left this paper, a new automatic indexing method will be present, which uses the external features to avoid the limitations of previous approaches.

III. Simple Approach Based on Term Frequency of Reference

In this section, a simple approach is introduced, which is simply based on term frequency. Intuitively, one term has more chance to be selected as the indexing term for a research paper if it appeared frequently as the indexing term in the references. So the direct way is to select the most frequent indexing terms of the references as the ones of the paper. Here we assume an indexing term appears only once in one reference. Based on such idea, the frequency of each indexing term is counted first. And then, the indexing terms are ranked by frequency. In order to make the rank more objective, citation frequency is incorporated, and the rank criteria of selecting indexing terms for a paper is as follows:

$$Rank(x)=\sum_{i=1}^{|references|} citation(ref_i)\bullet ref_i(x)$$

where |*references*| is the number of the references of the paper, *citation*(*ref*i) is the citation number of the *ith* reference, and *ref*$_i$(*x*) is a bool variable, where:

$$ref_i(x)=\begin{cases}1 & x\ in\ the\ ith\ reference\\ 0 & x\ not\ in\ the\ ith\ reference\end{cases}$$

However, Simple Approach is not effective and robust enough since frequency is not the most useful factor. In many scenarios, it would bring negative efforts instead. For example, in the fast developing subjects, the indexing terms those represent current progress need to be given more consideration on recent references, but frequent indexing terms are more likely appear past references, which are not representative enough. The experimental results reported in Section 5 will prove this fact.

IV. Automatic Indexing Based on Genetic Algorithm

In this section, a new approach is proposed based on genetic algorithm, which can eliminate the defects of Simple Approach. The idea of our proposed approach is similar with that of Wan et al[2,4], which generates indexing terms from neighbor documents. The main difference is that the neighbor documents of them are obtained by using document similarity search techniques, while ours are references.

A. Genetic Algorithm

Genetic Algorithms (GA) are adaptive heuristic search algorithms premised on the evolutionary ideas of natural selection and genetic. The basic concept of GA is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem. First pioneered by Holland (1975), Genetic Algorithms have been widely studied, experimented and applied in many fields[21,22]. GA provides an alternative method to solving problem that consistently outperforms other traditional methods in most of the problems link. Many of the real world problems involved finding optimal parameters that might prove difficult for traditional methods but ideal for GA. A population of individuals is maintained within search space for a GA, each representing a possible solution to a given problem. Each individual is coded as a finite length vector of components. These individuals are linked to chromosomes and the components are analogous to genes. Thus, a chromosome comprises several genes. A fitness score is assigned to each chromosome representing the abilities of an individual to compete. The individual with the optimal or generally near optimal fitness score is sought. The GA aims to use selective breeding of the solution to produce off springs better than parents do by combining information from the chromosomes. This process is repeated until the strings in the new generation are identical or certain termination conditions are met. GA can be more easily made scalable, concurrent and robust. The procedure of GA is described as follows:

i. Choose the initial population of individuals
ii. Evaluate the fitness of each individual in the population
iii. Population evolves until termination conditions are met (time limit, sufficient fitness achieved, etc.):
iv. Select the best-fit individuals for reproduction
v. Breed new individuals through crossover and mutation operations to give birth to offspring
vi. Evaluate the individual fitness of new individuals
vii. Replace least-fit population with new individuals

From the procedure above, three crucial components have to be well designed, which are coding for individuals, fitness function, and evolutionary process.

B. Coding for Individuals

The initial population is all the references $D=\{d_1,d_2,\ldots\ldots,d_n\}$ of paper d_0. Supposing the indexing term vector is $W=<w_1,w_2,\ldots\ldots,w_m>$ and the indexing term of $d_i(i\neq0)$ must be in W. Hence can be represented as one-dimensional vector $(x_{i1},x_{i2},\ldots\ldots,x_{im})$, where the value of x_{ij} is:

$$x_{ij}=\begin{cases}0 & w_j\notin d_i\\ 1 & w_j\in d_i\end{cases}$$

C. Fitness Function

GA is actually a process of survival of the fittest. The better genes(indexing terms) should have more possibility to be inherited to the next generation. Hence the individuals with more better genes must be assigned a larger fitness value. In the context of the problem studied in this paper, better genes can be considered as the indexing terms that can be representative of the paper.

To design the fitness function, several factors have to be taken into consideration. The first is the topic relevance. Intuitively, an indexing term would be far more representative if it appears frequently in the references and rare in the whole paper database. The second is the impact of the references. Simply, we use the citation number to express the impact of a reference. The third is the publication dates of the references. We argue that, in most cases, the innovation of a research paper is the improvement on the current progress. Usually one paper is more relevant to the current ones among the references, which is also embodied in indexing terms. Based on the three aspects above, the fitness function is designed below:

$$fitness(d_i)=lg\frac{\sum_{j=1}^{m}|w_j|*C(d_i)}{T(d_i)} \quad (1)$$

The formula includes three parts, where $|w_j|$ is the frequency of indexing term w_j that belongs to reference d_i, $C(d_i)$ is the citation number of reference d_i, and the time distance between the publication date of d_0 and its reference d_i.

D. Evolution Process

In GA, the population keeps evolving until the termination conditions are met. The population halves in each round. And the evolution process stops until there is only one individual left in the population. That is, the individuals of the current generation are paired off, and one of the next generation will be produced by two of the current generation. Hence, two questions should be answered in each evolution from the current generation to the next generation: first, how to make the individuals be paired off? Second, how to produce the individuals of next generation with the paired individuals of the current generation?

For the first question, we believe that the individuals that are relevant on indexing term should be paired off, so that the "better genes" have more chance to be inherited to the next generation. To this goal, an undirected graph model, indexing-term graph(ITG), is proposed to facilitate the pairing process. ITG is constructed as follows: for each distinct reference d_i, there exists a unique vertex $v\in V$. An undirected edge $e_{ij}\in E$ i.f.f. v_i and v_j share common indexing terms. For each e_{ij}, a number is assigned which is computed below

$$weight(e_{ij})=\sum|w_{ij}| \quad (2)$$

where w_{ij} is the shared indexing terms of d_i and d_j. By characterizing the references of one paper using ITG and the fitness, the algorithm of the evaluation for the initial generation P to the final generation P' is described in Fig. 1.

This is an iterative process. In each round, the next generation is produced according to the selecting operator and the individual fitness. The whole process stops until there is only one individual left in the current generation. The basic idea is described as follows. Firstly, undirected weighted graph G is constructed using the shared indexing terms between references during each iteration. And then, some references are selected to produce the

```
Population Evolution Algorithm
Input: the initial population P
Output: the final population P'   //only one individual in P'
1     if only one individual in P
2         P'=P, and the algorithm ends
3     else perform steps 4-17
4     construct ITG using P
3     if V is not null, perform steps 5-17
4         select vertex $v_x$ with the max fitness from ITG
5         select vertex $v_y$ that has the max weight with $v_x$
6       initialize an individual $d_{xy}$ as the next generation $d_x$ and $d_y$;
7       for each indexing term $w_j$ perform steps 7-
8         determine whether $w_j$ should be in $d_{xy}$ with probability $P(w_i)$
9         put $d_{xy}$ into P'
10        remove $v_x$ and $v_y$ from V
11            remove $e_{xy}$ from E
12        remove $d_x$ and $d_y$ from P
13      if P is not null
14        go to step 2
15      else
16        P=P'
17        go to step 1
18     End
```

Fig. 1. Population Evolution Algorithm

next generation according to the connection relation between references in G. As a result, two crucial problems have to be addressed: the construction of undirected weighted graph G and the selection of the next generation. The two issues will be discussed respectively.Using the shared indexing terms between references, the weighted undirected graph G is constructed as follows. V is the vertex set of G, and each vertex *v* in V corresponds to one conference. E is the edge set of G, and there is an edge two references if they shared indexing terms. The weight of one edge is the total number of the shared indexing between two references if they shared indexing terms. The weight of one edge is the total number of the shared indexing terms. To assure the fitness of the next generation, the higher-fitness individuals in the current generation should be selected to produce the next generation. In the algorithm in Fig. 1, the individual v_x with max fitness is selected from the left ones (see step 3), and then, the one v_y which is most similar to v_x is selected (see step 4). In step 5, the "good" genes have more probability to be passed to the next generation.

V. Experiments

To evaluate the proposed approach, a prototype system AIRPUR(**A**utomatic **I**ndexing for **R**esearch **P**aper **U**sing **R**eferences) is developed in C++ by modifying the open source code "Genetic Algorithm Library" downloaded from Website "Code Project"(www.codeproject.com). The key components "individual coding", "fitness function" and "evolution process" of the proposed approach replace the original parts of the open source code. AIRPUR runs on the computer with Intel Core 2 Duo, 2G memory and 500G hard disk.

The comprehensive experiments are conducted over four subjects. We first show the dataset for our experiments, and then present the experiments and discuss for them.

A. Dataset

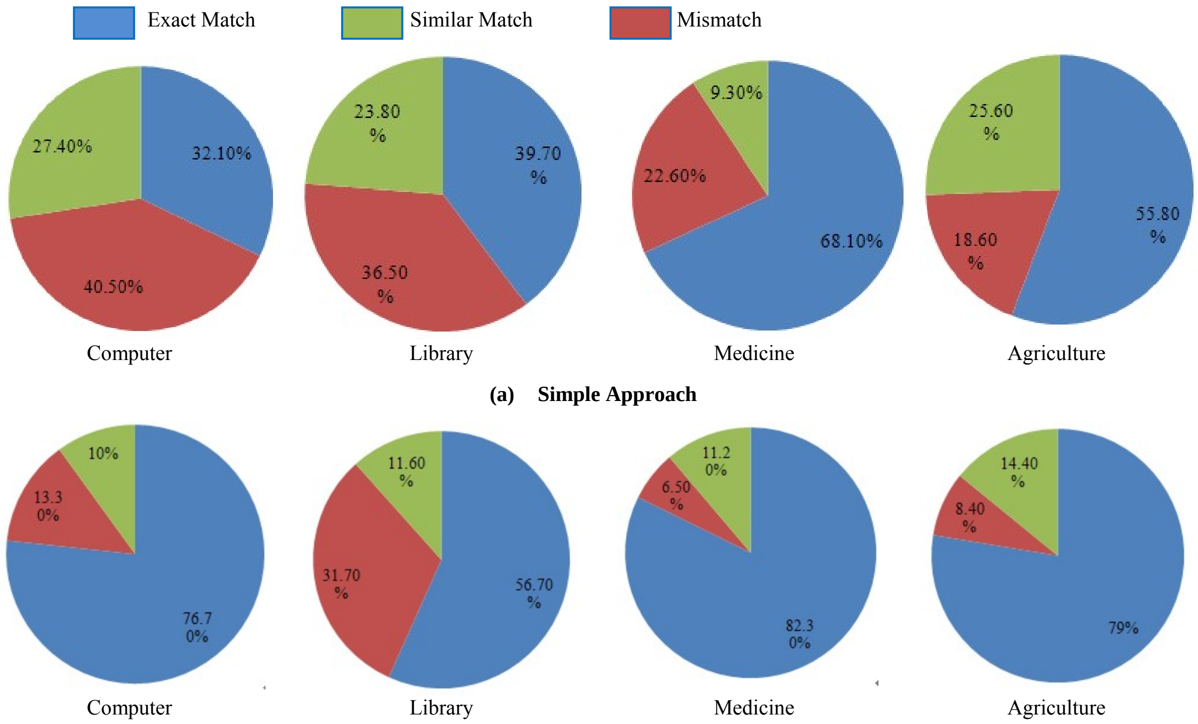
The dataset contains 40 research papers and their references from four subjects which are computer, library, medicine and agriculture. The 40 research papers are averagely selected from the 4 subjects and randomly selected between 2006 and 2010. All the papers and their references have been assigned 6 indexing terms by experts.

B. Evaluation Measures

To qualify the performance, the comparison results are classified into three parts: exact match, similar match and mismatch. In the experiments, we conducted evaluations in terms of exact match, similar match, and not match. Suppose *a* is an automatic indexing term and *b* is a manual indexing term, *a* and *b* being exact match means they are exactly the same, *a* and *b* being similar match means they are synonyms, and *a* and *b* being not match means they are fully different on semantic.

C. Performance Evaluation

We evaluated the performance of the two approaches on the dataset. The experiments are conducted over four subjects respectively, and the results are checked and compared with the indexing terms assigned by experts. By means of the evaluation measures, the experimental results are shown in Fig. 2 and Fig. 3.



(a) Simple Approach

(b) GA Appraoch

Fig. 2. The experimental results over four subjects

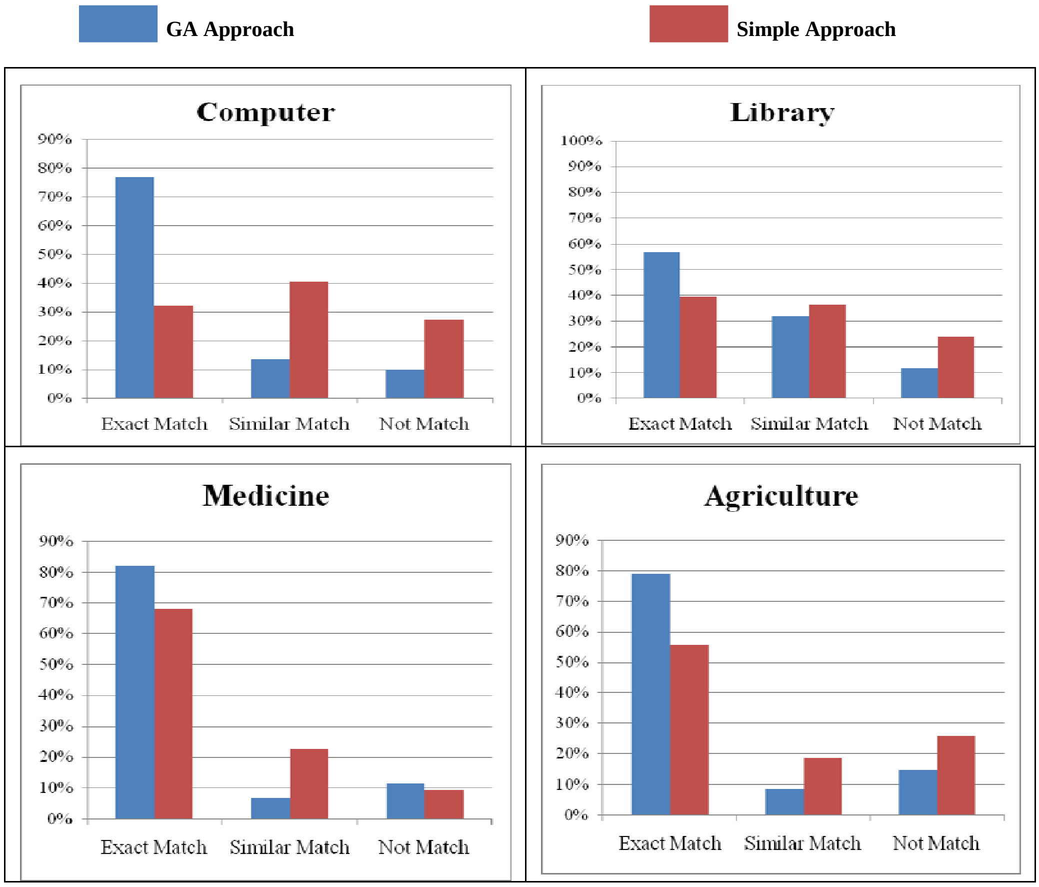


Fig. 3. Comparison between GA Appraoch and Simple Approach

Through comparing the experimental results in Fig. 2 and Fig. 3, we can reveal several interesting phenomena. First, the performance on “extract match” of the simple approach is very poor over most subjects except

Medicine. We also find that the performance on “extract match” of Simple Approach is good enough, and analytical results indicated that the indexing terms in Medicine are more formal and stricter than those in the other three subjects. Second, the performance of GA Approach is much better than that of Simple Approach, especially on “extract match”. This indicates simple statistics on frequency is not effective and robust enough; while using the idea of genetic selection, the performance can be greatly improved. Third, for GA approach, the “exact match” instances make up the majority(74% of average), which is much more than the “not match” instances(12% of average) and “similar match”(15% of average). This means GA algorithm performs better on overall performance. Forth, for GA approach, the experimental result on “exact match” in Medicine is much better than those over other three subjects. Fifth, the experimental result on “not match” in Library is much worse than those over other three subjects. From the comparisons among the four subjects, we think that the terms in science are more rigorous and clear than those in arts, which lead to fewer synonyms under each concept.

VI. Conclusion

Auto indexing is always a crucial issue in the field of digital library. However, most previous approaches suffer from the serious limitations by extracting indexing terms from title, abstract, main body, and so on, which. In this paper, a new auto indexing approach is proposed for research papers by selecting indexing terms from those of the references. In the approach, the genetic algorithm is improved to select appropriate indexing terms as the ones of the research paper. The experimental results over 4 subjects show the effectiveness of the proposed approach, and the performance difference among the 4 subjects is also discussed.

Acknowledgment

This work is partially supported by National Key Technology R&D Program Grant #2011BAH10B01 and the Advanced Research Program of ISTIC, Grant #1746.

References

[1] X. Jiang, Y. Hu, H. Li. A ranking approach to keyphrase extraction[C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009: 756-757.

[2] X. Wan, J. Xiao. Exploiting neighborhood knowledge for single document summarization and keyphrase extraction[J]. ACM Transactions on Information Systems (TOIS), 2010, 28(2): 8.

[3] S. N. Kim, O. Medelyan, M. Y. Kan, et al. Automatic keyphrase extraction from scientific articles[J]. Language resources and evaluation, 2013, 47(3): 723-742.

[4] X. Wan, J. Xiao. Single Document Keyphrase Extraction Using Neighborhood Knowledge[C]. In Proceedings of AAAI. 2008, 8: 855-860.

[5] G. Ercan, I. Cicekli. Using lexical chains for keyword extraction[J]. Information Processing & Management, 2007, 43(6): 1705-1714.

[6] Z. Liu, X. Chen and M. Sun. Mining the interests of Chinese microbloggers via keyword extraction[J]. Frontiers of Computer Science in China, 2012, 6(1): 76-87

[7] S. Marinai, B. Miotti and G. Soda. Digital Libraries and Document Image Retrieval Techniques: A Survey. Studies in Computational Intelligence, 2011, 375: 181-204

[8] W. Zhao, J. Jiang, J. He, et al. Topical keyphrase extraction from twitter[C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 379-388.

[9] F. Ferrara, C. Tasso. Extracting Keyphrases from Web Pages[C]. In Proceedings of Digital Libraries and Archives. Springer Berlin Heidelberg, 2013: 93-104

[10] W. You, D. Fontaine, J. P. Barthès. An automatic keyphrase extraction system for scientific documents[J]. Knowledge and information systems, 2013, 34(3): 691-724.

[11] W. Yih, J. Goodman, V. R. Carvalho. Finding advertising keywords on web pages[C]. In Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 213-222.

[12] C. Chen, Z. Jiun, C. Hsu. An Adaptive Reversible Image Watermarking Scheme Based on Integer Wavelet Coefficients[J]. Journal of Computers, 2013, 8(7).

[13] http://www.nstl.gov.cn/

[14] J. D. Cohen. Highlights: Language and Domain-independent Auto Indexing Terms for Abstracting[J]. Journal of American Society for Information Science. 1995, 46(3): 162-174

[15] K. Barker and N. Cornacchia. Using Noun Phrase Heads to Extract Document Keyphrases[C], In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, 2000: 40–52.

[16] Y. Matsuo, M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information[J]. Journal of Artificial Intelligence Tools, 2004, 3(1): 157-169.

[17] L. Chien. PAT-tree-based keyword extraction for Chinese information retrieval[C]. SIGIR 1997: 50-59.

[18] Ercan G, Cicekli I. Using lexical chains for keyword extraction[J]. Information Processing & Management, 2007, 43(6): 1705-1714.

[19] C. Zhang, H. Wang, Y. Liu. Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal of Computational Information Systems. 2008, 4(3):

[20] T. D. Nguyen, M. Kan. Keyphrase Extraction in Scientific Publications[C]. In Proceedings of ICADL 2007: 317-326

[21] J. Zhao W. Li. Intrusion Detection Based on Improved SOM with Optimized GA[J]. Journal of Computers, 2013, 8(6).

[22] D. Liu, X. Chen, J. Du. A Hybrid Genetic Algorithm for Constrained Optimization Problems[J]. Journal of Computers, 2013, 8(2).

Wei Liu was born in Shandong Province of China. He received the M.S. degree in Computer Science from School of Computer Science and Technology at ShanDong University in 2004, and the Ph.D. degree in Computer Science from School of Information at Renmin University of China in 2008. He did his Postdoc research with Prof. Jianguo Xiao in Institute of Computer Science & Technology at Peking University. His current research interests include Web data extraction, Deep Web data integration, and Science and technology information retrieval.

He is currently an associate research fellow at the information resource centre, Institute of Scientific and Technical

Information of China. He has published over 20 papers in some reputational international journals, such as IEEE Transactions on Knowledge and Data Engineering, Journal of Intelligent Information Systems, etc.

P4P Network Communication Components Based on Half-Sync/Half-Async and Pipe/Filter Patterns

[1,2] Cheng Wang
[1]College of Computer Science and Technology, HuaQiao University, Xiamen 361021, China
[2] Xi'an Jiaotong University, Xi'an 710049, China
wangcheng@hqu.edu.cn

Zhicong Liang
boxlzc@gmail.com

Abstract—**This paper describes P4P(Proactive network Provider Participation for peer-to-peer) network server based on the Half-Sync/Half-Async and Pipe/Filter design patterns, which implements the requirements of the P4P system. The P4P network server applies the Half-Async layer to listen to the specified network port and establishes network connections asynchronously; makes use of message queue layer to buffer established network connections; applies the Pipe/Filter pattern into the Half-Sync layer and takes the Half-Sync layer to receive data and send data concurrently. Thanks to these patterns and the design, it gains various levels of concurrency and flexibility.**

I. Introduction

With the rapid development of P2P(peer-to-peer) networks, some new modules and protocols are used to construct P2P systems, some people propose new architectures based on P2P called P4P to provide more effective cooperative traffic control between applications and network providers. As peer-to-peer (P2P) emerges as a major paradigm for scalable network application design, it also exposes significant new challenges to achieve efficient and fair utilization of Internet network resources[1]. To improve the feasibility, concurrency and effectiveness of distributed systems, more and more new architectures and modules will be proposed. P4P systems developed from and based on P2P systems. Consequently, P4P systems are becoming an important application of distributed software systems, and the researches on it are being of great significance. On account of the development of information technology, computers, mobile phones and various media terminals will continue to emerge, how to make these different terminals interact with each other is difficult and filled with challenges. How to resolve this difficulty depends on the development of the network communication components in these systems. As the diversity of operating systems and communication platforms, communications software developers often have to face so many problems as the performance of communication, the management of code, platform coverage, and so on. P2P applications may face various of requirements and challenges, there are many related researches, such as [2], [3], [4], [5], [6] and [7] try to conquer these challenges, in which some effective methods and systems are proposed to conquer these challenges. There are many design patterns used for different application fields, including classic design patterns[8] and distributed applications related design patterns[9][10][11][12][13], which help us conquer the design challenges. Some are used to create a specific types of software, such as the pattern languages for networked and concurrent computing[10] and enterprise application architectures[14]. The Half-Sync and Half-Async is one of these key patterns[8] in the domain of communication software. There are many common concurrency models for network server as follows:

A. Thread-Per-Connection Model.

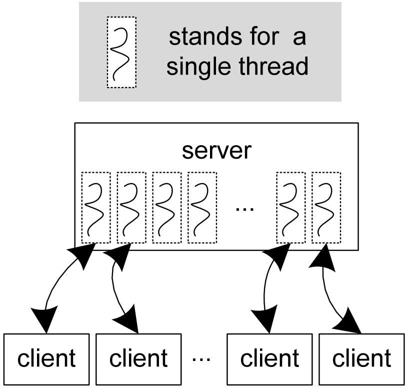


Figure 1. thread-per-connection model.

As is shown in figure 1, in this model, a new thread will be created and associated with the new connection request when a new connection request arrives. The new created thread is responsible for establishing connections, receiving network data, handling network data, sending result to clients, error handlings and shutdown the connection. This model takes the following disadvantages:

(a.) does not separate the business logic and network logic.

(b.) a huge number of threads will be created when a lot of networks connections arrive simultaneously, which costs much system resources and memory, the system may crash in the worst case.

doi:10.4304/jsw.9.11.2830-2836

(c.) developers are front with challenges from code maintain and business changes.
(d.) if all the threads were created in one process, the whole process will crash when any one of these threads crashes.

B. Process-Per-Connection Model.

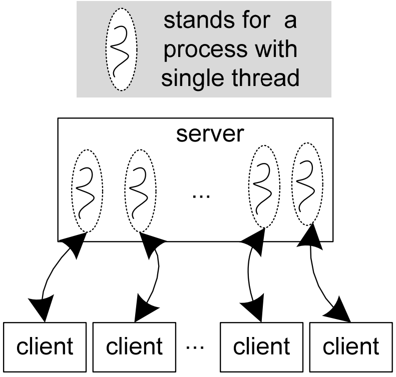


Figure 2. process-per-connection model.

As is shown in figure 2, this model is similar with the thread-per-connection model, but processes instead of threads are created when network connection requests arrived. This model takes the following disadvantages:
(a.) does not separate the business logic and network logic either.
(b.) a huge number of processes will be created when a lot of networks connections arrive simultaneously, which costs a lot of CPU time, system resource and memory, the system may crash when system load becomes bigger and bigger.
(c.) developers are front with challenges from code maintain and business changes.
(d.) frequent IPC(Inter-Process Communication) will happen, which costs much system resource and leads to bad performance.

C. Thread-pool Model.

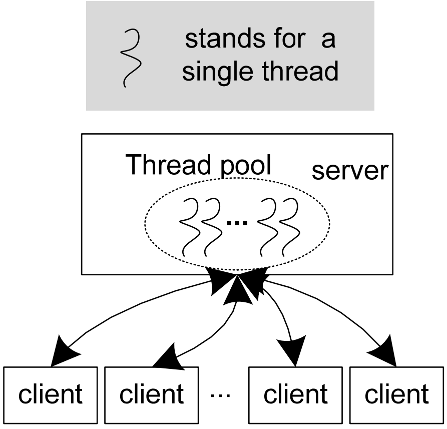


Figure 3. thread-pool model.

As is shown in figure 3, in this model, network server creates a fixed number of threads before connection requests arrive. The server selects an idle thread from thread-pool to provide services to clients when a new connection request arrives. In this model, the number of threads is fixed, one thread may provide services to different network connections during different time. This model costs fixed system resources, will not bring huge load to the system, but it is not able to handle all the network connections when the number of concurrent network connections is more than the number of threads in the thread pool, which makes some network connection requests blocked for a long time.

D. Leader/Follower Model[10].

The Leader/Followers architectural pattern provides an efficient concurrency model where multiple threads take turns to share a set of event sources in order to detect, demultiplex, dispatch, & process service requests that occur on the event sources[10]. There is no message queue in this model, it is not able to handle a lot of current network connections simultaneously.

E. Half-Sync/Half-async Model[10].

The Half-Sync/Half-Async architectural pattern decouples asynchronous and synchronous service processing in concurrent systems, to simplify programming without unduly reducing performance. The pattern introduces two intercommunicating layers, one for asynchronous and one for synchronous service processing[10]. This model includes three level layers, Half-Sync layer, message queue layer and Half-Async layer, which is shown in figure 4 as follows:

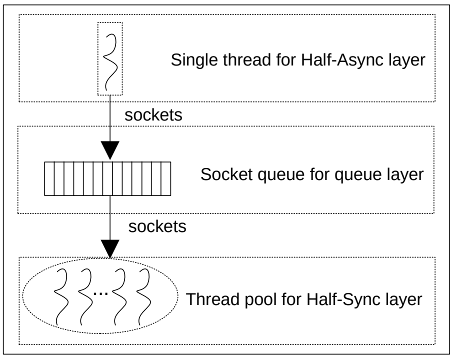

Single thread for Half-Async layer
sockets
Socket queue for queue layer
sockets
Thread pool for Half-Sync layer

Figure 4. Half-Sync/Half-Async architectural pattern[10].

Each layer`s responsibilities are as follows:
(a.) Half-Async Layer. This layer is responsible for handling asynchronous network connection requests from clients and establishing connections. Establishing a connection is a quick operation, which will not block other tasks or reduce the performance of the system. The concurrency strategy in this layer is various, including single thread strategy, thread pool strategy and so on. We choose single thread strategy here.
(b.) Queue Layer. This layer is responsible for buffering requests and provides communication mechanism between Half-Async Layer and Half-Sync Layer. This layer separates and decouples Half-Async Layer and Half-Sync Layer, which makes the strategies of these three layers independent and flexible.
(c.) Half-Sync Layer. This layer is responsible for handling requests in the queue layer. Generally speaking, there is a thread pool in this layer to handle

requests concurrently.

This model takes the following benefits:

(a.) higher-level tasks are simplified[10].

(b.) business logic and network logic is separated, which makes the design more flexible.

(c.) synchronization policies in each layer are decoupled[10].

(d.) inter-layer communication is localized at a single point[10].

(e.) performance is improved on multi-processors[10].

(f.) code maintain and business changes are easy to handle.

Each model owns its advantages and disadvantages, after comparing all the above models, we choose Half-Sync/Half-Async to build the P4P systems. With these advantages of Half-Sync/Half-Async pattern, it is easy to build a flexible network communication server with high performance and various levels of concurrency.

II. Requirements In The P4P System

The requirements of network communication components in the P4P system are as follows:

(1.) sending and receiving data.

(2.) good expansibility, concurrency and flexibility.

(3.) the server owns the capabilities to handle thousands of network connections concurrently.

(4.) implementing dynamic and exchangeable data handling flow to provide sufficient flexible protocol parsing strategies to fit for various requirements and businesses in the P4P system.

We should design the p4p systems flexible enough with high performance and various level of concurrency to fit all the requirements above, so it is very important to choose appropriate patterns to design the architecture.

III. Architecture And Design

When designing the network communication components of the P4P system, concurrency is a very important factor. How to maximize the number of concurrent connections is a problem front with developers.

As we stated above, Half-Sync/Half-Async model takes several advantages to bring good concurrency and performance. But this pattern is not flexible enough in front of the requirements of the P4P systems, especially for the Half-Sync layer. Because the data handling business in the P4P system is very complex and various, there are many different protocols need to parse. How do you implement a sequence of transformation modules so that you can combine and reuse them independently[15]? Fortunately, Pipe/Filter[16] pattern is designed to resolve this problem, which is shown in figure 5. This pattern requires the following[15]:

(1.) The output of the data source must be compatible with the input of filter 1.

(2.) The output of filter 1 must be compatible with the input of filter 2.

(3.) The output of filter 2 must be compatible with the input of the sink data.

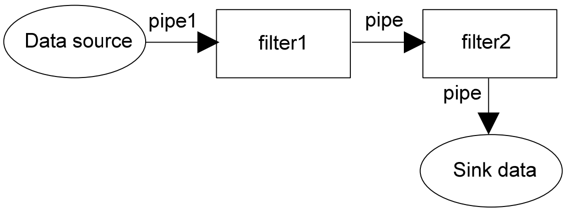


Figure 5. Pipe/Filter Pattern.

We use the Half-Sync/Half-Async and Pipe/Filter patterns to design and implement these network communication components in the P4P system.

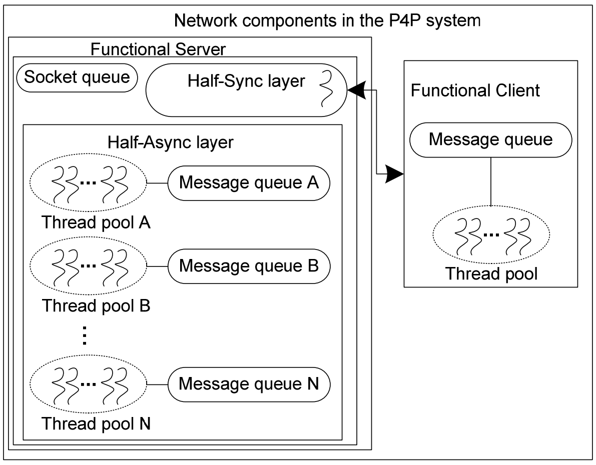


Figure 6. Architecture of network communication components.

As shown in figure 6, the network communication components in the P4P system includes two parts: one is Functional Server, the other is Functional Client. Functional Server provides functions such as monitoring the specified port, maintenance of passive connections, receiving, buffering, handling and sending data, while Functional Client provides functions such as initiating active connections, maintenance of active connections, receiving and sending data. Functional Server use a single thread to listen to the specified network socket and put the sockets into socket queue; creates several thread pools and corresponding message queues to buffer data. The Pipes and Filters architectural pattern divides the task of a system into several sequential processing steps[16]. In the P4P system, each sequential processing step is a data handling module which contains a thread pool and a message queue. Functional Client creates a thread pool to handle messages and creates a message queue to buffer data.

Data handling work flow

This section describes how to design the data handling work flow framework of the Functional server. This framework implements the Pipes and Filters pattern[16].

In the front of the requirements from current P4P applications, there are more and more data handling requirements appears, the general static data handling work flow is not flexible and robust enough. For example, a static and complex network data handling flow is shown in figure 7, if we wanted to add or delete some data handling module, we have to stop the system firstly, rewrite the whole data handling work flow or do much changes and rebuild the source code statically and then re-launch the system, which will cost us much time and make system unstable.

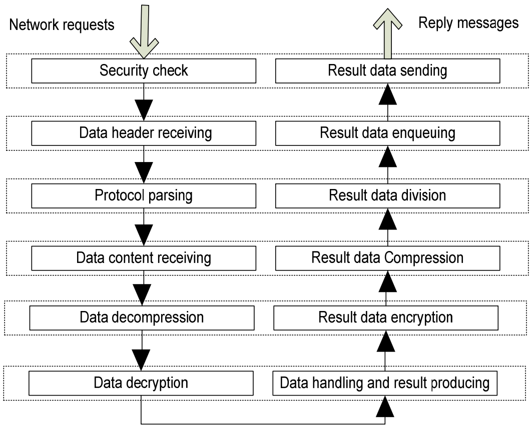


Figure 7. A complex network data handling flow.

So with the help of the Pipes and Filters pattern, we build a more flexible and dynamic data handling work flow model. In this model, we use a list to link all the data handling modules sequentially, when a new module is needed to add into the work flow, just inserts the new one into the list. Each module owns its private threads pool to handle data, owns its message queue to buffer data, which makes the whole system more flexible and gains good performance. The new data handling work flow is shown in figure 8.

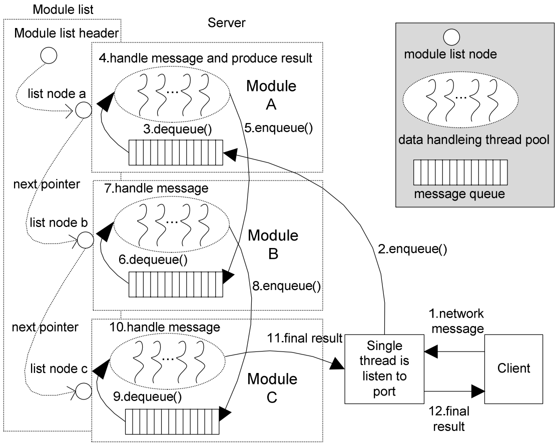


Figure 8. data handling work flow.

The whole procedure of data handling work flow is as follows:

(1.) The single thread listens to the specified network port, and establishes connections when connection request arrive.
(2.) The single thread which listens to the port puts network connections into the queue of module A.
(3.) Module A gets an item from its own queue.
(4.) Thread pool of Module A handles the item and gets the result.
(5.) According to the module list, module A finds out its next module which is module B and puts the result into the queue of module B.
(6.) (7.)(8.)(9.) Module B and the following modules will do the similar procedure with module A until the handled result arrives to the tail module.
(10.)(11.)(12.) The tail module of the module list handles the result passed from its previous module, produces the final result and sends it to the client.

Thanks to the design, each module owns its private thread pool, the data handlings are concurrent and independent in both each thread pool and each module. Each data handling module is a list node, when some new module is needed to add into the data handle flow during the system is running, it is just needed to insert this new module into the list, which is shown in figure 9 as follows. A lock is hold to protect the list during the inserting. Deleting a module is in the backward procedure. Because adding or deleting a module is not a frequent operation and is always done during wee hours, the cost of holding lock is acceptable.

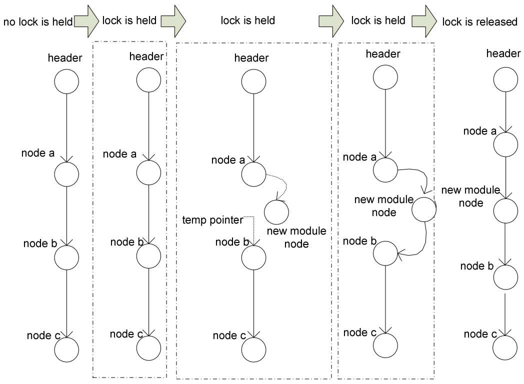


Figure 9. add a new module into data handling flow.

At the same time, to make this system more flexible, each module owns a state to indicate its current state. Each module is in one state of four states: idle, active, running and inactive. The translation among these states is shown in figure 10.

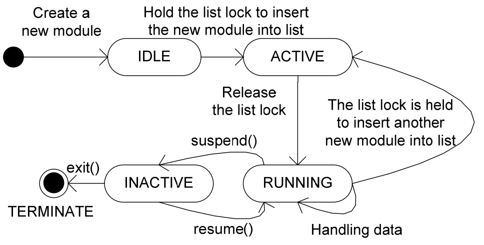

Create a new module
Hold the list lock to insert the new module into list
IDLE
ACTIVE
Release the list lock
The list lock is held to insert another new module into list
suspend()
exit()
TERMINATE
INACTIVE
RUNNING
resume()
Handling data

Figure 10. State translations of data handling modules.

Application Level Binary Exponent Backoff Strategy

As we known, network crowd may happen during the network is busy, which will cost much resources and time. Binary exponent back-off algorithm[17][18] is applied into Ethernet (802.3)[19] to reduce network crowd. Similarly, in network communication system, clients always send connection requests to the server simultaneously, but the handling abilities of server is limited, so it is needed to apply some strategy to coordinate the clients. The application level binary exponent back-off strategy is applied into the functional clients to decrease network crowd in the P4P system. The pseudo code of application level binary exponent back-off pseudo algorithm is shown in figure 11:

```
int backoff_reconnect(int NumberOfReconnect, int initialDelayTime,
string serverAddress) {
  int i, sock;
 ClientConnector connector;
 ClientBuffer buffer;
  int sock=socket(AF_INET, SOCK_STREAM,0);
 ClientTim timeout(initialDelayTime );
for(i = 0; i < NumberOfReconnect; i++) {
      if(i != 0) sleep(timeout);
             if (connector.connect (sock, serverAddress, &timeout) == -1)
 timeout *= 2; /* exponential backoff */
else{
connector.send(sock, buffer);
break;
        }
        }
     return (i == NumberOfReconnect)? -1 : 0;
}
```

Figure 11. Exponent backoff algorithm in clients.

The description of the application level exponential back-off algorithm is as follows:

(1.) Define the basic timeout time, in this communication system the basic timeout is defined by parameter initialDelayTime.
(2.) Define a parameter named NumberOfReconnect, which stands for the times of network reconnections.
(3.) New time out is equals to two multiplied by old timeout, the initial value of time out equals to initialDelayTime.
(4.) If connection still failed after NumberOfReconnect times, reports the error to high level applications of this communication system.
(5.) If connection was established, send messages to network.
(6.) After applying Half/Sync-Half/Async and Pipe/Filter patterns into the system, it fit for all the requirements, which not only gains high performance but also obtains enough flexible.

IV. The Run-Time Procedure Of Network Communication Components

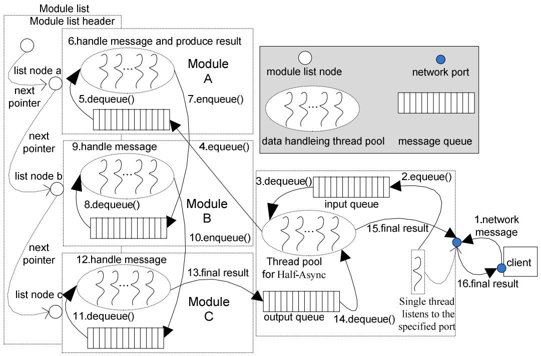


Figure 12. The run-time procedure.

Figure 12 shows the run-time process of network communication component in P4P system. Functional server runs a single thread to listen to the network socket, while each data handling module owns a private thread pool to handle data. The number of threads in each thread pool is specified by parameters to satisfy different performance of different machines. In order to communicate among threads, there is a message queue in each module. The run-time procedure of network communication component in P4P systems are as follows:

(1.) Clients initiate connection requests to the server and then send messages to the server. Both UDP[20] connections and TCP[21] connections are supported in the P4P systems to provide different services to various of businesses and requirements.
(2.) The single thread which listens to the network port detects the connection requests, establishes the connections and put the sockets into the input queue.
(3.) Thread pool for Half-Async gets these sockets items from the input queue, and calls select function to wait for network messages` coming. This function allows the process to instruct the kernel to wait for any one of multiple events to occur and to wake up the process only when one or more of these events occurs or when a specified amount of time has passed[22]. When messages are coming, the thread pool receives messages, handles these messages, and produces the results.
(4.) Thread pool for Half-Async puts the results into the message queue of the first data handling module of the module list, which is Module A in figure 12.
(5.) The thread pool of Module A gets messages from its message queue.
(6.)(7.) The thread pool of Module A handles messages and put the results into the message queue of Module A`s next Module, which is Module B in figure 12.
(8.)(9.)(10.)(11.)(12.) Module B and the following modules will handle the messages from their previous modules similar with Module A does until meets the final module.
(13.) The final module handles the messages, produces and puts the final results into the output queue of the thread pool for Half-Async.
(14.)(15.)(16.) The thread pool for Half-Async gets the final results from its output queue and sends them to clients by corresponding sockets.

From the run-time view of the P4P system, we got a clear understanding about how the system works and how these components coordinate with each other.

V. Performance Testing

We did the testing with the following hardware devices and configurations: Intel i7-4700HQ/4G platform with Ubuntu Server 12.04 and network with the speed of 1000Mps.

Testing in figure 13 shows high performance and advantages of network server based on Pipe/Filter and Half-Sync/Half-Async patterns compared with network server based on thread-per-connection and network server based on thread pool. With the number of network connections increasing, the performance of this system keeps at high level relatively.

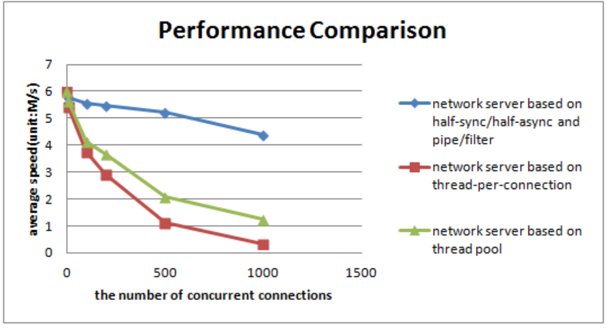


Figure 13. Network performance comparison.

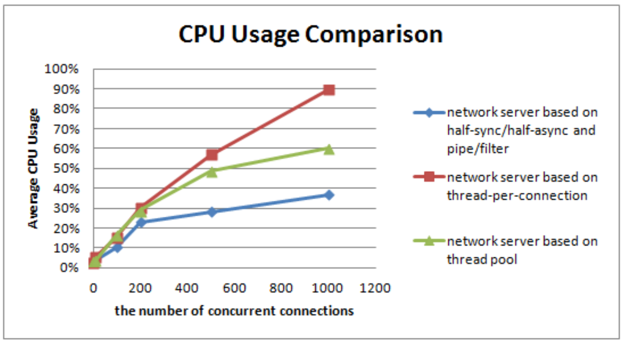


Figure 14. CPU Usage comparison.

As is shown in figure 14, with the increasing of the number of concurrent connections, the average cpu usage of network server based on half-sync/half-async and pipe/filter shows a smooth curve upward relatively, which means costing lower system resource and gains higher performance compared to others.

According to these above testing results, our system gains various levels of concurrency, which better meets the requirements and needs of the P4P system.

Future Work

Our future work will focus on the following aspects:

(1) The optimization of p4p protocols analysis. Because there are all kinds of application level protocols to support in this system, we need to optimize the protocol analysis components.

(2) The optimization of data handling components and flows. As we known, the data handling will cost much time and system resource, so it is necessary to improve the performance of data handling flow and components.

(3) Try to apply other network communication related design patterns into this system, which could improve the flexibility, extension further and make this system more maintainable.

(4) Do more stability related testing. Because the system is complex and will cover both Windows and Linux operating systems, it is necessary and important to do stability related testing.

VI. Conclusions

In this paper, we compared several network server models and stated the disadvantages and advantages of these models firstly. And then, we proposed the architecture of the P4P network communication components based on Half-Sync/Half-Async and Pipe/Filter Patterns. Finally, we give the testing results and related analysis about these models. The network communication components in this system take the Half-Sync/Half-Async pattern framework as its core framework, applies the Pipe/Filter pattern framework as its auxiliary framework and organizes all the data handling modules into a list to add and delete modules dynamically. Thanks to these frameworks and designs, it gains various levels of concurrency and flexibility which fits for the requirements of the P4P system.

Acknowledgements

This work was financially supported by National Natural Science Foundation of China (Grant No. 51305142) and introduction of talents Huaqiao University Scientific Research Projects (Project No. 12BS217).

References

[1] Haiyong Xie,Y. Richard Yang, Arvind Krishnamurthy,Yanbin Liu,Avi Silberschatz. P4P: Provider Portal for Applications.SIGCOMM'08, August 17–22, 2008, Seattle, Washington, USA.

[2] Adeela Bashiry, Sajjad A. Madaniy, Jawad Haider Kazmiy, Kalim Qureshi. Task Partitioning and Load Balancing Strategy for Matrix Applications on Distributed System, JOURNAL OF COMPUTERS, VOL. 8, NO. 3, MARCH 2013.

[3] Jinghua Wu,Yun Xu. A Decision Support System for Borrower's Loan in P2P Lending, JOURNAL OF COMPUTERS, VOL. 6, NO. 6, JUNE 2011.

[4] Zhengzhen Zhou, Yonglong Luo, Liangmin Guo, Meijing Ji. A Trust Evaluation Model based on Fuzzy Theory in P2P Networks, JOURNAL OF COMPUTERS, VOL. 6, NO. 8, AUGUST 2011.

[5] Choffnes, D. and F. Bustamante, "Taming the Torrent: A practical approach to reducing cross-ISP traffic in P2P systems", Proceedings of ACM SIGCOMM, August 2008.

[6] R. Bindal, P. Cao, W. Chan, J. Medval, G. Suwala, T. Bates and A. Zhang, "Improving Traffic Locality in BitTorrent via Biased Neighbor Selection". In IEEE International Conference on Distributed Computing System (ICDCS 2006).

[7] K. Shanahan and M. Freedman, "Locality Prediction for Oblivious Clients". International workshop on Peer-To-Peer Systems (IPTPS 2005).

[8] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software. Reading, MA: Addison-Wesley, 1995.

[9] Frank Buschmann, Regine Meunier, Hans Rohnert, Peter Sommerlad, and Michael Stal. Pattern-Oriented Software Architecture, A System of Patterns, Volume 1. Wiley and Sons, New York, 1996.

[10] Douglas C. Schmidt, Michael Stal, Hans Rohnert, and Frank Buschmann. Pattern-Oriented Software Architecture: Patterns for Concurrent and Networked Objects, Volume 2. Wiley & Sons, New York, 2000.

[11] Michael Kircher, Prashant Jain. Pattern-Oriented Software Architecture: Patterns for Resource Management, Volume 3. Wiley in 2004.

[12] Frank Buschmann, Kevlin Henney, Douglas C. Schmidt. Pattern-Oriented Software Architecture: A Pattern Language for Distributed Computing, Volume 4.Wiley & Sons in 2007.

[13] Frank Buschmann, Kevlin Henney, Douglas C. Schmidt. Pattern-Oriented Software Architecture: On Patterns and Pattern Languages, Volume 5. Wiley & Sons in 2007.

[14] M. Fowler, D. Rice, M. Foemmel, E. Hieatt, R. Mee, and R. Stafford, Patterns of Enterprise Application Architecture. Reading, Massachusetts: Addison-Wesley, 2002.

[15] http://msdn.microsoft.com/en-us/library/ff647419.aspx

[16] Frank Buschmann, Regine Meunier, Hans Rohnert, Peter Sommerlad, and Michael Stal. Pattern-Oriented Software Architecture, A System of Patterns, Volume 1. Wiley and Sons, New York, 1996.

[17] Larry L. Peterson, Bruce S. Davie, Computer Networks, Edition 4: A Systems Approach, pp.116-123, 2007, Elsevier, Inc.

[18] Kevin R. Fall, W. Richard Stevens. TCP/IP Illustrated, Volume 1: The Protocols, pp.114-116, 2012, Addison Wesley.

[19] "IEEE Standard 802.3-2008". IEEE. Retrieved 22 September 2010.

[20] RFC768; Postel, Jon. User Datagram Protocol, IETF, August 1980.

[21] RFC793; Postel, Jon. Transmission Control Protocol, IETF, September 1981.

[22] W. Richard Stevens, UNIX Network Programming Volume 1, Third Edition: The Sockets Networking API. pp.209-121, 2003, Addison Wesley.

Algorithms for Minimal Dependency Set and Membership Based on XML Functional Dependency and Multi-valued Dependency

Zhongping Zhang[1,2], Chunzhen Fang[1]
[1]The School of Information Science and Engineering, Yanshan University,
Qinhuangdao, Hebei, 066004, China
[2]The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province,
Qinhuangdao, Hebei, 066004, China
Email: zpzhang@ysu.edu.cn, fcz_chunzhen@163.com

***Abstract*—As the XML functional dependency and multi-valued dependency impact on the normalization design of semi-structured data, definitions of XML functional dependency, XML multi-valued dependency, path dependency base and the minimal dependency set are given in this paper. Algorithms for minimal dependency set and membership with path expression based on the coexistence of XFD and XMVD are then proposed. Finally, the correctness and termination of these algorithms are proved, and their time complexities are analyzed as well.**

***Index Terms*—XML functional dependency, XML multi-valued dependency, path dependency base, membership, minimal dependence set**

I. Introduction

XML (eXtensible Markup Language) inherits the powerful function of SGML(Standard Generalized Markup Language)[1] and makes up the deficiencies of HTML. It is a standard which is used for data expression and data exchange on the Internet, providing an effective means for data description and data exchange on the Internet applications. It is widely used in many fields [2-5]. XML schema is an important concept in the field of XML and it is the first step to build database applications. Currently, DTD develops well, and it is widely used in the practical applications of the XML document. However, because of some unusual data dependencies in XML database, it may result in data redundancies and abnormal operations due to design flaws for DTD [6].

In recent years, scholars have done a lot of exploration and research on the normalization of XML database: the normalization based on functional dependency [7-12] and multi-valued dependency [13-15]. These research literatures analyze the effect on XML data and free redundancy from a single point of view such as XML functional dependency or XML multi-valued dependency. However, they do not consider the effect on XML data and XML database normalization on the condition of coexistence of XML functional dependency and multi-valued dependency. Therefore, according to the inference rules based on the coexistence of XML functional dependency and XML multi-valued dependency [16], we propose algorithms for minimal dependency set (DEP-MINIMIZE algorithm) and membership (DEP-MEMBERSHIP algorithm) with path expression based on coexistence of XML functional dependency and multi-valued dependency. It simplifies the dependency set and ensures the simplification on analysis and calculation.

II. Preliminary Definitions And Notations

In this section, we present some preliminary definitions and notations that we need.

Definition 1: (XML Tree) An XML tree is defined as T=(V,lab,ele,att,val,root), it is said to conform to a DTD[9] D=(E_1,E_2,A,P,R, r), denoted by T⊨D[17], where

(i) T is the XML tree's name;

(ii) V is a finite set of nodes in T;

(iii) lab is a function from V to $E_1 \cup E_2 \cup A$,which assigns a identifier to each node in V. A node v in V is called a complex element node if lab(v)$\in E_1$; a simple element node if lab(v)$\in E_2$, and an attribute node if lab(v)$\in$ A;

(iv) ele is a function from V to a sequence of V nodes ,so that for any $v\in$ V, if lab(v)$\in E_1$, ele(v) is a set of some children of v. The node in ele(v) is element node, and if ele(v)=$\{v_1,...,v_m\}$, then {lab(v_1),...,lab(v_m)}$\in$ P(lab(v));

(v) att is a function from V to A. If att(v, l)=v_1 ,then lab(v)$\in E_1$ and lab(v_1)=l; if att(v)=$\{v_1,...,v_n\}$, then{lab(v_1) ,...,lab(v_n)}$\in$ R(lab(v)), where $v\in$ V, $l \in$ A;

Manuscript received December, 1, 2013; revised March 24, 2014; accepted April 21,2014.

doi:10.4304/jsw.9.11.2837-2843

(vi) val is a function that assigns a value to each node. If a node v is a leaf node or a simple element node of T, val(v) is a string value which is either the content of a text element or the content of an attribute; otherwise val(v) is the node's identifier of v.

(vii) root is the unique root node labeled with complex element name r.

Example 1: Describe a college's relationship among three entities Course, Student and Teacher and store data information in Relational database. Relational schema R of database is designed as following:

Course(Cno, Cname) Student(Sno, Sname)
Teacher(Tno, Tname)
Courses(Cno,Sno, Tno,Cname, Sname, Tname)

If the relational database need to be stored as an XML document, this document's DTD D is expressed as:

```
<!ELEMENT Courses (Course*)>
<!ELEMENT Course (Cname,Student*)>
    <!ATTLIST Course
Cno CDATA #REQUIRED>
<!ELEMENT Cname (#PCDATA)>
<!ELEMENT Student (Sname,Teacher)>
    <!ATTLIST Student
Sno CDATA #REQUIRED>
<!ELEMENT Sname (#PCDATA)>
<!ELEMENT Teacher (Tname)>
        <!ATTLIST Teacher
Tno CDATA #REQUIRED>
<!ELEMENT Tname (#PCDATA)>
```

There are some data instances in TABLE I:

TABLE I.
A PART OF STUDENTS' COURSE RECORDS OF ONELEGE COL

Cno	Cname	Sno	Sname	Tno	Tname
C01	XML	S001	Amily	T001	Mary
C01	XML	S002	Stephen	T001	Mary
C01	XML	S001	Amily	T002	Anne
C01	XML	S002	Stephen	T002	Anne

According to this course records, we can obtain an XML tree T which based on the DTD D. It is shown in Figure 1.

Definition2: (Path Instance on XML Tree) Let T be an XML tree that satisfied the given DTD D. A path instance [7] over an XML tree T is a sequence, $v_1 = v_r$ and for every v_i, $2 \le i \le n$, $v_i \in V$ and v_i is a child of v_{i-1}. A path $v_1.v_2. \ldots.v_{n-1}.v_n$ is defined as one path instance based on the path $p_1 .p_2 . \ldots .p_{n-1} . p_n$ if for all v_i, $1 \le i \le n$, lab(v_i)=p_i . All path instances based on a path p over a tree T form a set which denoted by Paths(p). All Path sets on D are defined as the Paths (D).

Example 2: As Figure 1 shows that, $v_1.v_2.v_4$ is a path instance of path Courses.Course.Cno, Paths(Courses.Course.Cno)={$v_1.v_2.v_4$, $v_1.v_3.v_6$}.

Definition 3:(XFD) Let T be an XML tree that satisfied the given DTD D. An XFD is a statement of the form $p_1, \ldots , p_k \rightarrow q_1 , \ldots ,q_m$, $k \ge 1, m \ge 1$. P={$p_1 ,. . .,p_k$ } and Q={$q_1, \ldots , q_m$ } are subsets in Paths(D).There are arbitrary two distinct path instances V={ $v_1^1 . v_2^1 . \ldots . v_{l-1}^1 . v_l^1 , \ldots , \ v_1^m . v_2^m . \ldots . v_{n-1}^m . v_n^m$ }

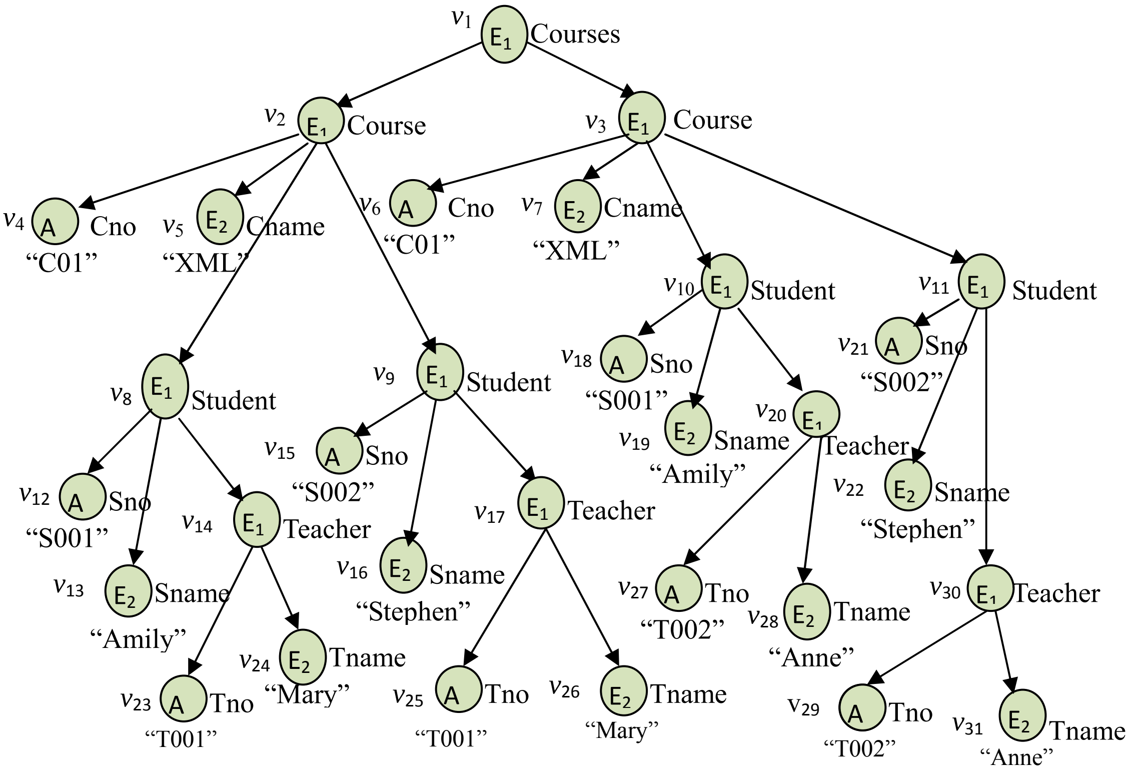


Figure 1. An XML tree formed from a part of students' course records

and W={ $w_1^1 . w_2^1 . \dots . w_{l-1}^1 . w_l^1$, … , $w_1^m . w_2^m . \dots . w_{n-1}^m . w_n^m$ } in Paths(Q), n≥1, l≥2, T satisfies the XFD: $p_1,\dots,p_k \to q_1,\dots,q_m$,where

(i) Q⊂P; or

(ii)For any two distinct path instances $v_1^i . v_2^i . \dots . v_{t-1}^i . v_t^i$ and $w_1^i . w_2^i . \dots . w_{t-1}^i . w_t^i$ in Paths(q_i), if Last(p_j)[16] ∈ E_1 and x_{ij} =y_{ij} , or Last(p_j) ∉ E_1 and val(Nodes(x_{ij} , p_j)[16])∩val(Nodes(y_{ij}, p_j)) ≠∅ then val(v_t^i)=val(w_t^i); where x_{ij}={v|v∈{ v_1^i , …, v_t^i } ∧ v∈N(p_j∩q_i)},y_{ij}={v|v∈{ w_1^i , …, w_t^i } ∧ v∈N(p_j∩q_i},1≤j≤k,1≤i≤m,t≥1.

We note that the path p_j ∩q_i is a prefix of q_i. There exists only one node in the set {v_1 , … ,v_t } also in N(p_j ∩q_i) , therefore x_{ij} contains only one node. y_{ij} .is similar to x_{ij} and it also contains only one node.

Definition 4: (XMVD) Let T be an XML tree that conforms to a DTD D. An XMVD is a statement of the form $p_1 , \dots ,p_k \to\to q_1 , \dots ,q_m |r_1 , \dots ,r_s$, 1≤k,1≤m,1≤s. P={p_1 ,· · · ,p_k },Q={q_1 ,· · · ,q_m } and R={r_1 , …, r_s } are subsets in Paths(D), {{p_1, ..., p_k}∪{q_1, ..., q_m}∪{r_1, ..., r_s}}⊂Paths(D). The tree T satisfies the XMVD if there exists a q_i, 1≤i≤m, and two distinct path instances $v_1^i . v_2^i . \dots . v_{t-1}^i . v_t^i$ and $w_1^i . w_2^i . \dots . w_{t-1}^i . w_t^i$ in Paths(q_i),1≤t, where

(i) val(v_t^i) ≠val(w_t^i);

(ii)There exists a r_j∈R,1≤j≤s, and two nodes z_1, z_2 ,where z_1 ∈Nodes(x_{ij} ,r_j) and z_2 ∈Nodes(y_{ij} ,r_j), such that val(z_1) ≠val(z_2);

(iii)For all p_h ,1≤h≤k, there exists two nodes z_3 and z_4 ,where z_3 ∈Nodes(x_{ijh} ,p_h)and z_4 ∈Nodes(y_{ijh} ,p_h), such that val(z_3)=val(z_4);

(iv)There exists a path instance $v'^i_1 . v'^i_2 . \dots . v'^i_{t-1} . v'^i_t$ in Paths(q_i), we have val(v'^i_t)=val(v_t^i), and there is a node z'$_1$ in Nodes(x'_{ij} , r_j) , such that val(z'$_1$)=val(z_2). There exists a node z'$_3$ in Nodes(x'$_{ijh}$, p_h)such that val(z'$_3$)=val(z_3);

(v)There is a path instance $w'^i_1 . w'^i_2 . \dots . w'^i_{t-1} . w'^i_t$ in Paths(q_i) making val(w'^i_t)=val(w_t^i), and there exists a node z'$_2$ in Nodes(y'$_{ij}$, r_j), we have val(z'$_2$)=val(z_1) and there exists a node z'$_4$ in Nodes(y'$_{ijh}$, p_h)such that val(z'$_4$)=val(z_4).

Where, x_{ij} ={v|v∈{ v_1^i , ... , v_t^i } and v∈N(r_j∩q_i)}, y_{ij}={v|v∈{ w_1^i , ... , w_t^i } and v∈N(r_j∩q_i)}, x_{ijh} ={v|v∈{ v_1^i , ... , v_t^i } and v∈N(p_h∩r_j∩q_i)}, y_{ijh}={v|v∈{ w_1^i , ... , w_t^i } and v∈N(p_h∩r_j∩q_i)}; x'$_{ij}$ ={v|v∈{ v'^i_1 , ... , v'^i_t } and v∈N(r_j∩q_i)}, y'$_{ij}$ ={v|v∈{ w'^i_1 , ... , w'^i_t } and v∈N(r_j∩q_i)}, x'$_{ijh}$ ={v|v∈{ v'^i_1 , ... , v'^i_t } and v∈N(p_h∩r_j∩q_i)}, y'$_{ijh}$ ={v|v∈{ w'^i_1, ... , w'^i_t } and v∈N(p_h∩r_j∩q_i)}.

We note that the path r_j ∩q_i is a prefix of q_i, there exists only one node in set {v_1^i, ... , v_t^i} also in N(r_j∩q_i), therefore x_{ij} contains only one node. y_{ij}, x_{ijh}, y_{ijh}, x'$_{ij}$, y'$_{ij}$, x'$_{ijh}$, y'$_{ijh}$ are similar to x_{ij}. The XMVD is symmetrical, i.e. the XMVD: $p_1, \dots , p_k \to\to q_1, \dots, q_m | r_1, \dots, r_s$ holds iff the XMVD: $p_1 , \dots ,p_k \to\to r_1 , \dots ,r_s |q_1 , \dots ,q_m$.

Definition 5: (Minimal Base) Let T be an XML tree that satisfied the given DTD D, Path set is P={P_1,...,P_k}. We assume that Paths(D)= P_1∪...∪P_k. The minimal base of P is denoted by MB(P) and it is a partition of Paths(D), S_1,...,S_q ,where

(i) Each P_i is a set of S_j by union operation;

(ii) There exists no partition that satisfies the condition (i) and the number of partition is less than q, where 1≤i≤k, 1≤j≤q.

Definition 6: (Dependency Base) Let ∑ be the set of XFD and XMVD over the complete instance document XML tree T which satisfied DTD D, P⊆Paths (D), the minimal base MB(P^+) of P^+ [16] is a path dependency base relative to ∑, denoted by DEP(P).

Definition7: (Logical Implication) Let Paths(D) be the path set of DTD D, and ∑ is the data dependency set of D. If each XML tree T of D satisfies ∑ and P→→Q|R or P→Q, we call P→→Q|R or P→Q implicated logically by ∑, and denotes ∑⊨P→→Q|R or ∑⊨P→Q.

III. Membership Algorithm On The Condition Of Coexistence Of Xml Functional Dependency And Xml Multi-Valued Dependency

A.1. Path Dependency Base Algorithm

Dependency base is a partition of attribute set of relational data schema in the relational data theory. The path dependency base can gain the logical implication of multi-valued dependency directly; in other words, when the dependency base based on given multi-valued dependency is confirmed, we can get all multi-valued dependencies implicated logically by some attribute set.

Lemma 1: Let T be an XML tree that satisfied the given DTD D. p_h, r_j, q_i∈D,1≤h≤k,1≤i≤m,1≤j≤s. If XML tree T satisfies XFD: $p_1, \dots , p_k \to q_1, \dots, q_m$, T will satisfy XMVD: $p_1, \dots, p_k \to\to q_1, \dots, q_m|r_1, \dots, r_s$, where R={$r_1$, ..., r_s}∈D.

B.1. Algorithm Description

We note that if there is one dependency: XMVD P→→Q(XFD P→Q), and dependency set ∑ implicates logically this dependency, so Q is the union set of some paths among DEP(P) according to the definition 6 and 7. The algorithm for path dependency base is given before solving the membership on the condition of coexistence of XML functional dependency and XML multi-valued dependency:

First, extend all the XFDs among dependency set ∑ to XMVD, then we can get a transformation from ∑ to ∑′ which only includes XMVD; the initial value of BASIS consists of all the single paths from path set P and Paths(D)-P. The initial value of change-flag is set as T, then execute "while" loop: initialize the value of change-flag as F, and then do the following operations for each MVD in $\sum'$: let P' be the non-empty set, all of its elements from the intersection of BASIS and P, Q'=Q-P'. At the first time, W is empty. For the g_i in BASIS, add g_i into W if $g_i \subset Q'$. If $Q' \neq \varnothing$and $Q' \neq W$, set change-flag as T, then add Q' into BASIS by the definition 5. Executing the "while" loop repeatedly before $Q'=\varnothing$ or Q' has been the union set of some elements from BASIS.

Algorithm 1: DEP-BASE(Path Dependency Base).

INPUT: mixed set $\sum$ consisting of XFD and XMVD, Paths(D), $P=\{p_1, \dots, p_k\}$;

OUTPUT: DEP(P).

```
DEP-BASE($\sum$,Paths(D),P)
Begin
(1) To transform $\sum$ into $\sum'$including XMVD only.
(2) BASIS:=$\{\{p_i\} \mid p_i \in P\} \cup \{Paths(D)-P\}$;
(3)change-flag:= 'T';
(4)while change-flag do
(5)    {change-flag:='F';
(6)      for every MVD $P \to\to Q \in \sum'$ do
(7)          {$P' := \cup\{R \mid R \in BASIS$ and $R \cap P \neq \varnothing\}$;
(8)           $Q' := Q-P'$;
(9)           $W := \varnothing$;
(10)            for $g_i \in$ BASIS do
(11)              {if $g_i \subset Q'$ then
(12)                 {$W := W \cup g_i$;}
(13)          if $Q' \neq \varnothing$ and $Q' \neq W$ then do
(14)                 {change-flag:= 'T';
(15)                   BASIS:= MB{BASIS$\cup Q'$};}}}
(16)return(BASIS)
End
```

C.1. Analysis of Algorithm

C.1.1. Termination of the Algorithm

After the initialization of step (2), step (3) and the step (4)'s loop operation, BASIS becomes a partition of Paths(D), because every subset after partitioning is non-empty, and the number of sets attained by partitioning the Paths(D) is at most the number of all the paths in Paths(D). After every time (except the last time) for executing step (4), the size of BASIS will always increase. We note that the size of BASIS is at most the number of all the Paths in Paths(D). Because the XMVD in $\sum'$ is finite, so that the "for" loop in step (6) is terminable. In conclusion, algorithm 1 is terminable.

C.1.2. Correctness of the Algorithm

Correctness of the algorithm is to prove that BASIS is equal to DEP(P) when the algorithm terminates. The proof process includes two parts. The first part is to prove BASIS$\subset$DEP(P), and the second part is to prove DEP(P) $\subset$BASIS.

First, to prove BASIS$\subset$DEP(P). When the algorithm terminates, BASIS=$\{\{P_1\},\dots,\{P_m\},\{p_1\},\dots,\{p_k\}\}$,where $\{P_1\},\dots,\{P_m\}$ is a partition of Paths(D)-P, and single paths of P form a set only including one path. According to the reflexivity inference rules, we note that $\sum$ implicates logically XMVD $P \to\to p_i$. The induction method is used to prove $P \to\to P_j \in \sum^+$, for each $\{P_j\} \in$BASIS, where $1 \leq j \leq m$ as following.

According to the reflexivity inference rules, we have $P \to\to Q \in \sum+$, so that $P \to\to$VPaths(D)-P-Q$\in \sum+$ holds. After executing the step(2), all elements in BASIS depend on P by multi-value, that is to say $P \to\to Pj \in \sum+$ hold , for each $\{Pj\} \in$BASIS, where $1 \leq j \leq m$.

After t($t \geq 0$) times loop, we assume $P \to\to P_j \in \sum^+$ hold, for each $\{P_j\} \in$BASIS, where $1 \leq j \leq m$. For the (t+1)th loop: if $P \to\to Q \in \sum$ is an XMVD during the (t+1)th loop, BASIS value will change(if BASIS values don't change, it will be the final value). Let P' be the non-empty set whose elements are from the intersection of BASIS and P. $P' \to\to Q$ holds in accordance with the augmentation inference rules, because BASIS is a partition of Paths(D) , $P \subset P'$. $P \to\to P_j \in \sum^+$ holds, for each $\{P_j\} \in$BASIS, where $1 \leq j \leq m$ before the (t+1)th loop, then $P \to\to P'$ holds under the union rules. At the same time, we note that $P \to\to Q-P' \in \sum^+$ holds according to transitivity rules. $Q'=Q-P'$ is set. If Q' is empty or Q' is a union set of some elements among BASIS, Q' is added into BASIS. Then solve the minimal base of BASIS. BASIS value doesn't change according to definition 6. We need to make a corresponding modification for BASIS when Q' is not empty and Q' isn't a union set of some elements among BASIS. Now, according to the difference rules, $P \to\to P_j$ holds after modifying, for each $\{P_j\} \in$BASIS. In conclusion, $P \to\to P_j \in \sum^+$holds, for every $\{P_j\} \in$BASIS, $1 \leq j \leq m$ during the (t+1)th loop, so that BASIS$\subset$DEP(P) holds in accordance with the definition 6.

Second, to prove DEP(P)$\subset$BASIS. Constructing a XML tree T that conforms to DTD D, where

(i) Each XMVD on T from $\sum$ is legitimate;

(ii) The necessary and sufficient condition of one XMVD $P \to\to Q$ on T holds is that Q' is a union set of some elements among BASIS.

Constructing an XML document tree T: there are 2^m tree tuples, and every tuple in $Tuple_T(D)$[16] has a group of corresponding sequence $\{a_1,\dots,a_m\}$, where $a_i \in [0,1]$. $\{P_1\},\dots,\{P_m\}$ is a partition of Paths(D)-P. The corresponding values of all paths in P^+ for every tuple in $Tuple_T(D)$ are 1 ,and the value of every path in P_i is a_i .

Properties of the XML document tree T:

Property 1: Each XMVD is valid when on the right of T is P_i.

Property 2: Some XMVD holds on T when on the right of P_i is a non-empty subset iff the left of XMVD has intersection with P_i.

Each multi-valued dependency on the T will be proved correct as following:

$P \to\to Q \in \Sigma$ holds, and P' is the set whose elements attained from the intersection of BASIS and P. We note that Q-P' is an empty set or the union set of some elements among BASIS in accordance with termination of the algorithm, so that $P \to\to Q-P'$ holds on T. According to Property 2, $P \to\to Q \cap P'$ holds on T, then $W \to\to R$ holds on T based on union rules among deriving.

We will prove $P\rightarrow\rightarrow Q$ hold on T iff Q is the union set of some elements in BASIS.

We note from definition 5: if Q is the union set of some elements in BASIS, $P\rightarrow\rightarrow Q$ holds on T, that is to say the sufficiency condition holds.

If $P\rightarrow\rightarrow Q$ holds on T, $P\rightarrow\rightarrow Q\cap P_i$ holds on T, for each i, $1\leq i\leq m$. Because the intersection of P and P_i is empty, $Q\cap P_i$ is empty or is P_i based on Property 1 and Property 2. Therefore, Q is the union set of some elements among BASIS.

In conclusion, BASIS=DEP(P), algorithm 1 is correct.

C.1.3. Time Complexity of the Algorithm

"m" expresses the total number of paths from Paths(D), and "n" expresses the number of dependencies from $\sum$.

Because XFD is a special case of XMVD, so we note that every XFD can be extended to the corresponding XMVD.

Partition the path set into g subsets, using matrix g×m for expressing. Each row with m digits expresses a set. Each column of the matrix has only one 1, other positions are 0. The $(ij)^{th}$ position of matrix is 1 iff path p_j belongs to the i^{th} set. Obviously, p_j and p_i are in the same path set iff the i^{th} column and the j^{th} column of the matrix are the same; in other words, when the set is on the h^{th} row, the values of h_i and h_j both are 1, then other positions of the i^{th} column and the j^{th} column of the matrix are 0. Therefore, to find the minimal path dependency base, we partition the same columns into one set whose row values are 1 in the same columns. Arbitrary two of all the columns are compared for g times. We note that the time cost of which solving the minimal path dependency base is $O(g\times m)$.

This algorithm initializes the path dependency base at first. Every single paths from path P set form a set respectively and all sets constitute one set, then path set Paths(D)-P forms one set. Since constituting two rows of the matrix after initialization, we get the time cost of the operation is $O(m)$.

The loop in step (4) executes at most m times. In every loop, for the given XMVD $P\rightarrow\rightarrow Q$, our goal is to find the union set of all the elements attained from the intersection of BASIS and Q. Since the size of BASIS is at most m, the time cost of the operation is $O(m^2)$. The time complexity of the "for" loop is $O(m^2)$ of step (10). In step (4), it is possible to examine whether n dependencies change the BASIS value or not. Therefore, the loop continuous until the BASIS value change, and it need to execute $O(n\times m^2)$ times. The time cost of solving minimal path dependency base is $O(m^2)$ and the time cost of executing step(4) once is $O(n\times m^2)$. After executing step (4) completely, the total time cost is $O(n\times m^3)$.

In conclusion, the total time complexity of this algorithm is $O(n\times m^3)$.

A.2. Membership Algorithm

The membership is to solve whether some dependency implicated logically by dependency set or not. On the basis of DEP-BASE algorithm, we provide the membership algorithm in the following.

B.2. Algorithm Description

First, we get the path dependency base, and the initial value of Q′ is set as empty. If the dependency is a multi-valued dependency, the following operations is executed: if the element in DEP(P) belongs to the right path set of this multi-valued dependency, then add this element into Q′. Whether Q′=Q hold or not examined. If holds, this dependency is implicated logically by Σ. If the dependency is functional dependency, and need to examine whether the right path set $Q\in P^+$ of this dependency hold or not; if holds, this dependency is implicated logically by Σ.

Algorithm 2: DEP-MEMBERSHIP(XFD and XMVD Membership).

INPUT: dependency set Σ, path set Paths(D), a dependency g:$P\rightarrow\rightarrow Q(P\rightarrow Q)$.

OUTPUT: if this dependency is implicated logically by Σ, the output is True; otherwise, the output is False.

```
DEP-MEMBERSHIP($\sum$, g)
Begin
(1)DEP(P):=DEP-BASE(Σ,Paths(D), P);
(2)$Q'=\varnothing$;
(3)if $P\rightarrow\rightarrow Q$
(4)    for every $P_i$ in DEP(P)do
(5)        if $P_i\subseteq Q$ then
(6)            $Q':= Q'\cup P_i$;
(7)if ($P\rightarrow\rightarrow Q$ and $Q'=Q$) or ($P\rightarrow Q$ and $Q\in P^+$) then
(8)    return(True);
(9)else
(10)    return(False);
End
```

C.2. Analysis of Algorithm

C.2.1. Termination Of the Algorithm

The step (1) of algorithm 2 calls the algorithm 1 to get the DEP(P), so the step(1) is terminable by algorithm 1. Because the number of elements from DEP(P) is finite and at most the number of paths in Paths(D), the "for" loop is terminable in step (4) of algorithm 2. Therefore, the algorithm 2 is terminable.

C.2.2. Correctness of the Algorithm

If the dependency needed to be examined is XMVD, it can be transformed from judging whether $\sum$ implicates logically the XMVD $P\rightarrow\rightarrow Q$ to judging whether the Q is the union set of some elements from DEP(P) by definition 6. The step (1) of algorithm 2 calls the algorithm 1 to attain the DEP(P), and we note that the step (1) is correct. The "for" loop in step(4) is to examine one by one whether every set belongs to Q, then Q′ is used for expressing the sets whose elements in DEP(P) also in Q. Finally, compare Q′ with Q. If they are same, Q is the union set of some elements from DEP(P), now return True, otherwise , return False. If the dependency needed to be examined is XFD, we note that algorithm 2 is correct by the definition of closure.

C.2.3. Time Complexity of the Algorithm

The number of paths from Paths(D) is expressed as m, and n expresses the number of dependencies from $\sum$. The time complexity of the step (1) in this algorithm is $O(n\times m^3)$. The "for" loop in step (4) mainly make a comparing among sets, and the number of elements from DEP(P) is at most m, so its time complexity is $O(m^2)$. Therefore, the total time cost of algorithm 2 is $O(n\times m^3)$.

IV. Minimal Dependency Set Algorithm

Definition 8: Let $\sum$ be the set of XFD and XMVD, if $\sum_{min}$ is a minimal dependency set of $\sum$, the $\sum_{min}$ should meet the following conditions:

(i) There is no redundancy path in $\sum_{min}$, that is to say there is no $P\to\to Q$(or $P\to Q$) that makes

$\sum_{min}=\sum_{min}-\{P\to\to Q\}$(or $\sum_{min}=\sum_{min}-\{P\to Q\}$).

(ii) There is no redundancy on the left of every dependency, namely for every XMVD $P\to\to Q$(or XFD $P\to Q)\in\sum_{min}$, there not exist XMVD $P'\to\to Q$ (or XFD $P'\to Q)\in\sum^+$ that makes $P'\subset P$ hold.

(iii) There is no redundancy on the right of every dependency, and the right of every dependency is single path, namely for every XMVD $P\to\to Q$(or XFD $P\to Q)\in\sum_{min}$, there not exist XMVD $P\to\to Q'$(or XFD $P\to Q')\in\sum^+$ that makes $\varnothing\subset Q'\subset Q$ hold.

A. Algorithm Description

Data dependency plays an important role in the normalization design of database [18]. Designing an XML database that can avoid data redundancy and abnormal operation is an important research subject in the XML field [19].

Some data dependency can be implicated logically by other data dependencies. Removing the redundancy dependency and redundancy path is to simplify the given dependency set and make sure it is simple during the analysis and computation, on the condition that the data dependency set closure doesn't change. This problem is about minimal dependency set.

First, we set change-flag as T, then execute "while" loop, let change-flag be F, and initialize the value of $\sum$ as $\sum'$. To perform the following operations for every d in $\sum$: judging whether the dependency is redundancy using DEP-MEMBERSHIP (algorithm 2) at first. If it is redundant, then remove it from $\sum$. If it is not redundant then to examine whether the left and right of this dependency exist the redundancy paths respectively, and if there is a redundancy path , it will be removed. After executing the operations, we need to examine whether $\sum\neq\sum'$ hold or not, if it holds, $\sum$ is not the minimal dependency set. Then set the change-flag as T and re-execute the "while" loop. If $\sum=\sum'$, $\sum$ is a minimal dependency set.

Algorithm3: DEP-MINIMIZE(Minimize Dependency).

INPUT: A set $\sum$ including XFD and XMVD.

OUTPUT: the minimal dependency set $\sum_{min}$ of $\sum$.

```
DEP-MINIMIZE($\sum$)
Begin
(1)change-flag:='T';
(2)while change-flag do
(3)change-flag:= 'F';
(4)$\sum'$:= $\sum$;
(5)for (every dependency d in $\sum$)   do
(6)   if DEP_MEMBERSHIP($\sum$-d, d) then
(7)         $\sum$:=$\sum$-d;
(8)   if (exist redundancy path on the left of d)
(9)      for every path $p\in P$ do
(10)        if d is a XFD, then $P\to Q$ do
(11)              if  DEP_MEMBERSHIP($\sum$,{P-p}$\to$Q)
                    then
(12)                $\sum$:=($\sum$-{ $P\to Q$ }$\cup${P-p}$\to$Q));
(13)       if d is a XMVD, then $P\to\to Q$ do
(14)            if  DEP_MEMBERSHIP($\sum$,{P-p}$\to\to$Q)
                   then
(15)            $\sum$:=($\sum$-{ $P\to\to Q$ }$\cup${P}-{p}$\to\to$Q));
(16)   if (exist redundancy attribute on the right of d)
(17)      for every path $q\in Q$ do
(18)        if d is a XFD, then $P\to Q$ do
(19)              $\sum$:=($\sum$-{ $P\to Q$ })$\cup${$P\to q_1$, ..., $P\to q_m$};
(20)        if d is a XMVD, then $P\to\to Q$ do
(21)          if DEP_MEMBERSHIP($\sum$,$P\to\to${Q}-{q})
                 then
(22) $\sum$:=($\sum$-{ $P\to\to Q$ })$\cup${ $P\to\to Q'$, $P\to\to${Q}-{q}}
(23)if $\sum\neq\sum'$
(24)   change-flag:= 'T';
(25)return($\sum$).
End
```

B. Analysis of Algorithm

B.1. Termination of the Algorithm

The number of elements from preliminary $\sum$ and from $\sum^+$ is finite. In addition, the size of $\sum$ is at most the size of $\sum^+$ after executing the step (5). The size of $\sum^+$ is finite which results in the "for" loop in step (5) is terminable. The left and right paths of every dependency are finite. We note that the "for" loop is terminable in step (9) and step (17) of algorithm 3. In conclusion, the algorithm 3 is terminable.

B.2. Correctness of the Algorithm

Examine whether every dependency conforms to the definition of minimal dependency set or not in "for" loop of step (5) after executing the step (1)~step (4) of this algorithm. step (6) and step(7) of algorithm calls DEP_MEMBERSHIP algorithm to remove the redundancy dependency; step(8)~step(15) calls DEP_MEMBERSHIP algorithm to remove the left redundancy path; step(18) and step (19) transforms the right path of XFD to single path; step (8)~step(15) calls DEP_MEMBERSHIP algorithm to remove the right redundancy path for XMVD. We note that algorithm 3 is correct in that the dependency meets the definition of minimal dependency set.

B.3. Time Complexity of the Algorithm

"m" expresses the total number of paths from Paths(D), and "n" expresses the number of dependencies from $\sum$.

The time cost of step (6) is $O(n\times m^3)$ based on calling the DEP_MEMBERSHIP algorithm; the time cost of step (8)~step(15) is equal to the time cost of the "for" loop, and it is $O(n\times m^4)$. The time complexity of step(16)~step(22) also depends on the "for" loop whose time complexity is $O(n\times m^4)$. In addition, the number of dependencies $\sum$ is n. Therefore, the time complexity of executing the step(5) once is $O(n^2\times m^4)$. In conclusion, the time complexity of algorithm 3 is $O(n^2\times m^4)$.

V. Conclusion

The design of database schema is the first step of the database application. If the database schema is well designed, abnormal data dependencies, data redundancies and abnormal operations can be avoided. This paper studies the membership problem from the view of the condition of coexistence of XML functional dependency and XML multi-valued dependency. At the same time, we propose DEP-BASE algorithm, DEP-MEMBERSHIP algorithm and DEP-MINIMIZE algorithm on the membership problem, not only proving the termination and correctness of these algorithms, but also analyzing their time complexities. They laid a foundation for designing a normal form which level of normalization is higher than others.

Acknowledgment

This work was financially supported by Hebei Provincial Natural Science Foundation of China (F2012203087), National Natural Science Foundation of China (61272124) and National Natural Science Foundation of China (61073060).

References

[1] Erik Naggam, "Standard Generalized Markup Language"[EB/OL].(1996-03-01)[2013-11-26] http://www.w3.org/MarkUp/SGML.

[2] Arash Termehchy, Marianne Winslett, "Using Structural Information in XML Keyword Search Effectively". ACM Transactions on Database Systems, 2011, vol.36, no.1, pp.4.

[3] Haiping Xu, Abhinay Reddyreddy, Daniel F. Fitch, "Defending Against XML-Based Attacks Using State-Based XML Firewall". Journal of Computers, 2011, vol. 6, no. 11, pp.2395-2407.

[4] Ren Li, Jianhua Luo, Dan Yang, Haibo Hu, Ling Chen, "A Scalable XSLT Processing Framework based on MapReduce". Journal of Computers, 2013, vol. 8, no. 9, pp.2175-2181.

[5] Shihan Yang, Jinzhao Wu, Anping He, Yunbo Rao, "Derivation of OWL Ontology from XML Documents by Formal Semantic Modeling". Journal of Computers, 2013, vol. 8, no. 2, pp. 372-379.

[6] Haiyan Huang, Ronghua Shi, Gaoshi Li, "The normalization of the algorithm on XML multi- valued dependency". Journal of hunan institute of science and technology, 2008, vol.29, no.12, pp. 126-129.

[7] Millist W.Vincent, JIXUE LIU, CHENGFEI LIU, "Strong Functional Dependencies and Their Application to Normal Forms in XML". ACM Transactions on Database Systems, 2004, vol.29, no.23, pp. 445-462.

[8] E. F. Codd, "Recent investigations in relational data base systems"[C]//Proceedings of IFIP Congress 74, Stockholm, Sweden, 1974, pp.1017-1021.

[9] Kamsuriah Ahmad, Ali Mamat, Hamidah lbrahim, Shahrul Azman Mohd Noah, "Defining Functional Dependency for XML". Journal of Information Systems, Research & Practices, 2008, vol.1, no.1, pp. 26-34.

[10] Xiangguo Zhao, Junchang Xin, Ende Zhang, "XML Functional Dependency and Schema Normalization"[C]//Proceedings of the 9th International Conference on Hybrid Intelligent Systems, Shenyang, China, 2009, pp.307-312.

[11] Marcelo Arenas, "Normalization Theory for XML". SIGMOD Record, 2006, vol.35, no.4, pp. 57-64.

[12] Tadeusz Pankowski, Tomasz Pilka, "Transformation of XML Data into XML Normal Form". Informatica (Slovenia). 2009, vol.33, no.4, pp.417-430.

[13] CATRIEL BEERI, "ON the Membership Problem for Functional and Multivalued Dependencies in Relational Database". ACM Transactions on Database Systems. 1980, vol.5, no.3, pp.241-259.

[14] Zhongping Zhang, "The logical implication algorithm research of XML multi-valued dependency". Computer Science, 2006, vol.33, no. 11(Supplement), pp.353-354.

[15] Wei Qiu, Lichen Zhang, "Study of Normalization Existing MVD in XML DTD". Computer Science, 2007, vol.34, no.2, pp. 149-185.

[16] Zhixiao Liu, "XML normalization research based on functional dependency and multi-valued dependency", yanshan university, qinhuangdao, MA, 2012.

[17] Millist W.Vincent, Jixue Liu, Chengfei Liu, Mukesh Mohania, "Mutivalued Dependencies and a 4NF for XML". CAiSE 2003, pp. 14-29.

[18] Jixue Liu, Jiuyong Li, Chengfei Liu, Yongfeng Chen., "Discover Dependencies from Data—A Review". IEEE Transactions on Knowledge and Data Engineering, 2012, vol. 24, no.2, pp.251-264.

[19] M. W. Vincent, J. Liu, M. Mohania, "The implication problem for 'closest node' functional dependencies in complete XML documents". Journal of Computer and System Sciences, 2012, vol.78, no. 4, pp. 1045-1098.

Zhongping Zhang, Male, Born in 1972, professor, Ph.D., post-doctoral, CCF Senior Member (E20-0006458S).His main research interests are the grid computing, data mining and semi-structured data etc. He has undertaken 1 project of provincial level and has participated 2 projects funded by national natural science foundation of China. He rewarded the provincial scientific and technological progress second-class Award. On the domestic and international academic conferences and journals, He published more than 80 papers, 15 of them were cited by EI.

Chunzhen Fang, Female, Born in 1987, Postgraduate student, the main research interest is the outlier detection in data mining.

Subject Perception Semantic Model for Information Retrieval in Tourism

Lingling Zi, Xin Cong
School of Electronic and Information Engineering, Liaoning Technical University, Huludao, China
Email: linglingziltu@126.com, chongzi610@163.com

Yaping Zhang
China Faculty of Computer Science & Information Technology, Yunnan Normal University, Kunming, Yunnan, China
Email: zhangyp.cs@gmail.com

***Abstract*—With the continued growth of tourism multimedia information made available through the World Wide Web, more efficient and accurate retrieval approaches are leading to an increasing demand. Most existing information retrieval approaches are based upon keywords and therefore provide limited capabilities to capture the query requirements. However, a complete understanding of search requirements is essential for improving the effectiveness of retrieval. To achieve this goal, we propose a novel subject perception semantic model (SPSM) for searching tourism information, which allows the customized query threshold to retrieve tourism information. In this model, we present the definition of Subject Hierarchy Graph to support semantic search capabilities, and propose computing methods of subject perception to quantify the semantics of query requirements. The experiments show that SPSM obtained encouraging performances.**

***Index Terms*—information retrieval; subject perception; tourism; quantification; query requirements**

I. Introduction

During the last decade, the rapid advance of search technology has made people acquire multimedia information easily. However, due to complex data formats of tourism information and the ambiguity of query requests, people have to spend much time to find out from query results exactly what they want. So how to understand accurately query requirements is a challenge for the development of information retrieval system. Nowadays, many technologies have been proposed, including ontology, schema summarization, semantic search and query expansion, and these technologies provide intelligent retrieval services [1-2]. For example, Avatar Semantic Search was used to express user semantic information through extracting facts and concepts [3]. KIM was provided to complete the service of automatic semantic annotation [4]. XSEarch was presented as a semantic search engine for XML [5]. A web search engine using page counts and texts was developed to measure semantic similarity between words [6]. A type of relevance feedback was proposed and more expansion words were associated with the user query [7]. A new pseudo-relevance feedback method was presented to retrieval information, and the words with the highest degree of query word were identified to new queries [8]. Nevertheless, there is still a problem of understanding difficulties. In order to better analyze query intention, we propose a subject perception semantic model (SPSM) from the perspective of requirement quantification. SPSM allows customized query thresholds to retrieval tourism information, stressing on the one hand the computation of subject perception, and on the other hand the consideration of label documents, which extract from multimedia resources as the target search space.

The development of SPSM presents the solutions to two key problems: one is how to quantify users' implication requirements and the other is how to integrate multimedia resources. With regard to the first problem, the key issue is to identify appropriate semantic information according to the requirements from users' vague queries. To address this problem, we introduce the definition of Subject Hierarchy Graph and subject perception computing to choose appropriate query subjects, and propose a new measuring method reflecting the ambiguity of users' requirements. The integration of multimedia resources is the second problem. It is important to make sure that multimedia search results be displayed in a comprehensive ranking. To address this, we propose the label documents method for tourism information retrieval. In conclusion, the novel contributions of this paper are shown as follows.

- We propose the definition of Subject Hierarchy Graph (SHG) to quantify the semantics of the query keywords, and it is the building block of Subject Perception.
- We present the computations of Subject Perception for measuring implication queries. And the application of Subject Perception in SPSM can also be demonstrated.
- Multimedia query results are displayed in a precise and comprehensive way, namely, not only text results, but also image results. In order to unify these multimedia resources, the label documents method is presented.

doi:10.4304/jsw.9.11.2844-2850

- Better solving of the semantic search problem, the implementation of SPSM is proposed for enhancing semantic retrieval. This system includes four parts 1) semantic annotation module 2) index module 3) subject perception module 4) result display module. SPSM is capable of improving the traditional problem of keyword search, so as to provide more accurate multimedia query results.

The rest of the paper is structured as follows. Section II discusses the related work. Section III shows the method of Subject Perception. Section IV illustrates the implementation of SPSM. Section V presents experimental works to demonstrate the effectiveness of SPSM. Section VI concludes the paper.

II. Related Work

Ontology has a good ability of semantic representation and it is widely used in the field of information retrieval [9-10]. So many models and methods using ontology technique have been proposed for efficient retrieval. Semantic retrieval based on concept designing [11] is proposed and it makes full use of ontological concepts. A concept map learning system based on domain ontology is presented [12] and it plays a very important role for education. A service search engine for industrial digital ecosystems [13] is developed and it achieves accurate semantic retrieval. In order to provide implication query results, graph-based query rewriting for knowledge sharing is proposed [14]. Inspired by the idea of ontology, we construct Subject Hierarchy Graph for tourism multimedia information, so as to complete the search task.

Schema summarization is of great help to concise overview of searching results, and an important advantage is that a user can determine query results in a short time. So this technology has been widely applied in system development. Dynamic summarization of bibliographic-based data is developed to accommodate relevant information [15]. A visual tool called VIREX for producing XML schema is designed and one of the attractive features is to support XML views update in a form of summary obtained from web-based database [16]. In the SPSM model, all the query results can be summarized in the form of subject and shown clearly to the users.

III. The Subject Perception Method

In this section, we elaborate the subject perception method to obtain query semantics. This method contains two parts, one is to construct Subject Hierarchy Graph (SHG) and the other is to calculate the values of subject perception based on SHG. Finally, we present the application of subject perception which can be used in SPSM.

A. Subject Hierarchy Graph

A subject hierarchy graph contains three parts: the subject layer, the concept layer and the instance layer. The subject layer is composed of subject nodes SN, which is denoted by $< sid, h, n_c, n_s >$. The concept layer is composed of concept nodes CN, which is denoted by $< cid, sort, n_i >$. And the instance layer is composed of instance nodes IN, which is an instance of a concept associated with the given subject. Table I shows the definition of each notation in SHG.

Consider the SHG in Fig. 1. Subject nodes in the subject layer are organized in the form of a tree. Take subject S5<S5,3,4,0> as an example. We get the following information: the subject id is S5, the subject lever is 3, the subject type is a leaf node, and four concept nodes containing the subject are placed at the concept layer, where c5 and c6 belong to the basic concept (BC for short), c7 belongs to the comment concept (CC for short) and c8 belongs to the association concept (AC for short). So the corresponding concepts can be represented as <C5,BC,1>, <C6,BC,1>, <C7,CC,3>, and <C8,AC,1> respectively. The third part of above angle brackets represents the number of instances associated with the concept. For example, c7 connects three instances (i.e. i8, i9 and i10) in the instance layer.

TABLE.I
NOTATIONS OF SHG

Notation	Definition
sid	The identity of SN
h	The level of SN
n_c	The concept number associated with SN
n_s	The number of child nodes of SN, including leaf nodes and connection nodes
cid	The identity of CN
sort	The type of CN, including basic concepts(BC), association concepts(AC), and comment concepts(CC)
n_i	The instance number associated with CN

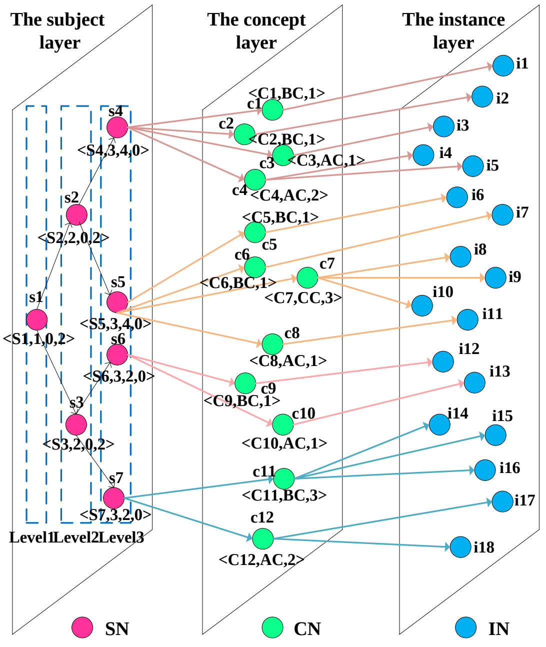


Figure 1.An example of SHG.

This work was supported by applied basic research programs of Yunnan Province (2010CD047) and the National Natural Science Funds (61262070).

B. The Computing of Subject Perception

An important contribution of this paper is to quantify the semantics of the query keywords. In order to achieve this goal, we compute the values of subject perception for each type nodes. Due to the type of nodes in SHG, the corresponding three methods are presented respectively, including the perception computing in the subject layer (P_S), the perception computing in the concept layer (P_C) and the perception computing in the instance layer (P_I).

P_S reflects the extent of subject concerned by tourists and the bigger the value of P_S is, the more attention this subject attracts. The computing formula of P_S is shown as (1).

$$P_S = w_1 * \vartheta(h) + w_2 * \frac{1}{n_s + 1} + w_3 * \frac{n_c + 1}{N_{\max} + 1} + w_4 * \varsigma * \varpi \quad (1)$$

In (1), $w_i (i = 1,2,3,4)$ is a weighting factor, which satisfies $w_1 + w_2 + w_3 + w_4 = 1$, $\vartheta(\mu) = (11 - \mu)/10$, $N_{\max}$ is the maximum number of concepts contained by the same subject, ϖ is the ratio of the subject resources to total resources, and ς is an amplification constant, here $\varsigma = 10$.

P_C is related to the concept type and the instance number n_i, shown as (2).

$$P_C = \vartheta(r) * \frac{n_i + 1}{\mathrm{I}_{\max} + 1} \quad (2)$$

Where r is the ranking number of concept type, and the descending order is BC, AC and CC. $\mathrm{I}_{\max}$ is the maximum number of instances with any concept contained by the same subject.

P_I indicates the attention degree of the instance, shown as (3).

$$P_I = \varepsilon_1 * P_C + \varepsilon_2 * \eta \quad (3)$$

Where ε_1 and ε_2 are adjustment coefficients, and η is the function of the linear conversion, shown as (4).

$$\eta = \frac{n_I - n_{\min}}{n_{\max} - n_{\min}} \quad (4)$$

Where n_I is the number of multimedia resources contained by the given IN, n_{min} is the minimal number of multimedia resources contained by any IN. Similarly, n_{max} is the maximal number.

C. The Application of Subject Perception

The proposed subject perception method is used to measure the implication query requests and this method is divided into leaf node measure (LNM) and connection node measure (CNM) according to the type of subject nodes in the SHG. The input parameters of each measure method contain subject keywords matched (*key* in short) and threshold σ (σ>0). Measure result is denoted by MR (*id*, *EK*), where *EK* represents expansion keywords of *key* and *id* is its corresponding sequence number. Furthermore, larger the value of σ (σ>1) is, the wider the range of subject is extended and with σ closer to 1, it indicates that *EK* is more important to the given *key*. Table II and Table III show the application of LNM and CNM respectively. The details of application way, including the method, condition and example, are demonstrated in the corresponding table.

TABLE.II
LNM

Case 1: 0<σ<1	Method	Find all the IN which satisfies the condition, rank IN, and output MR
	Condition	$P_I > \sigma$
	Example	Input: Imperial Palace and 0.9 Procedure: We find that the name of SN is the Imperial Palace and instance nodes satisfying P_I>0.9 are Hall of Supreme Harmony, Palace of Heavenly Purity and Palace of Earthly Tranquility. Output:(1,Hall of Supreme Harmony) (2,Palace of Heavenly Purity) (3,Palace of Earthly Tranquility)
Case 2: 1<σ	Method	Search all the SN whose parent node is the same with the parent node of *key*, find the SN which satisfies the condition, rank the SN, and output MR
	Condition	$P_{S_{key}} + \frac{1-\sigma}{e_1} < P_S < P_{S_{key}} + \frac{\sigma-1}{e_1}$, where $P_{S_{key}}$ denotes the value of P_S of *key* and e_1 is an amplification factor
	Example	Input: Great Wall Badaling and 1.5 Procedure: Subject nodes are Great Wall Badaling, Fragrant Hill and Xiayunling. Based on the computing of the above condition, results are Great Wall Badaling and Fragrant Hill. Output: (1,Great Wall Badaling) (2,Fragrant Hill)

TABLE.III
CNM

Case 3: 0< α <1	Method	Find all the SN whose child node is *key* and P_S of the obtained SN need satisfy the below condition, and output MR
	Condition	$\frac{P_S - P_{S_{\min}}}{P_{S_{\max}} - P_{S_{\min}}} > \sigma$, where $P_{S_{\min}}$ and $P_{S_{\max}}$ are the minimal and maximal values of P_S of subject nodes contained by the parent node *key*
	Example	Input: natural scenery and 0.7 Procedure: We find all the child nodes of natural scenery and get the subject node with the minimal value is Changping Huyun and the subject node with the maximum value is Great Wall Badaling. So the results returned are natural scenery, Great Wall Badaling and Fragrant Hill. Output：(1,natural scenery) (2,Great Wall Badaling) (3,Fragrant Hill)
Case 4: 1< α	Method	Search all the SN whose parent node is the same with the parent node of *key*, find the SN which satisfies the condition, and output MR
	Condition	$\lvert P_{S_{key}} - P_S \rvert * e_2 < (\sigma - 1)$, where e_2 is an amplification factor
	Example	Input: Old Town and 1.2 Procedure: To be similar with case 2, the range of P_S is computed under the setting of e_2=10. So the results are Old Town and street scene. Output：(1,Old Town) (2,street scene)

IV. The Subject-Aware Semantic Retrieval Model

In this section, we demonstrate the subject perception semantic model (SPSM). The model architecture is presented in Fig. 2, and it consists of four parts: semantic annotation module, index module, subject perception module, and results display module. It is emphasized that the subject perception module is the core of SPSM. This module is established according to the information of semantic annotation module and the output results of this module serve for the results display module. Next we take each part of the SPSM model and describe its corresponding function within the infrastructure.

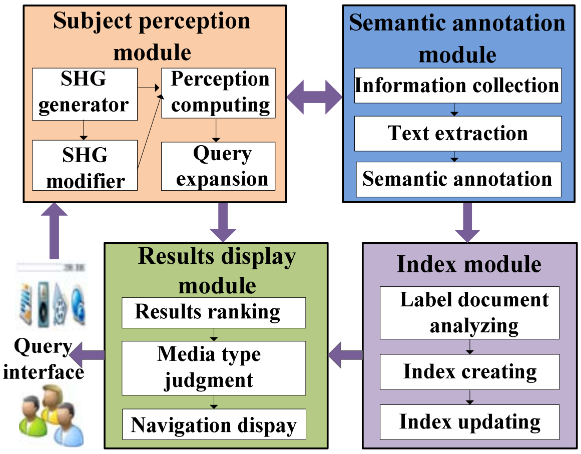


Figure 2 The architecture of SPSM.

A. Semantic Annotation Module

This model is responsible for collecting the massive tourism information from multiple sources of information, such as portal sites, travel forums, and blogs. It contains the following three parts. 1) Information collection. Firstly, we use a Meta search engine to search web sites which have a high correlation with tourism information and these web sites are put in the queue of URL, as a source set of information collection. Then a metadata judgment method using the technique of semantic analysis is used to extract new web links in this set, and keyword-based vector space model is adopted to exclude the web pages with useless information, so as to improve the accuracy of collected web pages. Finally, valid URL websites can be acquired through link filtering, and the corresponding images and texts can be captured. 2) Text extraction. Web crawler automatic captures multimedia information from URL websites and the obtained information are saved in the corresponding database. Then we adopt a series of operations, including feature extraction, structural analysis, and duplicate content elimination, to get information semantic [17]. And information semantic can be recorded in a file, which contains subject tags, the concept tags, instance tags, and label texts. 3) Semantic annotation. We establish a label document for each information file and the creation of this method is static, which is independent of the process of searching. The contents of label documents mainly contain document property information, resource collection information and semantic information. Through the method of label documents, images and texts can be unified from the perspective of semantic level.

B. Index Module

The index module is responsible for creating index fields according to the label documents, so as to lay the foundation for the searching of tourism information. This process mainly contains three steps (see Fig. 3):

Step1: The contents of label documents are analyzed and the corresponding index terms are extracted.

Step2: According to the obtained index terms, the index fields are constructed and the inverted index is created.

Step3: Both batch updating and incremental updating are adopted to complete constant renewal of index files.

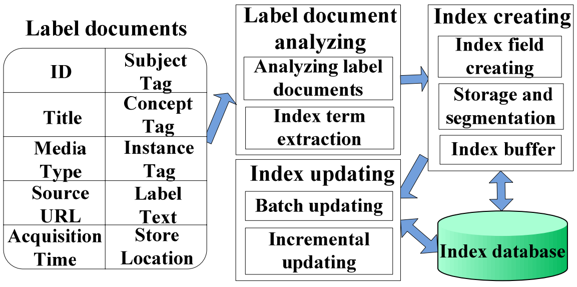


Figure 3 Index module.

C. Subject Perception Module

As a quantitative basis for semantic retrieval, the subject perception module is the core part of SPSM (see Fig. 4.). In this module, the values of subject perception are calculated according to SHG and on this basis, a list of expansion keywords are obtained according to the query keywords and the query thresholds. An overview of the process is shown as follows. 1) According to the keywords entered by users, preprocessing operations, such as null detection and Chinese word segmentation, are carried out. 2) The filtered keywords are matched in the SHG using the technique of word matching. 3) Using the proposed subject perception method, the appropriate query expansion lists are returned and saved in the hash tables.

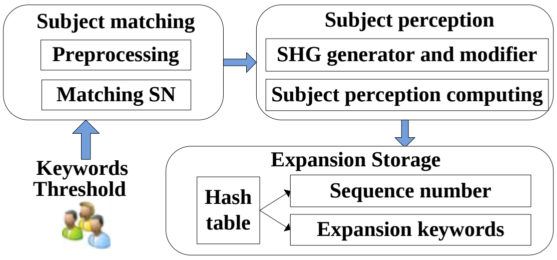


Figure 4 Subject perception module.

D. Results Display Module

As a large number of results obtained from the index module, the navigation method is used to display media results in an order way. Specifically, the expansion words are placed in the navigation bar, and all media results are classified and displayed according to the given expansion word. This module (see Fig. 5) consists of three parts. 1)

Results ranking. The type of ranking includes navigation ranking and content ranking. The former ranks expansion words according to the sequence number in the saving hash table and the latter ranks the results according to the correlation between the given expansion word and label texts in the label documents, i.e. the term frequency of label documents in the index file. 2) Media type judgment. Due to different display contents of media, the representation type of media results is determined. If the further details are need to be browed, users can only click the titles of returned results because the source URL is also saved the field of index file. 3) Navigation display. Using the navigation view, SPSM shows multi-faceted tourism information search results integrated with texts and images.

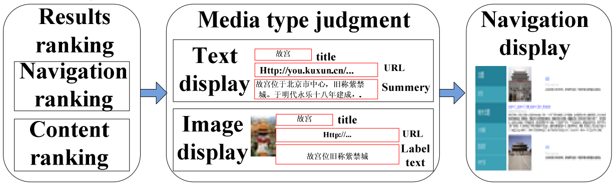


Figure 5 Results display module.

V. Experimental Results And Discussion

For the development of SPSM, we used Myeclipse8.5 platform, MySQL 5.1 and a PC with Intel Core(TM) 2 Duo T6570 processor and 2GB of main memory. Also, the open source full-text search engine Lucene and Web crawler Heritrix were also utilized. Using SPSM, we collected 6191 multimedia objects in the field of tourism, including 2934 texts and 3257 images. In this section, firstly we validate the efficiency and accuracy of the proposed approach by setting different weight values and then show the system performance from user's perspective. Finally, we conduct the comparison experiments.

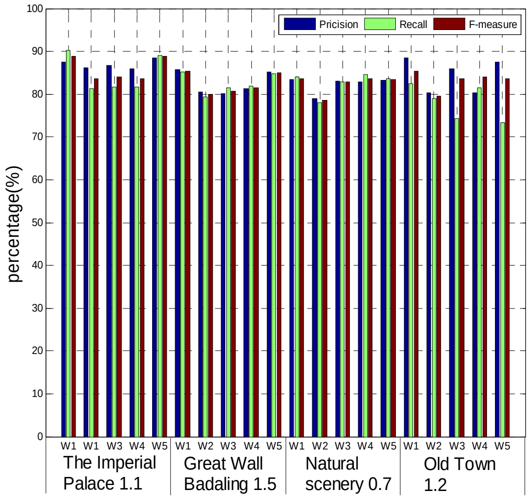


Figure 6 P/R/F results of different weights.

Three experiments were carried out to investigate the proposed method and it was measured by Precision, Recall and F-measure (P/R/F in short). Firstly, we set four query keywords and corresponding query thresholds using different weight values to evaluate P_S. The types of subject nodes of these query keywords contain both leaf nodes and connection nodes. The following weight value sets were studied: W1=<0.25,0.25,0,25,0.25>, W2=<0.1, 0.1,0,7,0.1>,W3=<0.7,0.1,0,1,0.1>,W4=<0.1,0.7,0,1,0.1> and W5=<0.1,0.1,0,1,0.7>. This experiment was helpful to find appropriate weights in the P_S formula. Fig. 6. summarizes the impact of different weights by presenting P/R/F results. More precisely, Precision values range from 79% to 88.5%, Recall values range from 73.4% to 90.3% and F-measure values range from 78.5% to 88.9%. We observe that W1 and W5 were the best weighting schemes. But for the last query case, Recall was only 73.4% using W5. In conclusion, W1 obtains better P/R/F results and this was due to a reasonable balance for all the factors of P_S. The reason of poor results produced by the other weighting schemes is because of highlighting only one factor contribution in the P_S formula.

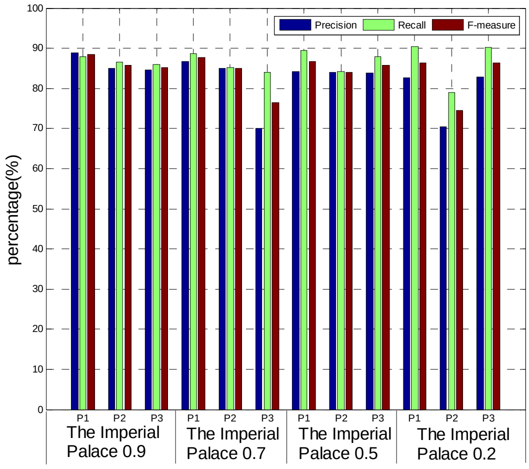


Figure 7 P/R/F results of different parameters.

Secondly, we investigated the P_I formula by setting different parameters. The following parameters were studied: P1: ε_1=0.5, ε_2=0.5; P2: ε_1=0.3, ε_2=0.7 and P3: ε_1=0.7, ε_2=0.3. Fig. 7 shows the P/R/F results in the case of the same query words under the different query thresholds and we can see the following three points. 1) Using the parameter P1, Precision values range from 82.7% to 88.9%, Recall values range from 87.9% to 90.5%, F-measure values range from 86.4% to 88.4%, and the average P/R/F values are 85.6%, 89.1%, 87.3% respectively. This shows that the results are relatively stable. (2) Using the parameter P2, Precision values range from 70.5% to 85%, Recall values range from 79% to 86.6%, F-measure values range from 74.5% to 85.8%, and the average P/R/F values are 81.1%, 83.8%, 82.3% respectively. It is noted that when query threshold is 0.2, Precision has the lowest value. Therefore, the results under this parameter are relatively unstable. (3) Using the parameter P3, Precision values range from 70.1% to 84.5%, Recall values range from 84.1% to 90.2%, F-measure values range from 76.5% to 86.3%, and the average P/R/F values are 80.4%, 87%, 83.4% respectively. It is also noted that when the query threshold is 0.7, Precision has the lowest value. Therefore, the results under this parameter are also relatively unstable. In view of these facts, we find the appropriate parameters in the P_I formula, namely ε_1=0.5 and ε_2=0.5.

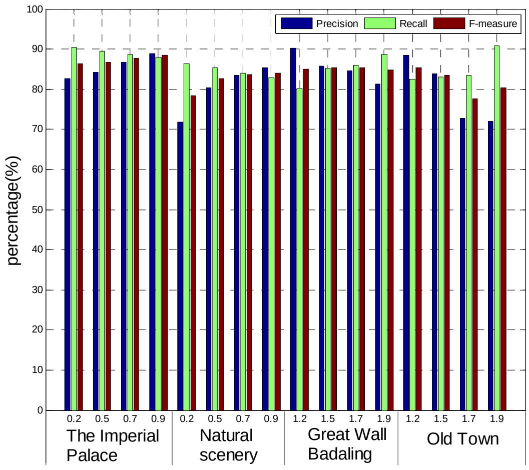


Figure 8 P/R/F results.

In the third experiment, four query keywords are selected and each query keyword contains four query thresholds in order to test the performance of SPSM. Fig. 8 depicts the results produced by W1 and P1. It can be seen from the figure that Precision values range from 71.8% to 90.3%, Recall values range from 80.1% to 90.8%, F-measure values range from 77.7% to 88.4% and the average P/R/F values are 82.6%, 85.9%, 84.1% respectively. It shows that SPSM has the encouraging results. Note that with regard to the same query keywords, we can find the following two points. 1) For the first two query keywords, as the decreasing of query thresholds, namely the value of query threshold gradually approaching zero, Precision values decrease whereas Recall values increase. (2) For the latter two query keywords, as the decreasing of query threshold, namely the value of query threshold gradually approaching one, Precision values increase whereas Recall values decrease. The reason of these facts is probably that different query thresholds get the different implicated query results.

TABLE.IV
THE QUERY CASE

Query ID	Query keyword	Query threshold
Q1	The Imperial Palace	0.9
Q2	The Imperial Palace	0.2
Q3	Great Wall Badaling	1.5
Q4	Great Wall Badaling	1.8
Q5	Natural scenery	0.7
Q6	Natural scenery	0.2
Q7	Old Town	1.2
Q8	Old Town	1.6

We demonstrate the performance of SPSM from the perspective of the users. The query cases are shown in Table IV and according to the results returned, ranking accuracy and satisfaction scores are displayed in Fig. 9. The criteria of satisfaction scores is set as follow: $0 < score \le 20$ represents slight satisfaction, $21 \le score \le 40$ represents fair satisfaction, $41 \le score \le 60$ represents moderate satisfaction, $61 \le score \le 80$ represents substantial satisfaction and $81 \le score \le 100$ represents almost perfect satisfaction. From Fig. 9, we find that users are satisfied with the query results.

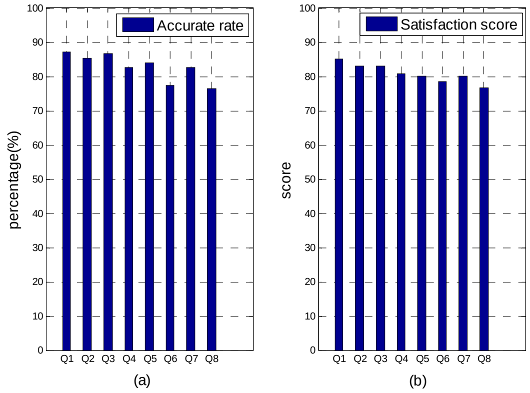


Figure 9 Performance evaluation by users.

Finally, we define the subject coverage measure to evaluate the integrity of query results. At the same time, the subject novelty measure is also presented to evaluate the expansibility of query results, shown as (5).

$$Coverage = \frac{N_{correct}}{N_{relevant}}, Novelty = \frac{N_{unknown}}{N_{known} + N_{unknown}} \qquad (5)$$

In (5), $N_{correct}$ denotes the number of correctly subjects of returned results, $N_{relevant}$ denotes the number of relevant subjects of returned results, N_{unknow} denotes the number of unknown subjects of returned results, and N_{known} denotes the number of known subjects. Table V shows the comparison results of *Coverage* and *Novelty* using SPSM and Mediapedia [18].

TABLE.V
COMPARISON OF COVERAGE AND NOVELTY

Query ID	*Coverage* of SPSM	*Coverage* of Mediapedia	*Novelty* of SPSM	*Novelty* of Mediapedia
Q1	0.27	0.66	0.07	0.26
Q2	0.77	0.66	0.33	0.26
Q3	0.33	0.17	0.16	0.16
Q4	0.5	0.17	0.28	0.16
Q5	0.25	0.81	0.09	0.23
Q6	0.88	0.81	0.29	0.23
Q7	0.5	0.25	0.5	0.5
Q8	0.75	0.25	0.66	0.5

The obtained values of *Coverage* and *Novelty* using SPSM are generally higher than those of Mediapedia. But for Q1 and Q5, the above values are lower than those of Mediapedia. That is due to the restriction of query thresholds. In a word, experimental results show that SPSM obtains good performance.

VI. Conclusions

This paper presents a novel method of measuring user implicated query intention, aiming at accurate searching tourism multimedia information. Based on the proposed approach, we construct the SPSM model which can quantify the relations between user query intention and query results. The experiments show that SPSM has encouraging performances. Future research lines will focus on extracting image semantics using the technology of transfer learning [19], so as to better improve the accuracy of information annotation.

Acknowledgment

This work was supported by applied basic research programs of Yunnan Province (2010CD047) and the National Natural Science Funds (61262070).

References

[1] G Huang, S Wang, X Zhang. “Query expansion based on associated semantic space,” *Journal of Computers*, vol. 6, no. 2, pp.172-177, 2011.

[2] X Peng, Z Niu, S Huang, et al. “Personalized web search using clickthrough data and web page rating,” *Journal of Computers*, vol. 7, no. 10, SPL.ISS., pp 2578-2584, 2012.

[3] E Kandogan, R Krishnamurthy, S Raghavan, et al. “Avatar semantic search: a database approach to information retrieval,” Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006: 790-792.

[4] A Kiryakov, B Popov, I Terziev, et al. “Semantic annotation, indexing, and retrieval,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, no. 1, pp. 49-79, 2004.

[5] S Cohen, J Mamou, Y Kanza, et al. “XSEarch: A semantic search engine for XML,” Proceedings of the 29th international conference on Very large data bases, Germany, 2003: 45-56.

[6] D Bollegala, Y Matsuo, M Ishizuka. “A web search engine-based approach to measure semantic similarity between words,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 977-990, 2011.

[7] J H Su, W J Huang, P S Yu, et al. “Efficient relevance feedback for content-based image retrieval by mining user navigation patterns,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 360-372, 2011.

[8] V Jalali, M R M Borujerdi. “Information retrieval with concept-based pseudo-relevance feedback in MEDLINE,” *Knowledge and information systems*, vol. 29, no. 1, pp. 237-248, 2011.

[9] J Li, P Shi, M Cheng. “Ranking Ontologies based on formal concept analysis,” *Journal of Computers*, vol. 9, no. 1, pp.215-221, 2014.

[10] L. Lin, Z. Xu, Y. Ding, “OWL Ontology Extraction from Relational Databases via Database Reverse Engineering,” *Journal of Software*, vol. 8, no. 11, pp. 2749-2760, 2013.

[11] R. Setchi, Q. Tang, I. Stankov, “Semantic-based information retrieval in support of concept design,” *Advanced Engineering Informatics*, vol. 25, no. 3, pp. 131-146, 2011.

[12] K. K. Chu, C. Lee, R.S. Tsai, “Ontology technology to assist learners’ navigation in the concept map learning system,” *Expert Systems with Applications*, vol. 38, no. 9, pp. 11293-11299, 2011.

[13] D. Hai, F.K. Hussain, E. Chang, “A service search engine for the industrial digital ecosystems,” *IEEE Transactions on Industrial Electronics*, vol. 58, no. 6, pp. 2138-2196, 2011.

[14] B. Qin, S. Wang, X. Du, Q. Chen, Q. Wang, “Graph-based query rewriting for knowledge sharing between peer ontologies,” *Information Sciences*, vol. 178, no. 18, pp. 3525-3542, 2008 .

[15] T.E. Workman, J.F. Hurdle, “Dynamic summarization of bibliographic-based data,” *BMC Medical Informatics and Decision Making*, vol. 11, no. 6, pp. 1-10, 2011.

[16] L. Anthony, O. Tansel, T. Radwan, K. Keivan, J. Jamal, A. Reda, “XML materialized views and schema evolution in VIREX,” *Information Sciences*, vol. 180, no. 24, pp. 4940-4957, 2010.

L. Kallipolitis, V. Karpis, I. Karali, “Semantic search in the world news domain using automatically extracted metadata files,” *Knowledge-Based Systems*, vol. 27, pp. 38-50, 2012.

[17] H. Richang, Z. ZhengJun, G. Yue, C. Tat-Seng, W. Xindong, “Multimedia encyclopedia construction by mining web knowledge,” *Signal Processing*, vol. 93, no. 8, pp. 2361-2368, 2013.

[18] S. J. Pan, Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2010.

Lingling Zi was born in Liaoning Province, China. She is currently a Ph.D. candidate in the school of Computer Science. Also she is a lecturer in School of Electronic and Information Engineering, Liaoning Technical University. Her research concentrates on multimedia systems and intelligent information systems.

Xin Cong was born in Liaoning Province, China. He is currently a Ph.D. candidate and a lecturer in School of Electronic and Information Engineering, Liaoning Technical University. His research concentrates on multimedia processing and converged network.

Yaping Zhang was born in Yunna Province, China. She received the MS degree in computational mathematics from Yunnan University in 2005, and PhD degree in Computer Science from Zhejiang University in 2010, China. She is currently an associate professor in the Computer Science Department of Yunnan Normal University. Her research concentrates on real-time computer graphics and parallel computing.

Factor Analysis Method for Text-independent Speaker Identification

Tingting Liu
University of science and technology of China, Hefei, China
Email:ltt1989@mail.ustc.edu.cn

Shengxiao Guan
University of science and technology of China, Hefei, China
guanxiao@ustc.edu.cn

***Abstract*—Factor analysis method offers state-of-the-art performance in speaker identification during the paper. The compact representations of speakers named i-vectors are extracted from the utterances in a new low dimensional speaker- and channel-dependent space, named a total variability space. LBG algorithm is combined with fuzzy theory in the initialization of speaker models,which improves the recognition rate of the system. Channel compensation techniques, such as Linear Discriminate Analysis (LDA), Principal Component Analysis (PCA), Nuisance Attribute Projection (NAP) and Within-class Covariance Normalization (WCCN) are compared during the experiment. It can be seen that LDA followed by WCCN achieves satisfying performance. In addition, several identification methods are contrasted in the experiments. One is through Support-Vector-Machine (SVM), another one directly uses the cosine distance similarity (CDS) as the final decision score, logarithmic likelihood and vector quantization are used to compare to above two methods. It demonstrates that CDS combined with score normalization obtains better result. The testing of mobile phone database shows the robustness of the system in complex channel environment. The graphical user interface of training and testing module is simulated on MATLAB in the end of the paper.**

***Index Terms*—factor analysis, total space, i-vector, channel compensation, cosine similarity, score normalization**

I. Introduction

Speaker recognition is becoming an increasingly significant biological authentication technology. Speaker identification mainly aims to identify the exact speaker from speaker utterance corpus. For the speech of high quality, such as the consistent training and testing background environment, general systems can achieve high recognition rate. Yet, the complexity and peculiarity of the transmission channels or other interference factors will make the performance of the system degrade sharply. Thus the systems cannot adapt to the development of practical environment, which leads to our research on the robustness of speaker recognition. Channel compensation or score normalization could be applied in order to isolate the target speakers.

Joint Factor Analysis (JFA), originally proposed by Kenny [1], has received considerable focus due to its successful application to the speaker recognition task. Both the channel effects and the information about speakers are contained in the speech. The first step of factor analysis is to train the Gaussian Mixture Models (GMMs). GMMs were obtained by adapting the Universal Background Model (UBM) with the use of the algorithm of Maximum A Posteriori (MAP) [2][3]. Then Gaussian Mixture Super Vectors (GSVs) are produced by concatenating the means of GMMs, which are used to train the eigenvoice and eigenchannel space. JFA assumes each GSV is composed of two independent components, one of them is associated with the speaker itself, and the other one has relationship with channel.

More recently, i-vector approach motivated by JFA was introduced. It only defines one variability space termed total variability space, instead of joint estimation of separate speaker and session spaces and factors. I-vectors stand for sessions of utterance in the new total space. Since the channel variability information is included in this total variability space, i-vectors need to be in conjunction with channel compensation methods, for example, within-class covariance normalization, linear discriminant analysis, principal component analysis and nuisance attribute projection. In the paper, we contrast these compensation techniques, and find that the best result comes from the combination of LDA and WCCN [4]. This algorithm makes full use of the advantage of LDA and WCCN, which has the maximum disparity as well as the minimum overall cost.

The main content of the paper is showed as follows. In section II, we describe the total variability space and i-vectors. Section III describes the channel compensation techniques. We analyze the recognition methods in section IV.The simulation experiments and results analysis are given in section V. Section VI is the conclusion of the paper.

doi:10.4304/jsw.9.11.2851-2860

II. Factor Analysis

This section mainly describes the following stages, including GMM training, total variability space training, i-vectors acquiring [5].

A. Training Gaussian Mixture Models

The first step is to initialize the clustering centers after extracting the feature parameters of all the utterances from different speakers. Then we apply LBG algorithm followed by fuzzy theory to get the new clustering centers. LBG algorithm was first proposed by Linda, Buzo and Gray in 1980. Since the approach does not take the prior information into account, it may lead to a negative influence on the final cluster centers. Thus fuzzy theory is used to update the clustering centers. The following formula is the definition of the function of overall cost.

$$J_m = \sum_{t=1}^{T}\sum_{j=1}^{M}(\mu_{jt})^m d^2(x_t, c_j) \tag{1}$$

$$d^2(x_t, c_j) = \left\| x_t - c_j \right\|^2 = (x_t - c_j)^T F_j^{-1}(x_t - c_j) \tag{2}$$

where M is the order of the GMM. μ_{jt} is the membership, which ranges from zero to one. Initial cluster centers are obtained when the partial derivatives are zero.

$$c_j = \sum_{t=1}^{T}(\mu_{jt})^m x_t \Big/ \sum_{t=1}^{T}(\mu_{jt})^m \tag{3}$$

In order to estimating the models accurately, we should utilize EM algorithm to obtain the optimal centers. Above procedures are used to acquire the UBM. UBM represents the speaker- and session-independent characteristics of all the speakers. Experimental results have shown that this initialization approach improves the accurate recognition rate to some degree.

To train GMMs corresponding to UBM is the next step. GMM is acquired by the adaptation of UBM. Then Gaussian Supervectors (GSVs) are produced by connecting means of GMMs. This self-adaption is only used to update mean vectors of GMMs; however, we assume that all the GMMs have the same weigh vector and covariance matrix.

B. Total Variability Space

In JFA system, we need to estimate both the channel and the speaker space. The speaker space is defined by the eigenvoice matrix V and the channel space is represented by eigenchannel matrix U. In the eigenvoice training, all the recordings of a given speaker are considered to belong to the same person. Meanwhile, in the eigenchannel training we should get the utterances in different channels. The training data is so complex that we begin the research on total space. The process of training total variability space is exactly the same as eigenvoice matrix training. Both the feature of speaker and the channel variability are included in the space.

It has to be mentioned that the entire utterances are produced by different speakers. The space assumes that an utterance can be represented by GMM supervector, which is concatenated by the mean vectors of each GMM. A Gaussian supervector M is defined as

$$M = m + T\omega \tag{4}$$

where m can be replaced by UBM supervector. The matrix T is the definition of the total space. Moreover w is a random vector, whose distribution can be described as $N(0, I)$. w is also called identity vector, which is referred to as i-vector in short. M is normally distributed. The mean vector of M is m and covariance matrix of M is TT^T. The model of i-vector can be seen as the reduction of dimension of the GMM supervector, which projects the GMM supervector onto the total variability space. Since the dimension of total variability space is far smaller than that of supervector space, the process makes the following manipulations such as intersession compensations, scoring, become tractable. The algorithm of training total variability space matrix T is given as follows

Step1: Compute the Baum-Welch statistics for the given utterance h. The statistics are extracted using UBM, similar to [6].

$$N_{c,h}(s) = \sum_t \gamma_t(c) \tag{5}$$

$$\widetilde{F}_{c,h}(s) = \sum_t \gamma_t(c)(Y_t - m_c) \tag{6}$$

$$S_{c,h}(s) = diag\left\{\sum_t \gamma_t(c) Y_t Y_t^T\right\} \tag{7}$$

where c is the Gaussian index. $\gamma_t(c)$ corresponds to posterior probability of mixture component c generating the vector Y_t. The centralized Baum-Welch first order statistics are showed in equation 3.

Step2: The initial value of matrix T is produced randomly. EM algorithm is used as an iterative method to estimate the total variability space matrix.

Step3: Compute the posterior distribution of the hidden variable $\omega_{s.h}$.

$$l(s) = I + T^T \Sigma^{-1} N_h(s) T \tag{8}$$

$$E\left[\omega_{s.h}\right] = l^{-1}(s) T^T \Sigma^{-1} \widetilde{F}_h(s) \tag{9}$$

$$E\left[\omega_{s.h}\omega_{s.h}^T\right] = E\left[\omega_{s.h}\right] E\left[\omega_{s.h}^T\right] + l^{-1}(s) \tag{10}$$

Where $N_h(s)$ is defined as the diagonal matrix of a given utterance h of speaker s, whose diagonal blocks are $N_{c,h}(s)I$. $\widetilde{F}_h(s)$ is a supervector obtained by concatenating all the $\widetilde{F}_{c,h}(s)$ for a given utterance h. Σ can be replaced by the covariance matrix of UBM.

Step4: Recalculate the statistics below through the data of training set and acquire the maximum likelihood.

$$\Phi_c = \sum_s \sum_h N_{c,h}(s) E\left[\omega_{s.h}\omega_{s.h}^T\right], (c = 1, 2, \cdots C) \tag{11}$$

$$\Omega = \sum_s \sum_h \widetilde{F}_h(s) E\left[\omega_{s.h}^T\right] \tag{12}$$

The updating formula of total variability space matrix T is written as follows.

$$T_i\Phi_c = \Omega_i \ (i = 1, 2, \cdots CP) \tag{13}$$

In general, the number of iterations of EM algorithm is about ten. In the paper, we decide to concatenate all the recordings of each speaker into one utterance. That is to say, each speaker has a whole session of speech. The value of the variable h is defined as one. In this case, the procedure of training the total variability space is simplified to some degree.

C. Identity Vector

Identity vector/total factor $\omega_{s.h}$ is a hidden variable, whose posterior distribution is also a Gaussian distribution. It has to mention that mean vector of total factor is the estimation of i-vector. Three steps are needed to obtain i-vectors, which is specifically described below

Step1: Calculate the Baum-Welch statistics of each target speaker.

Step2: Read the trained total variability matrix T.

Step3: Compute the mean value of $\omega_{s.h}$.

The estimation of i-vectors is showed in equation 9. The goal of factor analysis is to realize the extraction of low-dimensional features. Basically, i-vectors are the new features of each speaker. In the actual experiment of the paper, we don't carry out the process of subsection on the training utterances of each speaker. That is to say, consider the training speech of one speaker as one session. The following experiment proved that it doesn't affect the recognition rate of the system, instead, improves the efficiency of acquiring i-vectors.

III. Channel Compensation

As the total variability space, which is represented by matrix T , contains both speaker and channel variability, additional intersession/channel compensation techniques are required. It is an integral part of speaker recognition task, which can significantly reduce the classification errors. Three compensation techniques are described in this section as applied to i-vectors: Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Nuisance Attribute Projection (NAP), and Within Class Covariance Normalization (WCCN).

A. Linear Discriminant Analysis

Due to the reduction of dimension, LDA is extensively used in the scope of pattern recognition [7][8]. LDA helps better discriminate between separate classes through finding out the orthogonal axes of total space. The entire speech of one speaker is considered as one class. LDA algorithm attempts to simultaneously maximize the inter-speaker discrimination and minimize the intra-speaker variance. Therefore, the problem of optimization can be turned into maximizing the Rayleigh coefficient below

$$J(y) = \frac{y^T S_B y}{y^T S_w y} \tag{14}$$

The equation showed above describes the significance of the Rayleigh coefficient. And the between-class variance S_B and within-class variance S_W are calculated respectively as follows

$$S_B = \sum_{s=1}^{S} (\overline{\omega}_s - \overline{\omega})(\overline{\omega}_s - \overline{\omega})^T \tag{15}$$

$$S_W = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{h=1}^{n_s} (\omega_{s.h} - \overline{\omega}_s)(\omega_{s.h} - \overline{\omega}_s)^T \tag{16}$$

The mean vector of total factor is supposed to a zero vector since the identity vector is a random variable with a standard normal distribution. However, we use the actual computed global mean vector rather than assuming it was zero. The goal of maximization is to acquire a projection LDA matrix A which is constituted by the eigenvectors with the highest eigenvalues. The equation is written below

$$S_B y = \lambda S_w y \tag{17}$$

We can choose the k eigenvectors having the best eigenvalues given in equation 14 to construct the LDA matrix. LDA helps project the total factors onto a low-dimensional space.

B. Principal Component Analysis

Principal component analysis [9] is commonly used in data compression. The direction of projection is obtained by maximizing the projection value of feature. Detailed computing process is described as follows

Step1: Calculate the mean of i-vectors of every speaker. Then the intermediate value is obtained by difference between i-vector and the mean.

Step2: Acquire the covariance matrix of these above characteristics which are described in the following formula.

$$\mathrm{cov} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1R_\omega} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2R_\omega} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{R_\omega 1} & \sigma_{R_\omega 1} & \cdots & \sigma_{R_\omega R_\omega} \end{bmatrix} \tag{18}$$

Step3: Compute the eigenvalues and eigenvectors of covariance matrix. Then the next step is to normalize the eigenvectors.

Step4: Sort the eigenvalues in descending direction. Then the first K values corresponding to the eigenvectors are constructing the projection matrix.

C. Within Class Covariance Normalization

Attenuating the dimensions of high within-class variance is the purpose of WCCN [10]. However, it guarantees only the conservation of directions in space, and removes information about the between-class variability. We make the assumption that all the voices belong to one category of a given speaker. The matrix of WCCN is calculated as below

$$W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega_{s.h} - \overline{\omega}_s)(\omega_{s.h} - \overline{\omega}_s)^t \tag{19}$$

where $\overline{\omega_s}$ is the mean value of i-vectors of the speaker s, and S is the number of speakers. n_s is the number of utterances of speaker s. Considering with the form of cosine kernel, a feature projection function can be defined as follows

$$\varphi(\omega) = B^T\omega \tag{20}$$

Where B is obtained through Cholesky decomposition of the inverse of the within class covariance matrix $W^{-1} = BB^t$. WCCN applies the within class covariance matrix to normalize the cosine kernel function aim to realize channel compensation.

D. *Nuisance Attribute Projection*

NAP needs to construct a feature space that mainly describes the information of channel [11]. The purpose is to seek for the space which can best describe the characteristics of speakers. It allows us to project i-vectors onto speaker space in order to remove the nuisance direction. The projection matrix is showed in the following formula.

$$P = I - RR^T \tag{21}$$

where R is a matrix of low rank. The number of column vectors is k. The matrix is obtained by the eigenvectors of within-class covariance matrix with the first k large eigenvalues showed in equation 19. What is more, the channel space is assumed to be made up of all these eigenvectors.

IV. Recogniton Method

The system based on factor analysis applies several methods to estimate the similarity. Vector quantization is described in the previous research work. Thus we mainly describe support vector machine (SVM), cosine distance scoring (CDS) and logarithmic likelihood. Score normalization is used to reduce the difference caused by channel variability.

A. *Support Vector Machine (SVM)*

Using i-vectors to train the SVM has several advantages relative to GMM supervector, such as lowering the dimensions of features, providing more convenience for modeling, reducing the amount of computation in channel compensation. In general, SVM is mainly applied in two-class classification. We should design a series of two-class classifiers to solve this multi-class classification problem. For example, if N speakers are included in the speaker corpus, the number of classifiers is C_N^2. These multi-class classifiers are trained by all the utterances of the speakers.

The weakness of this approach is that these classifiers may interfere with each other and then have a negative impact on the recognition rate. Therefore we may attempt to take advantage of global traversal algorithm. That is to say, i-vectors of testing utterance are as the input to every classifier, and then calculate the score acquired by each classifier. As a result, the class having the highest score is the target speaker.

In this paper, we use a library of support vector machine (LIBSVM) [12], which is currently one of the most widely used SVM software. Two steps are included in the typical use of LIBSVM: firstly, to obtain a model by the training dataset and secondly, to predict information of testing data by the known model. The tool package provides us with different data formats and SVM parameters to choose.

B. *Cosine Distance Scoring*

Comparison between channel compensated i-vectors for speaker identification can also be accomplished using other approach. As both training and testing i-vectors undergo the same transformation，cosine distance [13] can be seen as a symmetric classification method. Given two i-vectors, $\omega_{t\arg et}$ from a known speaker, and ω_{test} from a unknown speaker, the cosine similarity is calculated as follows

$$score(\omega_{t\arg et}, \omega_{test}) = \frac{\langle \omega_{t\arg et}, \omega_{test} \rangle}{\|\omega_{t\arg et}\| \|\omega_{test}\|} \tag{22}$$

Note that the cosine distance scoring (CDS) only considers the angle between the two i-vectors and not their magnitudes. It is believed that channel/session information is included in the magnitude. Thus cosine distance scoring is reasonable to identify the target speaker in the utterance corpus and can improve the robustness of the system to some degree.

C. *Logarithmic Likelihood*

Logarithmic likelihood refers to the calculation of the probability of each speech frame. The object speaker is the one who gets the highest score. The probability value is computed below

$$P(\lambda_k | X) = P(\lambda_k | x_1, x_2, \cdots, x_T) \tag{23}$$

According to the Bayesian formula, the above formula can be converted to calculating $P(X|\lambda_k)$. If the probability is so small, the numerical value may beyond the scope of MATLAB. Therefore the logarithmic value is applied to remove the multiplicative noise and normalize the range of the value.

$$\ln(P(X|\lambda_k)) = \ln(\prod_{t=1}^{T} P(x_t|\lambda_k)) = \sum_{t=1}^{T} \ln(P(x_t|\lambda_k)) \tag{24}$$

As the length of testing speech is not consistent, the logarithmic likelihood should be normalized in order to reduce the impact of voice length on recognition rate. That is to say, the average logarithmic likelihood should be computed in the following formula.

$$score_k = \frac{1}{T}\sum_{t=1}^{T} \ln(\sum_{j=1}^{M} \omega_j P(x_t | j, \lambda_k)) \tag{25}$$

According to the similarity between the testing speech and each speaker in the database, the one corresponding to the maximum likelihood is the target speaker.

D. *Score Normalization*

Score normalization aims to counteract statistical variation in classification scores. This is accomplished by scaling all the scores to a global distribution with a zero mean and unit variance. Recently, score normalization technique is becoming inevitable in speaker recognition system [14].

In this paper, we decide to use Z-norm to normalize the scores. Z-norm is a kind of normalization technique; it aims to estimate the mean and variance of the score, and perform linear transformation to the score.

Compared to Z-norm, there are T-norm, and ZT-norm. Combination of different normalization techniques may not have a superimposed effect. Considering that T-norm needs to compute the mean value and covariance of several testing speech to normalize the score, thus the procedure will cost more time which may have a negative influence on the real-time requirement of the system. Therefore, the simple and convenient Z-norm is applied in the final score normalization process. The specific process of Z-norm is showed in the figure above.

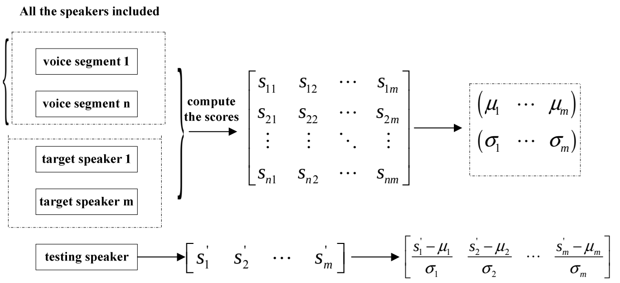

All the speakers included
voice segment 1
voice segment n
target speaker 1
target speaker m
compute the scores
s_{11} s_{12} $\cdots$ s_{1m}
s_{21} s_{22} $\cdots$ s_{2m}
s_{n1} s_{n2} $\cdots$ s_{nm}
$(\mu_1 \cdots \mu_m)$
$(\sigma_1 \cdots \sigma_m)$
testing speaker
$[s_1' \; s_2' \cdots s_m']$
$\left[\frac{s_1'-\mu_1}{\sigma_1} \; \frac{s_2'-\mu_2}{\sigma_2} \cdots \frac{s_m'-\mu_m}{\sigma_m}\right]$

Figure 1. The flow chart of score normalization.

V. Experiments And Results

A. Corpora

The experiments are performed on the corpus designed by the members of our laboratory to evaluate the text-independent speaker identification system. The corpus includes the recordings of 50 speakers in all (30 males and 20 females). These voices are recorded at an 8 KHz sampling rate with 16bit-quantization precision in the general laboratory environment on the same microphone. To each speaker, the length of the training data lasts 3 minutes, and the length of testing data is 30 seconds. Then we divide the training data into 40 sessions, and the testing data into 10 sessions. To make sure that each speaker has 40 samples to train the identification model and 10 samples to identify the target speaker. That is to say, we have 500 samples to test the accuracy of the system. It is mentioned that generally each sample will be divided into several sessions, each of which lasts about two seconds.

The experiments were carried out to compare the influence of different compensation techniques on recognition rate. On the other hand, we apply separate classification methods, such as SVM, cosine distance scoring, vector quantization and logarithmic likelihood. In the experiments, we will contrast the result of SVM with that of cosine distance. Despite of this point, we use the same channel compensation for these two conditions.

Another corpus used in our experiments is from a voice library of MIT mobile phone speaker recognition [15]. It is mainly conducted to test the performance of the system in different recording devices in order to show the advantage of i-vector based speaker identification system.

B. Experimental Configuration

Cepstral parameters, which are acquired by hamming window with the length of 30ms, are used in the experiments. The shifting is 15 ms. 12 Mel Frequency Cepstral Coefficients (MFCC) with appended delta coefficients were calculated. Two gender-dependent UBMs with 32 Gaussian Mixture Models were trained on microphone and mobile phone data using Expectation Maximum (EM) algorithm. The Random method is used to initialize the cluster centers. The training voices were truncated into several interval periods in order to increase the amount of the development data. Then GMM of each speaker was acquired by combining UBM with MAP. Concatenating the mean vectors of GMM is to acquire the GMM mean supervector, which is used to train the total variability matrix.

The dimension of i-vectors is generally 400. The corpora are also applied in training LDA, PCA, WCCN, NAP matrix to accomplish the channel compensation. Z-norm was utilized for the procedure of score normalization. LDA is the projection to reduce the dimension; however, WCCN and NAP are the equal dimension mapping. We assume the dimension of i-vector is 300 after the procedure of LDA. SVM training and classification was performed using i-vectors after session compensation. Cosine distance is used to compare the performance with SVM. Where applicable, score normalization is employed by Z-norm.

Evaluations in this paper focus primarily on three aspects: Firstly, the paper discusses the comparison between different compensation techniques as well as the impact of LDA and PCA dimension selection on recognition rate under the system of different orders. Then it comes to the influence of initialization method on accurate recognition result. Secondly, it contrasts separate identification approaches in speaker recognition system. In the end, the paper concludes the robustness of the speaker identification system based on factor analysis approach and its application on the utterances of mobile phone. It has to mention that all the experiments are realized on the software of MATLAB R2009b.

C. Results

All our experiments were carried out on the corpus of 50 speakers, 35 males and 15 females are included. Another corpus is the MIT mobile phone utterance library, normally 3 min in training and 30 seconds in testing.

1) The experiments carried in this section compare the recognition rates under several conditions. The goal of the first experiment is to contrast the performance of LDA with that of PCA in removing the nuisance directions. The results are reported in Figure 2. It's worth noting that during the stage of matching we make use of cosine distance scoring without score normalization.

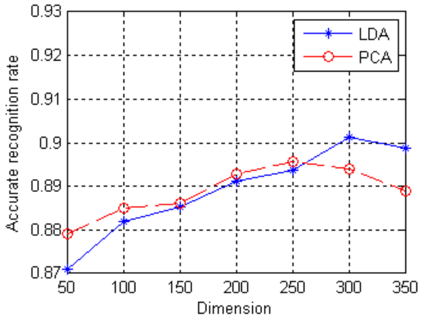


Figure 2. The effect on accuracy by .LDA and PCA in different dimension.

The results show that when the dimension is low, the effect of PCA is better than that of LDA. However as the increase of dimension, especially higher that 250, the system based on LDA behaves better. Considering both the recognition rate and experimental result, LDA is chosen and the optimal dimension of LDA space is 300 from the figure. LDA helps rotate the space in order to minimize the intra-speaker variance; and improve the performance in the case of cosine kernel. During the training step, we train the LDA projection matrix on all utterances used for training T matrix. The training data is projected onto the low-dimensional space to compensate for channel effects.

After determining the influence of dimension, it comes to the research on different order GMM in LDA based systems. The comparison result is showed in the below figure.

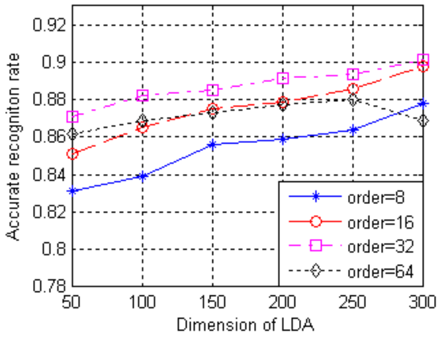


Figure 3. The effect on accuracy by LDA under the system of different order..

It can be seen in the figure that the recognition rate is on the rise as the order of the system becomes higher. The trend is embodied in the order of 8, 16 and 32. When the order rises to 64, there will be a certain degree of decline. The increase of the order will make the interval between different curves become reduced. That is to say, the model cannot well represent the characteristics when the dimension is too high or too low. If the dimension is too low, the models cannot distinguish among the speaker. However, the space dimension disaster will occur if the dimension is too high.

From the experiments, we can find the marked improvement with channel compensation. Moreover, single compensation technique will not appropriate for intersession compensation. Thus the comparison of accuracy between separate compensation techniques is showed in the following Table Ⅰ.

The experiments were carried on the same database. According to the results showed in the above table and figure， we note that applying WCCN in the LDA-projected space helps to improve the performance as compared to other single channel compensation techniques such as PCA, WCCN, NAP. Moreover, it is obvious that WCCN depicts better results than NAP, which makes us think over the combination of NAP and WCCN. Although, to some degree, the integration of WCCN and NAP improves the ratio, it does not obtain the best performance, so does the joint of three channel compensation techniques. Above all, the combination of LDA with WCCN achieves the best accurate recognition rate. The score of cosine distance after LDA and WCCN is showed in the following formula.

TABLE.I
COMPARISON BETWEEN DIFFERENT COMPENSATION TECHNIQUES

Compensation technique	Accuracy rate
LDA(dim=300)	90.1%
PCA(dim=300)	89.2%
WCCN	89.4%
NAP	88.9%
NAP+WCCN	90.8%
LDA(300)+WCCN	91.6%
LDA+NAP+WCCN	89.3%

$$\hat{\omega}_{LDA\_WCCN} = B^T A^T \omega \qquad (26)$$

TABLE.II
COMPARISON OF DIFFERENT INITIALIZATION APPROACHES

Initialization	random	LBG	LBG+Fuzzy
Accuracy	91.6%	92.1%	92.7%

The purpose of WCCN approach is to compensate the intra-speaker variability. Theoretically speaking, the application of WCCN in the two-dimensional LDA space reduces channel effects by minimizing the within-speaker variance. It is significant to note that the matrix of WCCN is a diagonal matrix after two dimensional LDA and WCCN projection. Thus the distribution of samples only executes the scale transformation without the change of rotation.

From the above experiments, it can be seen that LDA followed by WCCN acquires satisfying result. Then next experiment will show the contribution of initialization method on recognition rate, which increases by 1.1%. The following table II shows the result of different initialization methods including random, LBG algorithm, and LBG algorithm followed by fuzzy theory.

2) In this section, the thesis discusses difference between different identification approaches, including support vector machine, cosine distance scoring, logarithmic likelihood and vector quantization. It has to

mention that LBG algorithm combined with fuzzy theory is applied to generate the initial cluster centers.

Previously, SVM is always combined with JFA to constitute the system of JFA-SVM. The method of SVM is mainly through training SVM classifiers and then we have to calculate the score of each speaker. The target is the speaker with the highest score. We use the software package of LIBSVM as a library for support vector machines in the paper. In order to use this package, we have to change the data set into spectacular form. The form of training dataset and testing dataset is as follows

<label> <index> : <value1> <index>: <value2>…

Label is to identify the categories of the utterances. Index represents the numerical order of the exacted features. The index number starts from one. Values stand for every dimensional value of i-vectors after intersession/channel compensations. Then we should refer to three principal functions. One of them is libsvmread, which helps reading the testing and training data. The function of svmtrain is to produce a model for solving an optimization problem of SVM. The last one is svmpredict, which utilizes the trained model to get the category of the testing data. The advantage of the package is that we can choose different kernel functions and other parameters. [16].

The approach of cosine distance scoring is also applied in the total variability space. Compared to SVM, cosine distance scoring (CDS) provides a similar performance with a considerable increase in efficiency. The cosine similarity score operates by comparing the angles between the testing i-vector and the target i-vector after channel compensation without considering the amplitudes of i-vectors. The experiments are performed on the same dataset. The scores are normalized with Z-norm which further compensated for nuisance effects. The parameter to evaluate the system is the accurate recognition rate [17][18].

The method of logarithmic likelihood is to compute the probability scoring of testing speech sequences in each GMM. The target is the model with the best score. Vector quantization is to think i-vectors after channel compensation techniques as the parameters, which are used to generate the best codebook. The size of the codebook is 64. After quantizing the feature vectors of testing sequence, the relative difference between the testing speech and codebooks should be calculated. The speaker with minimum error is the identification result. The performance of the identification system using the above four method is displayed in table III.

TABLE.III
COMPARISON BETWEEN DIFFERENT IDENTIFICATION METHODS

Method	I-vectors	LDA	LDA+WCCN
VQ	79.4%	82.5%	84.2%
SVM	87.8%	86.5%	89.8%
LLR	87.5%	88.6%	91.2%
CDS	89.1%	91.6%	92.7%
CDS+Znorm	89.8%	92.3%	93.4%

From the above table, we can see that the system that combines LDA with WCCN using cosine distance scoring achieves the highest accurate recognition rate. Meanwhile, it is obvious that within class covariance normalization definitely optimizes the classification performance of SVM [19]. In a word, cosine distance scoring improves not only the accuracy of the system, but also the efficiency of algorithm to some degree [20].

TABLE.IV
AVERAGE TIME OVERHEAD OF DIFFERENT METHODS

Method	VQ	SVM	LLR	CDS
time	4.1s	6.7s	10.2s	8.6s

The above table IV shows the average time overhead of these identification methods. Vector quantization [21][22] method is the fastest approach relative to others. The second is support vector machine; logarithmic likelihood costs the highest time consumption. Because it needs to calculate the probability value of each frame of testing speech relative to different speaker, which increases the complexity of the process. It can be seen that cosine distance scoring is the modest weighing the algorithm complexity and identification rate.

3) This section illustrates that factor analysis method improves the robustness of the system in speaker identification. In practical application, a number of external factors have a negative effect on the recognition rate. For example, the training circumstance doesn't match the testing environment; the source of utterances comes from several ways. The experiments carried out in this section are used to verify that the system based on factor analysis can guarantee the recognition rate in spite of using separate sources. LBG algorithm combines with fuzzy theory in the process of initializing models; meanwhile cosine distance similarity with scoring normalization is used in obtaining the final score.

Channel compensation techniques exactly improve the accurate recognition rate of the system as expected in Figure 4. Firstly, as discussed in the paper, the accuracy achieves the highest rate when the dimension of LDA is 300. According to the process of getting i-vectors and the following channel compensations, the dimension of LDA is equal to that of recognition model. Secondly, for the sources come from microphone, the change of recognition rate is not very obvious under the circumstance of LDA and WCCN, which is increased by 3.6%. Nevertheless, for the sources come from microphone, the recognition ratio increased by 4.2% after channel compensations. Thus we can see that i-vectors compensated by LDA and WCCN can represent the characteristics of speaker preferably.

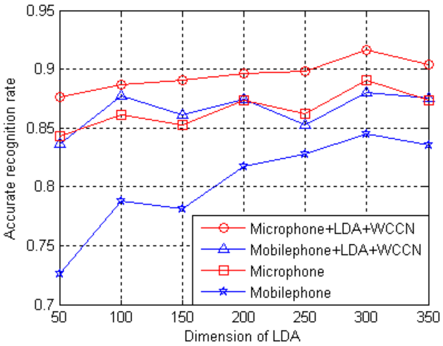
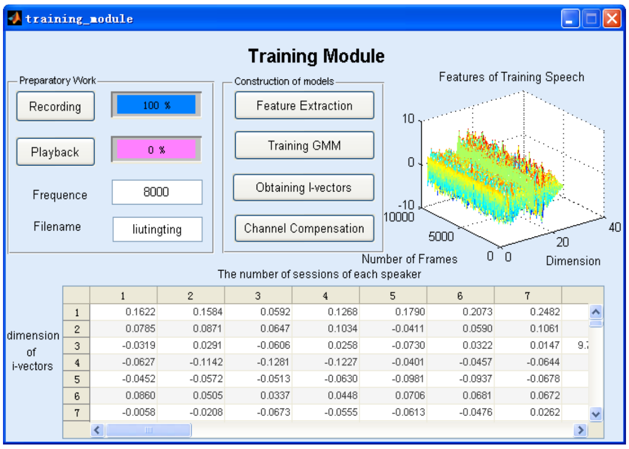


Figure 4. Comparison of accuracy between different sources.

4) The whole process of identification in the paper is simulated on MALAB. The system consists of training and testing module. The GUI of the system is showed in the below Figure 5 and Figure 6.

The first part of the system is training module showed in Figure 5. The speech of every speaker is analyzed at a 30 ms frame length and a 15 ms overlap. The feature parameters are included by 12 MFCCs together with first-order differential coefficients [23]. The GMM of each speaker is trained by the total features of one person. LBG algorithm and fuzzy theory are applied to initializing clusters, and EM algorithm is to update the GMM. GMM supervectos (GSV) are linked by the mean vectors of each GMM component. I-vectors are acquired by projecting the GSVs onto the total variability space. Then we move on to channel compensations. The process of training has ended so far.



Frequence 8000

Filename liutingting

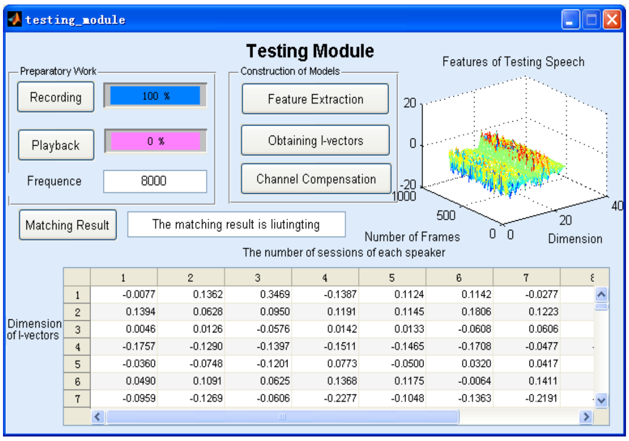
The number of sessions of each speaker

dimension of i-vectors

	1	2	3	4	5	6	7	
1	0.1622	0.1584	0.0592	0.1268	0.1790	0.2073	0.2482	
2	0.0785	0.0871	0.0647	0.1034	-0.0411	0.0590	0.1061	
3	-0.0319	0.0291	-0.0606	0.0258	-0.0730	0.0322	0.0147	9.7
4	-0.0627	-0.1142	-0.1281	-0.1227	-0.0401	-0.0457	-0.0644	
5	-0.0452	-0.0572	-0.0513	-0.0630	-0.0981	-0.0937	-0.0678	
6	0.0860	0.0505	0.0337	0.0448	0.0706	0.0681	0.0672	
7	-0.0058	-0.0208	-0.0673	-0.0555	-0.0613	-0.0476	0.0262	

Figure 5. Training Module.

In the training module, the training time of each speaker lasts for 3 minutes, which is divided into 40 sections. We can acquire 40 samples of every speaker as training data in this way. The button of playback it to test that if the microphone is working normally. In addition, the number of components of GMM is selected as 32 based on the above experiments. In the below part of the figure, we can see the table that contains i-vectors of trained speaker. The column of the table represents for the number of i-vectors, while the row of it stands for the dimension of i-vectors. Channel compensation includes the optimized combination techniques discussed in the above experiments.



Frequence 8000

The matching result is liutingting

The number of sessions of each speaker

Dimension of I-vectors

	1	2	3	4	5	6	7	8
1	-0.0077	0.1362	0.3469	-0.1387	0.1124	0.1142	-0.0277	
2	0.1394	0.0628	0.0950	0.1191	0.1145	0.1806	0.1223	
3	0.0046	0.0126	-0.0576	0.0142	0.0133	-0.0608	0.0606	
4	-0.1757	-0.1290	-0.1397	-0.1511	-0.1465	-0.1708	-0.0477	-
5	-0.0360	-0.0748	-0.1201	0.0773	-0.0500	0.0320	0.0417	
6	0.0490	0.1091	0.0625	0.1368	0.1175	-0.0064	0.1411	
7	-0.0959	-0.1269	-0.0606	-0.2277	-0.1048	-0.1363	-0.2191	-

Figure 6. Testing Module.

Another part is testing module showed in Figure 6. I-vectors of testing utterances are obtained during this section according to the trained total variability space, GSVs of testing data, and projection matrix of channel compensations. For example, if the button of channel compensation is pushed down, the projection matrix will be loaded. Then it compares the processed i-vectors of testing data to that of training data using cosine distance scoring (CDS). The speaker with the highest score after score normalization [24][25] is the target speaker.

It is significant to mention that the speech of tested speaker is recorded for 30 seconds, which is changed into 10 partitions. The sum score of these samples will help us find the target speaker displayed in the edit blank showed in the above figure. The right part of the figure displays the feature parameters of the tested speaker. Contrast the feature with that of the target speaker in Figure 5, we can find out that a certain similarity has existed.

VI. Conclusions

A text-independent speaker identification system based on factor analysis is represented in the paper. The method is to define a total variability space, which contains both the speaker and complex channel information. LBG algorithm combined with fuzzy theory is used to initialize the mean vectors of the models. All the GMM supervectors are projected onto this space to obtain the evaluation value of i-vectors. Compensation techniques are applied on i-vectors to remove the channel inference in order to improve the robustness of the system. The paper contrasts several channel compensation techniques, which are respectively nuisance attribute projection (NAP), linear discriminant analysis (LDA), and principal component analysis (PCA), within-class covariance normalization (WCCN). During the experiments in the paper, we can find out that the combination of LDA and WCCN may achieve the satisfying performance. As to the design of the classifiers, in contrast to SVM, LLR and VQ, cosine distance scoring not only provides higher recognition rate, but also makes the decision process less complicated. However, score normalization is inevitable, the goal of which is to counteract statistical variation in classification scores.

From the experiments, the results demonstrated that the system obtains the best performance when the dimension of LDA is 300. The initialization approach mentioned in the paper exactly increases the recognition rate by 1.1% compared to randomly generating the clustering centers or only using LBG algorithm. The technique of LDA may remove the nuisance directions, maximization of the variance between the speakers and minimization of the variance within the speakers. Experiment uses correct recognition rate as the assessment of the systems. Comparing to other approaches, LDA followed by WCCN has shown over 3.6% improvement during the experiment. In addition, the system guarantees the recognition rate in spite of using cellphone utterances, which increase the recognition rate by 4.2%. In the future we intend to optimize the algorithms in order to increase the recognition rate of telephone channel voice. Then the proposed method can be applied in an extended field.

REFERENCES

[1] P. Kenny, P. Quellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” IEEE transactions on Audio, Speech and Language Processing, vol.16, no.5, pp.980-988, 2008.

[2] D. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using Adapted Gaussian Mixture Models,” Digital Signal Processing, vol.15, no.4, pp.19-41, 2000.

[3] Gengzhao, Xufei Li, “Identity authentication scheme expansion based on speaker verification,” in Journal of computers, vol.8, no.8, pp:2027-2033, 2013.

[4] A. Hatch, S. Kajarekar, and A. Stolcke, “Within Class Covariance Normalization for SVM-Based Speaker Recognition,” in international Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 2006.

[5] N. Dehak, R. Dehak, P. Kenny, P. Dumouchel, and P. Quellet. “Front-end factor analysis for speaker verification,” In print IEEE trans. Audio, Speech and Language Processing, 2010.

[6] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” IEEE Trans. Speech Audio Processing, vol.13, no.3, May 2005.

[7] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, “An i-vector extractor suitable for speaker recognition with both microphone and telephone speech,” in Proc. Odyssey Speaker and Language Recognition Workshop, 2010, pp28-33.

[8] Bo Yang, Yingyong Bu, “A comparative study on vector-based and matrix-based linear discriminant analysis,” in Journal of computers, vol.6, no.4, pp:818-824, 2011.

[9] Ruoyu Yan, Ran Liu, “Principal component analysis based network traffic classification,” in Journal of computers, vol.9, no.5, pp:1234-1240, 2014.

[10] D. Matrouf, N. Scheffer, B. Fauve, and J. F. Bonastre, “A straight-forward and efficient implementation of the factor analysis model for speaker verification,” in International Conference on Speech Communication and Technology, 2007.

[11] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, “SVM Based Speaker Verification using a GMM supervector Kernel and NAP Variability Compensation,” in IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, vol.1, pp.97-100, 2006.

[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] Najim Dehak, Rda Dehak, Partick Kenny, Niko Brummer, Pierre Quellet, and Pierre Dumouchel, “Support Vector Machine versus Fast Scoring in the low-dimensional Total Variability Space for Speaker Verification,” in INTER-SPEECH, Brighton, UK, September, 2009.

[14] R. Aukenthaler, M. Carey, and H. Lloyd-Thomas, “Score Normalization for text-independent speaker verification systems,” Digital Signal Processing, vol.10, no.1, pp.42-54, 2000.

[15] http://www.datatang.com/data/4143

[16] P. M. Bousquet, D. Matrouf, and J. F. Bonastre, “Intersession compensation and scoring methods in the i-vectors space for speaker recognition,” in Interspeech, 2011.

[17] A.Kanagasundaram, R. Vogt,, D.Den, S. Sridharan, and M. Mason, “i-vectors Based speaker recognition on Short Utterances,” in Interspeech, 2011.

[18] Mitchell Mclaren, David van Leeuwen, “Source-Normalized LDA for Robust Speaker Recognition Using i-vectors From Multiple Speech Sources,” IEEE trans.Audio, Speech and Language Processing, 2012.

[19] J.Weston, C.Watkins, “Multi-class support vector machines,” In proceedings of Seventh European Symposium On Artificial Neutral Network, 1999.

[20] N. Dehak, “Discriminative and generative approaches for long- and short-term speaker characteristics modeling,” Application to speaker verification, Ph.D. thesis.

[21] D. Burton, “Text-dependent speaker verification using vector quantization source coding,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol 35, no.2, pp: 133-143, 1987.

[22] F. K. Soong, A. E. Rosenberg, et al, “A vector quantization approach to speaker recognition,” AT&T Technical Journal,vol 66, pp:14-26, 1987.

[23] N. Farvardin, R. Laroia, “Efficient encoding of speech LSP parameters using the discrete cosine transformation,” International Conference on Acoustics, Speech, and Signal Processing, 1989, pp:168-171.

[24] O.Glembek, L.Burget, N.Brummer, et al, “Comparison of scoring methods used in speaker recognition with joint factor analysis,” in IEEE Conference on Spoken Languish, Pittsburgh, PA, USA, September 2006.

[25] N.Dehak, R.Dehak, J.Glass, D.Reynolds, et al,“ Cosine similarity scoring without score normalization techniques,” in Proc. Odyssey Speaker and Language Recognition Workshop, 2010.

Tingting Liu graduate student studies in laboratory of audio-visual Information Processing and Pattern recognition Dep. of Automation, university of science and technology of China (USTC). She received her bachelor’s degree in Northeast Forestry University (NEFU) in China. Her research interest is speaker identification. Previously, she mainly focuses on the method of vector quantization. The current study is about speaker recognition system based on i-vectors.

Shengxiao Guan associate professor received PhD degree in engineering from department of Automation in University of Science and Technology of China (USTC). Currently, his research areas concentrate on pattern recognition and intelligence systems, new energy resources and industrial design. His laboratory of audio-visual Information Processing and Pattern Recognition mainly research on intelligent robot, embedded systems, the control of new energy, speaker identification, image tracking and processing.

An Automatic Software Requirement Analysis Approach Based on Intelligent Planning Technology

Hong He, Dongbo Liu
School of Computer and Communication, Hunan Institute of Engineering, Xiangtan, 411104, China
Email: hhong1970@126.com

Abstract—**With the development of information technology, the scale of current software is growing dramatically. This motivates the needs of techniques for intelligent software requirements engineering, which allows for modeling and analyzing requirements formally, rapidly and automatically, avoiding mistakes made by misunderstanding between engineers and users, and saving lots of time and manpower. In this paper, we propose an approach to acquiring requirements automatically, which adopts automated planning techniques and machine learning methods to convert software requirement into an incomplete planning domain. By this approach, we design an algorithm called Intelligent Planning based Requirement Analysis (IPRA), to learn action models with uncertain effects. Furthermore, we obtain a complete planning domain by applying this algorithm and convert it into software requirement specification.**

Index Terms—**intelligent planning, quality of service, requirement analysis, software engineering**

I. Introduction

Software requirement is an abstract concept, which is represented as software requirement specification. Requirements serve to tie the implementation world of the developers to the problem world of the stakeholder [1, 2, 21, 22]. Most empirical studies of requirements have shown that misunderstanding and changing requirements cause the majority of failures and costs in software [19, 20, 23, 24]. Since software engineers usually have limited knowledge about related field, they have to focus on analyzing obtained business process, and possibly neglect some uncertain factors. That is the reason why some software can not be applied in practice. On the other hand, it is usually difficult for users to express their demands accurately and completely without necessary hint.

Therefore, more and more attention is paid on how to acquire requirement rapidly and accurately in software requirement engineering [2, 22, 24]. For example, acquisition of software requirements based on ontology [1] is one of hot topics, which focuses on inducing users to offer system information with normal situation examples. Since those examples are collected randomly, it is difficult to make sure that a group of situation examples can cover the whole system, and induce users to offer requirement information completely and exactly, therefore this method can not be applied generally.

In this paper, we focus on applying intelligent methods to acquire software requirement specification automatically, which will make great difference in practice to avoid incomplete information and misunderstanding. In traditional planning research, we normally assume that action models with conditional effects and probabilistic effects could be built manually by experience, but in fact, it is difficult even for experts. It requires that experts not only should grasp logic of domain, but also have enough prior knowledge. Therefore, we propose an algorithm called Intelligent Planning based Requirement Analysis (IPRA) to learn action models with conditional effects and probabilistic effects and apply this algorithm to acquire software requirement automatically. Compared with previous action model learning algorithms, IPRA make the following contributions: (1) obtained action models by IPRA could have uncertain effects, including conditional effects and probabilistic effects. In practice, effects of actions are usually uncertain and conditional, with multiple possibilities; (2) state information of the planning traces could be incomplete. It is difficult to obtain complete state information in reality. IPRA can be applied with incomplete state information.

The rest of this paper is organized as follows. In section 2, we introduce related work. In section 3 and section 4, we make problem definition and present the steps of algorithm IPRA in detail. In section 5, we construct experiments in four planning domains to estimate the error rates of learned action models by IPRA, and apply IPRA algorithm to acquire software requirement specification. In section 6, we summarize this paper and discuss our future works.

II. Related Work

Automated planning systems achieve goals by producing sequences of actions from given action models that are provided as input. In 1971, Fikes and Nils designed STRIPS system [3] to introduce definitions of STRIPS operators, which made significant difference in the research of automated planning. In 1991, Soderland and Weld [4] designed the first nonlinear planning system SNLP of the world. In 1996, Kautz [5] converted planning into SAT problem, which effectively solved partial planning problem and showed new direction of

doi:10.4304/jsw.9.11.2861-2867

automated planning. In 1995, Avrim and Merrick [6] designed the first graph planner system Graphplan to solve planning problem, and proposed concept of graph plan. In 1998, Malik proposed Plan Domain Definition Language (PDDL) [7], then PDDL gradually became a general standard of representing domain models and was applied broadly in international planning competitions.

In recent ten years, researchers have proposed a series of planning algorithms to solve problems with uncertainty. Planning problems with uncertainty have become one of the most important research topics in artificial intelligent field. Artificial intelligence magazine organized a special version to introduce planning problems with uncertainty. As a direction of planning problem with uncertainty, probabilistic planning problems have contracted more and more attention. In international intelligent planning competition of 2004, researchers organized the first probabilistic planning competition. Younes and Littman [8] proposed PPDDL1.0 to solve probabilistic planning problems with uncertain effects and was applied in competition.

Recently, researchers have proposed some algorithms to learn action models. According to whether state information is complete, these algorithms could be divided into two parts. Some algorithms are, to learn action models from plan races with complete state information [9-16], which means for each action, we obtain the state information before and after it happens in advance, and then learn preconditions and effects of action model by statistics and reasoning. Gil et al. [9] build EXPO system, bootstraped by an incomplete STRIPS-like domain description with the rest being filled in through experience. Oates et al. [10] use a general classification system to learn preconditions and effects of actions. Schmill et al. [11] learn action models by approximate computation in relative domains. Wang et al.[12] propose an approach to learn action model automatically by observing planning traces and refine the operators through practice in a learning-by-doing paradigm. Pasula et al. [13, 14] present how to learn stochastic action models without conditional effects. Holmes et al. [15] model synthetic items based on experience to build action models. Walsh et al. [16] propose an efficient algorithm to learn action models for describing Web services.

III. Problem Description And Definition

A STRIPS-like planning problem with conditional effects and probabilistic effects can be defined as a four-tuple $<S,s_0,s_g,O>$, where S represents a set of states, and each state is a set of propositions; s_0 represents the initial state and s_g represents the goal state which is the final state following with a series of states transition, starting with initial state; O represents a set of action models with conditional effects and probabilistic effects. In this paper, we note O as a three-tuple <*a*, *PRE*, *CPEFF*>, where *a* represents an action schema with action name and parameters, *PRE* represents preconditions, *CPEFF* represents conditional effects and probabilistic effects.

Normally, *CPEFF* can formally be expressed as $<(p_{i1},c_i,e_{i1})...(p_{ij},c_i,e_{ij})\ldots(p_{in},c_i,e_{in})>$, where c_i represents the i^th^ condition composed of literal and conditions $c_i\ (1 \le i \le k)$ are mutually exclusive, the corresponding j^th^ effect is represented by e_{ij} with probability p_{ij}, which is a conjunction of literal and $\sum_{j=1}^{n} p_{i_j} = 1, p_{ij} \ge 0$. In the case when condition c_i is empty, conditional effects are exactly equal to probabilistic effects. If preconditions of an action are satisfied in state s, then the action can be applied in state *s*, and its effects can be selected according to conditions and probabilities. A possible action sequence is denoted as $<a_1,a_2,\ldots,a_n>$, transferring from initial state s_0 to goal state s_g. Furthermore, we call $(s_0,a_1,s_1,a_2,\ldots,s_n,a_n,s_g)$ as a planning trace, where the middle state s_i might be null, and a_i represents action schema.

Action model learning with conditional effects and probabilistic effects can be described as follows. Given planning traces set *T*, propositions set *P* as input, algorithm IPRA outputs all the action models with conditional effects and probabilistic effects in *A*. We show an example of action model learning with conditional effects and probabilistic effects in Table 1, which is is chosen from the domain slippery-gripper, an indeterminate planning domain.

TABLE I
AN EXAMPLE INPUT IN IP RA

Input: Predicates *P* *(block ?b) (gripper ?g) (gripper-dry ?g) (holding-block ?b) (block-painted ?b) (gripper-clean ?g)*			
Input: Action Schemas *A* *(pickup ?b ?g) (dry ?g) (paint ?b ?g)*			
Input: Plan Traces *T*			
	Trace 1	Trace 2	Trace 3
Initial state	*(gripper G)* *(block B)* *(gripper-clean G)* *(gripper-dry G)*	*(gripper G)* *(block B)* *(gripper-clean)*	*(gripper G)* *(block B)* *(gripper-clean)*
Action 1	(paint B G)	(pickup B G)	(pickup B G)
Observ 1		*not(holding-block B)*	*(holding-block B)*
Action 2	(pickup B G)	(dry G)	(paint B G)
Observ 2			
Action 3		(pickup B G)	
Observ 3			
Action 4		(paint G)	
Goal state	*(gripper-clean G)* *(holding-block B)* *(block-painted B)*	*not(gripper-clean G)* *(holding-block B)* *(block-painted B)*	*(gripper-clean G)* *(holding-block B)* *(block-painted B)*

IV. Framework Of Algorithm Ipra

The motivation of our algorithm IPRA is to transform the action model learning problem into weights learning problem in MLNs, and obtain action models with conditional effects and probabilistic effects. The frameworks of algorithm IPRA is shown as following: (1) Encode each plan trace as a set of propositions; (2) Generate candidate formulas, using *A* and *P*; (3) Apply

MLNs to learn weights of all the candidate formulas; (4) Choose some of candidate formulas according to given threshold, and convert weighted candidate formulas to action models with conditional effects and probabilistic effects as output. In the following subsections, we will show a detailed description of each step of the algorithm IPRA.

A. Encode Plan Traces

In the first step of algorithm IPRA, we encode all the plan traces as a set of proposition databases *DBs* with plan traces *T* as input. Firstly, we use propositions to represent each state of plan traces. For example, consider domain slippery-gripper in table 1, which includes two objects *B* and *G*. Present state s_1, describing that *B* is a block, *G* is a gripper, and *G* is clean, can be represented as *(block B s_1)* ∧ *(gripper G s_1)* ∧ *(gripper-clean G s_1)*. Secondly, we can consider an action as transition of states, then action can be encoded as the conjunction of propositions. For example, the action *(pickup B G s_1)* in table 1 can be treated as transition from the state *(block B s_1)* ∧ *(gripper G s_1)* ∧ *(gripper-clean G s_1)* to the state *(holding-block B s_2)*, then the action *(pickup B G s_1)* can be encoded as: *(block B s_1)* ∧ *(gripper G s_1)* ∧ *(gripper-clean G s_1)* ∧ *(pickup B G s_1)* ∧ *(holding-block B s_2)*.

According to the above method, we can encode each plan trace into a conjunction of grounded literals, and then convert them into a database(*DB*), where each record in a *DB* is a ground literal, and records are related as conjunction. For the sake of simplicity, we use *i* to denote the state symbol s_i. As an example, we encode the plan traces in Table 1 as database, and the results are shown in Table 2. In the paper, we make open world assumption, which means the grounded literal not shown in Table 2 is considered as unknown.

TABLE II
ENCODINGS OF PLAN TRACES AS DATABASES

DB1	DB2	DB3
(gripper G 0) *(block B 0)* *(gripper-clean G 0)* *(gripper-dry G 0)* *(paint B G 0)* *(pickup B G 1)* *(gripper-clean G 2)* *(holding-block B 2)* *(block-painted B 2)*	*(gripper G 0)* *(block B 0)* *(gripper-clean G 0)* *(pickup B G 0)* *not(holding-block B 1)* *(dry G 1)* *(pickup B G 2)* *(paint G 3)* *not(gripper-clean G 4)* *(holding-block B 4)* *(block-painted B 4)*	*(gripper G 0)* *(block B 0)* *(gripper-clean G 0)* *(pickup B G 0)* *(holding-block B 1)* *(paint B G 1)* *not(gripper-cleanG 2)* *(holding-block B 2)* *(block-painted B 2)*

B. Generate Candidate Formulas

In STRIPS model, if a predicate is a negative effect of an action, then the predicate should be a precondition of the action; and a predicate can not be both positive effect and negative effect of an action. Considering the two characteristics, we describe an action model in two parts:

(1) Preconditions. If predicate *p* is a precondition of action *a*, then *p* must be satisfied when the action *a* is executed, which can be described formally as:

$$\forall i,\bar{x},\bar{y},a(\bar{x},i)\rightarrow p(\bar{y},i), \tag{1}$$

where $\bar{x},\ \bar{y}$ are parameters, and *i* is the state symbol. In formula (1), since $p(\bar{y},i)$ is a necessary condition, not a sufficient condition, we choose $p(\bar{y},i)$ from candidate formulas with weights bigger than some threshold as preconditions in action model.

(2) Conditional effects. If predicate *p* is a positive effect of action *a* with condition *c*, then *p* should be added to the next state after the action *a* when condition *c* is satisfied, which can be described formally as:

$$\forall i,\bar{x},\bar{y},a(\bar{x},i)\rightarrow p(\bar{y},i)\hat{\ }p(\bar{y},i+1)\hat{\ }c(\bar{z},i) \tag{2}$$

where $\bar{x},\ \bar{y},\ \bar{z}$ are parameters, and *i* is the state symbol.

If predicate *q* is a negative effect of action *a* with condition *c*, then *q* is satisfied when *a* is executing and condition *c* is satisfied, but not satisfied after action *a*, which can be described formally as:

$$\forall i,\bar{x},\bar{y},a(\bar{x},i)\rightarrow q(\bar{y},i)\wedge\neg q(\bar{y},i+1)\wedge c(\bar{z},i) \tag{3}$$

where $\bar{x},\ \bar{y},\ \bar{z}$ are parameters, and *i* is the state symbol.

Similarly, suppose action *a* has a positive effect *p* and *a* negative effect *q* with condition *c*, then it can be described formally as

$$\begin{aligned}\forall i,\bar{x},\bar{y},a(\bar{x},i)\rightarrow\neg p(\bar{y},i)\wedge p(\bar{y},i+1)\wedge\\ q(\bar{y},i)\wedge\neg q(\bar{y},i+1)\wedge c(\bar{z},i)\end{aligned} \tag{4}$$

where $\bar{x},\ \bar{y},\ \bar{z}$ are parameters, and *i* is the state symbol, which means that effects of an action can be described as conjunction of some atomic formulas.

Applying formula (1) and (4), we can acquire candidate formulas of preconditions and conditional effects. For example, in slippery-gripper domain, candidate formulas of preconditions and conditional effects of action *pickup* are shown in Table 3 and Table 4.

TABLE III
CANDIDATE FORMULAS OF PRECONDITIONS BY (1)

ID	Formulas
1	$\forall i,b,g,(pickup\ \ b\ \ g\ \ i)\rightarrow(gripper\ \ g\ \ i)$
2	$\forall i,b,g,(pickup\ \ b\ \ g\ \ i)\rightarrow(block\ \ b\ \ i)$
3	$\forall i,b,g,(pickup\ \ b\ \ g\ \ i)\rightarrow(gripper\text{-}dry\ \ g\ \ i)$
4	$\forall i,b,g,(pickup\ \ b\ \ g\ \ i)\rightarrow(holding\text{-}block\ \ b\ \ i)$

TABLE IV
CANDIDATE FORMULAS OF CONDITIONAL EFFECTS BY (4)

ID	Formulas
1	$\forall i.b.(pickup\ \ b\ \ g\ \ i)\rightarrow(gripper\text{-}dry\ \ g\ \ i)$ $\hat{\ }\neg(holding\text{-}block\ \ b\ \ i)\hat{\ }(holding\text{-}block\ \ b\ \ i+1)$
2	$\forall i.b.(pickup\ \ b\ \ g\ \ i)\rightarrow(gripper\text{-}dry\ \ g\ \ i)$ $\hat{\ }(holding\text{-}block\ \ b\ \ i)\hat{\ }\neg(holding\text{-}block\ \ b\ \ i+1)$
3	$\forall i.b.(pickup\ \ b\ \ g\ \ i)\rightarrow\neg(gripper\text{-}dry\ \ g\ \ i)$ $\hat{\ }\neg(holding\text{-}block\ \ b\ \ i)\hat{\ }(holding\text{-}block\ \ b\ \ i+1)$
4	$\forall i.b.(pickup\ \ b\ \ g\ \ i)\rightarrow\neg(gripper\text{-}dry\ \ g\ \ i)$ $\hat{\ }(holding\text{-}block\ \ b\ \ i)\hat{\ }\neg(holding\text{-}block\ \ b\ \ i+1)$
5	$\forall i.b.(pickup\ \ b\ \ g\ \ i)\rightarrow(holding\text{-}block\ \ b\ \ i)$ $\hat{\ }(gripper\text{-}dry\ \ g\ \ i)\hat{\ }\neg(gripper\text{-}dry\ \ g\ \ i+1)$
…	…

C. Learn Weights of Candidate Formulas

According to reference [17], Markov Logic Networks *L* consists of a set of pairs (F_i,ω_i), where F_i is a formula in first-order logic and ω_i is a real number. With a finite set of constants $C=\{c_1,c_2,\cdots,c_n\}$, it defines a Markov network $M_{L,C}$ as following steps: (1) $M_{L,C}$ contains one binary node for each possible grounding of each predicate appearing in L. The value of the node is 1, if the grounded predicate is true, and 0 otherwise; (2) $M_{L,C}$ contains one feature for each possible grounding of each formula F_i in L. The value of this feature is 1 if the ground formula is true, and 0 otherwise. The weight of the feature is ω_i associated with F_i in L.

We apply Alchemy system [18] to learn weights of candidate formulas, by using weighted optimized pseudo log-likelihood. For each atomic formula, if it appears in DBs, then it corresponds to $x_i = 1$, otherwise 0. As mentioned in step 2, we can obtain the candidate formulas of preconditions and effects of actions by (1), (4), then learn weights of all the candidate formulas by MLNs. For example, the weights of candidate formulas in Table 3 and 4, are shown in Table 5 and Table 6.

TABLE V
WEIGHTS OF CANDIDATE FORMULAS FOR PRECONDITIONS

ID	Weights	Formulas
1	0.3	$\forall i,b,g,(pickup\quad b\quad g\quad i)$ $\rightarrow (gripper\quad g\quad i)$
2	0.5	$\forall i,b,g,(pickup\quad b\quad g\quad i)$ $\rightarrow (block\quad b\quad i)$
3	-0.4	$\forall i,b,g,(pickup\quad b\quad g\quad i)$ $\rightarrow (gripper\text{ - }dry\quad g\quad i)$
4	-0.2	$\forall i,b,g,(pickup\quad b\quad g\quad i)$ $\rightarrow (holding\text{ - }block\quad b\quad i)$

TABLE VI
WEIGHTS OF CANDIDATE FORMULAS FOR EFFECTS

ID	Weights	Formulas
1	0.77	$\forall i.b.(pickup\quad b\quad g\quad i) \rightarrow (gripper\text{ - }dry\quad g\quad i)$ $\wedge\neg(holding\text{ - }block\quad b\quad i)\wedge(holding\text{ - }block\quad b\quad i+1)$
2	0.12	$\forall i.b.(pickup\quad b\quad g\quad i) \rightarrow (gripper\text{-}dry\quad g\quad i)$ $\wedge(holding\text{-}block\quad b\quad i)\wedge\neg(holding\text{-}block\quad b\quad i+1)$
3	0.44	$\forall i.b.(pickup\quad b\quad g\quad i) \rightarrow \neg(gripper\text{-}dry\quad g\quad i)$ $\wedge\neg(holding\text{-}block\quad b\quad i)\wedge(holding\text{-}block\quad b\quad i+1)$
4	0.47	$\forall i.b.(pickup\quad b\quad g\quad i) \rightarrow \neg(gripper\text{-}dry\quad g\quad i)$ $\wedge(holding\text{-}block\quad b\quad i)\wedge\neg(holding\text{-}block\quad b\quad i+1)$
5	-0.3	$\forall i.b.(pickup\quad b\quad g\quad i) \rightarrow (holding\text{-}block\quad b\quad i)$ $\wedge(gripper\text{-}dry\quad g\quad i)\wedge\neg(gripper\text{-}dry\quad g\quad i+1)$
…	…	…

D. Obtain Action Model

In the candidate formulas of preconditions, we choose those formulas with weights bigger than some threshold as a set and convert the set into the preconditions of action model. Similarly, we can choose some candidate formulas of conditional effects and calculate their corresponding probabilities. Finally, we can obtain action model with probabilistic conditional effects. Weight of a formula in MLNs reflects the level of truth, which means the higher weight, the more formulas with true value after instantiation. At the beginning, we need to decide a threshold of the weights. For example, we set the threshold to be 0, then we can choose all the formulas with weights bigger than 0 in Table 7, as shown below.

$$\begin{cases}\forall \text{i,b,g,(pickup}\quad \text{b}\quad \text{g}\quad \text{i}) \rightarrow (\text{gripper}\quad \text{g}\quad \text{i}) \\ \forall \text{i,b,g,(pickup}\quad \text{b}\quad \text{g}\quad \text{i}) \rightarrow (\text{block}\quad \text{b}\quad \text{i})\end{cases}$$

Therefore, predicates (gripper g i),(block b i) are the preconditions of action (pickup b i).

Similarly, we choose those formulas under the same condition, with weights bigger than 0 in Table 6, and calculate their corresponding probabilities, then we can acquire the action model of (pickup b i) with probabilistic and conditional effects as shown in Table 7.

TABLE VII
THE ACQUIRED ACTION MODEL

Action	pickup(?b ?g)
Preconditions	*block(?b), gripper(?g)*
Probabilistic conditional effects:	*<(0.87 (gripper-dry ?g) (and (holding-block ?b))),* *(0.13(gripper-dry?g)* *(and(not(holdingblock ?b)))) >* *<(0.48(not(gripper-dry?g))(and(holdingblock? b))), (0.52 (not (gripper-dry ?g)) (and(not(holding-block ?b))))>*

V. EXPERIMENTS EVALUATION

A. Datasets and Evaluation Criteria

To evaluate the algorithm IPRA, we collected plan traces from the following planning domains: slippery-gripper, blocks-world, zenotravel, logistics-strips. These domains have the characteristics we need to evaluate in IPRA algorithm: all the four domains have uncertain effects. Using probabilistic planner Probabilistic-FF, we generated 20-100 planning traces from the three domains, as training data of learning action models with probabilistic effects. We consider the given action models in the above web-page as correct ones, and then use the correct action models to evaluate the error rates of learned action models.

We define the error rates of our algorithm as follows:

(1) Error rates of preconditions: let the number of all the possible preconditions in action models be $\mathrm{N_{pre}}$,the set of preconditions of learned action models be $\mathrm{T'_{pre}}$,and the set of preconditions of correct action models be $\mathrm{T''_{pre}}$ .If a precondition belongs to $\mathrm{T'_{pre}}$, not $\mathrm{T''_{pre}}$, then the number of errors in preconditions denoted by $\mathrm{E_{pre}}$,adds one; similarly if a precondition belongs to $\mathrm{T''_{pre}}$,not $\mathrm{T'_{pre}}$, $\mathrm{E_{pre}}$ adds one. Then the number of errors in preconditions can be expressed as $\mathrm{n_{pre}} = \left|\mathrm{T'_{pre}} \cup \mathrm{T''_{pre}} - \mathrm{T'_{pre}} \cap \mathrm{T''_{pre}}\right|$ .Thus error rate of preconditions can be calculated as $\mathrm{P_{pre}} = \dfrac{\mathrm{n_{pre}}}{\mathrm{N_{pre}}}$.

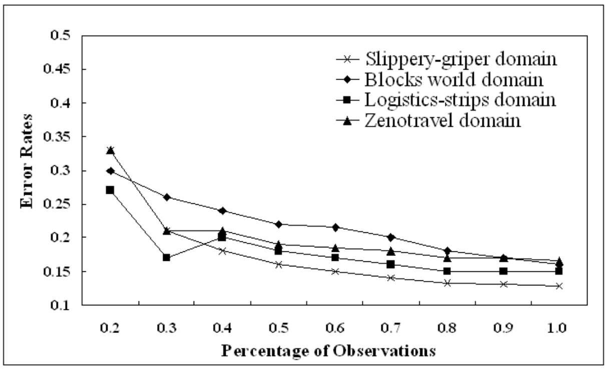
(2) Error rates of effects: since the learned action models have probabilistic effects, then we calculate the error rates of effects in a different method. Suppose for action a, the correct action model with probabilistic effects has m effects, and the corresponding probability

is $p_i, 1 \le i \le m$,and $\sum_{i=1}^{m} p_i = 1$ .Suppose the learned action model has n effects, and the corresponding probability is $q_j, 1 \le j \le n$,and $\sum_{j=1}^{n} q_j = 1$. We compare the ith effect e_i in the correct action model with the jth effect f_j in the learned one. Let the number of atomic formulas belonging to e_i , not to f_j ,be n_{miss} ; let the number of atomic formulas belonging to f_j , not to e_i ,be n_{extra} .The number of errors is denoted by $n_{ij} = n_{miss} + n_{extra}$ .Therefore, we can calculate the average number of errors of action a as $n_{effect} = \sum_{i=1}^{m} p_i \sum_{j=1}^{n} n_{ij} q_j$ .Let the number of possible errors be N_{effect} ,then the error rate of action a can be calculated as $P_{effect} = \frac{n_{effect}}{N_{effect}}$.

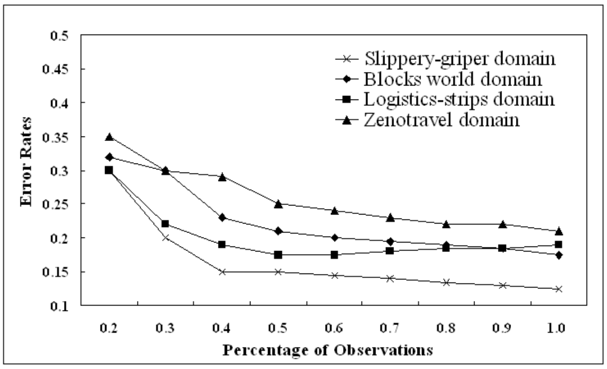
Furthermore, for action a, the error rate of action model with probabilistic effects can be defined as $R(a) = \frac{1}{2}(P_{pre} + P_{effect})$ .Here we assume that the error rates of preconditions and effects were equally important, and the range of error rate $R(a) \in [0,1]$ .Moreover, the error rate of all the action models A in a domain is defined as $R(A) = \frac{1}{|A|}\sum_{a \in A} R(a)$,where $|A|$ is the number of A's elements.

B. Accuracy and the Observed Intermediate States

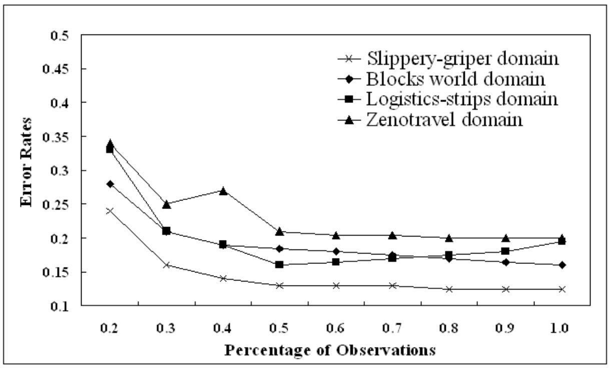
To simulate partial observation between two actions in a plan trace, from the plan traces, we randomly select observed states with specific percentage of observations 1/5, 1/4, 1/3, 1/2, 1. For each percentage value, e.g. 1/3, we randomly select an observation within three consecutive states in a plan trace. We run the selection process three times. IPRA generates learned action models each time, and meanwhile error rates are calculated. Finally, we calculate an average error rate on the plan traces. The results of these tests are shown in Figure 1.



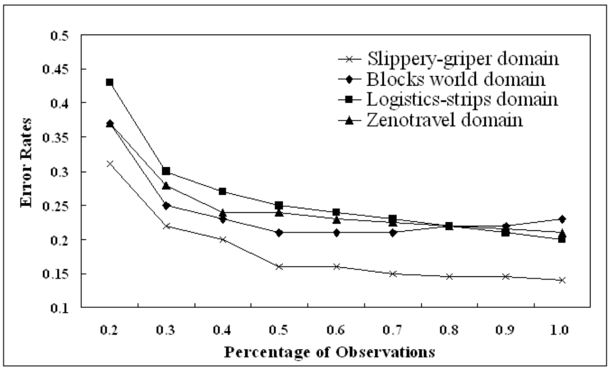
(a) Threhold t = 0.5



(b) Threhold t = 0.01



(c) Threhold t = 0.5



(d) Threhold t = 0.001

Figure 1. Error Rates of Learned Action Models in Different Domains

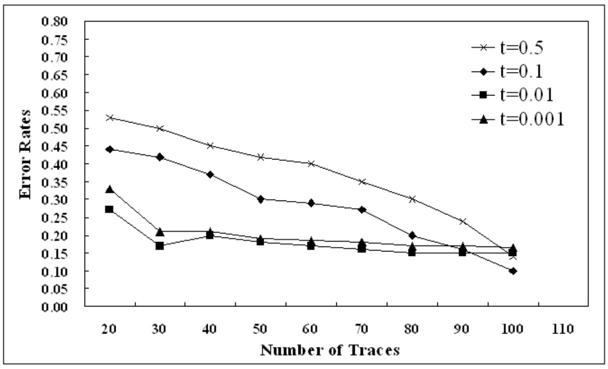
Figure 1 shows the performance of the IPRA algorithm with respect to different threshold values t used to select the candidate formulas, which are set to be 0.001, 0.01, 0.1 and 0.5, respectively. From the results, we find that error rate is sensitive to the choice of threshold. Generally, thresholds shall not be set to be extremely smaller or bigger. A bigger threshold will miss out some useful formulas, meanwhile a smaller threshold will cover some formulas with noise. From these experiments, it is shown that when the threshold is set to be 0.1, the mean average accuracy is optimal. Furthermore, the error bars representing the confidence intervals, show that our algorithm performance is stable.

The result also shows the relationship between the accuracy of learned model and percentage of observed intermediate states. In most cases, the more observations we have, the lower the error rate will be, which is consistent with our intuition. However, there are some cases, e.g., when threshold *t* is set to be 0.5, and there are only 1/4 of the states observed, the error rate is lower than the case when 1/3 of the states are given. These cases are not consistent with our intuition, but they are possible, since when more observations are obtained, the weights of their corresponding formulas go up and the weights of other formulas may go down in the whole learning process. Thus, if the threshold *t* is still set to be 0.5, some formulas which were chosen before are missed out, and the error rate will be higher. Thus, we conclude that in these cases, we need to reduce the value of the threshold correspondingly to make the error rate lower.

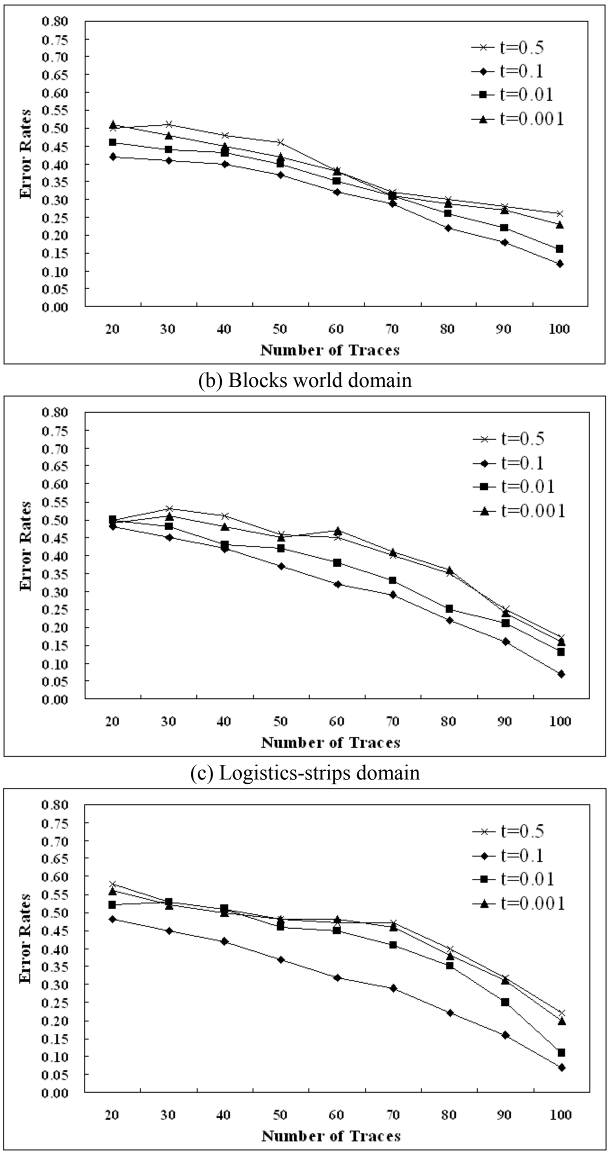
When threshold is set to 0.1, comparing the error rates of the four domains, it is obviously observed that the error rate of more complicated domain (with more predicates and actions) is generally higher than that of other domains, while with the increase of the number of plan traces, the error rate will decrease to about 10%. The reason is that in those complicated domains, a large number of predicates and actions will result in more candidate formulas of preconditions and conditional effects. In this case, if we don't have enough number of plan traces, then the noise in the experimental result will be quite serious. Therefore, in those complicated plan domains, the number of plan traces should be at least 100.

C. Plan Traces in Action-model Learning

To see how error rate are affected by the number of plan traces, we used different number of plan traces as the training data to evaluate the performance. In experiments, we assume that each plan trace had 1/5of fully observed intermediate states. These observed states were randomly selected. The process of generating state observations is repeated five times, where each time an error rate is generated under different selections. Figure 2 shows that error rates are affected by the number of given plan traces.



(a) Slippery-griper domain



(b) Blocks world domain

(c) Logistics-strips domain

(d) Zenotravel domain

Figure 2. Error Rates of different Number of Plan Traces (Observed Intermediate States is 1/5)

Generally, error rate decreases when the number of plan traces increase. When the number of plan traces is smaller, the error rate is higher. When the number of plan traces increases to some extent, the error rate decreases rapidly, but eventually it goes down slowly. It means that the difference between learned action models and correct ones is obvious when information is limited, but the difference will decrease when enough information is available. It can be speculated that learned action models will be approximate to correct ones, when enough number of plan traces is available.

When threshold is set to 0.1, comparing the error rates in the four domains, it is obviously observed that the error rate of the more complex domain (with more predicates and actions) is generally higher than the others. With the increase of the number of plan traces, the error rate will decrease to lower than 10%. The reason is that in those complex domains, a large number of predicates and actions will result in more candidate formulas of preconditions and effects. In this case, if we have not enough number of plan traces, then the noise in the experimental result will be quite serious. Therefore, in those complicated plan domains, the number of plan traces will be more than 100.

VI. Conclusion

In this paper, we adopt methods of automated planning and machine learning to translate software requirements into partial planning domain, formally described by PDDL language. Then we build up an action model learning algorithm to obtain complete planning domain and requirements specification. The proposed method can be used to acquire software requirement automatically. In future, we are planning to improve IPRA algorithm to apply it in the problem of system re-configuration at runtime.

References

[1] Zhao Y, Dong J, Peng T. Ontology Classification for Semantic-Web-Based Software Engineering. IEEE Transactions on Services Computing, 2009, 2(4):303~317.

[2] Drouin N, Badri M, Touré F. Analyzing Software Quality Evolution using Metrics: An Empirical Study on Open Source Software. Journal of Software, 2013, 8(10): 2462~2473.

[3] Fikes R, Nils J. N. Strips: A new approach to the application of theorem proving to problem solving. Artificial Intelligence, 1971,2(3):189~203.

[4] Soderland S, Weld D. Evaluating nonlinear planning. Technical Report TR 91-02-03. University of Washington CSE, 1991.

[5] Kautz H, McAllester D, Selman B. Encoding plans in propositional logic. In Proceedings of the 5th International Conference of Principles of Knowledge Representation and Reasoning, 1996.1084~1090.

[6] Avrim L. B, Merrick L. F. Fast planning through planning graph analysis. In Proceeding of the 14th International Joint Conferences on Artificial Intelligence, 1995:1636~1642.

[7] Malik G, Adele H, Craig K, Drew M, Ashiwin R, Manuela V, Daniel W, David W. PDDL-the planning domain definition language, http://www.informatik.uni-ulm.de/ki/Edu/Vorlesungen/GdKI/WS0203/pddl.pdf,1998.

[8] Smith D, Weld D. Confor m ant graphplan. Proceeding of 15th National Conference on Artificial Intelligence, 1998.

[9] Weld D, Anderson C, Smith D. Extending graphplan to handle uncertainty and sensing actions. Proceedings of 15th National Conference on Artificial Intelligence, 1998.

[10] Chen Y. Constrained partitioning in penalty formulations for solving temporal planning problems. Artificial Intelligence, 2009, 170(3):187~231.

[11] Philipple L. Algorithms for propagating resource constraints in AI planning and scheduling: existing approaches and new results. Artificial Intelligence, 2009,143(2):151~188.

[12] Omid M, Steve H, Anne C. On the undecidability of probabilistic planning and related stochastic optimization problems. Artificial Intelligence, 2013, 147:5~34.

[13] Younes H, Littman M. L, Weissman D. The first probabilistic track of the international planning competition. Journal of Artificial Intelligence Research, 2010,24:851~887.

[14] Zhou JP, Yin MH, Gu WX, Sun JG. Research on Decreasing Observation Variables for Strong Planning under Partial Observation. Journal of software, 2009, 20(2):290~304.

[15] Yan SY, Yin MH, Gu WX, Liu XF. Research and advances in probabilistic planning. CAAI Transactions on Intelligence Systems, 2008, 3(1):9~22.

[16] Yolanda G. Learning by experimentation: Incremental refinement of incomplete planning domains. In Proceeding of the Eleventh International Conference on Machine Learning(ICML 1994), 1994. 87~95.

[17] Thomas J. W, Michael L. L. Efficient learning of action schema and web-service descriptions. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI 2008), 2008. 714~719.

[18] Stanley K, Parag S, Matthew R, Pedro D. The Alchemy System for Relational AI. University of Washington, Seattle, 2005.

[19] Tao C, Li B, Gao J. A Systematic State-Based Approach to Regression Testing of Component Software. Journal of Software, 2013, 8(3):560~571.

[20] Pervez Z, Khattak A M, Lee S, Lee Y K. Achieving Dynamic and Distributed Session Management with Chord for Software as a Service Cloud. Journal of Software, 2012, 7(6):1403~1412.

[21] Wu K D, Liu W, Jin Z. Managing Software Requirements Changes Based on Negotiation-Style Revision. Journal of Computer Science and Technology, 26(5):890~907, 2011.

[22] Peng X, Yu Y, Zhao W. Analyzing evolution of variability in a software product line: From contexts and requirements to features. Information and Software Technology, 53(7):707~721, 2011.

[23] Perini A, Susi A, Avesani P. A Machine Learning Approach to Software Requirements Prioritization. IEEE Transactions on Software Engineering, 39(4):445~461, 2013.

[24] Portillo-Rodriguez J, Vizcaino A, Piattini M. Tools used in Global Software Engineering: A systematic mapping review. Information and Software Technology, 54(7): 663~685, 2012.

Hong He received his B.S. degree at Wuhan University of Technology in 1996, and M.S. degree at Xiangtan University in 2006. Currently, he works in Hunan Institute of Engineering as an associate professor. His research interesting is grid computing, cloud computing, distributed resource management.

Dongbo Liu received his master degree in Hunan University in 2004. Now he works in Hunan Institute of Engineering and is a Ph.D candidate in Hunan University. His research interests include distributed intelligence, multi-agent systems, high-performance application. He is now a student member of CCF in China, and worked as Senior Engineer in HP High-performance Lab.

Research on Automated Software Test Case Generation

Daisen Wei
College of computer science and technology, Shandong University, Jinan, China
Inspur Genersoft Co., Ltd. Jinan, China
E_mail: weids@inspur.com

Longye Tang
Inspur Genersoft Co., Ltd. Jinan, China
E_mail:tangly@inspur.com

Xueqing Li
College of computer science and technology, Shandong University, Jinan, China
E_mail: xqli@sdu.edu.cn

Ling Shang
Lifl, University of Science and Technology of Lille, Lille, France
E_mail: ling.shang@lifl.fr

***Abstract*—To improve the efficiency of software testing, a model-driven method is proposed to automatically generate test cases from UML design model. In it, PITCs (platform-independent test cases) are generated first from a UML design model. And then, according to the predefined rules, a process is implemented to transform PITCs into the corresponding PSTCs (platform-specific test cases). The experiment and comparison had showed that the method proposed in this paper was easier to be understood and implemented by users to generate test cases than the ones existed.**

***Index Terms*—software testing, test case, PITC, PSTC, transformation rule**

I. Introduction

Testing is one of key steps in software development and even runs through the whole software lifecycle. And now, it is regarded as one of the most important and effective ways to improve the quality of software by trying to find the potential faults that may exist in source codes. The first step of implementing it is to design and generate test cases, a set of data generally including input and the expected output that satisfy the design or testing requirements of SUT (system under test).

In recent decades, UML has been one of the most popular design tools widely used in some key subprocesses of software lifecycle, especially design and testing. So, UML based testing has already been paid more attention by researchers from both academia and industry. UML models, however, are generally in the form of diagrams, so it is very hard to directly generate test cases from them. One approach adopted to deal with it was to create the corresponding test models from UML design ones, from which test cases were generated. A specification, UML2.0 test profile, had also been released by OMG in 2004[1] to support this model-based testing. And on the basis of it and the methodology of MDD (model driven development), model-riven testing were also widely researched [2]. One of the advantages of this method is to provide a good way to automatically generate test cases through model transformation.

Model-riven testing, however, is also a challenge for testers due to some reasons, one of which is that how the input space of SUT is to be defined and then from which the appropriate testing data are selected to form test cases. This is one of the important factors to influence that whether test cases can automatically be generated or not. And this problem can also cause that a test case would usually be defined as a form, in most methods existed, being different from that of the executable one with the real input/expected output data used in practical testing. Obviously, all these can make the process of test case generation low efficiency and time consuming. So, research on automated test case generation from UML models is necessary and very worthy of doing. More attention had also been paid to it in academia and industry and some achievements had been obtained in recent years [3].

In this paper, one model-driven method was proposed to generate software test cases, the executable ones, from a UML design model. And state diagram was selected in a case given in section V below. The idea of this method

Manuscript received January 20, 2014; revised May 22, 2014; accepted May 23, 2014.

doi:10.4304/jsw.9.11.2868-2876

originated from MDD, in which object source codes were automatically generated by model transformation from UML design models. The basic process to implement it was that PITCs were generated from a UML design model, PIM (platform independent model), and then data mapping rules from PITC to PSTC were defined to guide the PSTCs generation from the corresponding PITCs.

The remainder of this paper is organized as follows. In section II, some concepts such as PITC, PSTC and transformation from PITC to PSTC etc. are defined. In section III, the processes including PITC generation, PITC-to-PSTC transformation and PSTC generation are described respectively. In section IV, a case is given to show the whole PITC-to-PSTC process. In section V, the method is analyzed and compared to some related ones from two perspectives. And the conclusions are given in section VI.

II. Some Concepts

A test case in software engineering is a set of conditions or variables under which a tester will determine whether an application, software system or one of its features is working as it was originally established for it to do. In this paper, a software test case is defined as follows:

Definition 1: Test Case (TC). A test case is a 3-tuples: (Id, InitState, Data), where: (1) the "Id" is an unique and numbered string assigned to each test case, and (2) the "InitState" is the current state of SUT, followed by the execution of the test case selected, and (3) the "Data" is a set in the form of {<ini, eouti>} in which the "ini" represents the input and the "eouti" the expected output.

Note that: the data pair <ini, eouti> implements an atomic testing step. The "atomic" means that the <ini, eouti> is the minimum input-output data pair, that is, there are no other input or output data between the "ini" and the "eouti".

Definition 2: Platform-Independent Test Case (PITC). A PITC is a test case generated from system design model, in which all data including parameters' types, values and syntax are not bound to any programming language such as JAVA or a platform specification such as .NET and JSP. For example, the following data string is a PITC being designed to verify the validity of user's identity before he or she tries to login and enter a web system.

(tc1, login page, < [*name, password*], *main page of the system>)*

In the PITC tc1, the "login page" means the page loaded for users to login and also represents the current system state before the users' account are entered. The "[name, password]" represents the user's account as input. The "main page of the system" means the loaded page as output, that is, the new system state after the user's name and password are entered and then submitted. As we can see, all data involved in tc1 are platform-independent. In this paper, this type of data is called as platform-independent data (PID), the form of data without definite values in a PIM.

Definition 3: Platform-Specific Test Case (PSTC). A PSTC is the refined version of a PITC, in which all data involved are platform-specific and the syntax of them conforms to a specific programming language or platform specification. The following PSTC tc11 corresponding to the above PITC tc1 is given as follows:

(tc11, UserLogin.jsp, < ["administrator","12@abMN67"], default.jsp >)

Note that: the syntax of tc11 complies with the object specification JSP. In it, the NO. of test case, tc11, can be changed as required. The "*UserLogin.jsp*" is a JSP page that means the concrete login page of SUT. The pair "*['administrator','12@abMN67']*" represents the login account including the user's name and password. And the "default.jsp" represents the main page of SUT as expected output after a user's account is verified to be true.

From the definition 4 and 5 above, it can be seen that a PSTC can be executed manually but a PITC cannot. In this example, the syntax of all data in tc11 conforms to the platform specification JSP. The tc11 is also a test case that can be executed manually. And if required, it can further be transformed automatically into a script that can directly be executed by test tools. In this paper, this type of data is called as platform-specific data (PSD), the form of ones with definite values in a PSM (platform specific model). In MDD, a PSM is always transformed from a specific PIM.

The figure 1 below shows the test case generation process marked with the directed real lines. And it consists of two subprocesses, one of which is the PITCs generation from PIM and the other is the PSTCs generation through the data transformation from the corresponding PITCs.

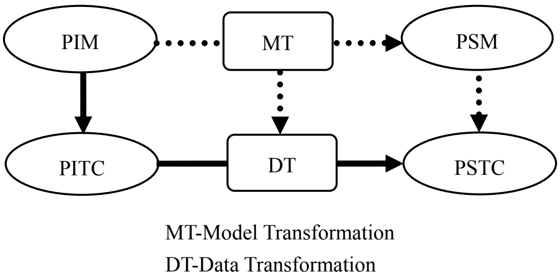


Figure 1 Model-driven test case generation

Note that: in figure 1, the process "MT" means the model transformation from PIM to PSM, which is implemented according to the predefined rules. And this transformation process involves two subprocesses of refinements, logic structure or syntax and data object, from PIM to PSM. The process "DT" represents the refinement of data objects from PIM to PSM. So, DT is only one part of MT.

The data object is defined as follows:

Definition 4: Data Object (DO). A DO is an object with attributes that appears in test cases for SUT.

In object-oriented methodology, an object usually consists of attributes and methods or functions. If an object appears in a test case, only the values assigned to the corresponding attributes of it are involved generally.

So, the DO is defined here just from the perspective of test cases, that is, only the attributes and their values of an object are focused and used.

And in this paper, a DO has two forms, one of which is PIDO (platform independent data object) and the other PSDO (platform specific data object). The only difference between them is that PIDO is defined or used in PIM and the corresponding PSDO in PSM. Correspondingly, the attributes of a PIDO are named as PIA (platform independent attributes) and the ones of a PSDO as PSA (platform specific attributes). The states of a PIDO are called as PIS (platform independent states) and the ones of a PSDO as PSS (platform specific states). And all these terms are used in the following figure 2 and definition 5.

For instance, in the PITC tc1 given above, two attribute parameters, name and password, together define a DO user account. The same DO user account is described in the form of [name, password] in tc1 and correspondingly, that in the form of [“administerator”,”12@abMN67”] in the PSTC tc11. Obviously, it shows that a PITC and all data objects in it are platform independent and a corresponding PSTC and all the same data objects in it platform specific.

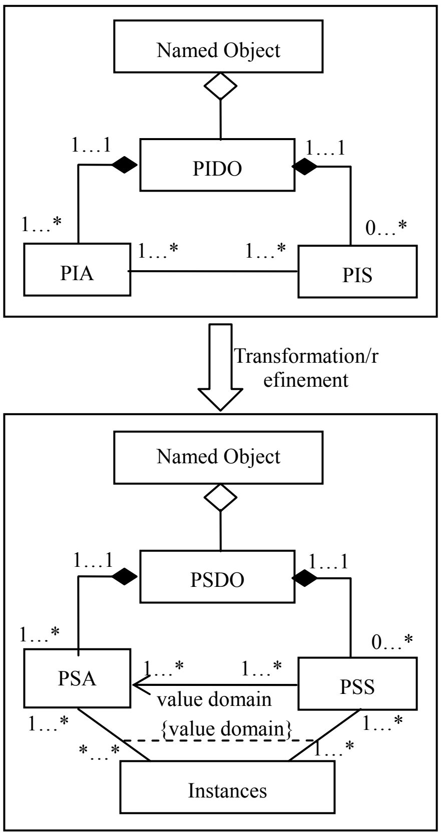


Figure 2 Transformation from PIDO to PSDO

Definition 5: ⓣ. The ⓣ is defined as the operation of transforming a PITC into the corresponding PSTC (s).

Because each attribute variable can be assigned to different values that may represent different states of a data object, the ⓣ implements a one-to-many function. That is, one PITC can be transformed into many PSTCs. One example is that each variable should be assigned at least to two constant values, the valid one accepted by SUT and the invalid one failed in SUT.

The detailed process of transforming a PITC into the PSTC (s) was given in section III(C) below.

In essence, the operation ⓣ implements the process of data refinement between PIM and PSM, in which only constant values from the data space of PSM are assigned to the corresponding attributes of PIDO in PIM. Because the ⓣ does not change the semantic of these attributes, it should keep the property preservation in this transformation from a PITC to the corresponding PSTCs.

For example, the transformation from the PITC tc1 to the PSTC tc11 given above can be correspondingly described as the following table I.

TABLE I

AN EXAMPLE: TRANSFORMATION FROM PIDO TO PSDO

DO	PIDO in the PITC *tc1*	PSDO in the PSTC *tc11*
page	login page	*UserLogin.jsp*
user account	name, password	*“administrator”, ”12@abMN67”*
page	main page of the system	*default.jsp*

III. PITCS AND PSTCS GENERATION

A. Generating the Executable Paths

A test case always corresponds to an executable path in SUT. In this paper, the approach to generating test cases is on the basis of UML design models such as activity and state diagram. So, in order to be retrieved easily, the UML diagram used must be described as a correspondingly directed graph. After that, all executable paths can be generated by retrieving this graph. And such a graph is named as UML Graph (UG).

In the case given in section IV below, an UG was created from UML state diagram. In this UG, a node represents a system state and a directed edge a transfer between two adjoining states.

Definition 5 Executable Path (EP). An EP is a path with one unique start node and one tail node in UG.

In some cases, UG may involve loop. Generally, a loop appears in an executable path only zero and 1 time in the path coverage of software testing. So, before the graph-retrieved algorithm is implemented, this should be configured as a constraint condition.

Note that: the graph-retrieved algorithm adopted in this paper is general and common to that we study in the course of data structure.

In this paper, a set named as PATH, {p1, p2, p3, p4…}, is defined to store all executable paths generated and each pi in it corresponds to an executable path. A detailed case will be given in section V below.

B. PITCs Generation

After the set PATH including all executable paths of SUT is generated, PITCs can be generated from it. In order to complete this process, the contents of each node and edge of a path pi in PATH should be determined and given. And the following table II is defined to do it for providing the information needed.

In such table, the state with input and output corresponding to each node is given clear. The table should be created in advance and the description of each item in it should be given accurately. The table is named as SET (state and event table).

TABLE II

THE STATE AND EVENT TABLE (SET)

State NO.	Current state	Input	Expected output

In table II above, the column State NO. corresponds to the NO. of each node in UG. The column Current state means the system state followed by the test case execution. The column Input represents the data that may be entered in the current state and Expected output the ones appeared in next system state. And the Input and Expected output together determine a corresponding data object. Exactly, each pair of the input and output required in a state of SUT is given manually by analyzing and determining the input boundary of each attribute of one data object. All input and excepted output are in the form of platform- independent data.

The following process is given to implement PITCs generation from the PATH according to the table SET created in advance. And each executable path included in PATH is processed one by one.

Step1: take the ith path pi from PATH, and then determine the initial state of the first node of pi. The initial state is the content of the column Current state in the table SET.

Step2: determine the <inj, eoutj>, the input and expected output of the jth state node in the current path pi. Here, the inj is the content of the column input corresponding to the jth state in SET, and it may be null; the eoutj is the content of the column Expected output corresponding to the jth state in SET.

Step3: continue to take the next adjoining node in the current path pi and then go to the Step2 until all nodes in pi have been processed eventually. Note that, the last node in each path pi is the end node of UG and that is marked with “END”.

Step4: according to the processed order of each executable path in PATH, each PITC generated is numbered with a string, for example tc1, tc2….

The above process from step1 to step4 will be repeated until all executable paths in PATH have completely been transformed at last.

A PITC generated according to the above process is not an executed test case because in it all attribute parameters involved are not assigned to concrete values that conform to a high programming language or platform specification. So, it must be transformed into the corresponding PSTC, one type of executable test case defined in this paper.

C. *PSTCs Generation*

In this section, the work is just to identify all data objects and their input variables involved in each PITC and then choose appropriate values from the input space for all variables.

The value space of an attribute variable generally consists of a valid subspace, in which values are expected to be accepted by SUT, and a failure one, in which values are invalid and expected to cause the SUT to produce some kind of failure response.

To implement this process, the table PISDMT (platform independent-specific data mapping table) is defined in the form of the table III below which conforms to the figure 2 and Definition 5 given in section II. Exactly, the PISDMT is used to describe the information about PIDO and PSDO and the mapping relation between them, which is very essential for the generating process from PITCs to PSTCs.

In Table III below, the column DO is used to identify each unique data object. The BSF (basic state feature) is to describe the state features, valid or invalid, of the value space of the current DO. The PSA refers to the platform specific counterpart of the current DO. The PSS (platform specification) is to describe the valid or invalid values assigned to the current DO under the final application platform.

TABLE III

PLATFORM INDEPENDENT-SPECIFIC DATA MAPPING TABLE

(PISDMT)

DO	BSF	PSA	Value Space	PSS

According to the contents of PISDMT, the detailed transformation process is defined as follows:

PITC ⓣ PSTC ≡ PIDO: (PIA, PIS) ⓣ PSDO:(PSA, PSS）

In it, the “≡” represents “being defined”. It means that the transformation process from a PITC to the corresponding PSTC(s) is equal to that from the PIAs and PISs of each PIDO in a PITC to the PSAs and PSSs of the corresponding PSDO in table PISDMT. In fact, it completes the process in which all variables in PITC are assigned to the concrete values that comply with the syntax of the final application platform determined.

The main contents of a test case usually include the initial state and a set of data pair including input and expected output. For each part of it, one corresponding transformation rule is described as follows:

(1) Transforming the initial state in a PITC into the one in a PSTC

Rule1: the initial state transformation

IF (∀PITC (∃do∈PITC.InitState Λ do==PISDMT.DO Λ ∃pss∈ PISDMT.PSS)) THEN

PITC.InitState.do← pss

Here, the “←” represents “being replaced” and the “∈” “being included”. It means that if a data object do exists in the InitState of a PITC, the do is to be replaced by the corresponding data pss included in the current row of the table PISDMT.

(2) Transforming the input and output in a PITC into the ones in a PSTC.

Rule2: the input and output data transformation

IF (∀PITC (∃do ∈ PITC.Data Λ do==PISDMT.DO Λ ∃pss∈ PISDMT.PSS)) THEN

PITC.Data.do←pss

It means that if a data object do exists in the <ini, outi> of a PITC, the do is to be replaced by the corresponding data pss included in current row of the table PISDMT.

Note that: two points are very important for the PITC-to-PSTC transformation process and should be further elaborated as follows.

(1) Correctness of TRs (transformation rules). TRs in the form of the above table-transformation process should keep consistent with the mapping relation showed in figure 2. And they implement the PIA-to-PSA transformation process for each data object. Because this process only assigns platform-specific values to the attributes of a data object but not changes the semantic of them, it holds the property preservation. That is, the semantic of an object in PIM/PITC cannot be changed and continues to be preserved in the corresponding PSM/PSTC.

(2) Traceability of transformation process. According to figure 2 and table III above, between the platform -independent data and the corresponding platform -specific ones is one-to-many relation. And that is also the relation between a PITC and the corresponding PSTC (s). Therefore, the PITC can be traced uniquely from a PSTC.

According to the transformation rules defined above, each PITC can be processed to be transformed into some PSTCs as follows:

Step1: According to Rule1, take a PITC tc_i from the set PITCs, then replace the DO in PITC.InitState with the corresponding *pss* in PISDMT.PSS.

Step2: Take the first data pair $<in_1,eout_1>$ of tc_i, then replace the DO in $<in_1,eout_1>$ with the corresponding *pss* in PISDMT.PSS. If the DO has many values in PISDMT.PSS, respectively generate a correspondingly new test case by using each *pss* to replace the DO until the PISDMT.PSS becomes empty.

Step3: Go on to process next $<in_j,\ eout_j>$ of tc_i by replacing all DOs in it. And repeat Step2 until all input-output data pair has completely been processed at last.

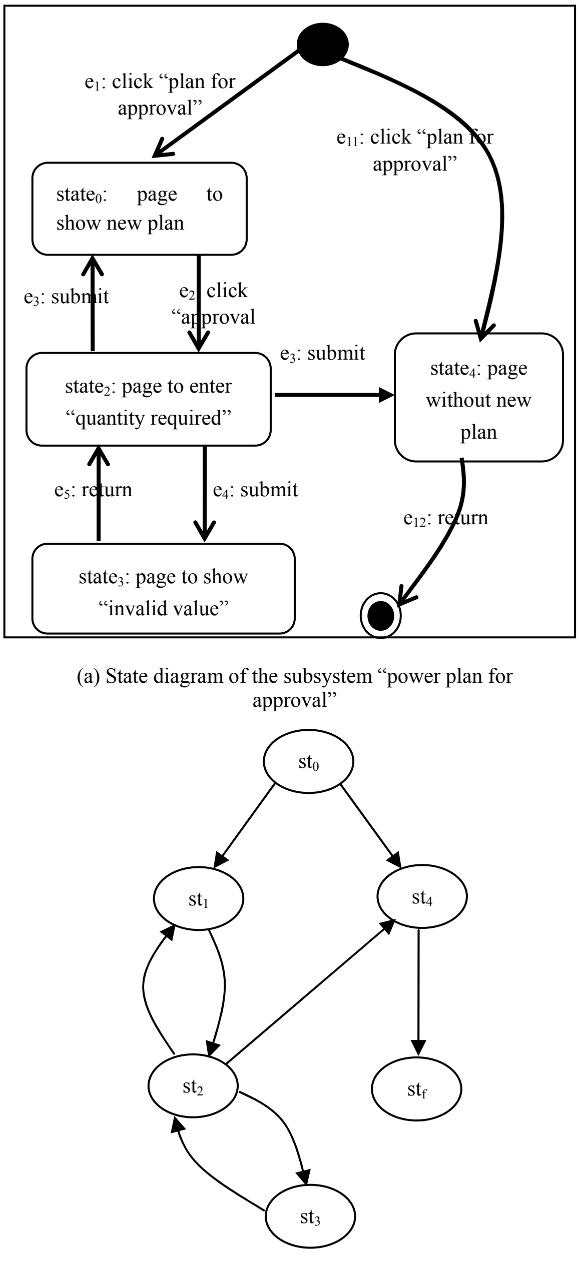
Step4: Go to Step1 to take and process next PITC tc_{i+1} of the set PITCs, and repeat from Step1 to Step3 until the set PITCs becomes empty.

In the above process, the main work is to replace each DO in each PITC with all valid and invalid data, the corresponding value pss in PISDMT.PSS. After this process, the set PSTCs can be generated and test cases in it can be executed manually.

IV. A Case Study: State Diagram-Based Unit Test Case Generation

The following case is about a subsystem "power plan for approval" which can be used to online submit the power quantity for next month to the administration and apply for approval.

Step1: Create the directed graph for retrieval from the state diagram of the subsystem "power plan for approval", seen in figure 3 (b) below. Note that: each event in figure 3(a) is to be viewed as one part of the input of one source state node corresponding to it. In this abbreviated graph, the st0 corresponds to the start node in state diagram and the stf the unique end node.



e1: click "plan for approval"
e11: click "plan for approval"
state0: page to show new plan
e3: submit
e2: click "approval
e3: submit
state2: page to enter "quantity required"
state4: page without new plan
e5: return
e4: submit
e12: return
state3: page to show "invalid value"
st0
st1
st4
st2
stf
st3

(a) State diagram of the subsystem "power plan for approval"

(b) The corresponding graph for retrieval

Figure 3 A case: state diagram and the corresponding directed graph from it

Step2: Retrieve the graph in figure 3(b) to generate all executable paths, a set PATH, of the subsystem. The set PATH is *{p1, p2, p3, p4}*, where:

p1= (st_0, st_1, st_2, st_1, st_2, st_4, st_f) // Check and approve two plans continually, that is, including the loop path one time between st1 and st2.

p2= (st_0, st_1, st_2, st_4, st_f) // no loop between st_1 and st_2.

p3= (st_0, st_1, st_2, st_3, st_2, st_4, st_f) // loop one time between st_2 and st_3.

p_4= (st_0, st_4, st_f) // no new plan for approval.

Step3: define the state and event table (SET) as table IV below, and then, according to it and the PATH generated in step2 above, implement the process given in section III(B) to generate the corresponding set PITCs.

The PITCs generated for this case is the set {*tc_1*, *tc_2*, *tc_3*, *tc_4*}, where

(tc_1, Initial system page, {< Click “check and approve”, page to show new plans >, < Click “approval”, page to enter “quantity required”>, <enter Valid quantity and click “submit”, page to show new plans >, < Click “approval”, page to enter “quantity required”>, <enter Valid quantity and click “submit”, page without new plans >, < Click “Return”, Initial system page >});

(tc_2, Initial system page, {< Click “check and approve”, page to show new plans >, < Click “approval”, page to enter “quantity required”>, <enter Valid quantity and click “submit”, page without new plans >, < Click “Return”, Initial system page >});

(tc3, Initial system page, {< Click “check and approve”, page to show new plans >, < Click “approval”, page to enter “quantity required”>, <enter Invalid quantity and click “submit”, Page to show “Invalid value”>, < Click “approval”, page to enter “quantity required”>, <enter Valid quantity and click “submit”, page without new plans >, < Click “Return”, Initial system page >}); // given that the maximum is 1000, so 1001 is an invalid number.

TABLE IV

THE STATE AND EVENT TABLE (SET)

State NO.	Current state	Input	Expected output
st_0	Initial system page	Click “check and approve”	page to show new plans
st_0	Initial system page	Click “check and approve”	page without new plan
st_1	page to show new plans	Click “approval”	page to enter “quantity required”
st_2	page to enter “quantity required”	Valid quantity, click”submit”	page to show new plans
st_2	page to enter “quantity required”	Valid quantity, click”submit”	page without new plans
st_2	page to enter “quantity required”	Invalid quantity, click”submit”	Page to show “Invalid value”
st_3	Page to show “Invalid value”	Click “Return”	page to enter “quantity required”
st_4	page without new plan	Click “Return”	Initial system page

(tc_4, Initial system page, {< Click “check and approve”, page without new plans >, < Click “Return”, Initial system page >}).

Step4: define the table PISDMT. According to definition 4 given in section II, a data object is one with attributes that determine the input space of SUT. The table PISDMT for the subsystem “power plan for approval” can refer to Table V defined below.

TABLE V

PISDMT FOR THE SUBSYSTEM “POWER PLAN FOR APPROVAL”

DO	BS	PSA	Values space	PSS
quantity	valid	Valid quantity	[0,1000]	50/150/1000
quantity	invalid	Invalid quantity	(1000,+∞)/(-∞,0)	1001/-1/sgh123

Step5: on the basis of PISDMT, generate the set PSTCs from the set PITCs according to the process given in the subsection C of the former section III.

Note that: in this paper, the concrete values assigned to one corresponding attribute variable of a data object can be manually defined in advance after the table PISDMT is created. Of course, they can also be generated temporarily according to the value space but this can cause some performance problems such as time consuming. Generally, all valid values for data objects can be included a test case named as a success one. And each invalid value outside of the input space should individually correspond to a test case called as a failure one. So, the following set PSTCs generated in this case includes 6 executed test cases, and that is a set {tc_1, tc_2, tc_3, tc_4, tc_5, tc_6 }, where

(tc_1, Initial system page, {< Click "check and approve", page to show new plans >, < Click "approval", page to enter "quantity required">, <enter "50" and click "submit", page to show new plans >, < Click "approval", page to enter "quantity required">, <enter "150" and click "submit", page without new plans >, < Click "Return", Initial system page >}); //continue to check and approve two plans

(tc_2, Initial system page, {< Click "check and approve", page to show new plans >, < Click "approval", page to enter "quantity required">, <enter "50" and click "submit", page without new plans >, < Click "Return", Initial system page >}); //only check and approve two plans

(tc_3, Initial system page, {< Click "check and approve", page to show new plans >, < Click "approval", page to enter "quantity required">, <enter "1001" and click "submit", Page to show "Invalid value">, < Click "approval", page to enter "quantity required">, <enter "150" and click "submit", page without new plans >, < Click "Return", Initial system page >}); // given that the maximum is 1000, so 1001 is an invalid number.

(tc_4, Initial system page, {< Click "check and approve", page to show new plans >, < Click "approval", page to enter "quantity required">, <enter "-1" and click "submit", Page to show "Invalid value">, < Click "approval", page to enter "quantity required">, <enter "1000" and click "submit", page without new plans >, < Click "Return", Initial system page >}); // given that the maximum is 0, so -1 is an invalid number.

(tc_5, Initial system page, {< Click "check and approve", page to show new plans >, < Click "approval", page to enter "quantity required">, <enter "sgh123" and click "submit", Page to show "Invalid value">, < Click "approval", page to enter "quantity required">, <enter "50" and click "submit", page without new plans >, < Click "Return", Initial system page >}); // given that the value is only a number between 0 and 1000, so a string including letter is invalid.

(tc_6, Initial system page, {< Click "check and approve", page without new plans >, < Click "Return", Initial system page >}). // have no plan for approval

V. Experiments And Analysis

A.Z. Javed etc. [4] proposed an approach to model-driven component testing. According to it, the meta-models corresponding to PIM and PSM and the transformation rules from PIM to PSM ware defined respectively, and then the process to generate test cases was implemented on the basis of them. But in [4], only an idea was given and the detailed implementing method was absent. Moreover, it "generates" test cases through PIM and PSM and the transformation rules between them. This was also different from the method proposed in this paper, in which test cases were generated by the means of transformation from PITC to PSTC.

Another approach to implementing model-driven testing was to create test models corresponding to PIM and PSM respectively and then to define the transformation rules between elements of them. This way is also adopted by most researchers [2] [5]-[11]. UML 2.0 test profile had also been released as a specification by OMG in 2004. Being different from these methods proposed, the one in this paper was to generate test cases through creating meta-models of platform-independent data and platform-specific one and the mapping relationship between them. The mapping and transformation rules in it were defined in the form of relation table, which made the process of test case generation easy to be understood and implemented by users. For instance, all data involved it can easily be handled by database or other forms.

Tcases [12] is an open-source tool for black-box test case generation. With Tcases, users can define the input space for the SUT and the level of coverage that they want. Then Tcases can generate a minimal set of test cases that meets testing requirements. Tcases was guided by the coverage of the input space of SUT. In Tcases, the input space and the functions of SUT were defined and described in two individual XML files respectively. It used input values to generate two types of test cases — "success" cases, which use only valid values for all variables, and "failure" cases, which use a failure value for exactly one variable.

PItoPSTcases is a simple tool to generate test cases, which implements the method proposed in this paper. It was developed by our team using Eclipse (SDK 3.7). In it, all related tables were described in the form of databases of SQL Server. The architecture of PItoPSTcases is illustrated in the figure 4 below.

The main differences between two tools are listed in the table VI below. According to the content of it, the input space and functions of SUT must be created in the form of XML file respectively before Tcases runs. And the functions here are used to describe the logic of SUT.

But, for PItoPSTcases, only data object and the mapping relation between PIDO and PSDO are to be described as tables, and the logic of SUT can directly be obtained from the selected UML design model and the graph form it. So, on the basis of the table SET and PISDMT created in advance, it can generate the executed test cases, PSTCs. The values of an attribute variable were selected and configured manually in the form of table. To some extent, this improves the accurateness of test cases generated. Additionally, the section of input data is also easier to be implemented than that in Tcases. However, the accurateness of test cases generated by using Tcases are heavily dependent on that of XML files, in which all variables and the conditions or constraints related to them must be recorded accurately.

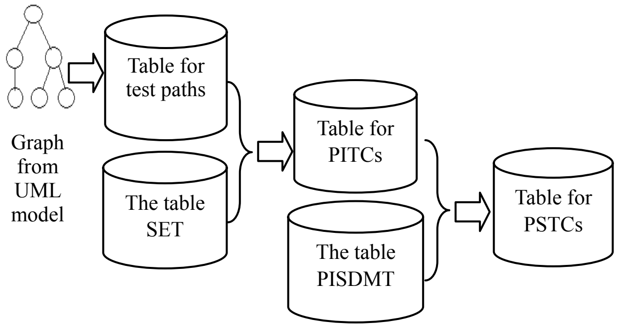


Figure 4 The architecture of PItoPSTcases

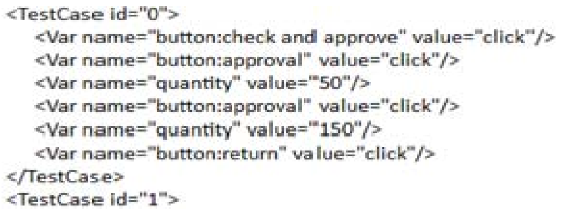
TABLE VI
DIFFERENCES BETWEEN TCASES AND PITOPSTCASES

Comparison item	Tcases	PItoPSTcases
model type	XML file	UML/graph from UML
Input	(1)Xml for input space of SUT (2)Xml for functions of SUT	(1)Table for describing data objects (2)Table for the mapping from PITC to PSTC
Output	Xml for test cases	Table(Text string) for test cases

Other tools given in some researches were to generate test cases just by selecting data from the input and output space in a random way. This can also make users face the problem that the selected data cannot satisfy the testing requirements well.

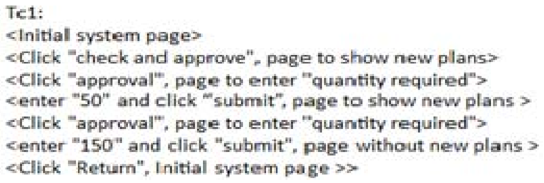
Both Tcases and PItoPSTcases were implemented at the same environments as follows: (1) OS: windows 7 core 64; (2) hardware: 10 computers with CPU (Intel(R) Core (TM) i5-3320M, 2.60GHz) and RAM 8GB.

Two fragments of screen shot of their output are respectively given in the following figure 5 and figure 6.



```
<TestCase id="0">
    <Var name="button:check and approve" value="click"/>
    <Var name="button:approval" value="click"/>
    <Var name="quantity" value="50"/>
    <Var name="button:approval" value="click"/>
    <Var name="quantity" value="150"/>
    <Var name="button:return" value="click"/>
</TestCase>
<TestCase id="1">
```

Figure 5 Output format for Tcases



```
Tc1:
<Initial system page>
<Click "check and approve", page to show new plans>
<Click "approval", page to enter "quantity required">
<enter "50" and click "submit", page to show new plans >
<Click "approval", page to enter "quantity required">
<enter "150" and click "submit", page without new plans >
<Click "Return", Initial system page >>
```

Figure 6 Output format for PItoPSTcases

The following figure 7 showed that two tools, Tcases and PItoPSTcases, were respectively compared, from two perspectives, the average time (run 10 times) spent to create the input and the one spent to generate test cases for the same subsystem "power plan for approval".

From the figure 7 below, it can be concluded that the average time spent in preparing the input by PItoPSTcases was lower about 37.5% than that spent by Tcases. And the average time spent by PItoPSTcases to generate test cases is lower about 43% than that spent by Tcases.

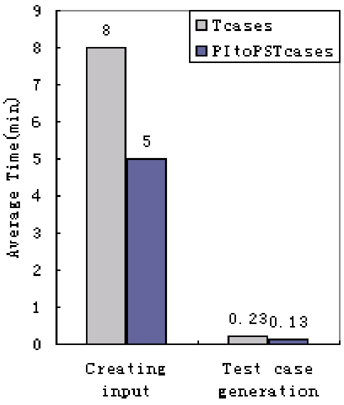


Figure 7 The average time comparison: that of creating input and that of generating test cases

VI. CONLUSIONS

To design and generate test cases is one of the most important steps to implement software testing. And Based on the idea of MDD, one method was proposed in this paper to generate executed test cases. All input data are described in the form of table which can be created and used easily. And a simple experiment given in section V showed that the method had a larger advantage of efficiency in time spent.

Of course, the type of test case defined in this paper is only executed by hand now. The next work for us is to improve the tool to generate test cases in the form of script which can directly be executed by some test tools such as xUNIT.

REFERENCES

[1] OMG. *Ptc/04-04-02: UML 2.0 Testing Profile*, Finalized Specification.

[2] R. Zhen. Model-Driven Testing with UML 2.0 [C]. In: Proceeding of Second European Workshop on Model

Driven Architecture (MDA), Canterbury, Kent, University of Kent (2004), pp. 179-187.

[3] S. Anand, E. Burke and T. Chen *et al*. An orchestrated survey of methodologies for automated software test case generation. *Journal of Systems and Software*. 2013, Vol. 86, issue 8, pp. 1978-2001.

[4] A. Javed, P. Strooper and G. Watson. Automated Generation of Test Cases Using Model-Driven Architecture[C]. In: Proceeding of Second International Workshop on Automation of Software Test, 2007. AST '07. pp. 1-7.

[5] Q. Yuan, J. Wu, C. Liu and L. Zhang, A model driven approach toward business process test case generation[C]. In Proc. of the 10th International Symposium on Web Site Evolution (WSE), 2008, pp. 41-44.

[6] M. Mussa, S. Ouchani, W. Sammane and A. Hamou-Lhadj. A Survey of Model-Driven Testing Techniques [C]. In: Proceeding of 9th International Conference on Quality Software, 2009, QSIC '09. pp. 167-172.

[7] N. Li, Q. Ma and J. Wu *et al*. A Framework of Model-Driven Web Application Testing[C]. In: Proceeding of 30th Annual International Computer Software and Applications Conference, 2006. COMPSAC '06. Vol. 2, pp. 157-162.

[8] J. Gutierrez, M. Escalona and M. Mejias *et al*. An approach for Model-Driven test generation[C]. In: Proceeding of Third International Conference on Research Challenges in Information Science, 2009. RCIS 2009. pp. 303-312.

[9] D. Mathaikutty, S. Ahuja, A. Dingankar and S. Shukla. Model-driven test generation for system level validation[C]. In: Proceeding of IEEE International High Level Design Validation and Test Workshop, 2007. HLVDT 2007.pp. 83-90.

[10] F. Wang, S. Wang, and Y. Ji. An Automatic Generation Method of Executable Test Case Using Model-Driven Architecture[C]. In: Proceeding of 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC), 2009, pp. 389-393.

[11] M. Felderer, P. Zech and F. Fiedler *et al*. Model–driven System Testing of Service Oriented Systems [C]. In: Proceeding of the 9th International Conference on Quality Software (QSIC'2009), 2009. pp. 1-8.

[12] tcases - A model-driven test case generator. http://code.google.com/p/tcases/ (May, 2013)

D. Wei Now he is a PHD candidate of college of computer science & technology in Shandong University. His interests are software development and testing.

L. Tang He is a PHD of Inspur Genersoft Co., Ltd. His interests are ERP software development &testing.

X. Li He is a Professor of college of computer science & technology of Shandong University and also the corresponding author of this paper. His interests are software development & testing.

L. Shang He is a PhD of University of Science and Technology of Lille, Lille, France. His interests are distributed computing and software quality.

A Study of Dependency Features for Chinese Sentiment Classification

Pu Zhang[1,2,*], Zhongshi He[1],Lina Tao[3]
[1] College of Computer Science, Chongqing University, Chongqing ,China
[2] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing ,China
[3] Chongqing Communications Research and Design Institute Co.,Ltd. ,China Merchants,Chongqing, China
[*]Corresponding author Email: zhangpu1977@yahoo.com

***Abstract*—Syntactic dependency features, which encode long-range dependency relations and word order information, have been employed in sentiment classification. However, much of the research has been done in English, and researches conducted on exploring how features based on syntactic dependency relations can be utilized in Chinese sentiment classification are very rare. In this study, we present an empirical study of syntactic dependency features for Chinese sentiment classification. First, we consider two types of feature sets (word unigrams and word-dependency relations), three commonly-used feature weighting schemes (term presence, term frequency, and TF-IDF), and two well-known learning methods (Naive Bayes and SVM) to evaluate the performance of different classifiers. Then, we use ensemble technique to combine different types of features and classification algorithms. Specifically, two types of ensemble methods, namely average combination method and meta-learning combination method, are evaluated for two ensemble strategies. Through a wide range of comparative experiments conducted on two widely-used datasets in Chinese sentiment classification, finally, some in-depth discussion is presented and conclusions are drawn about the effectiveness of dependency features for Chinese sentiment classification.**

***Index Terms*—sentiment analysis, sentiment classification, dependency features, ensemble learning**

I. Introduction

Sentiment analysis, also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, attitudes, and emotions from written language [1], and it has received increasing attention from academics and practitioners in recent years. Among various sentiment analysis tasks, an important sub-task is sentiment classification, which aims to classify an opinionated piece of text as expressing an overall positive or negative polarity.

A very popular technique for sentiment classification is supervised machine learning approach. The performance of the approach is heavily dependent on the choice of algorithms and the representation of documents. Since the pioneering work of Pang et al. [2], various supervised classification algorithms including Naive Bayes (NB), Maximum Entropy (ME), Winnow, KNN, ANN and support vector machines (SVM) have been employed as machine learning methods [3-6]. Apart from the choice of algorithms, the representation of documents also plays a critical role in supervised machine learning approaches for sentiment classification. Early work in [2] showed that using word unigrams as features and encoding each word feature in the set by its presence or absence in the document performs quite well in classification. In this case, a document is represented as a binary vector.

In subsequent research works, various kinds of features such as word Ngrams [4,8], character Ngrams [4,7,9-10], POS based features [11-12], substring-group features and sentiment word features [10] have been exploited.

With the attempt to capture word relation information behind the text, dependency-based features were also utilized in sentiment classification [13-16]. However, these works only focused on English documents, and the research results of sentiment classification of English are often unable to be directly applied to Chinese ones owing to unique way of emotional expression in Chinese and a larger variety of syntactic dependency and a higher degree of ambiguity in sentences than English [17]. To our best knowledge, the studies on using dependency-based features for Chinese sentiment classification are very rare and no extensive evaluation has been carried out to systematically analyze the impact of syntactic dependency features in Chinese sentiment classification. Furthermore, while the above-mentioned works focused on extending the document representation with dependency-based features, no systematic study has been carried out on feature weighting to explore whether different feature weighting schemes can result in different classification accuracy.

Therefore, the primary goal of this study is making an intensive study of the use of syntactic dependency features for Chinese sentiment classification, and our attempt is to seek answers based on empirical evidence to the following questions:

1) Are dependency features suitable for Chinese sentiment classification?

2) By jointly using word unigrams features and dependency-based features, can the performance of a sentiment classification model benefit from the addition

doi:10.4304/jsw.9.11.2877-2885

of dependency-based features over a feature space that only includes traditional word unigrams?

3) By combining different types of features (word unigrams and dependency-based features) with classification algorithms, can the performance of a sentiment classification model benefit from the ensemble technique?

4) In topical text classification, feature weighting scheme plays an important role, does this also hold for Chinese sentiment classification?

In this study, we first investigate two machine learning methods (NB and SVM) on datasets using word unigrams and dependency relations features with three feature weighting schemes. Then, we apply two types of ensemble methods (average combination method and meta learning combination method) with two ensemble strategies (ensemble of feature sets and ensemble of both feature sets and classification algorithms) to integrate feature sets with three different feature weighting schemes including term presence (TP), term frequency (TF), and term frequency – inverse document frequency (TF-IDF) respectively. A number of extensive experiments are conducted on two widely-used datasets in Chinese sentiment classification, and we make in-depth discussion and answer the above four questions

The rest of this paper is organized as follows. Related work and the machine learning methods are described in Section 2 and 3 respectively. In Section 4, we present the ensemble framework for sentiment classification. Sections 5 and 6 describe the experimental setup and results respectively. Finally, we draw conclusions and outline directions for future work in Section 7.

II. Related Work

As a special case of text classification for opinionated texts, in recent years, sentiment classification has become increasingly important due to more and more opinionated information appearing on the Internet. Pang et al. [2] firstly applied NB, ME, and SVM to classify movie reviews into positive and negative classes. The experimental results demonstrated that using word unigrams (a bag of words) as features in classification performed well. Cui et al. [3] studied multiple classification algorithms for sentiment classification on large-scale data set. Liu et al. [18] explored various lexical features and different classification strategies for opinion analysis on blog data. In subsequent work [19], they compared different linguistic features for both blog and review sentiment classification. Arora et al. [20] used an efficient frequent subgraph mining algorithm to extract subgraph features for sentiment classification.

For Chinese sentiment classification, various methods such as lexicon based method [21] and ontology based method [22] have been explored in recent years. Besides that, there have been many works based on the supervised machine learning techniques. Li et al.[4] compared four machine learning methods(NB,SVM, MaxEnt,and ANN) using different feature representations including Word-Based Unigram (WBU), Bigram (WBB), Chinese Character-Based Bigram (CBB), and Trigram (CBT) with different feature weighting schemes on a review corpus which is made up of 16000 reviews. Tan et al. [5] used word unigram as feature with TF-IDF weighting scheme and investigated five supervised machine learning methods(NB,SVM,KNN, centroid classifier, and winnow classifier) and four feature selection methods(MI,IG, CHI and DF) on a Chinese sentiment corpus. Zhai et al. [10] used the SVM method and exploited more complex features including substrings, substring-groups, and key-substring-groups on two Chinese review datasets in different domains.

As a traditional text representation method in sentiment classification, bag-of words (BOW) model is quite efficient and simple. However, word order is disrupted and syntactic structures are broken, and a great deal of information from original text is discarded [15]. Therefore, some works have tried more sophisticated syntactic features such as dependency relations for the task of sentiment classification. Joshi et al. [13] used a transformation of dependency relation triples as additional features to unigrams and yielded a better performance than using purely lexicalized dependency relations. Xia et al. [15] conducted experiments on both the movie reviews dataset used in [2] and the E-product dataset used in [13] and found that individual word dependency relations features (WR-DP) are inferior to unigrams. Furthermore, they noted that the performance of ensemble model integrating different types of features is significantly better than joint features. In subsequent work, they also took advantage of ensemble frameworks for integrating different feature sets and classification algorithms to boost the overall performance of classification model on the five datasets [16]. By using mined frequent dependency subtree patterns as features for SVM, Matsumoto et al. [23] attained significant improvement in the performance of sentiment classification on movie reviews dataset. Dave et al. [24] used adjective-noun dependency relationships as additional features to word unigrams and found that it is ineffective to improve the performance. Ng et al. [14] observed that the addition of dependency relationships does not improve performance over a feature space that includes unigrams, bigrams and trigrams.

However, all the above-mentioned works which used dependency relationships as features only focused on English, and the studies on using dependency-based features for Chinese sentiment classifications are very rare.

Besides features, the weight of each feature in feature vector is also the key component of the representation of a document. In sentiment classification, term presence has been widely used as feature weighting method [2,15-16,25] and has become the most frequently used feature weighting scheme. Other feature weighting schemes were also used to calculate feature weights and achieved good performance in sentiment classification. For example, term frequency was used as feature weighting scheme in the works of [6, 12]. Standard TF-IDF and variants of it were used in the works of [4-5,10,26-27]. However, despite the fact that the use of these feature weighting schemes is commonplace, there has been little research into the effects of different feature weighting schemes in sentiment classification.

III. Machine Learning Methods

A. Naive Bayes

As a probabilistic generative model, NB treats each document as a bag of words and assumes the words are mutually independent. It classifies each test document using Bayes rule by calculating the posterior probability that the document belongs to different classes and assigns the document to the class with the highest posterior probability. Assume a test document ***d*** is represented by [$w_1, . . . , w_m$] ,where w_k is the kth word appearing in the document, and ***C*** denotes the class label set of documents, which is represented by [$c_1, . . . , c_n$],where c_k denotes the kth class label appearing in the documents. By using Bayes rule and conditional independent assumption, Naive Bayes decision can be describe as the following :

$$argmax_{j=1,\dots,n} \prod_{t=1}^{m} P(c_j)P(w_t|c_j) \qquad (1)$$

There are two commonly used models for Naive Bayes, namely multinomial model and multi-variate Bernoulli model [28]. In the multinomial model, a document is represented by the set of word occurrences. And in the multi-variate Bernoulli model, a document is represented as a vector of binary attributes indicating the presence or absence of the word. Based on the multinomial model, Rennie et al. [29] proposed transformed weight-normalized complement Naive Bayes (TWCNB) model with some of the modifications including TF-IDF conversion and document length normalization. Specifically, the multi-variate Bernoulli model, the multinomial model, and the TWCNB model use TP,TF and TF-IDF feature weighting schemes respectively.

B. Support Vector Machines

As a discriminative model, SVM is based on the structural risk minimization principle from the computational learning theory. It seeks the maximal margin decision boundary to separate the data points into positive and negative examples. As stated in [30], SVM can be categorized into linear SVM and nonlinear SVM.

Given the training data set as {($\mathbf{x}_1$, $\mathbf{y}_1$), ($\mathbf{x}_2$, $\mathbf{y}_2$), …,($\mathbf{x}_n$, $\mathbf{y}_n$)}, where $\mathbf{x}_i$ = (x_{i1}, x_{i2}, …, x_{ir}) is a r-dimensional input vector in a real-valued space, $\mathbf{y}_i$ ∈ {1, -1},the optimization problem of finding maximized margin is as following:

$$\begin{cases} \text{Minimize}: \frac{1}{2} < \mathrm{W}^{\mathrm{T}}\mathrm{W} > +C\sum_{i=1}^{N}\xi_i \\ \text{Subject to}: \mathrm{y_i}\,\mathrm{g(X_i)} \geq 1-\xi_i\,, i=1,2,\dots,N \\ \xi_i \geq 0,\ \mathrm{i}=1,2,\dots,N \end{cases} \qquad (2)$$

Where W = (w_1, w_2, …, w_r) is called as the weight vector, C is the penalty coefficient, and ξ_i denotes the slack variable. The linear SVM uses $g(X_i)=W^TX_i+b$ as the discriminant function. To deal with nonlinearly separable data, the same formulation and solution techniques as for the linear SVM are still used for the nonlinear SVM. It uses $g(X_i)=W^T\ \emptyset(X_i)\ +b$ as the discriminant function, where $\emptyset(.)$ is a nonlinear mapping which maps the data in the input space to a feature space.

IV. The Ensemble Model

In recent years, there has been a growing interest in using ensemble learning techniques in sentiment classification. Xu et al. [7] proposed an ensemble learning algorithm based on random feature space division method for sentiment recognition of Chinese movie reviews. Abbasi et al. [9] proposed a correlation ensemble method for affect analysis. Whitehead et al. [31] used ensemble methods including bagging, boosting, and random subspace for sentiment classification. Wang et.al [32] also conducted a comparative assessment of the performance of these three popular ensemble methods on ten public sentiment analysis datasets to verify the effectiveness of ensemble learning for sentiment analysis. Rather than an ensemble of different data re-sampling methods such as bagging and boosting, recently, the ensemble method is further enriched by considering various feature sets and learning models, and it has been applied in sentiment classification [15-16,33-34].

In this study, we also adopt ensemble of feature sets and classification algorithms. In the ensemble framework, different participants can be generated by different contributing classifiers on component feature sets. And two types of feature sets are employed by us for sentiment classification, namely word unigrams and word-dependency parsing pairs. The process of feature extraction is as follows. Each review document is split by punctuation mark into sentences, then, the features can be obtained from sentences. Since Chinese does not segment words by spaces in sentence, each sentence is needed to be segmented into words and word unigram features can be obtained. Taking the sentence ‘外形确实不错 (Appearance is really good)’ as an example, ‘外形’(‘appearance’), ‘确实’(‘really’) and ‘不错’(‘good’) are considered as word unigram features. Similarly, dependency features can be obtained from a given sentence by using parser. As a structured representation, the word dependency parsing pairs for a given sentence are essentially a set of triples, each of which expresses the dependency relation between words, and the triple is composed of a grammatical relation and the pair of words from the sentence. For example, the dependency triples of the sentence ‘外形确实不错 (Appearance is really good)’ are demonstrated in Fig.1.

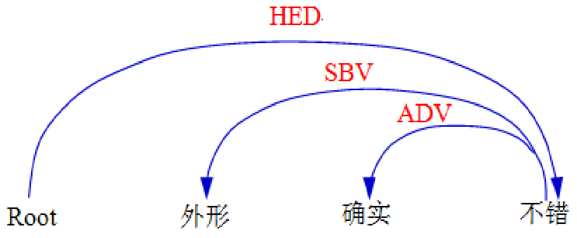


Figure 1. A demonstration of dependency parsing tree

In the Fig.1, each dependency parsing pair has the form of <w_j, w_k, rel_i>, where rel_i is the dependency relation between words w_j and w_k. For instance, dependency parsing pairs for the example sentence include '<外形,不错,SBV>', '<确实,不错,ADV>' and '<-1, 不错,HED>' , where 'SBV' indicates that '外形'('appearance') is a subject modifier of the target '不错'('good') , 'ADV' indicates that '确实' ('really') is an adverbial modifier of the target '不错'('good') , and 'HED' indicates that '不错'('good') is the head word in the sentence. In this study, we straightforwardly use dependency parsing pairs as dependency features.

For our ensemble task, the output values generated by base classifiers are taken as inputs of the ensemble methods to form an integrated output. And we apply the following ensemble methods.

1) Average

It is the most intuitive combination method and the new semantic orientation value of a document is the average of the output values from constituent classifiers.

2) Meta-Learning

The meta-learning method [35] has been used in sentiment classification [15-16]. The key idea behind it is to train a meta-classifier with the outputs of the base classifiers as input attributes. Development data is usually needed for meta-learning to generate the meta-training data. In this study, instead of using extra development data, we perform stacking [36] with 5-fold cross-validation to generate the meta-training data. The stacking is a two-phase framework that is concerned with combining multiple classifiers. In the first phase, a set of base-level classifiers are generated. And in the second phase, a meta-level classifier that combines the outputs of the base-level classifiers is learned.

Taking a dataset as an example, in each loop of the 5-fold cross validation, we consider the probabilistic outputs of the test fold as test samples for meta-leaning. To generate a training set for meta-leaning classifier, we apply an inner 4-fold leave-one-out procedure to the training data. In each of the four fold, samples are trained on the remaining three folds to obtain the probabilistic outputs which are treated as training samples for meta-learning classifier.

It should be noted that the outputs of different classifiers should be transformed to a uniform measure when evaluating the degree of decision confidence [37]. Therefore, for NB model, we use the posterior probability obtained from the classifier as the output. And the posterior probability can be obtained by the following formula:

$$\mathrm{p}(y|\boldsymbol{d}) = \frac{\exp\{\mathrm{o_j}\}}{\sum_{\mathrm{k}=1}^{C}\exp\{\mathrm{o_k}\}} \qquad (3)$$

Where $\mathrm{o_j} = \log(p(y = j)\mathrm{p}(\boldsymbol{d}|y = j))$, it denotes the output belonging to class *j, C* is the number of class label set, $\boldsymbol{d}$ denotes a document and y is the class label of $\boldsymbol{d}$.

As the output of the SVM classifier for a document is a real number score, to convert the score into the polarity probability, we use the following formula which is presented in [38].

$$\begin{cases} \mathrm{Pr_{pos}}(d) = \begin{cases} 1 & s \geq \xi \\ 0.5 + \frac{s}{2\xi} & s \in (-\xi, \xi) \\ 0 & s \leq -\xi \end{cases} \\ \mathrm{Pr_{neg}}(d) = 1 - \mathrm{Pr_{pos}}(d) \end{cases} \qquad (4)$$

In the formula, *s* denotes the output of the polarity classifier for a document *d,* it is a real-number score. And ξ is a threshold value, we use an empirical threshold ξ = 2.

V. Experiments Setup

A. Datasets and Evaluation Metrics

We conduct experiments on the two Chinese datasets. The first is *ChnSentiCorp-2000* dataset [39] from hotel domain. The other dataset is from IT product domain and it is introduced in [40], for convenience, we name it as "Mobile" dataset. A brief summary of the two datasets is shown in Table 1.

Table I
The Summary Of The Datasets

	Labeled positive reviews	Labeled negative reviews
Mobile	1159	1158
ChnSentiCorp-2000	1000	1000

We use accuracy metric to measure the overall performance. And the metric is calculated by using the following formula:

$$\text{accuracy} = \frac{\text{number of correctly identified reviews}}{\text{number of labelled reviews}} \qquad (5)$$

B. Pre-processing

Unlike English, Chinese does not segment words by spaces in sentence. Therefore, to get the word unigram features and dependency relations, pre-processing steps such as word tokenization and dependency parsing should be taken. We use the ICTCLAS toolkit [41] for word tokenization. And LTP- an integrated Chinese processing platform described in [42] is chosen as the tool to extract dependency parsing features.

C. Implementation

For the two datasets, each dataset is evenly divided into 5 folds and all the following experimental results are obtained with a 5-fold cross validation, where each test fold contains all the reviews of one of the fold, and the reviews of the remaining 4 folds are used for training. The performance results reported in all of the following tables are in terms of the average classification accuracy.

We experiment with two standard classification algorithms: NB and SVM. For NB classification model, three types of naive bayes classifiers provided in WEKA [43] , namely NaiveBayes, NaiveBayesMultinomial and ComplementNaiveBayes are employed. Specifically, these three classifiers are implementations of the multi-variate Bernoulli model, the multinomial model, and the TWCNB model respectively. The SVMLight toolkit [44] is chosen as the SVM classifier. The tool of SVMLight is chosen with the linear kernel and default parameter values. NB and SVM are also used as base level classification models of the ensemble model.

Ⅵ. Experiment Results And Analysis

In this section, we first show the results of individual classifiers, and then report the results of two ensemble models which adopt average ensemble method and meta-learning ensemble method respectively with two different ensemble strategies, namely ensemble of feature sets and ensemble of both feature sets and classification algorithms.

A. Results of Individual Classifiers

We evaluate performances of two learning methods on the two datasets using different feature representations combined with three different feature weighting schemes. The feature representation schemes include Word-Based Unigram (WU), dependency parsing pair (DEP), and joint feature (WU+DEP). The feature weighting schemes include TP, TF, and TF-IDF. The results are presented in Table 2-3 respectively. In order to show the comparative results more clearly, we also give the average accuracies of the two datasets in Table 4.

(1) Comparison of different features

On the ChnSentiCorp-2000 dataset, as shown in Table 3, under NB or SVM model with three different feature weighting methods, DEP is always superior to WU. For example, the accuracy of NB using DEP feature with TF weights is 90.75%, while the accuracy of NB using WU with TF is 88.15%.

TABLE II.
ACCURACIES (%) OF INDIVIDUAL CLASSIFIERS ON MOBILE DATASET

Feature	Classifier	Feature weighting		
		TP	TF	TF-IDF
Dep	NB	83.25	91.66	91.79
	SVM	86.66	86.18	93.98
WU	NB	87.35	90.37	90.28
	SVM	91.70	91.49	94.43
WU+Dep	NB	89.42	89.51	93.95
	SVM	90.29	89.60	**95.59**

TABLE III.
ACCURACIES (%) OF INDIVIDUAL CLASSIFIERS ON CHNSENTICORP-2000 DATASET

Feature	Classifier	Feature weighting		
		TP	TF	TF-IDF
Dep	NB	81.5	90.75	91.55
	SVM	88.89	88.44	92.05
WU	NB	76.20	88.15	88.60
	SVM	88.45	86.85	90.35
WU+Dep	NB	87.85	86.65	92.00
	SVM	86.95	85.39	**92.20**

TABLE IV.
AVERAGE ACCURACIES (%) OF INDIVIDUAL CLASSIFIERS ON TWO DATASETS

Feature	Classifier	Feature weighting		
		TP	TF	TF-IDF
Dep	NB	82.38	91.21	91.67
	SVM	87.78	87.31	93.02
WU	NB	81.78	89.26	89.44
	SVM	90.08	89.17	92.39
WU+Dep	NB	88.64	88.08	92.98
	SVM	88.62	87.50	**93.90**

On the Mobile dataset, as we can see from Table 2, for NB and SVM, when using TP as feature weighting, WU consistently outperforms DEP, while using the other two feature weighting schemes, this is not always the case.

Another observation from Table 2 and 3 is that when using TP as feature weighting scheme, for NB classification model, joint features (WU+DEP) provide a significant improvement over any individual feature set. For example, there is 11.65% absolute improvement in accuracy over the case using WU features on the ChnSentiCorp-2000 dataset (87.85% vs. 76.20%) in Table 3. In most other cases, compared with the best results achieved for individual feature sets, however, WU+DEP leads to no significant improvements, sometimes it even yields worse results. For example, on the ChnSentiCorp-2000 dataset, we can see from Table 3 that with TF feature weighting method, the performance of NB using DEP is 90.75%, while the performance of NB using WU+DEP is 86.65%, with a decrease of 4.1% in accuracy.

So we can see that adding DEP as extra feature does not always provide benefit over a simple bag-of-words based feature space when using TP or TF feature weighting. On the other side, when using TF-IDF feature weighting scheme, as indicated by the last column of Table 4, improvement in performance can be obtained for NB and SVM by adding DEP as extra feature in addition to WU (92.98% vs. 91.67% & 89.44% and 93.90% vs. 93.02% & 92.39%).

(2) Comparison of different feature weighting schemes

In each row of the Table 2-4, we can see that among the three feature weighting schemes, TF-IDF consistently outperforms the other two schemes once the classifier and the feature sets are selected. So, the results demonstrate that TF-IDF turns out to be the most effective among all the three feature weighting schemes. Meanwhile, with performances respect to TP and TF feature weights, we can see that in Table 2-4, for the SVM model, no matter what form of feature sets are used, TP is always better than TF. While for the NB model, when using DEP or WU feature sets, TF is always better than TP.

(3) Comparison of different classifiers

With TP feature weighting scheme, on individual feature sets (WU or DEP), we can see from Table 2-3 that SVM always exceeds better performance than NB. This is in accordance with the results reported by Pang et al. [2]. But it turns out this is not always the case when using TF or TF-IDF feature weighting schemes. For example, when using DEP feature sets with TF weights, from Table 4, we can see that the average accuracies on the two datasets of NB always outperform SVM. However, in most other cases, the performance of SVM is superior to NB.

Besides, we can also see from Table 2-3 that the highest performance of the individual classification model is achieved by the SVM classifier using WU+DEP with TF-IDF weights across the two datasets. Specifically, the results are 95.59% for Mobile dataset and 92.20% for ChnSentiCorp-2000 dataset respectively.

B. Results of Ensemble Models

The results of ensemble models on the two datasets are reported in Table 5-6. And there are two ensemble methods, namely average combination method and meta-learning combination method, both of which use two ensemble strategies (ensemble of different feature sets and ensemble of both feature sets and classifiers). For convenience, we name the former strategy as Strategy-1, and the latter strategy as Strategy-2. In Table 7, we also give out the average results of the ensemble models on two datasets.

In Table 5-7, we use denotations for convenience. For example, 'Ave-(SVM@WU)&(SVM@DEP)' denotes an ensemble model which uses the average combination method with Strategy-1 as the ensemble strategy, and the strategy integrates two kinds of feature sets(WU and DEP) with the individual SVM classifier. While 'Meta-(SVM@WU)&(SVM@ DEP)' denotes an ensemble model which uses the meta-learning combination method with Strategy-1. Similarly, 'Ave-(SVM@WU)&(NB@ DEP)' denotes an ensemble model which uses the average combination method with Strategy-2 as the ensemble strategy, and the strategy integrates two kinds of feature sets(WU and DEP) with different classifiers (SVM and NB) , where SVM uses WU as feature sets and NB uses DEP as feature sets. 'Meta-(NB@DEP)&(SVM@WU)' denotes an ensemble model which uses the meta-learning combination method with Strategy-2. The other denotations are similar, and so on.

TABLE V.
ACCURACIES (%) OF ENSEMBLE MODELS ON MOBILE DATASET

Ensemble strategy	Classification model	Feature weighting		
		TP	TF	TF-IDF
	NB@(WU+DEP)	89.42	89.51	93.95
	SVM@(WU+DEP)	90.29	89.60	**95.59**
Strategy-1	Ave-(SVM@WU)&(SVM@DEP)	**94.71**	91.30	95.19
	Ave-(NB@WU)&(NB@ DEP)	90.02	**93.30**	93.09
Strategy-2	Ave-(NB@WU)&(SVM@DEP)	88.86	91.29	93.89
	Ave-(SVM@WU)&(NB@DEP)	88.51	92.65	94.77
Strategy-1	Meta-(SVM@WU)&(SVM@DEP)	92.92	92.48	94.71
	Meta-(NB@WU)&(NB@ DEP)	89.38	93.13	94.17
Strategy-2	Meta-(NB@WU)&(SVM@DEP)	91.96	91.47	94.36
	Meta-(SVM@WU)&(NB@DEP)	92.39	92.65	94.90

TABLE VI.
ACCURACIES (%) OF ENSEMBLE MODELS ON CHNSENTICORP-2000 DATASET

Ensemble strategy	Classification model	Feature weighting		
		TP	TF	TF-IDF
	NB@(WU+DEP)	87.85	86.65	92.00
	SVM@(WU+DEP)	86.95	85.39	92.20
Strategy-1	Ave-(SVM@WU)&(SVM@DEP)	91.29	90.09	**92.35**
	Ave-(NB@WU)&(NB@ DEP)	80.75	**91.85**	92.10
Strategy-2	Ave-(NB@WU)&(SVM@DEP)	77.70	89.00	91.90
	Ave-(SVM@WU)&(NB@DEP)	83.85	91.20	92.25
Strategy-1	Meta-(SVM@WU)&(SVM@DEP)	**91.60**	91.34	92.14
	Meta-(NB@WU)&(NB@ DEP)	81.65	91.00	92.10
Strategy-2	Meta-(NB@WU)&(SVM@DEP)	89.39	90.49	92.30
	Meta-(SVM@WU)&(NB@DEP)	88.55	91.00	92.10

TABLE VII.
AVERAGE ACCURACIES (%) OF ENSEMBLE ALGORITHMS OF THE TWO DATASETS

Ensemble strategy	Classification model	Feature weighting		
		TP	TF	TF-IDF
	NB@(WU+DEP)	88.64	88.08	92.98
	SVM@(WU+DEP)	88.62	87.50	**93.90**
Strategy-1	Ave-(SVM@WU)&(SVM@DEP)	**93.00**	90.70	93.77
	Ave-(NB@WU)&(NB@ DEP)	85.39	**92.58**	92.60
Strategy-2	Ave-(NB@WU)&(SVM@DEP)	83.28	90.15	92.90
	Ave-(SVM@WU)&(NB@DEP)	86.18	91.93	93.51
Strategy-1	Meta-(SVM@WU)&(SVM@DEP)	92.26	91.91	93.43
	Meta-(NB@WU)&(NB@ DEP)	85.52	92.07	93.14
Strategy-2	Meta-(NB@WU)&(SVM@DEP)	90.68	90.98	93.33
	Meta-(SVM@WU)&(NB@DEP)	90.47	91.83	93.50

For comparison purposes, we also report the results of the classifiers which use joint feature sets (WU+DEP) in Table 5-7, where 'NB@(WU+DEP)' denotes the NB model which uses WU+DEP as feature representation. Similarly, 'SVM@(WU+DEP)' denotes the SVM classification model on WU+DEP feature sets.

(1) Comparison with joint features

From Table 7, we can see that when using TF feature weighting scheme, all of the ensemble models can consistently outperform the individual classification model which uses joint features, and the performance improvements range from 2.07% (90.15% vs. 88.08%) to 5.08% (92.58% vs. 87.50%). For the other two feature weighting schemes, when the highest performance of the individual classification model and that of the ensemble models are selected for comparison, we can see the performance of the ensemble model gains improvement by 4.36%(93.00% vs. 88.64%) for TP weighting scheme, while with little drop (93.77% vs. 93.90%) for TF-IDF weighting scheme. So we can conclude that when using TF-IDF weights, in most cases, the performance do not benefit from using multiple classifiers combination.

(2) Comparison with ensemble models

Among the different ensemble models, the model which uses average combination method is the most attractive. As shown in Table 7, on the average, regarding to the three feature weighting methods: TP,TF,and TF-IDF, the highest performance is achieved by 'Ave-(SVM@ WU)&(SVM@DEP)', 'Ave-(NB@WU)&(NB@ DEP)', and 'Ave-(SVM@ WU)&(SVM@DEP)' respectively.Furthermore,among all the ensemble models, in general, 'Ave-(SVM@WU)&(SVM@DEP)' yields the best results. From the above results, we can conclude that the average combination method is preferred since its overall performance is better than the meta-learning combination method, furthermore, the computational cost of the average combination method is less than the meta-learning combination method.

When considering the ensemble strategies, it can be seen from Table 5-7, the best performances are always achieved by the Strategy-1 across both ensemble methods and datasets. It is thus concluded that the Strategy-1 is more effective than the Strategy-2.

(3) Comparison with feature weighting schemes

As we can see from each row of Table 7, the results of TF-IDF consistently exceed the other results. In more detail, the same conclusion holds for each row of Table 5-6 with only one exception in Table 5, that is 'Ave-(NB@WU) & (NB@DEP)'.

The reason for this phenomenon lies in the fact that in Table 2-4, we can see that TF-IDF always achieves the best performance once the individual classifier and the feature sets are selected. Naturally, when the multiple classifiers using TF-IDF weights are combined into an ensemble model, the results also attain the highest performance. This is in accordance with the conclusions drawn from Table 2-4.

VII. Conclusion And Future Work

In this study, we aim to explore the use of dependency features for Chinese sentiment classification. We conduct a range of comparative experiments on two widely-used Chinese datasets by considering two types of feature sets, two schemes of ensemble methods, and two ensemble strategies. Based on the experimental results, questions in the Section 1 are answered respectively as following:

1). As we can see from Table 3, on the ChnSentiCorp-2000 dataset from hotel domain, DEP feature is more effective than traditional WU feature when the classifier and feature weighting scheme are selected. On the Mobile dataset from digital products domain, from Table 2, we can see that when using TF and TF-IDF schemes, WU is superior to DEP. Therefore, in general, we can see that DEP feature can still be treated as a candidate effective feature depending on the domain and the feature weighting scheme. Furthermore, when using dependency features as feature representations, to achieve better performance, if TP or TF-IDF are selected as feature weight schemes, we prefer using SVM as supervised learning method. But when using TF feature weighting scheme, we prefer using NB as learning method.

2). When using dependency features as additional supplement features to word unigrams, with TF-IDF weighting scheme, it is effective for Chinese sentiment classification. While with TP feature weighting scheme, the effects are not so obvious. Furthermore, when using TF as feature weighting scheme, the performances of the classification models on the two datasets even degrade. So, we can conclude that the effectiveness of using joint feature sets (WU+DEP) should be related to feature weighting scheme.

3). As described in the Section 6, from the last column of Table 7, we can see that when using TF-IDF feature weighting scheme, compared with the results of individual classifiers on joint feature sets, in most cases, the performances do not benefit much from using multiple classifiers combination. Whereas using TF weighting scheme, all of the ensemble models which use different ensemble methods with different ensemble strategies consistently outperform the performances of individual classifier models which use joint feature sets, as is shown in Table 7. When using TP feature weighting scheme, we can see that 'Ave-(SVM@WU)&(SVM@DEP)' performs well compared with the results of individual classification models which use joint features, whereas for the other ensemble models, this is not always the case. Generally, when using ensemble technique for Chinese sentiment classification, we prefer average combination method with the strategy that is an ensemble of feature sets using a single classification model.

4). From the above points, we can conclude that feature weighting scheme does indeed impact on the performance of classification model and choosing the proper feature weighting scheme can be crucial to the performance of Chinese sentiment classification. Overall, TF-IDF is the most successful feature weighting scheme for both SVM and NB.

Future work will include evaluating feature selection methods and finding out the reason of why the effectiveness of dependency features depends on domain and feature weighting scheme , it is also worth utilizing corpus in other domains (e.g. restaurant reviews) for this work.

Acknowledgment

This work was supported in part by the National Nature Science Foundation of China (No. 61075019) and the Fund of Nature Science of Chongqing University of Posts and Telecommunications (No.A2011-31).

References

[1] Liu, B., *Sentiment Analysis and Opinion Mining*, San Rafael, CA: Morgan and Claypool,2012.

[2] Pang, B., Lee, L., and Vaithyanathan, S., "Thumbs Up? Sentiment Classification Using Machine Learning Techniques", in *Proceedings of the Conference on*

Empirical Methods in Natural Language Processing, Philadelphia, PA, 6–7 July, 2002, pp. 79–86.

[3] Cui, H., Mittal, V., and Datar, M., “Comparative Experiments on Sentiment Classification for Online Product Reviews”, In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, July 16 - 20, 2006, Boston, Massachusetts.

[4] Li, J., and Sun, M., “Experimental Study on Sentiment Classification of Chinese Review Using Machine Learning Techniques”. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2007,pp.393–400.

[5] Tan, S., and Zhang, J., “An Empirical Study of Sentiment Analysis for Chinese Documents”, *Expert Systems with Applications*, 34(4), 2622–2629,2008.

[6] Zhang, Z., Ye, Q., Zhang, Z., and Li, Y., “Sentiment Classification of Internet Restaurant Reviews Written in Cantonese”, *Expert Systems with Applications*, 38(6), 7674-7682,2011.

[7] Xu, W., Liu, Z., Wang, T., & Liu, S., “Sentiment Recognition of Online Chinese Micro Movie Reviews Using Multiple Probabilistic Reasoning Model”. *Journal of Computers*, *8*(8),1906-1911, 2013.

[8] Riloff, E., Patwardhan, S., and Wiebe, J., “Feature Subsumption for Opinion Analysis”, In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp. 440–448,2006.

[9] Abbasi, A., Chen, H., Thoms, S., and Fu, T. “Affect Analysis of Web Forums and Blogs Using Correlation Ensembles”, *IEEE Transactions on Knowledge and Data Engineering*, 20(9), 1168-1180,2008.

[10] Zhai Z, Xu H, Kang B and Jia P, “Exploiting Effective Features for Chinese Sentiment Classification”, *Expert Systems with Applications*, 38(8): 9139–9146,2011.

[11] Subrahmanian, V. S., and Reforgiato, D., “AVA: Adjective-verb-adverb combinations for sentiment analysis”, IEEE Intelligent Systems, 23(4), 43-50,2008.

[12] Xu, J., Ding, Y. X., and Wang, X. L., “Sentiment Classification for Chinese News using Machine Learning Methods”, *Journal of Chinese Information Processing*, 21(6), pp: 95-100,2007.

[13] Joshi, M., and Penstein-Rosé, C., “Generalizing Dependency Features for Opinion Mining”, In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics(ACL-IJCNLP)*, pp. 313-316,2009.

[14] Ng, V., Dasgupta, S., and Arifin, S. M., “Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews”, In *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 611-618, 2006.

[15] Xia, R., and Zong, C., “Exploring the Use of Word Relation Features for Sentiment Classification”, In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pp. 1336-1344, 2010.

[16] Xia, R., Zong, C., and Li, S., “Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification”, *Information Sciences*, 181(6), 1138-1152, 2011.

[17] Zhang, C., Zeng, D., Li, J., Wang, F. Y., and Zuo, W., “Sentiment analysis of Chinese documents: From sentence to document level”, *Journal of the American Society for Information Science and Technology*, 60(12), 2474-2487, 2009.

[18] Liu, F., Li, B., and Liu, Y., “Finding Opinionated Blogs using Statistical Classifiers and Lexical Features”, In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*,2009.

[19] Liu, F., Wang, D., Li, B., and Liu, Y., “Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods”, In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (HLT-NAACL)*, pp. 309-312, 2010.

[20] Arora, S., Mayfield, E., Penstein-Rosé, C., & Nyberg, E., “Sentiment classification using automatically extracted subgraph features”, In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* ,pp.131-139,2010.

[21] Liu, L.,Lei, M., & Wang,H., “Combining Domain-Specific Sentiment Lexicon with Hownet for Chinese Sentiment Analysis”, *Journal of Computers*, 8(4),878-883,2013.

[22] Wang, H., Nie, X., Liu, L., & Lu, J. , “A Fuzzy Domain Sentiment Ontology based Opinion Mining Approach for Chinese Online Product Reviews”, *Journal of Computers*, 8(9),2225-2231,2013.

[23] Matsumoto, S., Takamura, H., and Okumura, M., “Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees”. In *Advances in Knowledge Discovery and Data Mining, 9th Pacific-Asia Conference, PAKDD*, Hanoi, Vietnam ,pp. 301-311,2005.

[24] Dave, K., Lawrence, S., and Pennock, D. M., Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews”, In *Proceedings of the 12th international conference on World Wide Web(WWW)*, pp. 519-528, 2003.

[25] Pang, B., and Lee, L., “A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts”. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, pp.271–278,2004.

[26] Martineau, J., and Finin, T., “Delta tfidf: An Improved Feature Space for Sentiment Analysis”, In *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media* ,pp. 258-261,2009.

[27] Paltoglou, G., and Thelwall, M., “A Study of Information Retrieval Weighting Schemes for Sentiment Analysis”. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1386-1395, 2010.

[28] McCallum, A.,and Nigam, K., “A Comparison of Event Models for Naive Bayes Text Classification”, In *AAAI-98 workshop on learning for text categorization*, pp. 41-48,1998.

[29] Rennie, J. D., Shih, L., Teevan, J., and Karger, D., “Tackling the Poor Assumptions of Naive Bayes Text Classifiers”, In *Proceedings of the 20th International Conference on Machine Learning (ICML)* , Washington DC, USA, 2003.

[30] Liu, B., *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Second Edition, Springer Verlag , 2011.

[31] Whitehead,M., Yaeger,L. “Sentiment mining using ensemble classification models”, *Innovations and Advances in Computer Sciences and Engineering*, Springer, pp. 509–514,2010.

[32] Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. , “Sentiment classification: The contribution of ensemble learning”, *Decision Support Systems*, 57, 77-93, 2014.

[33] Wan X, “Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis”, In *Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp.553-561,2008.

[34] Xia, R., Zong, C., Hu, X., & Cambria, E., “Feature ensemble plus sample selection: A comprehensive approach to domain adaptation for sentiment classification”, *IEEE Intelligent Systems*, 28(3), pp. 10-18, 2013.

[35] Vilalta, R., and Drissi, Y., “A Perspective View and Survey of Meta-learning”, *Artificial Intelligence Review*, 18(2), pp.77-95,2002.

[36] Dzeroski,S., and zenko,B., “Is combining classifiers with stacking better than selecting the best one?”, *Machine Learning* ,54(3),pp.255–273,2004.

[37] Hao, H., Liu, C. L., and Sako, H., “Confidence Evaluation for Combining Diverse Classifiers”, In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*,2003.

[38] Wang, X. , Wei,F., Liu,X., Zhou,M., Zhang,M., “Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach”, In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*,2011.

[39] ChnSentiCorp-2000 dataset: http://www.searchforum.org.cn/tansongbo/corpus/ChnSentiCorp_htl_ba_2000.rar

[40] Zagibalov, T., and Carroll, J., “Unsupervised Classification of Sentiment and Objectivity in Chinese Text”, In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, India, pp. 304-311,2008.

[41] ICTCLAS: http://www.ictclas.org

[42] Che, W., Li, Z., and Liu, T., “Ltp: A Chinese Language Technology Platform”, In *Proceedings of the 23rd International Conference on Computational Linguistics*: Demonstrations, pp.13-16, 2010.

[43] WEKA: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[44] SVMLight: http://svmlight.joachims.org

Pu Zhang is a PhD candidate in College of Computer Science , ChongQing University, China. He received M.S. degree in Computer Science Department from Sichuan University, China. He is an Associate Professor in College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His current research focuses on nature language processing, sentiment analysis and text mining.

Zhongshi He received Ph.D. degree in Computer Science from Chongqing University in 1996. He is now a Professor and PhD supervisor of Chongqing University, Chongqing, China. His research interests include nature language processing, machine learning and pattern recognition.

Lina Tao is now a Research Associate of Chongqing Communications Research and Design Institute Co.,Ltd., Chongqing, China. Her current research interests include data mining, and scientific computing.

A Construction Model of Ancient Architecture Protection Domain Ontology Based on Software Engineering and CLT

Zhenbo Bi
Architecture School, Xi’an University of Architecture and Technology, Xi’an, China
School of Mathematics, Physics & Information Science, Zhejiang Ocean University, Zhoushan, China
E-mail: bzb136@sina.com

Huiqin Wang and Ying Lu
Information and Control Engineering School , Xi’an University of Architecture and Technology, Xi’an, China
E-mail: {hqwang, luying}@xauat.edu.cn

Abstract—**In allusion to the lack of formal mathematical methods and engineering characteristics on the construction of AAPDO (ancient architecture protection domain ontology, AAPDO), a construction model on AAPDO based on software engineering and CLT (Concept Lattice Theory, CLT) is proposed, and the application steps and methods of the model are also elaborated on. The model uses a guiding ideology of engineering, a formal way of expression and the standardized work steps to complete the construction of AAPDO. Finally, taking Chinese Traditional Roof as an example, an instance on the construction of the field ontology of Chinese Traditional Roof based on software engineering and CLT is given.**

Index Terms—**Concept lattice, Software engineering, Ontology construction, Ontology for ancient architectures protection**

I. Introduction

Ontology construction in ancient architecture protection is the premise of all the knowledge engineering activities based on its ontology. Improving the formalization and standardization level of ontology construction in ancient architecture protection is the foundation for enhancing interoperability, knowledge expression ability and the ontology inferential capability on the basis of a semantic ancient architecture software intersystem. Ontology construction in ancient architecture protection has attracted increasing attention in the ancient architecture and other related research fields. For example, Nadezhda Govedarova, et al.[1] put forward architecture knowledge management based on ontology in Bulgaria cultural heritage BULCHINO network directory project, for searching and browsing; Liu Qifeng, et al.[2] proposed a design method for ontology repository based on Ontology Definition Metamodel(ODM), to build framework ontology by recognizing the key concepts and their incidence relation in ancient architecture protection; Bai Weijing, et al.[3] designed and implemented an ancient architecture repository using semantic web technology, thus established an automatic animation system; Song Yu, et al.[4] established an ontology model integrating architecture components, culture and structure, thus explored a retrieval system based on the ancient architecture component ontology. These studies achieve their objectives through the establishment of ancient architecture ontology. However, one common feature is that their ontology construction in ancient architecture basically relies on personal knowledge or experience of researchers, developers and experts, which easily causes the ontology concept deletion or repetition, even the ambiguity and confusion among the concepts due to its strong subjectivity. Particularly, when faced with large-scale data organization, it is time-consuming, laboursome and inefficient. Therefore, ontology construction in ancient architecture by formalization mathematical method is of great significance.

CLT is a branch of applied mathematics, which is a kind of knowledge description and data analysis tools built on the hierarchy of concept based on the mathematics. With the help of CLT, the concepts composed of the extent, the intent and the hierarchical relationships among these concepts can be discovered, constructed and demonstrated. It is a formal and standardized domain ontology construction method. CLT and domain ontology are the two kinds of methods on formal knowledge representation, and the literature [5] pointed out the links and differences between them. There are a lot of published literatures about CLT used in the construction of domain ontology [6-10]. These documents have laid the foundation for the produce of the theory of domain ontology building based on CLT. But when they used CLT to build domain ontology, they mostly focused on some aspects of domain ontology construction or a specific stage, for example, access to formal context, concept lattice constructing algorithm and the improvement of concept lattice. There are no

Corresponding author: Zhenbo Bi

doi:10.4304/jsw.9.11.2886-2894

engineered features in domain ontology building process based on CLT. In fact, the ultimate development trend of the construction of AAPDO must be an engineering way to solve the problem on its building.

In allusion to the lack of formal mathematical methods and engineering characteristics on the construction of AAPDO, we propose a construction model on AAPDO based on software engineering and CLT. The model adopts a guiding ideology of engineering [11-13], a formal way of expression and standardized working steps to complete the construction of AAPDO, which improves the engineering degree and degree of formalization of the construction of AAPDO. This will provide a theoretical basis and new ideas of application for the work and some related research.

II. CLT Relevant Definitions

Definition 1. A binary relation R on set M is called an order relation (or shortly an order), if it satisfies the following for all elements $x, y, z \in M$: 1) xRx ; 2) xRy and $x \neq y \Rightarrow$ not yRx ; 3) xRy and $yRz \Rightarrow xRz$.

Definition 2. A formal context $K=(G,M,I)$ consists of two sets *G* and M and a relation *I* between *G* and *M*. The elements of *G* are called the objects and the elements of *M* are called the attributes of the context. In order to express that an object *g* is in a relation *I* with an attribute *m*, we write $g \text{ Im } \in I$ and read it as “the object *g* has the attribute *m*”.

Definition 3. For a set $A \subseteq G$ of objects, we define $A'=\{m\in M \mid \forall g\in A, gIm\}$ (the set of attributes common to the objects in A). Correspondingly, for a set B of attributes, we define $B'=\{g\in G \mid \forall m\in B, gIm\}$ (the set of objects which have all attributes in B).

Definition 4. A formal concept of the context (G, M, I) is a pair (A, B) with $A\subseteq G, B\subseteq M$, $A'=B, B'=A$. We call A the extent and B the intent of the concept (A, B).

Definition 5. If (A,B) and (C,D) are concepts of a context. (A,B) is called a super concept of (C,D) (which is equivalent to that (C,D) is a subconcept of (A,B)), provided that $B\subseteq D$ (which is equivalent to $C\subseteq A$), we write $(C,D)\leq(A,B)$, namely, $(C,D)\leq(A,B)\Rightarrow B\subseteq D(\Leftrightarrow C\subseteq A)$. The relation $\leq$ is called the hierarchical order (or simply order) of the concepts. The set of all concepts of *K* ordered in this way is denoted by $(\mathrm{B}(K), \leq)$, $(\mathrm{B}(K), \leq)$ is called the concept lattice of the context *K*.

Definition 6. An order relation $\leq$ is decided by a pair of elements with covering relations (for example, $a\prec b$) in a finite set $(S,\leq)$, if each element in set S is shown with a small circle in the same plane, when b cover a, a and b be connected with a line. Thus obtained graph is called Hasse diagram of order set $(S,\leq)$. Hasse diagram can be used to express vividly the relationship among elements in order set.

III. Construction Model Of Aapdo Based On Software Engineering And Clt

As the construction of AAPDO is a systematic project, common domain ontology construction methods, such as "skeleton method", "Assessment method", "Bernaras", "Methontology method" and the "Sensus method", which have not formed a unified standard of domain ontology construction are application-specific ontology construction methods. Therefore, according to the characteristics of ancient architecture protection domain knowledge and the current difficulties, the construction of domain ontology should adopt the development method of increment model [13], as shown in Fig.1. Each linear sequence in the Fig.1 represents a delta; the result would have an ontology which can be used. Among them, the first increment is a core ontology which meets the basic needs (Performance and features); the second increment is an expansion of the previous incremental, as time goes on, a completed ontology will be eventually formed which meets all application requirements in the field. The results of user feedback or comments are the cause of the next increment plan. Incremental plan illustrates the need for core ontology changes, also illustrates the need to increase the features and performance.

In Fig.1, the process flow of each increment can adopt waterfall development paradigm or prototype development paradigm, here is given priority to with waterfall model, the process is divided into six stages: learning from the idea of the waterfall model of software engineering, the domain ontology description model of ancient architecture protection shown in Fig.1 has an important practical significance. Here its build process is divided into six stages: Ⅰ. Fields and related fields definition;II. Feasibility study; III. requirement analysis; IV. Domain ontology design; V. Domain ontology implementation; VI. Application and maintenance of domain ontology. If there are problems involving a preceding stage with a vague description at each stage, you can return to perfect, as shown in fig.2. In Fig.2, the phase division of software development, the key problems and the standard of the end of the software lifecycle methodology have been the model for a reference. Considering some characteristics of ontology engineering, some improvements have been made to the waterfall model, such as the testing of ontology is omitted and tested on the application. In Fig.2, CLT's role is mainly reflected in the design stage of AAPDO. The detailed corresponding model of CLT combined with software engineering is shown in Fig.3.

A. Fields and Related Fields Definition

The priority of AAPDO engineering is to fully understand the domain knowledge. To understand the domain knowledge, first, we must have a definition on domain knowledge to determine what knowledge is to be studied in the field of knowledge, at the same time, the cross of one domain knowledge and other fields of knowledge must also be defined. Finally, specifications of fields' definition need to be written.

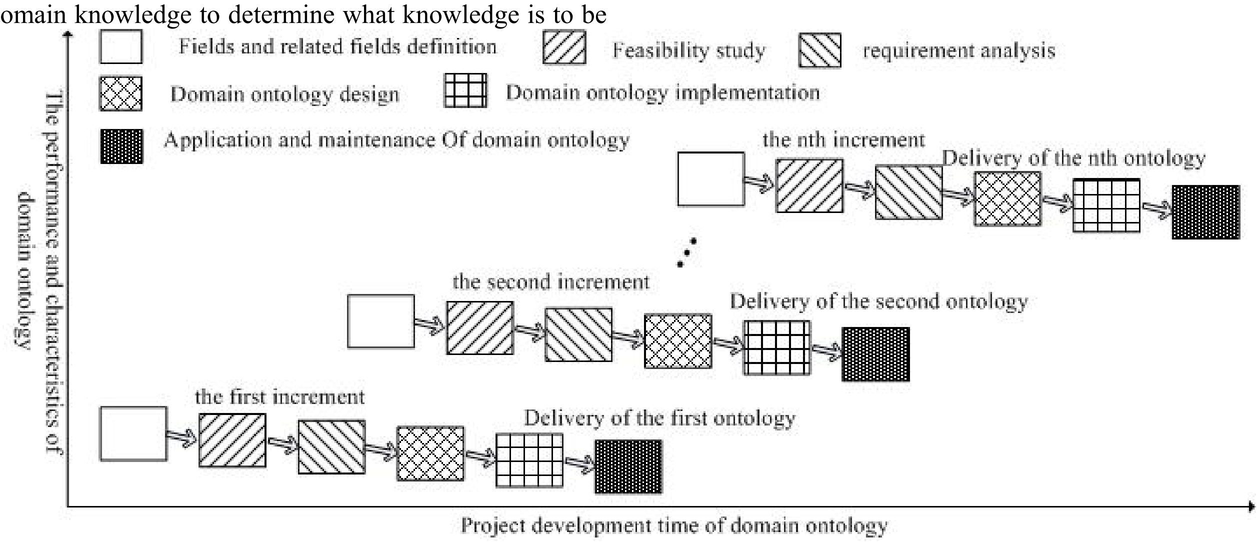

Fields and related fields definition
Feasibility study
requirement analysis
Domain ontology design
Domain ontology implementation
Application and maintenance Of domain ontology
The performance and characteristics of domain ontology
the first increment
Delivery of the first ontology
the second increment
Delivery of the second ontology
the nth increment
Delivery of the nth ontology
Project development time of domain ontology

Figure 1. Incremental model of ancient architecture protection ontology

B. Feasibility Study

After the scope of range of fields and related fields definition is defined, we need to analyze whether the existing resources can live up to modeling the defined domain knowledge and whether implementing the entire project is feasible, and we also need to estimate the future workload and difficulty of the domain ontology development. These require knowledge engineers, domain experts and the related personnel to decide through their comprehensive evaluation. If possible, the work can be turned to the next step; otherwise, it is necessary to discover and find the factors that may lead to failure of the project and to determine whether to proceed with the project. At the end of the stage, a feasibility study report needs to be written.

C. Requirement Analysis

According to the problems existing in the ancient architecture protection, we should explicitly put forward the goal to be achieved for the knowledge base in this phase, namely, an AAPDO constructed by the five elements of the conceptual classes, relations, functions, axiom and instances is the goal of the phase, which build a bridge of exchange and communication to achieve transparent access and interoperability between the applications systems and the users and between the different applications based on the ontology. This is the premise of knowledge development based on the ontology. In practice, we should begin from understanding application to solve the core problem, and the core issues of the applications and the core issues of the areas need to be distinguished here, the former involves the concepts of multiple areas, there can be or not a link between each other; the latter involves the core concepts carefully selected of specific areas. Analysis of application requirements should begin from the decomposition of application problems, and then we can draw lessons from the top-down method of software engineering to do it. Finally, for specific sub-applications, according to the core problems to be solved, we need to use the corresponding techniques [14] to extract the core term set of areas from data sources, such as an existing glossary, documents, databases and domain experts according to the specific types of data sources.

D. Domain Ontology Design

The module's central task is to design a complete concept lattice, including four subtasks:

Subtask 1: According to the Binary relation of “Object – attributes”, a context of "object as a line head" and "attribute as the column name" is created through the domain core terms got in the requirements analysis phase, namely, a context is represented by a matrix, a head of each row of the matrix represents an object O_i, a head of each column of the matrix represents an attribute P_j, when $o_i, p_j \in I$, it represents that "the object O_i have the attribute P_j", then we draw a "*" at the corner of the row *i* and the column *j*. After finishing the work, we need to determine whether the context is a standard context, if not, we need to analyze the reasons (such as multi-value context, non-purification context), and take corresponding measures (such as a multi-valued context into a single-valued context, context purification) to standardize the context.

A multi-valued context (*G*, *M*, *W*, *I*) is composed of sets *G*, *M*, *W* and the ternary relationship *I* among the three sets, namely, $I \subseteq G \times M \times W$,moreover, $(g, m, w) \subseteq I$ and $(g, m, v) \subseteq I$, always implies $w = v$, the elements of *G* are called the objects, the elements of *M* are called the multi-valued attributes of the context and the elements of W are called the attributes of the context. $(g, m, w) \subseteq I$ is read as "The value of attribute m of object g is w". if W has n elements, (*G*, *M*, *W*, *I*) is called a n-value context. According to the different types of attribute value of a multi-valued context, it mainly includes numerical multi-valued context, interval multi-value context and language multi-value context [14]. For different types of multi-valued contexts, we need to take a different approach (such as membership degree transformation method, concepts scaling transformation method and Language scale segmentation transformation method, etc.) for the conversion of single-valued context.

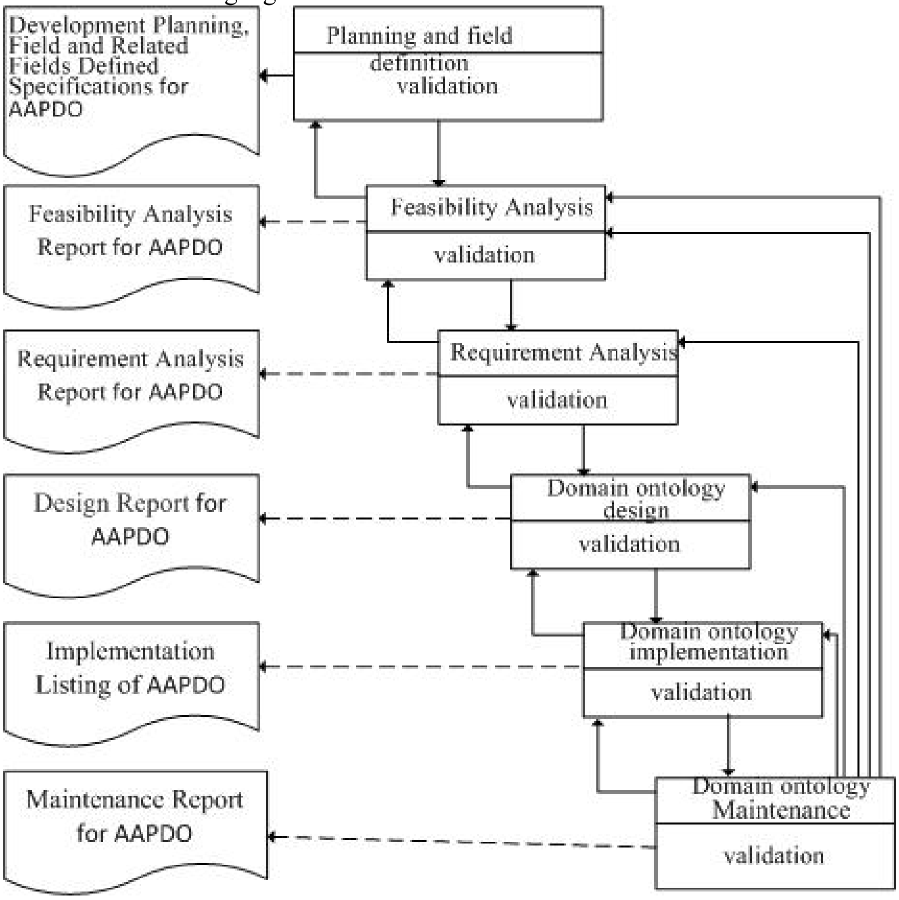

Development Planning, Field and Related Fields Defined Specifications for AAPDO
Planning and field definition
validation
Feasibility Analysis Report for AAPDO
Feasibility Analysis
validation
Requirement Analysis Report for AAPDO
Requirement Analysis
validation
Design Report for AAPDO
Domain ontology design
validation
Implementation Listing of AAPDO
Domain ontology implementation
validation
Maintenance Report for AAPDO
Domain ontology Maintenance
validation

Figure 2. Waterfall model for AAPDO construction

For $k = (G, M, I)$, if $\exists g, h \in G$ and $g' = h'$, there must be $g = h$, then the object *g* and *h* are redundant, the lines with G and h can be reduced; Similarly, $\exists m, n \in M$ and $m' = n'$, there must be $m = n$, then the attribute *m* and *n* are redundant, the columns with G and h can be reduced. From the perspective of object and attribute redundancy, we combine the same objects and attributes in the context to meet the basic simple purification of the context here. In the same context, if $\exists g, h, t \in G$, $t' \subset g'$, $t' \subset h'$, and $g' \cap h' = t'$, then the object *t* is redundant, the line with t is reduced [15]. From the perspective of mutual expression of objects and attributes, we remove the extents (the intents) expressed with the intersection of all other objects (attributes) which can be used here.

Through the above basic operation, the standardization of the context can be realized.

Subtask 2: Through concept lattice construction algorithm (algorithm for constructing a batch or incremental construction algorithm) [15-17], the standard context can be converted into a concept lattice, and through the Hasse diagram, the concept lattice can be visually represented. And then domain experts and knowledge engineers need to determine whether the concept lattice is reasonable on the basis of visualization. For unreasonable concept lattice, we make it a more complete satisfaction concept lattice through adding or removing objects, adding or removing attributes, editing attributes by certain rules. Concept lattice can generate new objects which are not in the table of concept, we can add these objects; The whole process is repeated continuously until the concept lattice is reasonable and perfect.

Subtask 3: The conversion of the edited concept lattice complete mainly includes the naming of top nodes, the labeling of intermediate nodes, the removal of bottom nodes and the conversion of Relationship between nodes to the relationship between concepts in Hasse diagram, the conversion result is a domain ontology prototype.

Subtask 4: According to the actual situation, attribute expansion, instance expansion, axioms expansion and relationships expansion on the basis of the domain ontology prototype can be done under the participation of experts; finally the expanded domain ontology prototype has been got. Among them, attribute expansion is used to improve the intent of the ontology concept and instance expansion is used to improve the extent of the ontology concept, relationships expansion aims to perfect the relationship of the domain ontology concepts in addition to the classification relationship, axioms expansion is in order to achieve the ontology reasoning.

E. Domain Ontology Implementation

For the expanded domain ontology prototype, it needs to be formalized description through the appropriate ontology description tools and language, namely, an encoding process of ontology is completed, and the resulting domain ontology is got. The encoding process includes coding of various aspects, such as domain concepts, relationships between concepts, attributes, instances, axioms and inference rules. Next we need to use an ontology reasoning machine to achieve the ontology knowledge reasoning which is based on the concept lattice with a mathematical theory support. Using the concept lattice will effectively help knowledge engineers to complete the logical description of the domain knowledge. Here the ontology reasoning comprises detecting conflicts, optimizing expression and the tacit knowledge acquired by reasoning based on the explicit knowledge.

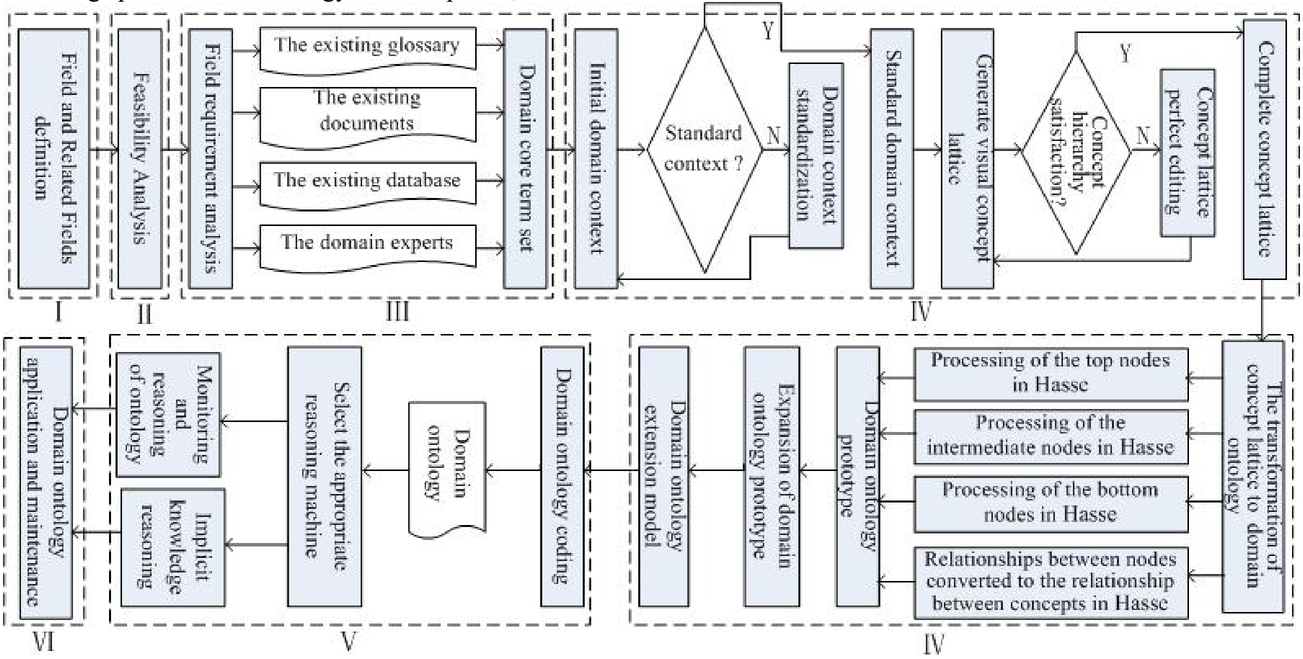

Field and Related Fields definition
Feasibility Analysis
Field requirement analysis
The existing glossary
The existing documents
The existing database
The domain experts
Domain core term set
Initial domain context
Standard context ?
Y
N
Domain context standardization
Standard domain context
Generate visual concept lattice
Concept hierarchy satisfaction?
Y
N
Concept lattice perfect editing
Complete concept lattice
I
II
III
IV
The transformation of concept lattice to domain ontology
Processing of the top nodes in Hasse
Processing of the intermediate nodes in Hasse
Processing of the bottom nodes in Hasse
Relationships between nodes converted to the relationship between concepts in Hasse
Domain ontology prototype
Expansion of domain ontology prototype
Domain ontology extension model
Domain ontology coding
Domain ontology
Select the appropriate reasoning machine
Monitoring and reasoning of ontology
Implicit knowledge reasoning
Domain ontology application and maintenance
VI
V
IV

Figure 3. AAPDO construction model based on CLT and software engineering

F. Domain Ontology Application and Maintenance

For the implemented domain ontology, it can be put into practical application. However, domain knowledge is constantly evolving, with new knowledge creation and added in, so the implemented domain ontology cannot be fixed and unalterable, according to the development and changes of domain knowledge, we often need to improve maintenance for the ontology, in this way can we ensure universal applicability of the domain ontology.

IV. Example: The Construction Of Traditional Building Roof Protection Ontology

The purpose of the example aims to verify the practical effect of the proposed model. Therefore, the choice of architecture field does not be too complicated to be able to clarify the correctness, the availability and ease of use of the theory.

A. Step 1: Fields and Related Fields Definition

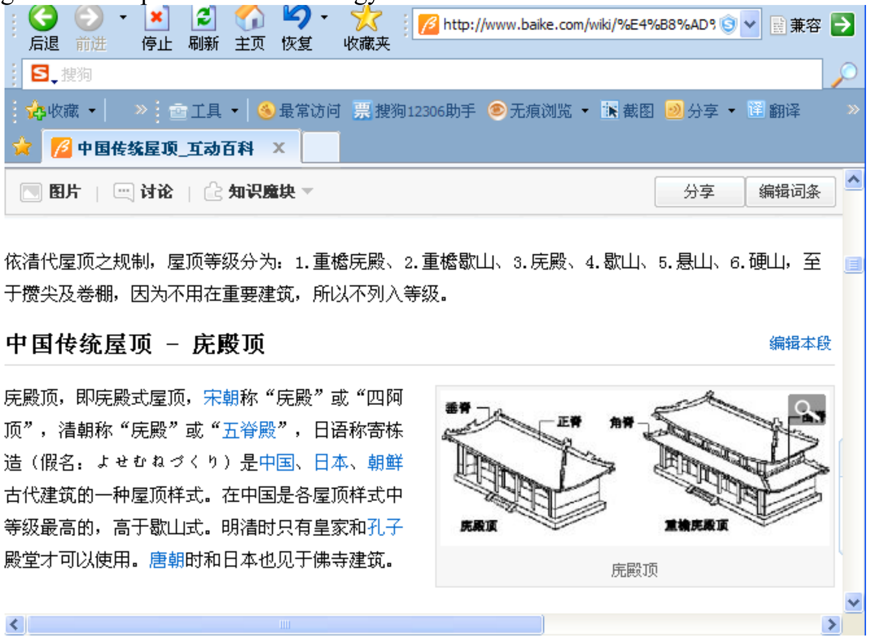
The field of protection of ancient architecture is a large and complex cross-cutting field which is composed of multiple sub-domain knowledge. To build the entire AAPDO, we need to divide it into many sub-domain ontologies to build in turn. First, we build a top ontology, and then we built sub-domain ontology. Using the method of the top-down, applications decomposition and divide and conquer contributes to simplify the construction of the ontology. For example, Xi'an is an ancient city with a long history and culture, since thousands of years, many excellent ancient buildings are preserved, such as ancient city wall, ancient gate tower, Dayan Pagoda, palace buildings and Bell Tower, etc. Due to the long time, these ancient buildings urgently needed to be protected through the use of modern advanced technology. Therefore, here we have identified Xi'an 'ancient architecture protection areas and the corresponding sub-fields according to the ancient architecture types. First, we have built the top ontology Xi'an ancient architecture protection ontology, and then we have also built sub-domain ontology and their sub domain ontologies.

B. Step 2: Feasibility Study

The resources of the construction of Xi'an ancient architecture protection ontology consist of economic and technical resources. Economic aspects, the project has gained national and provincial research funding support, and the related research funds also provide support for the project research; technical aspects, the project team has a large number of ancient architecture protection experts and technical staff who are associated with the research, most of them have rich and practical experience, in addition, there are also a large number of knowledge management engineers, they also have a wealth of knowledge management and application development experience. Therefore, the project implementation of the ontology is feasible, and after the discussion of knowledge engineers, domain experts and relevant personnel, the future workload and technical difficulty of the research projects is also predictable and controllable.

C. Step 3: Requirement Analysis

By analyzing ancient architectural history of china, ancient architecture academic literature, ancient architecture domain knowledge and ancient architecture protection technical literature, we have understood the characteristics of knowledge in the field of ancient architecture protection and extracted key concepts such as ancient architectural complexes, single ancient building, components, damaging state and protection technology [1]. In protection engineering, we need to use the corresponding protection technology according to the specific damaged status of components. Wooden architecture in Single ancient building is the main; with its roof style divided into flush gable roof, overhanging gable roof, gable and hip roof, hip roof, pyramidal roof, etc. Protection repair documentation for structural description is described through specific technology. In allusion to the damaging of specific components, the corresponding detailed protection technical description is given. To illustrate the problem, here only "Chinese Traditional Roof" as an example to tell, and the "Chinese traditional roofs" is seen as a special component (a combination of many atomic components), which avoid the disadvantages of the context that is too large to be easy to express, at the same time we also necessarily simplify the attributes of the related objects, and added some attributes from the perspective of protection. In this paper, the vocabulary entry "Chinese traditional roofs"[18] in interactive encyclopedia is taken as domain unstructured data, and on the basis of which the ontology of Chinese traditional roof will be built. "Chinese Traditional roofs" is shown in Fig. 4.



依清代屋顶之规制，屋顶等级分为：1.重檐庑殿、2.重檐歇山、3.庑殿、4.歇山、5.悬山、6.硬山，至于攒尖及卷棚，因为不用在重要建筑，所以不列入等级。

中国传统屋顶 – 庑殿顶

庑殿顶，即庑殿式屋顶，宋朝称“庑殿”或“四阿顶”，清朝称“庑殿”或“五脊殿”，日语称寄栋造（假名：よせむねづくり）是中国、日本、朝鲜古代建筑的一种屋顶样式。在中国是各屋顶样式中等级最高的，高于歇山式。明清时只有皇家和孔子殿堂才可以使用。唐朝时和日本也见于佛寺建筑。

Figure 4. Chinese traditional roof initial data

After the text in the vocabulary entry "Chinese traditional roofs" is preprocessed, we use natural language processing techniques (such as Pao Ding Xie Yang Chinese Word Segmentation) combined with manual analysis techniques to get the domain unstructured data (text), and on the basis of which a core terms set has been extracted. The core terms set includes an attribute set and an object set, the attribute set: {P_1 main ridge, P_2 double eave roof, P_3 slope face, P_4 palace architecture, P_5 civil architecture, P_6 leaking, P_7 damaged tile, P_8 weeding}; the object set: {O_1 hip roof, O_2 gable and hip roof, O_3 pyramidal roof, O_4 overhanging gable roof, O_5 flush gable roof, O_6 round ridge roof}.

D. Step 4：Domain Ontology Design

An initial context based on the above attributes and object sets is shown in Table 1. At this time, the context isn’t a regularization context, the attribute P_2 is no corresponding object, so the object O_1 is changed to the object O_{11} single eave hip roof and the object O_{12} double eave hip roof; the attributes P_6, P_7 and P_8 are important public attributes, we put them into the attribute set of the father object “Chinese traditional roofs” of the objects with these attributes. The objects O_4 and O_5 should be combined into one object according to Subtask 1 of Section 2.4, but according to the ancient architectural knowledge, the objects O_4 and O_5 are actually two different objects, here is the same, because some attributes are simplified, to make the difference, the simplified attribute "P' protruding gable wall" needs to be added in the context shown in table 2. P_3 is a multi-valued attribute, it needs to be changed to Single-valued attributes, so P_3 is changed to the attributes P_{31} 1-slope face, P_{32} 4-slope face、P_{33} 6-slope face and P_{34} 8-slope face，at the same time, in order to make the context more rationalized, the object O_3 is changed to the objects O_{31} round pavilion roof , O_{32} 4-angle pavilion roof, O_{33} 6-angle pavilion roof and O_{34} 8-angle pavilion roof. The final context is shown in Table 2.

TABLE I.
INITIAL CONTEXT OF CHINESE TRADITIONAL ROOF

G M	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
O_1	*		*	*		*	*	*
O_2	*		*	*		*	*	*
O_3	*		*			*	*	*
O_4	*		*		*	*	*	*
O_5	*		*		*	*	*	*
O_6			*		*	*	*	*

And when building them, we had communication with domain experts to ensure that the concept lattice can meet the practical requirements, or they are easy to deviate from the field as they really are. Among them, the context of Chinese traditional roof was converted into a concept lattice shown in Fig. 5 by using Concept Explorer.

TABLE 2.
CONTEXT OF CHINESE TRADITIONAL ROOF

G M	P_1	P_2	P'	P_{31}	P_{32}	P_{33}	P_{34}	P_4	P_5	P_6	P_7	P_8
O_{11}	*				*			*		*	*	*
O_{12}	*	*					*	*		*	*	*
O_2	*	*			*			*	*	*	*	*
O_{31}		*		*						*	*	*
O_{32}		*			*					*	*	*
O_{33}		*				*				*	*	*
O_{34}		*					*			*	*	*
O_4	*		*						*	*	*	*
O_5	*								*	*	*	*
O_6			*						*	*	*	*

There are 16 concepts in Fig. 5, in addition to the circular nodes with two colors, and the concept represented by the other nodes are actually hidden, which are domain concepts generated by automatic clustering, which represents objects with a number of attributes. The concept lattice achieves automatic classification of domain concepts or object, the node hierarchy presents hierarchy or hyponymy of concepts or classes.

For each node in Fig. 5, if domain experts think that the concept lattice can't accurately describe the domain knowledge, the concept lattice can be edited in the help of knowledge engineers in accordance with the relevant rules, here to skip this step. After obtaining a complete concept lattice, we need to deal with nodes and relationships between nodes in the concept lattice to get the prototype of the domain ontology. First, with the help of domain experts, the corresponding concepts of the nodes are named; Second, all the attributes (including inherited properties from the ancestor nodes) and instaces of the nodes are labeled, namely, the intents and the extents of the concepts need to be defined; third, the relationships between the nodes are converted into relationships between the corresponding concepts. The following are the three typical nodes:

Node 1: Chinese traditional roof（{ O_{11}，O_{12}，O_2，O_{31}，O_{32}，O_{33}，O_{34}，O_4，O_5，O_6 }，{ P_6 , P_7 ,P_8}）, This node contains all instance and the common attributes of all instances in the domain, and represents a chinese traditional roof with P_6, P_7 and P_8 .

Node 10: single eave hip roof ({ O_{11} }，{ P_1 ,P_{32} , P_4 , P_6 , P_7 , P_8}).

Node 16: (Φ ,{ P_1 ,P_2 ,P',P_{31} , P_{32} , P_{33} , P_{34} , P_4, P_5 ,P_6 , P_7 , P_8})，the bottom node represents that no roofs meet all the attributes, the concept needs to be removed.

After completing the conversion of concept lattice, we have gotten the ontology prototype of "Chinese traditional roof", the ontology prototype removed the bottom node is shown in Fig. 5.

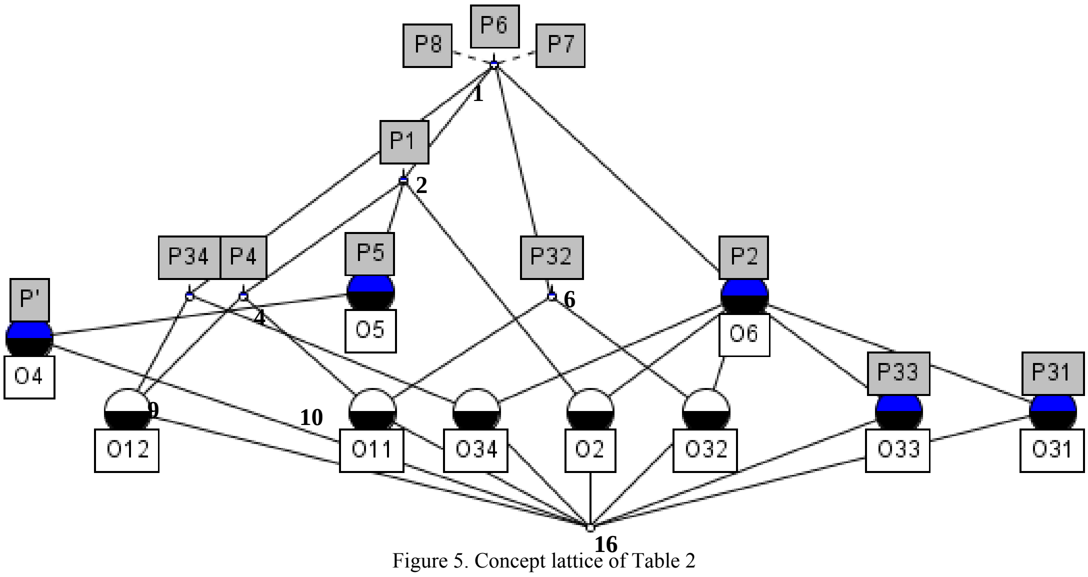


Figure 5. Concept lattice of Table 2

Attributes, instances and axioms in the prototype may appear imperfect, so their expansions need to be done with the help of domain experts and knowledge engineers. For example, for the concept 10 (node 10): single eave hip roof（{ O_{11} }，{ P_1 ,P_{32} , P_4 , P_6 , P_7 , P_8}）, the axiom {single eave hip roof class (P_4. P_1) ∀(P_{32}. P_1)} is added.

E. Step 5: Implementation of the Domain Ontology

We use protégé and ontology language RDF/OWL to describe the ontology, in face, protégé can automatically generate the RDF/OWL code of the ontology, the ontology of Chinese traditional roof contains 15 concepts and 11 attributes, which clarify the relationship between the concepts, the attributes of concepts and instances. Parts of concepts and their relationships in Chinese traditional roof ontology represented in RDF are as follow:

```
    ……
    <rdfs: Class rdf: ID=" O11"
            <rdfs: subClassOf    rdf: resource="# O1" />
        </rdfs: Class>
    ……
        <rdfs: Class rdf: ID=" O1"
            <rdfs: subClassOf    rdf: resource="#Croof" />
        </rdfs: Class>
        <rdfs: Class rdf: ID=" Croof "
            <rdfs: subClassOf    rdf: resource="# Ccomponent
" />
        </rdfs: Class>
          ……
        <rdfs: Class rdf: ID=" Ccomponent "
            <rdfs:comment>    component    class    </rdfs:
comment>
        </rdfs: Class>
    ……
    <rdfs: Property rdf: ID=" P8"
              <rdfs: domain rdf: resource="# Croof " />
              <rdfs: range rdf: resource="&rdf; Literal" />
    </rdfs: Property >
    ……
```

Ontology reasoning is based on RDF file using reasoning machine RacerPro to achieve the ontology reasoning process. Description logic is the basis of ontology reasoning, therefore, it is particularly important how to accurately obtained logical relationships between the domain ontology concepts from domain ontology model. Combining with Fig. 5, we summarized the actual situation of the use of concept lattice to improve description logic. For example, node 2$\sqsubseteq$node 1，this mean that node 2 is its sub concept of node 1; with $Has\Pr operty(1,P_6)$, we can get $kindOf(1,2)\wedge Has\Pr operty(1,P_6)\Rightarrow Has\Pr operty(2,P_6)$, among them, $Has\Pr operty(1,P_6)$ represents that node 1 has attribute P$_6$, $kindOf(1,2)$ represents node 2$\sqsubseteq$node 1; node 4$\cap$node 6 represents the intersection of all attributes of the two concepts; node 9$\cup$node 10，represents the union of all attributes of the two concepts；$\neg P_1$ represents the roof without a main ridge; $\exists P_4.P_1$ represents that there are at least a subsequent attribute for node 4； $\forall P_4.P_1$ represents that there are any a subsequent attribute for node 4; single eave hip roof class ≡($\forall P_4.P_1$) ∩ ($\forall P_{32}$. P_1)，represents the two are equivalent.

F. Step 6: Domain Ontology Application and Maintenance

For the ontology has been achieved, we have intended to apply it to Xi'an ancient architecture protection knowledge management systems and ancient architecture protection knowledge retrieval systems based on the ontology. Those systems are developed based on a Web platform, which make full use of the advantages of internet. In the course of the systems ‘running, if we find that some parts of the ontology in the bottom of the systems cannot meet the actual needs, or new knowledge appears, we will keep upgrading the ontology to ensure its effectiveness.

V. Conclusions

The construction of AAPDO is a type of large and complex system engineering. Our work reasonably applied software engineering and CLT to the process of the domain ontology building, which is with engineering characteristics of ontology building and rigorous mathematical theory; the proposed model supports all the phases of the domain ontology of ancient architecture protection. And it has effectively improved the standardization level, engineering level and formal level on the domain ontology description of ancient architecture protection. In consideration of the small scope of cases in the work, the related studies should be taken as the focus in the next step to ensure the effective domain ontology building of ancient architecture protection in a wide range.

Acknowledgements

This work is partially funded by the specialized research fund for the doctoral program of higher education of the ministry of education (20126120110008) and Xi'an Construction Science and Technology Project (2011008). Thanks for anonymous reviewers' valuable comments.

Reference

[1] Govedarova, Nadezhda, Stanimir Stoyanov, Ivan Popchev. An ontology based CBR architecture for knowledge management in BULCHINO catalogue. Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing. ACM, 2008.

[2] Liu, Qifeng, Bingguo Chang. ON ODM-BASED DESIGN APPROACH FOR ONTOLOGY-BASED KNOWLEDGE BASE OF ANCIENT BUILDINGS PROTECTION TECHNOLOGY. Jisuanji Yingyong yu Ruanjian 4 (2010) 71-73.

[3] BAI, Wei-jing, Song-mao ZHANG, Chun-nian LIU. A knowledge base of traditional Chinese architecture and its efficient implementation demonstrated through. CAAI Transactions on Intelligent Systems 6 (2010) 012.

[4] Song Yu, Qian Liping. Ontology Model of Ancient Chinese Architecture Members and Culture Design and Application. Journal of Beijing University of Civil Engineering and Architecture 4(2012)48-52.

[5] HUANG Mei-Li, LIU Zong-Tian. Research on Domain Ontology Building Methods Based on Formal Analysis. computer science 1(2006) 210-212.

[6] Cimiano P. Staab S, Tane 1. Automatic acquisition of taxonomies form text: FCA meets NLP. In: Proceeding of the International Workshop on Adaptive Text Extraction and Mining,2003.

[7] Haav H M. A Semi-automatic Method to Ontology Design by Using FCA. In: Snasel V. Belohlavek R, eds. Concept Lattices and their Applications. Proceedings of the 2nd International CLA Workshop, TU of Ostrava (2004) 13-25.

[8] Tho, Quan Thanh, et al. Automatic fuzzy ontology generation for semantic web. Knowledge and Data Engineering, IEEE Transactions on 18.6 (2006): 842-856.

[9] Li, Man, Xiao-Yong Du, and Shan Wang. Learning ontology from relational database. Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. Vol. 6. IEEE, 2005.

[10] Li, Junli, et al. A geographic ontology fusion method based on granular theory. Geomatics and Information Science of Wuhan University 38.4 (2013).

[11] Li M, Zhou J, Liang X. Modeling and Description of Organization-Oriented Architecture[J]. Journal of Software, 2014, 9(4): 867-872.

[12] Jiang B, Zhu M. Ontology-Based Information Extraction of Crop Diseases on Chinese Web Pages[J]. Journal of computers, 2013, 8(1).

[13] Kruchten P. Contextualizing agile software development[J]. Journal of Software: Evolution and Process, 2013, 25(4): 351-361.

[14] ZHANG Yunzhong. Research on Domain ontology Construction Method Based on Formal concept Analysis.Jilin University (2009).

[15] YANG Li. The research on concept lattice and reduction method based on dynamic multi-value context. Southwest Jiaotong University (2006).

[16] Thomas Tilley. formal concept analysis applications to requirements engineering and design.The University of Queensland (2003).

[17] WANG Shaofei. Concept lattice construction algorithm and application based on the ontology[D].Dalian Jiaotong University (2006).

[18] Interactive encyclopedia. (2013). Chinese traditional roof. Available at: http://www.baike.com/wiki/%E4%B8%AD%E5%9B%BD%E4%BC%A0%E7%BB%9F%E5%B1%8B%E9%A1%B6.

Zhenbo Bi is a University Lecturer at School of Mathematics, Physics & Information Science, Zhejiang Ocean University, Zhejiang, P.R. China. He has received B.S. degree in Computer Information Management at Xidian University and M.S. degree in Software Engineering at Xi'an Jiaotong University. Now, He is working for PH.D degree in Digital Architecture at Xi'an University of Architecture and Technology. His main research interests are information processing technology in digital architecture and digital ocean.

Huiqin Wang is a professor at Information and Control Engineering School, Xi'an University of Architecture and Technology, Shanxi, P.R. China. Her current interests include signal and digital image processing.

Ying Lu is a student at Architecture School, Xi'an University of Architecture and Technology, Shanxi, P.R. China. Now, she is working for PH.D degree in Digital Architecture. Her current interest is information processing technology in digital architecture.

Socio-Technical Dependencies in Forked OSS Projects: Evidence from the BSD Family

M.M. Mahbubul Syeed[a], Imed Hammouda[b]
[a] Department of of Pervasive Computing, Tampere University of Technology, Finland.
Email: mm.syeed@tut.fi
[b] Chalmers and University of Gothenburg, Sweden.
Email: imed.hammouda@cse.gu.se

Abstract— **Existing studies show that open source projects may enjoy high level of socio-technical congruence despite their open and distributed character. Such observation is yet to be confirmed in the case of forking, where projects originating from the same root evolve in parallel and are typically lead by different development teams. In this paper, we empirically investigate the endogenous and exogenous characteristics of BSD family projects related to socio-technical congruence. Our motivation is that BSD family, as a representative example of forked projects, share a common development ground for both the code-base and the development community, which may influence their evolution from a socio-technical perspective. Our study results show that the BSD family maintain a certain level of collaboration throughout the project history, mainly due to a shared portion of the community. This partly explains the relative harmony of socio-technical congruence levels in the BSD projects.**

Index Terms— **Open Source Software, Evolution, Conway's Law, Socio-Technical Congruence, Forking**

I. Introduction

SOFTWARE development requires effective communication, coordination, and collaboration among developers working on interdependent modules of the same project. The need for coordination is even more evident in Open Source Software (OSS) projects where development is often more dispersed and distributed [1]. As argued in the literature, such coordination and communication may be influenced and guided by the cooperation needs devised by the design of the software [2]. This suggests that there might exist a two way mapping between the communication patterns of the developer community and the architectural dependencies among the components of the software, in which one can be used to approximate the other.

This collaboration can effectively be examined and verified through the notion of socio-technical congruence which defines the match between the coordination needs established by the technical domain (i.e., the architectural dependencies in the software) and the actual coordination activities carried out by project members (i.e., within the members of the development team) [3].

In fact, socio-technical congruence provides an empirical verification of a well-known but insufficiently understood phenomenon known as Conway's Law [4] and describes to which extent the law is enforced in a given software development project [3] [5]. Such empirical verification has been a primary motivation for many research efforts in the realm of socio-technical congruence [6] [7]. However, research on the topic has always assumed that development of a software project is performed by the same organization or group of developers. In the case of open source, projects may evolve in parallel, lead by different development teams. This is known as "forking" [8]. To the knowledge of the authors, no research has been performed yet on socio-technical congruence in the context of forked projects.

In an earlier work we have studied socio-technical congruence, and the significance of Conway's Law, in the FreeBSD open source project [9]. Our previous study showed that the congruence measure is significantly high in FreeBSD and that the congruence value remains stable as the project matured. In this work, we extend our earlier study to cover the BSD project family, empirically investigating the endogenous and exogenous characteristics of BSD projects. BSD projects are popularly known in the research community. For instance, in [10], change history information is extracted from BSD projects for the visualization of change dependencies. Similarly, FreeBSD project has been studied to verify the viability of incremental development approach, and to identify the common characteristics of successful OSS development process in relation to their quality, in [11] and [12], respectively.

Within the endogenous characteristics we investigated the notion of socio-technical dependency through the measure of socio-technical congruence in the individual projects. In the technical domain, the architectural dependencies have been constructed out of source code syntactic information such as functional dependency, attribute referencing and header file inclusion dependency. On the social side, the coordination network has been built out of email conversations between developers.

Among the exogenous characteristics, we examined to what extent the forked projects collaborate and communicate with each other. As a measuring criteria of such

This work was supported in part by TiSE Graduate School, Tampere, Finland and Nokia Foundation Grant, Finland.

doi:10.4304/jsw.9.11.2895-2909

collaboration, we quantitatively measured the alignment among the source code and the developer community of the forked projects. The rationale here is that forked projects hold the same root for both the code-base and the community, thus sharing a common development ground. Hereof it is worth to empirically investigate the extent to which such common ground is maintained during projects evolution.

The remaining of the paper is organized as follows. Section II introduces a number of key concepts that this study uses. Section III introduces the research questions explored and Section IV presents our study design. Results are reported and discussed in Section V, followed by final discussion and related work in Section VI. The overall impact of missing data on the reported results and the replication guidelines are presented in Section VII. Possible limitations and threats to validity are highlighted in Section VIII. Finally, Section IX concludes the paper and sheds light on future research.

II. Definitions

In this section we define a set of concepts used in this study.

A. Conway's Law

Conway's Law in its purest form states that "organizations which design systems are constrained to produce systems which are copies of the communication structures of these organizations" [4]. In other words, the software product architecture reflects the organizational structure of its development team [4] [3]. In [13], Conway's Law is considered homomorphic and thus claimed to be true in reverse as well. This means the communication pattern within a developer community should reflect the architectural dependency in the developed software. Thus, Conway's Law can effectively be interpreted as the basis for studying the social and technical interdependency within a software project [14].

B. Socio-technical congruence

The contemporary phenomenon "Socio-technical congruence" is actually the conceptualization of Conway's Law. Socio-technical congruence can be defined as the match between the coordination needs established by the technical domain (i.e., the architectural dependency in the software) and the actual coordination activities carried out by project members (i.e., within the members of the developer community) [3]. This coordination need can be determined by analyzing the assignments of people to a technical entity such as a source code module, and the technical dependencies among the technical entities [3]. Accordingly, developers within the community should communicate if there exists a communication need. For example, developers working on the same module or on the interdependent modules should be coordinating.

C. Developer Contribution

In this work, developer contribution to a software project can be defined as code contribution or any form of commit made to the code base.

D. Explicit Architecture

The explicit architecture of a software presents the relationship among components of a software (e.g., modules, files or packages) based on the actual design and implementation. For this work, functional dependency, attribute referencing and header file inclusion dependency at code file level are used to derive the Explicit Architecture of a software product.

E. Explicit Coordination Network

The explicit coordination network is a social network in which two developers have a relationship if they have direct communication history as seen by the mailing archives representing the social and technical interactions among the developers.

F. Implicit Architecture

The implicit architecture defines an architecture of the software where any two components (e.g., packages or code files) are related if there are developers who have either (a) contributed to both components, or (b) have direct communication at organizational level (e.g., a one to one email conversation). For instance, consider that developer D1 has contributed to packages P1 and P2, and developer D2 has contributed to package P3. Also consider that both developers have direct communication at organizational level as shown in Fig. 1(a). Thus according to the definition, packages P1, P2 and P3 are linked to each other in the Implicit Architecture (Fig. 1(b)).

G. Implicit Coordination Network

The implicit coordination network is the developer relationship network in which two developers have a relationship if they have contributed either (a) to a common code file or (b) to the code files that have direct relationships in the Explicit Architecture. For instance consider that developers D1 and D2 have contributed to package P1 and developer D3 has contributed to package P2. Also consider that P1 and P2 have a functional dependency (i.e., a direct relationship) as shown in Fig. 2(a). Then, according to the definition, developers D1, D2 and D3 are linked to each other in the Implicit Coordination Network as shown in Fig. 2(b).

H. Forking

In the context of open source development, *forking* occurs when a part of a development community (or a third party not related to the original project) starts a completely independent line of development based on the source code of the original project [8] [15]. To be considered as a fork, a project should have:

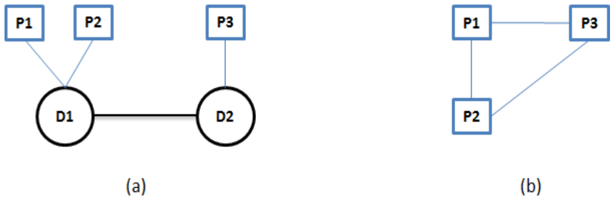


Figure 1. (a) Explicit Coordination Network with contribution to code base (b) Corresponding Implicit Architecture

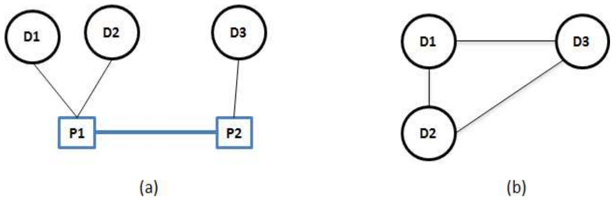


Figure 2. (a) Explicit Architecture with contributing developers (b) Corresponding Implicit Coordination Network

- A new project name.
- A branch of the software.
- A parallel infrastructure (web site, version control system, mailing lists, etc.).
- And a new developer community.

Based on this definition, we propose the following set of relationships within a pair of forked projects: (a) parent-child, in which one project is forked from the other, (b) siblings, if two projects are forked from the same parent project, and (c) lineages, for all descendant relationships in which (a) and (b) do not hold. For example, in Fig. 3, NetBSD and OpenBSD have a parent-child relationship, FreeBSD and NetBSD are sibling projects, whereas FreeBSD, and OpenBSD are the lineages of 386BSD.

III. Research Questions

Our choice of research questions is motivated by our agenda to measure the exogenous and endogenous characteristics of forked OSS projects related to socio-technical congruence.

(RQ1) How does the software architecture compare and evolve across forked OSS projects?

When a project is forked, the source code of the parent project is copied [8]. Thus it is natural that at the initial stage the source code, and hence the architecture, of both systems are similar. Based on this observation, we are interested in exploring the extent to which the forked projects share common architectural structure during their evolution. In doing so, we calculated and compared the architectures of the forked projects at three abstraction levels. Namely, at package level, at first directory level and at n^{th} directory level (i.e., the last directory where the files reside). We argue that this three level comparison can provide a holistic view of the architectural overlapping. For instance, it might be possible that forked projects maintain homogeneous architectural design at higher level of abstraction (e.g., in package level), yet getting liberated at detailed architectural level (e.g., n^{th} directory level).

(RQ2) How does the community compare and evolve across forked OSS projects?

Traditionally, the developer community divides when a project is forked [8]. This is typically followed by a community rebuild and restructuring process in both projects. Community members in both projects might communicate and coordinate in such circumstances in making both projects survive. Thus, our intention here is to examine how these fragmented communities act in building the projects: do they contribute to both projects? Does such collaboration sustain during the evolution of the projects?

(RQ3) How does the socio-technical congruence evolve within the forked OSS projects?

Socio-technical congruence is a natural consequence and a desired property for collaborative development activities, like OSS projects [16]. Conventional wisdom suggests that correspondence between the social and technical domain of a project may reduce the communication overhead and may increase productivity [7]. Furthermore, lack of collaboration is classified as a negative stimuli to performance [17] and has an influence on lowering productivity [5]. Consequently, socio-technical congruence can be a decisive property of a successful project [3] [5] [16]. With strong congruence measure projects can get more cohesive, organized, and self-dependent with higher productivity. Thus our intention here is to stress these reported observations in forked OSS projects by examining the extent to which Socio-Technical Congruence holds in forked projects.

IV. Study Design

This section presents in detail our study design, covering discussion on the case study selection, required data sets, data acquisition, cleaning, and analysis process.

A. Case and Subject Selection

To explore the three research questions, we performed a case study with three large (more than 1,414,641 LOC [18]), long-lived (around 20 years of evolution history) OSS projects that were forked from their predecessors. These case study projects are FreeBSD [19], NetBSD [20] and OpenBSD [21]. All three projects originate from the 386BSD project, which is the version of UNIX developed at the University of California, Berkeley. FreeBSD and NetBSD were directly forked from 386BSD during late 1993, and therefore have a sibling relationship. OpenBSD was forked from NetBSD in 1995, thus having a parent-child forking relationship. Whereas, FreeBSD and OpenBSD are lineages of 386BSD. As a consequence, the core of these projects encompass the code base of 386BSD. The forked relationship among these projects are shown in Fig. 3 and the lifetime of these projects till 2013 are shown in Fig. 4.

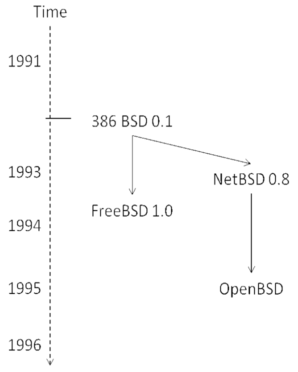


Figure 3. Rough time line of the forked BSD projects

Our selection of the BSD project family was influenced by the following factors: (a) the code base of these projects have undergone continuous development, improvement, and optimization for twenty years [19], (b) these projects have been developed and maintained by a large team of individuals [20], (c) the properties of a forked project hold for these projects, (d) these projects have extensively been used in earlier research on the evolution of OSS projects [22] [23] [18], and (e) results reported in this study can be stressed to OSS projects having similar properties, e.g., forking history, domain, community structure, and size.

B. Data Sets

OSS projects often consist of a number of software development repositories. These repositories contain a plethora of information on both the underlying software and the associated communication and development process [24] [25]. In the literature [26] a great emphasis was given to leveraging these repositories for deriving technical dependencies as well as developers' coordination patterns. The repository data are often longitudinal, allowing for analysis along the whole project evolution phases. Such data sources are highly accepted and utilized medium for empirical studies on OSS projects [27] [28] [29]. In this study we utilized the following repositories.

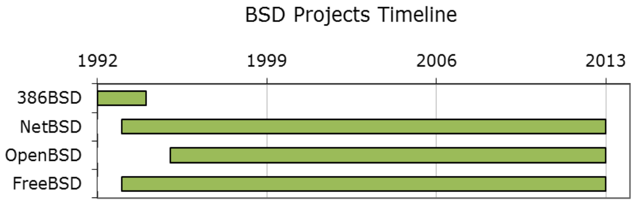


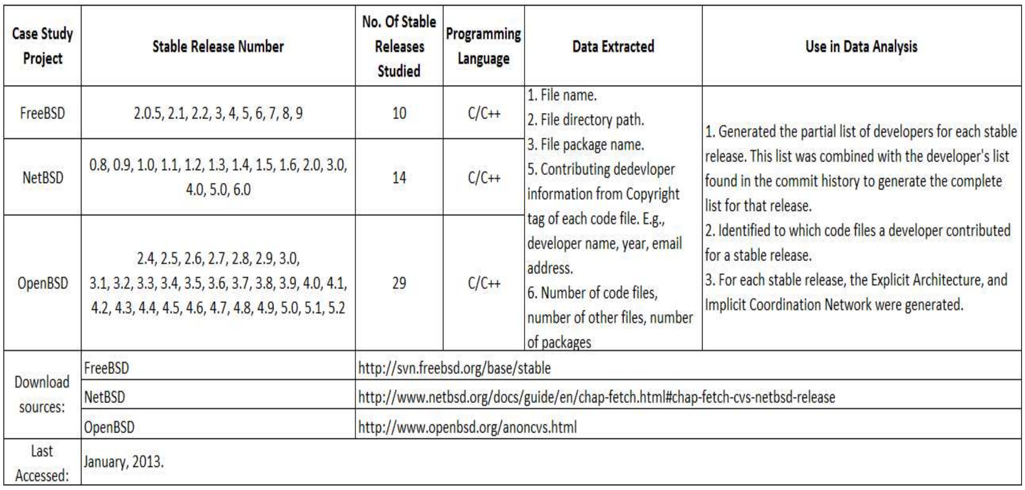
Figure 4. Life time of the BSD projects

Source code repository: We downloaded the source code of each stable release of the three projects. FreeBSD maintains its source code in Subversion version control system, whereas NetBSD and OpenBSD use CVS. In Fig. 5 we provide the details of the stable releases, the data collected from each release, and the corresponding download sources.

Mailing list archive: In OSS projects, email archives provide a useful trace of task-oriented communication and co-ordination activities of the developers during project evolution [30]. In the studied projects, email archives are categorized according to their purpose including commit records, stable release planning, chat, user emails, and bug reports. The archives contain the commit history and the email conversations since the initiation of the projects. In this study we used a complete list of commit records and email conversations from the beginning of each studied project. Consequently, data from relevant email archives was extracted and refined from each project, detail of which is presented in Fig. 6.

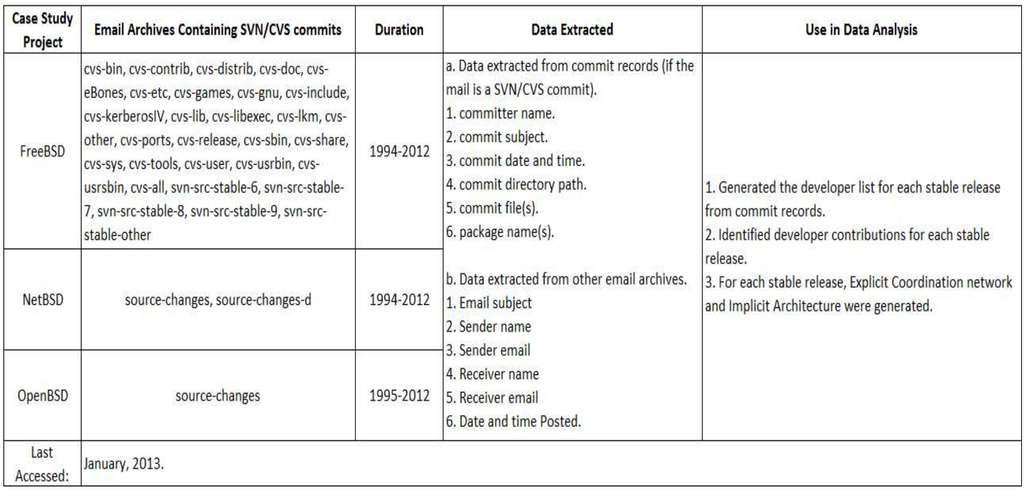
C. Data Collection

From source code repositories: The source code of each stable release of the selected projects was downloaded to a local directory. Fig. 5 lists the stable releases that were downloaded for each project. To extract data from each of the releases, a parser was written in Java. The parser searched through each directory of a stable release, read through the files in a directory and parsed relevant data. Each code file in a release contains a copyright directive. Under this directive the contributing developer name, email, and the copyright year is mentioned. The developers that were found in the process were considered as the initial contributors to that file. To get a complete list of contributors for a stable release, developers names were extracted from the commit history log and were merged with this contributor list. This process is described in following



Case Study Project	Stable Release Number	No. Of Stable Releases Studied	Programming Language	Data Extracted	Use in Data Analysis
FreeBSD	2.0.5, 2.1, 2.2, 3, 4, 5, 6, 7, 8, 9	10	C/C++	1. File name. 2. File directory path. 3. File package name. 5. Contributing dedevloper information from Copyright tag of each code file. E.g., developer name, year, email address. 6. Number of code files, number of other files, number of packages	1. Generated the partial list of developers for each stable release. This list was combined with the developer's list found in the commit history to generate the complete list for that release. 2. Identified to which code files a developer contributed for a stable release. 3. For each stable release, the Explicit Architecture, and Implicit Coordination Network were generated.
NetBSD	0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 2.0, 3.0, 4.0, 5.0, 6.0	14	C/C++		
OpenBSD	2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1, 5.2	29	C/C++		
Download sources:	FreeBSD	http://svn.freebsd.org/base/stable			
	NetBSD	http://www.netbsd.org/docs/guide/en/chap-fetch.html#chap-fetch-cvs-netbsd-release			
	OpenBSD	http://www.openbsd.org/anoncvs.html			
Last Accessed:	January, 2013.				

Figure 5. Stable Releases of BSD Projects (FreeBSD, NetBSD and OpenBSD)



Case Study Project	Email Archives Containing SVN/CVS commits	Duration	Data Extracted	Use in Data Analysis
FreeBSD	cvs-bin, cvs-contrib, cvs-distrib, cvs-doc, cvs-eBones, cvs-etc, cvs-games, cvs-gnu, cvs-include, cvs-kerberosIV, cvs-lib, cvs-libexec, cvs-lkm, cvs-other, cvs-ports, cvs-release, cvs-sbin, cvs-share, cvs-sys, cvs-tools, cvs-user, cvs-usrbin, cvs-usrsbin, cvs-all, svn-src-stable-6, svn-src-stable-7, svn-src-stable-8, svn-src-stable-9, svn-src-stable-other	1994-2012	a. Data extracted from commit records (if the mail is a SVN/CVS commit). 1. committer name. 2. commit subject. 3. commit date and time. 4. commit directory path. 5. commit file(s). 6. package name(s). b. Data extracted from other email archives. 1. Email subject 2. Sender name 3. Sender email 4. Receiver name 5. Receiver email 6. Date and time Posted.	1. Generated the developer list for each stable release from commit records. 2. Identified developer contributions for each stable release. 3. For each stable release, Explicit Coordination network and Implicit Architecture were generated.
NetBSD	source-changes, source-changes-d	1994-2012		
OpenBSD	source-changes	1995-2012		
Last Accessed:	January, 2013.			

Figure 6. FreeBSD, NetBSD and OpenBSD email archives

sections. Information that was extracted using the parser is listed in Fig. 5, column 5. The parsed data for each stable release was then stored in a spreadsheet for further analysis.

From email archives: Data that is maintained in the email archives can be broadly classified into two groups, (a) email archives that maintain CVS/SVN commit records, and (b) archives that store general community discussions (e.g., on stable release planning, chat entries). Fig. 6 presents the total number of email archives that were extracted for each project along with specific names of archives containing the commit records, data collection period, collected data, and their analysis purpose.

For extracting data from each email entry, a data extraction program was written in Java. This data extractor used the web interface of the email archives. Thus each email was read as an HTML page and the data was extracted using the Jsoup HTML parser [31]. Data extracted from each email entry is listed in Fig. 6, column 4. This data was then stored in spreadsheets according to the archive name and year. After that, email data was sorted according to each stable release as follows: (a) emails and commit records were categorized into a specific release if the release number was mentioned in email subject (e.g., SVN commit emails provide release number in email subject for FreeBSD) and (b) other emails for which the release numbers were not mentioned (e.g., freeBSD-stable, freeBSD-chat and some of the CVS commit emails), the posting dates were checked. In this case, for instance, an email was categorized to stable release 3 if its posting date falls between the release date of stable release 2 and 3. The rationale here is that developers would commit to the code base and discuss on its release strategy before it is officially released.

For the CVS/SVN commit email, we parsed the commit path to the repository. The commit path was either mentioned in the subject or in the email body (in specified format). We extracted information like the directory path, package name, and if provided, the name of the modified code file(s) and the stable release number. The name of the committer for each of these CVS/SVN commit emails was considered as a contributor to the code base. Contributors found in this process were combined with the contributors found in the code base to get a complete list of contributing developers for each stable release.

Data preprocessing: Data that was extracted and parsed following the above process contained anomalies in many cases. For instance, developer names and email addresses might contain punctuation characters like semi-colons, inverted comas, brackets, unnecessary white space, and hyphens. Furthermore, parsers may have parsed data inappropriately in some cases. For example, the text *copyright rights reserved* can be treated as part of developer name while parsing copyright directive from a code file. To clean such anomalies data and punctuation characters, data cleaning programs were written in Java. To ensure the correctness of this process, we performed a manual checking on a randomly selected data to verify their correctness.

D. Data Analysis

This section is focused on topics related to the construction of the communication networks, architectures, and their use in measuring the socio-technical congruence utilizing the collected data.

Data analysis is restricted to the stable releases of the projects. This means, analysis point of this study is the stable release dates for a project. This choice of analysis point (instead of discrete time stamps) is made due to the following reasons: (a) a stable release reflects clear milestone for a project, which can also be counted as a step towards successful evolution, and (b) the source code for this study is available for stable releases only, which makes it obvious choice to take release dates as analysis points.

Developer Contribution: Developer contributions were measured release-wise in two ways: (a) from the copyright information provided in each source code file of a release and (b) from the commits made by a developer for a release. Fig. 7(a) shows a sample contribution made by developer *John Birrell* in FreeBSD stable release 3.

Explicit Architecture: The Explicit Architecture of a stable release was constructed based on functional dependency, attribute referencing, and header file inclusion dependency at code file level. For doing this, we used a tool named Understand [32]. This tool takes a source code repository as input and generates the corresponding Explicit Architecture. This tool has been used in previous research, e.g., in [33] [34]. The explicit architecture for each stable release of a project was derived at two abstraction levels, e.g., at code file level and at package level. An example of these two architectures for FreeBSD release 3 is shown in Fig. 8.

Explicit Coordination Network: Following the definition in Section II-E, the Explicit Coordination Network was derived for each stable release of a project. Email conversations for each stable release were used for this purpose. Fig. 7(b) shows example relationships in the Explicit Coordination Network of FreeBSD stable release 3. The weight column in this figure shows the number of email conversations that took place between two developers.

Implicit Architecture: The implicit architecture was generated following the definition in Section II-F. A partial snapshot of the package level Implicit Architecture for FreeBSD stable release 3 is shown in Fig. 9(a). In this architecture, a link weight between two packages designates the number of times the conditions (from Section II-F) hold. The significance of this network lays in the fact that developer communication patterns within the community may simulate the actual architectural dependency. That is, two developers should have communication if they are contributing to same or interrelated components of the software.

Implicit Coordination Network: This network was generated according to the definition presented in Section II-G. A snapshot of this network for FreeBSD stable release 3 is shown in Fig. 9(b). The network shows the actual communication need among developers, based on the design of the software (i.e., the Explicit Architecture). This network is essential due to the fact that if two subsystems exchange information, it is likely that communication among the developers of the two subsystems exists [4].

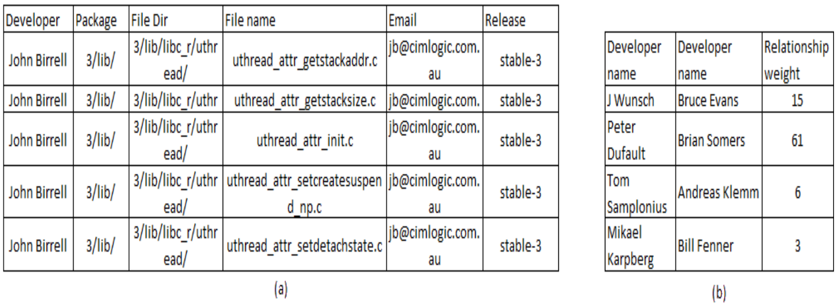
Measuring concurring and congruence among architectures and networks: Comparison among the architectures and communities was measured for two purposes: (a) to measure how the software architectures and communities compare and evolve across forked projects, and (b) to identify how the socio-technical congruence evolve within each forked project.

We applied the following similarity measure to serve both purposes. This approach is analogous to the fit measure used in organizational theory method [5]. An identical approach was applied in [9] for measuring the congruence in FreeBSD project.

$$Concurring/Congruence = \frac{Ref_{A/N} \bigcap Analogous_{A/N}}{\left|Ref_{A/N}\right|} \times 100) \quad (1)$$

In the above equation, $Ref_{A/N}$ is the reference architecture or network (either explicit or implicit), and $Analogous_{A/N}$ it the analogous architecture or network (either explicit or implicit) with which concurring or congruence will be measured.

This equation measures concurring between the two architectures or networks with respect to the reference one, $Ref_{A/N}$. Therefore, the numerator of equation (1) identifies the commonalities between the two given ar-



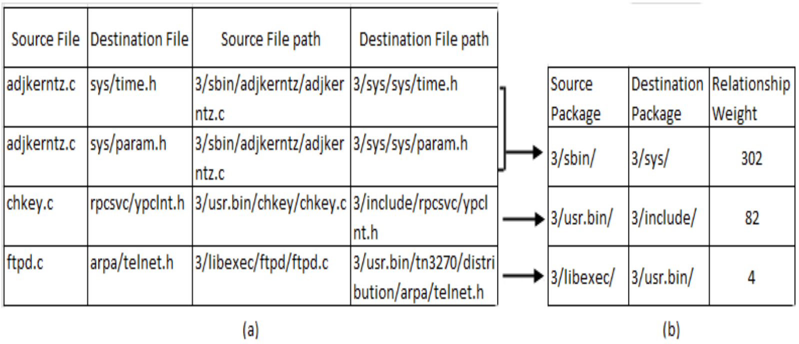
Developer	Package	File Dir	File name	Email	Release
John Birrell	3/lib/	3/lib/libc_r/uthread/	uthread_attr_getstackaddr.c	jb@cimlogic.com.au	stable-3
John Birrell	3/lib/	3/lib/libc_r/uthr	uthread_attr_getstacksize.c	jb@cimlogic.com.	stable-3
John Birrell	3/lib/	3/lib/libc_r/uthread/	uthread_attr_init.c	jb@cimlogic.com.au	stable-3
John Birrell	3/lib/	3/lib/libc_r/uthread/	uthread_attr_setcreatesuspend_np.c	jb@cimlogic.com.au	stable-3
John Birrell	3/lib/	3/lib/libc_r/uthread/	uthread_attr_setdetachstate.c	jb@cimlogic.com.au	stable-3

(a)

Developer name	Developer name	Relationship weight
J Wunsch	Bruce Evans	15
Peter Dufault	Brian Somers	61
Tom Samplonius	Andreas Klemm	6
Mikael Karpberg	Bill Fenner	3

(b)

Figure 7. (a) Sample contributions made by developer John Birrell (b) Sample relationships in Explicit Coordination Network



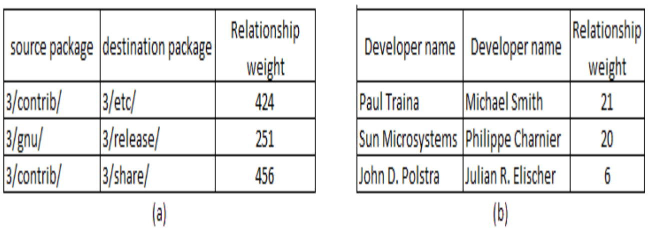
Source File	Destination File	Source File path	Destination File path
adjkerntz.c	sys/time.h	3/sbin/adjkerntz/adjkerntz.c	3/sys/sys/time.h
adjkerntz.c	sys/param.h	3/sbin/adjkerntz/adjkerntz.c	3/sys/sys/param.h
chkey.c	rpcsvc/ypclnt.h	3/usr.bin/chkey/chkey.c	3/include/rpcsvc/ypclnt.h
ftpd.c	arpa/telnet.h	3/libexec/ftpd/ftpd.c	3/usr.bin/tn3270/distribution/arpa/telnet.h

(a)

Source Package	Destination Package	Relationship Weight
3/sbin/	3/sys/	302
3/usr.bin/	3/include/	82
3/libexec/	3/usr.bin/	4

(b)

Figure 8. (a) Code file level Explicit Architecture (b) Package level Explicit Architecture



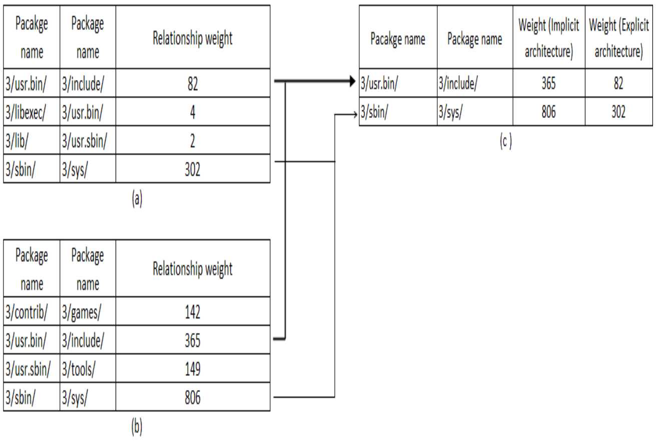
source package	destination package	Relationship weight
3/contrib/	3/etc/	424
3/gnu/	3/release/	251
3/contrib/	3/share/	456

(a)

Developer name	Developer name	Relationship weight
Paul Traina	Michael Smith	21
Sun Microsystems	Philippe Charnier	20
John D. Polstra	Julian R. Elischer	6

(b)

Figure 9. (a) Implicit Architecture (b) Implicit Coordination Network



Pacakge name	Package name	Relationship weight
3/usr.bin/	3/include/	82
3/libexec/	3/usr.bin/	4
3/lib/	3/usr.sbin/	2
3/sbin/	3/sys/	302

(a)

Package name	Package name	Relationship weight
3/contrib/	3/games/	142
3/usr.bin/	3/include/	365
3/usr.sbin/	3/tools/	149
3/sbin/	3/sys/	806

(b)

Pacakge name	Package name	Weight (Implicit architecture)	Weight (Explicit architecture)
3/usr.bin/	3/include/	365	82
3/sbin/	3/sys/	806	302

(c)

Figure 10. (a) Explicit Architecture (b) Implicit Architecture (c) Congruence

chitectures or networks, then is divided by the size of the reference architecture and expressed in a scale of 100.

The application of Equation (1) to specific cases is presented next.

Comparing the Architecture: To measure and compare the architectural concurring among the three projects, we performed a stable release wise comparison of the explicit architectures for each pair of forked projects. Thus in this case, both $Ref_{A/N}$ and $Analogous_{A/N}$ represent two comparable explicit architectures taken from two projects. To be comparable, the stable releases of two projects should be released around the same time period. For instance, consider the stable releases of FreeBSD and NetBSD projects. FreeBSD has 10 stable releases whereas NetBSD has 14 (Fig. 5, column 2). Thus to compare two releases, each taken from the two projects, we determined the release date-wise correspondence. Therefore, FreeBSD release 6 and NetBSD release 3.0 have a correspondence as they were released in November, 2005 and December, 2005, respectively.

The intersection operation in numerator of equation (1) is calculated at three abstraction levels of the explicit architectures, namely, package level (p), first directory level (d_1) and code file directory level (d_n). For package level, the intersection operation results in the number of packages that are common (by comparing the names of the packages) between two releases. On the other hand, for the directory level, e.g., d_1 and d_n, the intersection operation provides the total number of directories that have the complete match in their directory paths. As an illustrative example, consider FreeBSD release 6 and NetBSD release 3.0 which have 19 and 22 packages, respectively. Thus $|FreeBSD - release - 6|$ = 19 and $|NetBSD - release - 3.0|$ = 22. The intersection operation between these two explicit architectures resulted in 16 packages having the same names.

Finally, the concurring value was calculated taking each of these architectures as a reference architecture. This value depicts the extent to which each of these stable releases coincide with the other. In continuation to the above example, FreeBSD release 6 has 84.21% (16/19*100) and NetBSD release 3.0 has 72.72% (16/22*100) concurring with each other. These values were then plotted in a trend chart to visualize how such concurring evolves with the projects. An example of this process is presented in Fig.11 and discussed in Section V-A.

Comparing the Community: To compare the communities among the three forked projects using the similarity measure in Equation (1), we carried out the following: first, the release wise developer list was generated for each project. This step was discussed in section IV-C. Second, for a given pair of releases, the union operation in the numerator identifies the number of contributors in both releases whose names are lexically identical. Finally, for each of the stable releases, concurring value was calculated considering each as a reference network. These values were then plotted in a trend chart. An example of this process is presented in Fig. 14 and discussed in Section V-B.

Socio-technical Congruence: To measure socio-technical congruence using the similarity measure in (1) the following approach was applied: the intersection operation in numerator was carried out between (a) Explicit Architecture and Implicit Architecture, and between (b) Explicit Coordination Network and Implicit Coordination Network. This operation identifies the number of edges (or relationships) that are identical for both the architectures or the networks.

The former measure (in (a)) illustrates the match between the architectural dependency and the architecture produced due to the communication structure of the community. The latter measure (in (b)) in turn depicts the match between the actual coordination activities in the community and the coordination need established by the architectural dependency of the software. These measures verify Conway's Law and the reverse Conway's Law, respectively. Both the measures were determined for each stable release for all three projects. A partial snapshot of the congruence between Explicit and Implicit Architectures of FreeBSD stable release 3 is shown in Fig. 10.

Then to identify the extent to which the implicit architecture and implicit network approximate the corresponding explicit one, we calculated the similarity measure in (1), taking each of the explicit architecture and network as the reference one. The resulting values were plotted in a trend chart for each project to conceptualize their evolution pattern. An example of this analysis is presented in Fig. 17 and discussed in Section V-C.

E. Implementation and Verification

Tools Used In the Study: A number of existing tools and OSS packages were used in this work. For instance, we used the tool *Understand (version: 3.1.659)* [32] to generate the Explicit Architectures. To read/write excel files Apache POI [35] was used. Also, Jsoup HTML parser [31] was used to parse the HTML files.

Implementation and Verification of the Developed Programs: We implemented several data extraction, cleaning, and analysis programs in Java for this work. Data extraction programs were used to extract data from relevant sources and cleaning programs were used for removing the anomalies in the collected data. To verify the correctness of these programs, a two pass evaluation were conducted. First, the programs were tested with a limited number of data samples taken from each of the projects. Notified bugs (e.g., errors in the parsed data for an HTML tag) were fixed accordingly. Second, a manual checking on a random sample of the actual collected data was done. The accuracy of collected data in the second pass was reported to be over 97%.

Additionally, analysis programs were written for generating the architectures, communication networks, release-wise comparisons, and for measuring congruence. These programs in turn were tested following a similar method as stated above.

V. Result Analysis

The target of this study is three-fold. First, we verify the extent to which the forked projects collaborate in both technical and social domain. Second, we measure the socio-technical congruence in each project to conceptualize the socio-technical dependencies. Finally, we study the projects' pattern of evolution during their maturation.

A. Pattern of Architecture Evolution

In this section we present the results of the evolution of the architectural design for each forked project in relation to the other projects. In verifying this, pairwise comparison of the architectural designs (each taken form the compared projects) were made at three abstraction levels. This action was performed according to the procedure presented in Section IV-D. The result of this comparison is presented in Figs. 11, 12 and 13, one for each pair of projects. These figures show the concurring of architectures (plotted in the Y-axis) for each comparable stable release pair (plotted in the X-axis) of the projects.

Overall architectural evolution revealed similar patterns for all three types of forking relationships, e.g., sibling projects, parent-child projects, and lineages. At higher abstraction level (e.g., package level) the architectures of the forked projects maintain high correspondence between them, which remains consistent as the projects evolve. However, at the detailed architectural level (e.g., at directory levels d_1 and d_n), the design and implementation became more disjoint and independent.

For instance, in Fig. 11, the package level concurring between the architectures of FreeBSD and NetBSD projects remain high throughout their release history. For FreeBSD it remains between 61,9% and 84,21%, whereas for NetBSD it is between 57,69% and 80% with slight drifts between the ranges. Contrary to this, directory level overlapping (d_1 and d_n) point out a different trend. In both of these cases, a consistent decrease in concurring can be noticed. For example, for NetBSD and FreeBSD the overlapping at d_1 directory level begins with 82,81% and 56,1% respectively, which gradually decreases to 37,39% and 41.72% respectively. Likewise, at d_n level, the overlapping goes down to 3,63% and 3,34% from 29,77% and 10,82% respectively.

For the other two cases (Fig. 12 and 13), a similar trend was noticed with minor distinction during the early stages of the projects. For instance, in Fig. 13 the overlapping of all three architectural level starts with a very low ratio, which however had a sharp rise in the next release. For the subsequent releases, the pattern remains similar to the observations stated earlier.

Additionally, at any given point of the comparison, the adherence to common architectural design falls off significantly from abstract to detail level of the design. For instance, in 2012, the FreeBSD package level overlapping is 75%, which is however around 41,72% and 3,34% for directory level overlapping d_1 and d_n, respectively. This observation holds for all the three projects.

These observations indicate that the BSD forked projects preserve a common structure at higher level of design, which are however, get liberated progressively at the detailed architectural design. However, thorough analyses of architectural design need to be conducted to fully affirm this claim.

B. Pattern of Community Evolution

Forking of a project causes a split in the community. The fragmentation of the community is typically followed by a rebuild and restructuring phases in both projects (the original and the fork). However, both projects share the same source of code-base, which could stimulate the development communities of the two projects to contribute to both. This observation lead us to investigate the extent to which the community members (from each project) contribute during the evolution of both projects.

The investigation was done according to the process defined in Section IV-D. The results are presented in Fig. 14, 15, and 16, one for each pair of projects. The findings reveal that the level of participation of the community members in the compared projects remains consistent within a given range. Also, a similar pattern of participation is noticed for the three types of forking, confirming the earlier observation in Section V-A.

Relating these observations to individual cases show that for the FreeBSD and NetBSD projects (Fig. 14), the community overlapping remains between 23,49% and 44,9%, whereas for NetBSD it is between 26,47% and 44,23%. Within this range of participation there exist several drifts. For instance, in 1999 and 2007 (Fig. 14), a decrease in participation can be observed.

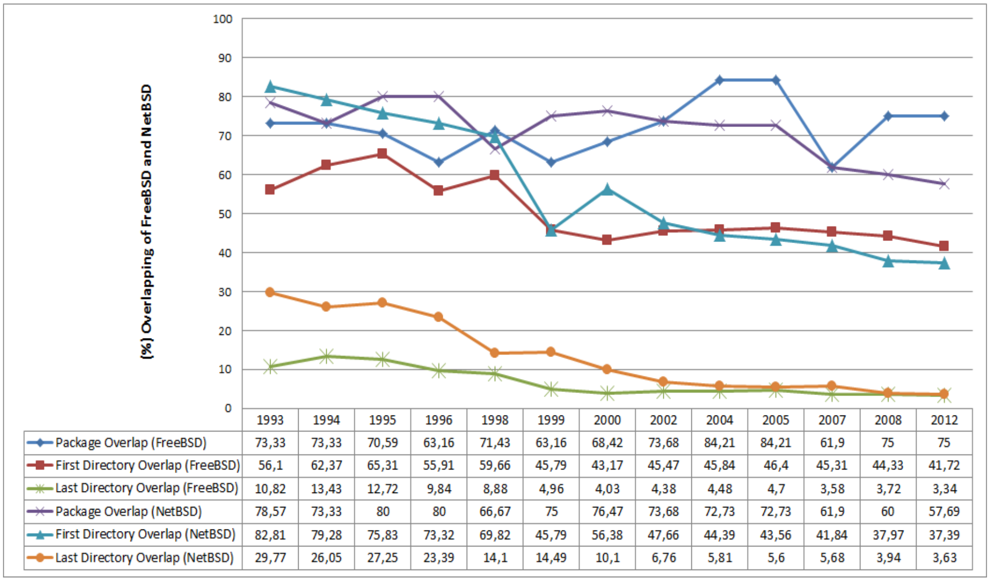
For the other two cases (Fig. 15, and 16), the pattern of overlapping follows a similar trend, except for the first two releases. This observation is similar to that discussed in Section V-A. For instance, the level of contribution rises sharply after having a low participation at the early release. Apart from this, the participation level (in Fig. 15) for NetBSD remains between 42,69% and 50,3%, and for OpenBSD between 34,42% and 38,31%. Similarly, for FreeBSD and OpenBSD (Fig. 16) it is 30,58%-35,05% and 27,18%-30,38%, respectively.

These results lead to the point that a certain group of community members maintain contributions to all the projects. The number of participation also remains stable throughout the evolution.

C. Evolution pattern of Socio-technical Congruence

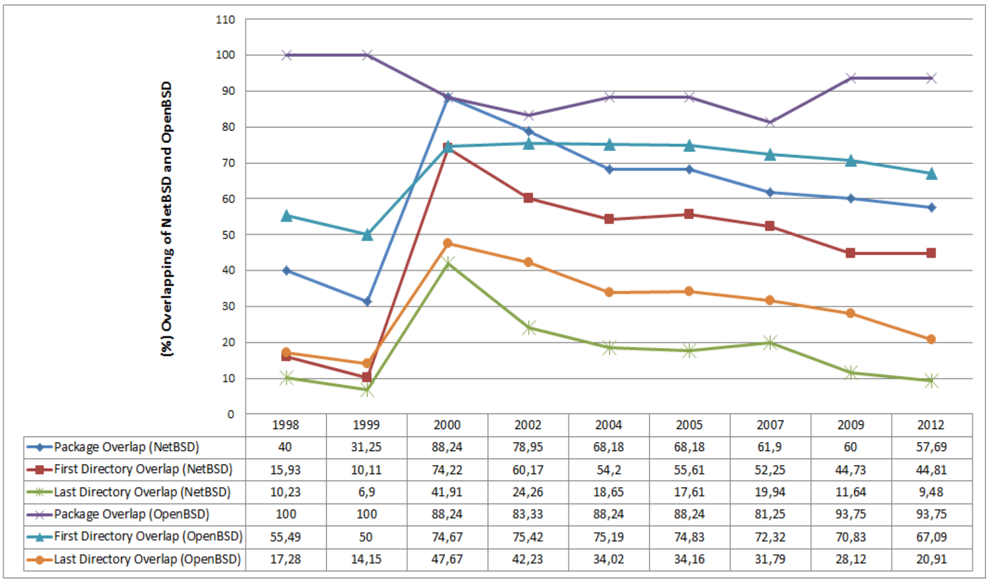
The measurement of Socio-technical Congruence for a project is a two step process. First, the extent to which the communication patterns of the members of the developer community resemble the actual architectural dependencies is verified. And then, the resemblance of the architecture to the community communication is investigated. In doing so, we derived both the implicit and explicit architectures and community collaboration networks, and measured the corresponding congruence. This process was discussed in detail in Section IV-D.

The evolution of congruence at architectural level for the three projects is shown in a trend chart in Fig. 17. In this figure, the congruence approximation is plotted in the



	1993	1994	1995	1996	1998	1999	2000	2002	2004	2005	2007	2008	2012
◆ Package Overlap (FreeBSD)	73,33	73,33	70,59	63,16	71,43	63,16	68,42	73,68	84,21	84,21	61,9	75	75
■ First Directory Overlap (FreeBSD)	56,1	62,37	65,31	55,91	59,66	45,79	43,17	45,47	45,84	46,4	45,31	44,33	41,72
✳ Last Directory Overlap (FreeBSD)	10,82	13,43	12,72	9,84	8,88	4,96	4,03	4,38	4,48	4,7	3,58	3,72	3,34
× Package Overlap (NetBSD)	78,57	73,33	80	80	66,67	75	76,47	73,68	72,73	72,73	61,9	60	57,69
▲ First Directory Overlap (NetBSD)	82,81	79,28	75,83	73,32	69,82	45,79	56,38	47,66	44,39	43,56	41,84	37,97	37,39
● Last Directory Overlap (NetBSD)	29,77	26,05	27,25	23,39	14,1	14,49	10,1	6,76	5,81	5,6	5,68	3,94	3,63

Figure 11. Architectural evolution between the sibling forked projects (FreeBSD and NetBSD)



	1998	1999	2000	2002	2004	2005	2007	2009	2012
◆ Package Overlap (NetBSD)	40	31,25	88,24	78,95	68,18	68,18	61,9	60	57,69
■ First Directory Overlap (NetBSD)	15,93	10,11	74,22	60,17	54,2	55,61	52,25	44,73	44,81
✳ Last Directory Overlap (NetBSD)	10,23	6,9	41,91	24,26	18,65	17,61	19,94	11,64	9,48
× Package Overlap (OpenBSD)	100	100	88,24	83,33	88,24	88,24	81,25	93,75	93,75
▲ First Directory Overlap (OpenBSD)	55,49	50	74,67	75,42	75,19	74,83	72,32	70,83	67,09
● Last Directory Overlap (OpenBSD)	17,28	14,15	47,67	42,23	34,02	34,16	31,79	28,12	20,91

Figure 12. Architectural evolution between parent-child forked projects (NetBSD and OpenBSD)

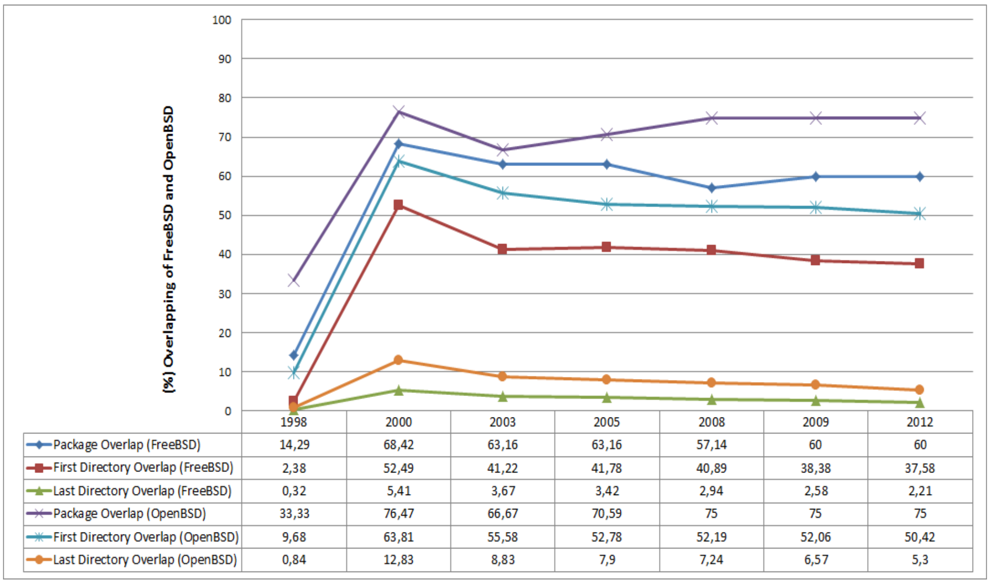
Y-axis (in percentile value) against each stable release of the projects (plotted in the X-axis).

For FreeBSD (the blue line in Fig. 17), the approximation of the congruence consistently has risen starting from 60,5% at the first stable release and has gone up to 89,4%. It had a sharp rise during the early five releases and got stabilized for the later six releases. During this period the congruence level remained between 84,83% and 89,4%. We considered the first four congruence values as outliers as a project usually goes under considerable restructuring and reformation after it is being forked.

For OpenBSD (the green line in Fig. 17) we observed a similar trend of congruence to that of FreeBSD. For the initial two releases the approximation of congruence were around 75%, that increased sharply to 88,38% on the third stable release. Till then onwards it remained stable within the range 85,56% and 88,78%.

In contrast to these two projects, NetBSD (the maroon line in Fig. 17) had a different pattern. In NetBSD the congruence approximation started with 85% and remained stable around 80,77% to 87,5% for the first twelve releases. Nevertheless, for the recent releases (e.g., the last two stable releases), the project experienced a decrease in congruence which has gone bellow 80%.

Accumulation of these results portrays that the approximation of the Explicit Architecture by the congruence is considerably high in all these three projects, which remains stable throughout the evolution. This implies that the architecture derived from the communication pattern of the developer community effectively represents the actual architecture of the software. That is, to a considerable extent the communication of the contributing developers



	1998	2000	2003	2005	2008	2009	2012
Package Overlap (FreeBSD)	14,29	68,42	63,16	63,16	57,14	60	60
First Directory Overlap (FreeBSD)	2,38	52,49	41,22	41,78	40,89	38,38	37,58
Last Directory Overlap (FreeBSD)	0,32	5,41	3,67	3,42	2,94	2,58	2,21
Package Overlap (OpenBSD)	33,33	76,47	66,67	70,59	75	75	75
First Directory Overlap (OpenBSD)	9,68	63,81	55,58	52,78	52,19	52,06	50,42
Last Directory Overlap (OpenBSD)	0,84	12,83	8,83	7,9	7,24	6,57	5,3

Figure 13. Architectural evolution pattern between lineage forked projects (FreeBSD and openBSD)

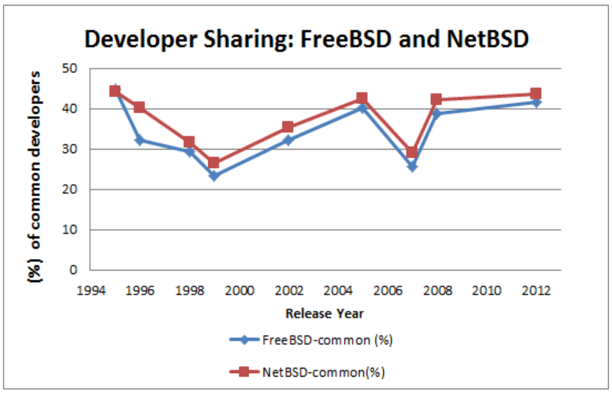


Figure 14. Community concurring pattern between FreeBSD and NetBSD projects

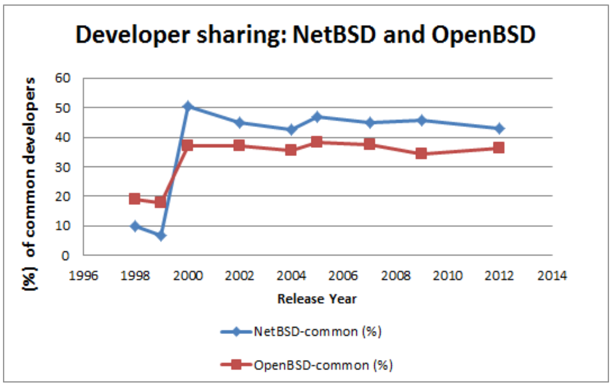


Figure 15. Community concurring pattern between NetBSD and OpenBSD projects

in the community may actually be due to the coordination needs as identified by the architectural dependencies.

On the other hand, the approximation level of the congruence to that of the Explicit Coordination Network reveals a similar pattern for the three projects. Fig. 18

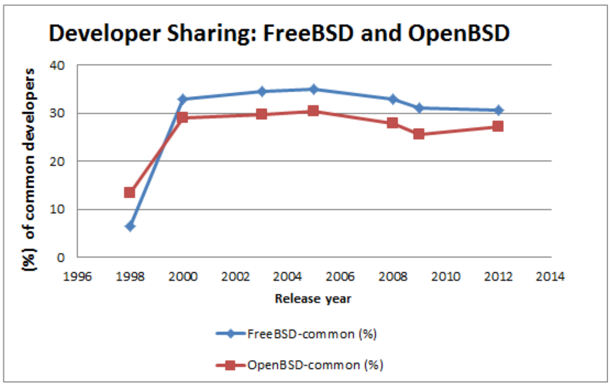


Figure 16. Community concurring pattern between FreeBSD and OpenBSD projects

shows the evolution of approximation against each stable release of the projects.

For FreeBSD (the blue line in Fig. 18), the approximation of the congruence remained between 70,63% and 87,31% from the fourth stable release onwards. A few drifts in congruence in the early three releases were noticed, which can be justified with the same reasoning as before. Yet, there was a decreasing trend of congruence noticed for the last two stable releases.

In the case of OpenBSD (the green line in Fig. 18), the approximation of the congruence to that of Explicit Coordination Network started with 80%, and remained stable between the value 73,35% and 87,77% during the entire evolution of the project. Only for the last release the congruence value went down to 39,58%, which is mainly due to missing data.

For NetBSD (the maroon line in Fig. 18) the congruence approximation started with a high value of 98,87% and remained stable between 8139% and 98,87% as the project progressed. Only for the tenth release (May 2005 in the chart) the congruence has gone as bellow as

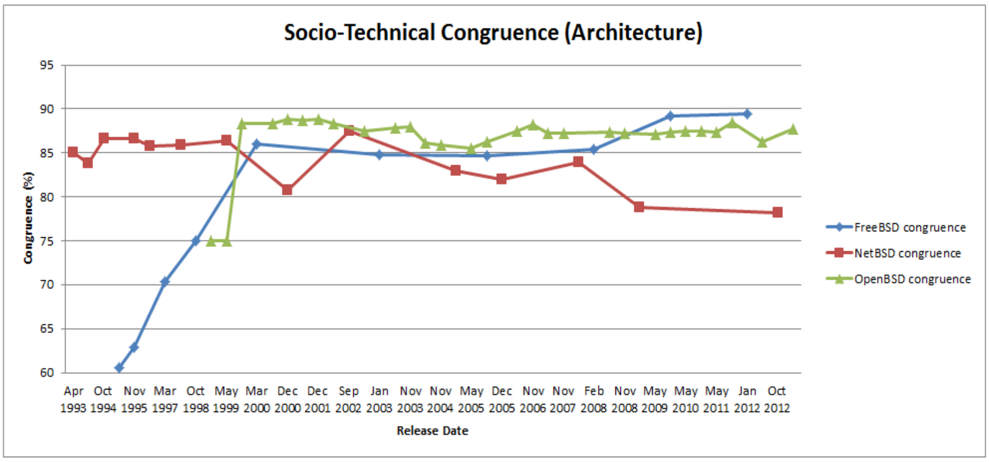


Figure 17. Evolution of Congruence at Architectural Level of the BSD Projects

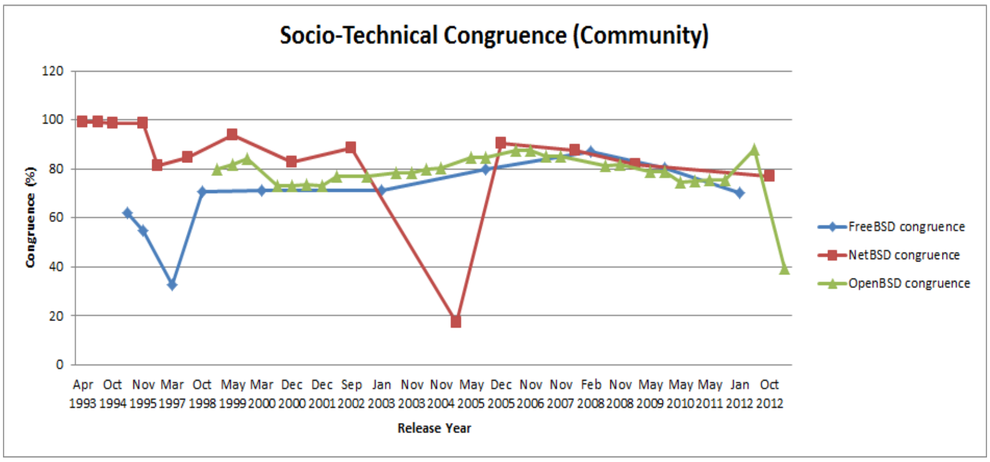


Figure 18. Evolution of Congruence at Community Level of the BSD Projects

17,23%. But it can be treated as an outlier due to missing data. Yet there was a slight decrease noticed for the last three stable releases.

To summarize these results, it can be conceived that the congruence approximation to that of Explicit Coordination Network is considerably high for the three projects. That is, the communication pattern of the developer community derived from the architectural dependency of the components effectively resembles the actual communication pattern. Thus, the communication pattern of contributing developer community can be used to simulate the underlying architectural dependency of the software to a great extent.

VI. Discussion

In this section we hereby summarize the findings of this study and possible implications in relation to prior works.

A. *Research Questions Revisited*

The evidence presented provides a strong indication that each forked project in the BSD family enjoys a high level of Socio-technical congruence throughout their evolution history. Thus, it can be affirmed that to a considerable extent the communication of the contributing developers in the BSD communities might be due to the coordination needs as identified by the technical dependency, and vice-versa. This observation is in-line with the prior work that reported congruence as a desired property and a natural phenomenon of collaborative development works [16] [36].

Alongside these observations, communities of the forked BSD projects have maintained a certain level of collaboration throughout the project history. Our reported model of collaboration shows that a portion of the community is mutual for both the projects. In literature, this group of community members are termed as the bridge between the projects [37], and a means of information flow and collaboration [37] [38].

Moreover, the architectural design at higher abstraction level has remained homogeneous among the forked projects. This might have supported the developer community with better understanding of the overall system designs and have created a common ground for collaboration and contribution. However contrary to this, at detail architectural level these projects are progressively getting liberated. This could be explained by the fact that the developer community of each fork has adopted their own implementation strategies when it comes to fine grained design decisions.

Finally, it was noticed that the pattern of community and architectural evolution for all the three forking relationships (e.g., siblings, parent-child and lineages) have followed similar patterns. This observation highlights the point that forked projects that have originated from the same root project would ideally share a common architectural design and a healthy inter-project collaboration.

B. Implications

It can be argued here that Socio-technical congruence plays a pivotal role in forming cohesive and organized community driven projects, which eventually leads to their successful evolution with high quality. This argument is also affirmed in earlier literature conducted on in-house projects: Higher congruence influences project success [3] [5] [16], with improved productivity [39] [6], maintainability [40], and quality [7].

This measure of socio-technical congruence would better serve the purpose of software development process and organization. Because it provides a quick index of how well the organization is actually aligned with the current and planned sub-division of responsibility in the project [41]. Additionally, the Implicit Architecture can be used as a complementary to the traditional reverse engineering process [42] [43] to derive and validate the recovery of the Explicit Architecture of legacy systems.

The identified pattern of collaboration among the three projects could be one way to explain the sustainability of the forked projects [44], particularly during their early formation stages. Additionally, further study could be initiated to verify the impact of such collaboration on cross project porting and code cloning [45] [46].

Overall, based on our study results, we claim that the traditional perception of forking in OSS projects, which is thought to have negative stimuli for sustainable evolution of the projects [8], can be effectively remedied though (a) maintaining a consistent and cohesive abstract architectural design to form a common ground of collaboration among the forked projects, (b) adopt a collaboration model in which members of a project could participate in other forks, and (c) maintain a consistent and high socio-technical congruence within the project.

None-the-less, this study puts a step forward in reasoning about the successful evolution of forked OSS projects, as this perspective has rarely been studied in current literature on OSS evolution analysis [8] [47].

VII. On the Missing data and Replication of the Study

Data collection process for this study sufferers from some missing data. The missing data constitutes the general communication emails stored in the email archives. Missing email conversations are encountered for NetBSD and OpenBSD projects. To be specific, email conversations during the period of April, 2005 to May, 2005 can not be extracted fully for NetBSD project. Whereas for OpenBSD project, missing emails are noticed during the period of September and October, 2012. In case of NetBSD it is mainly due to broken links to the archives, and for OpenBSD it is probably due to unavailability of the data during that time period.

However, the volume of such missing data is not massive, and thus, have little impact on the overall results. Only at the two points of congruence measure (as discussed in Section V-C), such missing data injected drifts, which however, do not hamper the overall trend of the congruence.

Replication of the study depends on addressing several issues, which includes, (a) data collection from the relevant sources, (b) cleaning and representation of the data and finally, (c) carrying out the analysis. In what follows, a guideline to accomplish these tasks.

Data is collected from two sources, SVN/CVS repositories and email archives. A detail discussion on downloading and extracting data from these sources are presented in Sections IV-B and IV-C. However, to ease this process of data collection for interested researchers, we make available the extracted data in the link given bellow[1]. Further instructions on how to interpret and use the data in replicating this study is discussed in the given link.

Finally, generating the architectures and networks, and carrying out the congruence measure are done thorough the implementation of scripts. There scripts are directly derived from the definitions and analysis methods discussed in Sections II and IV-D, respectively. Tools and packages listed in Section IV-E are used for script implementation. All the packages are open source and are available online for free downloading. However, the scripts used in this study are not made available in the given link. If researchers require assistance in implementing the scripts, we could provide adequate guidelines and the scripts upon request[2].

VIII. Threats to Validity

The following aspects have been identified which could lead to threats to validity of this study.

External validity (how results can be generalized): As case study subject, projects from the BSD family were selected, which are FreeBSD, NetBSD and OpenBSD. All these projects belong to the operating system domain, have large developer and user communities, and have over twenty years of evolution history. Additionally, OSS evolution studies often used these projects as case study. Thus it might be possible to stress the results reported in this article to the population of OSS projects having similar properties, e.g., domain, project size, evolution history. Yet, we cannot claim complete external validity of the results.

[1] http://msyeed.weebly.com/replication-package.html

[2] Contact: $rajit.cit@gmail.com$

Internal validity (confounding factors can influence the findings): Missing historical data - the study has been able to make use only of available data. It is possible, for instance, that there are commit records and developer chat entries other than that recorded in the emails. Additionally, we encountered several broken URL links for emails that could not be retrieved. Thus, we make no claim on the completeness of the email entries with relevance to this study target.

Construct validity (relationship between theory and observation): There exist a few issues that concern the construct validity of the study. First, part of the email entries were categorized to a specific stable release according to their date of post. The reasoning here is that developers commit and discuss on release planning before the product is officially released. Yet, we do not claim the perfection of this approach. Second, the data extraction programs written for this study provided an accuracy of 97%, which was measured with random sample of the collected data. This may affect the construct validity.

IX. Conclusions

The current study provides empirical evidence that successful OSS forked projects that are lineages of an ancestor project may follow similar evolution patterns in terms of (a) technical and social dependencies and (b) achieving a high level of congruence that sustains throughout their evolution. Though from a technical perspective the forked projects get more and more independent by time, they may enjoy a sustainable level of cross project collaboration. Keeping in line with prior evidence [9], we can argue that congruence is an implicit characteristic of successful forked OSS projects, and combining it with inter project collaboration would portray the reason behind the success of such projects. This claim however needs further empirical evidence. As an alternative to the qualitative argumentation approach taken in our study, one could frame our research questions as hypotheses and perform statistical analysis to evaluate them. This constitutes our future work.

References

[1] A. Mockus, R. Fielding, and J. Herbsleb, “Two case studies of open source software development: Apache and mozilla,” *Journal of TOSEM*, vol. 11, no. 3, pp. 309–346, 2002.

[2] G. Valetto, S. Chulani, and C. Williams, “Balancing the value and risk of socio-technical congruence,” *Workshop on Sociotechnical Congruence*, 2008.

[3] I. Kwan, A. Schrter, and D. Damian, “Does socio-technical congruence have an effect on software build success? a study of coordination in a software project,” in *IEEE Trans. Software Eng.*, vol. 37, no. 3, 2011, pp. 307–324.

[4] M. E. Conway, “How do committees invent?” *Datamation*, vol. 14, no. 4, pp. 28–31, 1968.

[5] M. Cataldo, P. A. Wagstrom, J. D. Herbsleb, and K. M. Carley, “Identification of coordination requirements: Implications for the design of collaboration and awareness tools,” in *ACM CSCW*, 2006, pp. 353–362.

[6] L. Colfer and C. Baldwin, “The mirroring hypothesis: Theory, evidence and exceptions,” in *working paper, Harvard Business School*, 2010.

[7] N. Nagappan, B. Murphy, and V. Basili, “The influence of organizational structure on software quality: an empirical case study,” in *ICSE ’08 Proceedings of the 30th international conference on Software engineering*, 2008, pp. 521–530.

[8] G. Robles and J. Gonzalez-Barahona, “A comprehensive study of software forks: Dates, reasons and outcomes,” in *OSS, IFIP AICT 378*, 2012, pp. 1–14.

[9] M. Syeed and I. Hammouda, “Socio-technical congruence in oss projects: Exploring conways law in freebsd oss evolution,” in *Proceedings of 9th International Conference of Open Source Systems (OSS), Springer*, 2013.

[10] M. Fischer, J. Oberleitner, J. Ratzinger, and H. Gall, “Mining evolution data of a product family,” *ACM SIGSOFT Software Engineering Notes*, vol. 4, no. 30, pp. 1–5, 2005.

[11] J. Niels, “Putting it all in the trunk: incremental software development in the freebsd open source project,” *Information Systems Journal*, vol. 11, no. 4, pp. 321–336, 2001.

[12] T. Dinh-Trong and J. Bieman, “The freebsd project: A replication case study of open source development,” *Software Engineering, IEEE Transactions on*, vol. 31, no. 6, pp. 481–494, 2005.

[13] J. Han, C. wu, and B. Lee, “Extracting development organization from open source software,” in *16th Asia-Pacific Software Engineering Conference, IEEE.*, 2009, pp. 441–448.

[14] E. S. Raymond, “The new hacker’s dictionary (3rd ed.),” in *Cambridge, MA, USA: MIT Press*, 1996.

[15] L. M. Nyman and T. Mikkonen, “To fork or not to fork: Fork motivations in sourceforge projects,” in *Source Systems: Grounding Research : IFIP Advances in Information and Communication Technology*, 2011, pp. 259–268.

[16] T. Browning, “Applying the design structure matrix to system decomposition and integration problems: a review and new directions,” in *Engineering Management, IEEE Transactions on*, vol. 48, no. 3, 2001, pp. 292–306.

[17] M. E. Sosa, S. D. Eppinger, and C. M. Rowles, “The misalignment of product architecture and organizational structure in complex product development,” in *Management Science*, vol. 50, no. 12, 2004, pp. 1674–1689.

[18] I. Herraiz, J. Gonzalez-Barahona, G. Robles, and D. German, “On the prediction of the evolution of libre software projects,” in *ICSM*, oct. 2007, pp. 405 –414.

[19] FreeBSD, “http://www.freebsd.org/,” 2013.

[20] NetBSD, “http://www.netbsd.org/about/,” 2013.

[21] OpenBSD, “http://www.openbsd.org/,” 2013.

[22] J. Wu, R. Holt, and A. Hassan, “Empirical evidence for soc dynamics in software evolution,” in *Software Maintenance, 2007. ICSM 2007. IEEE International Conference on*, oct. 2007, pp. 244 –254.

[23] I. Herraiz, “A statistical examination of the evolution and properties of libre software,” in *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*, sept. 2009, pp. 439 –442.

[24] J. C. JE, L. V. LG, and A. Wolf, “Cost-effective analysis of in-place software processes,” in *IEEE Transactions on Software Engineering*, vol. 24, no. 8, 1998, pp. 650–663.

[25] D. Atkins, T. Ball, T. Graves, and A. Mockus, “Using version control data to evaluate the impact of software tools,” in *Proceedings 21st International Conference on Software Engineering*, vol. 24, no. 8, 1999, pp. 324–333.

[26] I. Kwan, M. Cataldo, and D. Damian, “Conway’s law revisited: The evidence for a task-based perspective,” *IEEE Software*, vol. 29, no. 1, pp. 90–93, 2012.

[27] M. Goeminne and T. Mens, “A framework for analysing and visualising open source software ecosystems,” in *Proceeding IWPSE-EVOL ’10*, 2010, pp. 42–47.

[28] D. M. German, “Using software trails to reconstruct the evolution of software,” in *JOURNAL OF SOFTWARE MAINTENANCE AND EVOLUTION: RESEARCH AND PRACTICE*, vol. 16, 2004, pp. 367–384.

[29] Y. Wang, D. Guo, and H. Shi, “Measuring the evolution of open source software systems with their communities,” in *ACM SIGSOFT Software Engineering Notes*, vol. 32, no. 6, 2007.

[30] W. Zhang, Y. Yang, and Q. Wang, “Network analysis of oss evolution: An empirical study on argouml project,” in *IWPSE-EVOL11*, 2011.

[31] jsoup: Java HTML Parser, “http://jsoup.org/,” 2013.

[32] U. S. C. Analysis and Metrics, “http://www.scitools.com/,” 2013.

[33] D. Darcy, S. Daniel, and K. Stewart, “Exploring complexity in open source software: Evolutionary patterns, antecedents, and outcomes,” in *Proceedings of the 43rd Hawaii International Conference on System Sciences*, 2010, pp. 1–11.

[34] M. Simmons, P. Vercellone-Smith, and P. Laplante, “Understanding open source software through software archaeology: The case of nethack,” in *Proceedings of the 30th Annual IEEE/NASA Software Engineering Workshop*, 2006, pp. 47–58.

[35] A. P.-J. A. for Microsoft Documents, “http://poi.apache.org/,” 2013.

[36] J. Herbsleb and R. Grinter, “Architectures, coordination, and distance: Conway’s law and beyond,” in *Journal IEEE Software*, vol. 16, no. 5, 1999, pp. 63–70.

[37] M. Weiss, G. Moroiu, and P. Zhao, “Evolution of open source communities,” in *IFIP International Federation for Information Processing, Volume 203, Open Source Systems*, 2006, pp. 21–32.

[38] J. Gonzalez-Barahona, L. Lopez, and G. Robles, “Community structure of modules in the apache project,” in *Workshop on Open Source Software Engineering*, 2004.

[39] C. Baldwin and K. Clark, “Design rules: The power of modularity,” in *MIT Press*, 2000.

[40] F. P. Brooks, “The mythical man-month,” in *Anniversary Edition: Addison-Wesley Publishing Company*, 1995.

[41] G. Valetto, M. Helander, K. Ehrlich, S. Chulani, M. Wegman, and C. Williams, “Using software repositories to investigate socio-technical congruence in development projects,” in *ICSE Workshops MSR*, 2007, pp. 25–25.

[42] H. Dayani-Fard, Y. Yu, J. Mylopoulos, and A. Periklis, “Improving the build architecture of legacy c/c++ software systems,” in *8th FASE*, 2005.

[43] R. Kazman and S. Carrire, “Playing detective: Reconstructing software architecture from available evidence,” in *Technical Report CMU/SEI-97-TR-010, Carnegie Mellon University*, 1997.

[44] J. Gamalielsson and B. Lundell, “Sustainability of open source software communities beyond a fork: How and why has the libreoffice project evolved?” *Journal of Systems and Software*, vol. 89, pp. 128–145, 2014.

[45] B. Ray and M. Kim, “A case study of cross-system porting in forked projects,” in *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*. ACM, 2012, p. 53.

[46] D. German, M. D. Penta, Y.-G. G. éhéneuc, and G. Antoniol, “Code siblings: Technical and legal implications of copying code between applications,” in *MSR’09*. IEEE, 2009, pp. 81–90.

[47] M. Syeed, I. Hammouda, and T. Systa, “The evolution of open source software projects: a systematic literature review,” *Journal of Software*, vol. 8, no. 11, pp. 2815–2829, 2013.

M.M. Mahbubul Syeed received his B.Sc degree in Computer Science and Information Technology from Islamic University of Technology, Bangladesh in September, 2002 and his M.Sc degree in Information Technology from Tampere University of Technology, Finland in April, 2010. He is currently working towards his Ph.D. degree and working as a researcher in the same university. His current research interest includes study of Open Source Software ecosystem, ecosystem enabling architecture, project evolution, experimental software development, and big data mining and knowledge extraction.

Dr. Imed Hammouda joined University of Gothenburg in September 2013. Before that, he was Associate Professor of software engineering at Tampere University of Technology (TUT), Finland. At TUT, he was heading the international masters programme at the Department of Pervasive Computing. He got his Ph.D. in software engineering from TUT in 2005. Dr. Hammouda’s research interests include open source software, software architecture, software development methods and tools, and variability management. He was a founding member and leader of TUTOpen - TUT research group on open source software. He has been the principal investigator of several research projects on various open initiatives. Dr. Hammouda’s publication record includes over fifty journal and conference papers.

Optimal Selection System of Internal Fixation Methods for Femoral Neck Fracture

Monan Wang
Robotics Institute, Harbin University of Science and Technology, Harbin, China
Email: qqwmnan@163.com

Abstract—**Objective: In order to avoid the adverse consequences which cause by anthropogenic factor in the femoral neck fracture surgery and meet the requirements of selecting the most reasonable internal fixation way, from the viewpoint of biomechanics, an optimal selection system of internal fixation methods for femoral neck facture are established which integrates preprocessing, solving and post-processing. Method: A reasonable simplified femur model is presented by analyzing a precise model of femur standing on one leg. After special patient's femur model is set up, different internal fixation model are analyzed using finite element method (FEM). The interactive communication between APDL and object-oriented programming language is realized. Following evaluation parameters are got based finite element analysis: sinking displacement, lateral displacement, torsional displacement, sum displacement of femur head and gap displacement between fracture surfaces. According to the results, the evaluation algorithm is complied to select the optimal internal fixation method. This study is completed in 3/20/2012 at Harbin University of Science and Technology. Result: As the application of optimal selection program of internal fixation methods for femoral neck fracture, the best fixation way is got when fractures angle is 70°. The result show, without considering bone substance, materials and complications, only from the displacement changing trends of femoral head and fracture interfaces, the internal fixation method with two erect screws is the most reasonable one. Conclusion: The system can provide the number of tightening screw, the fixed angle and combinatorial way of the optimal result to doctors to perform operations.**

Index Terms—**Femoral neck fracture, Internal fixation, Optimal selection, Finite element method, Evaluation algorithm**

I. Introduction

The treatment of femoral neck fracture is divided into conservative therapy and aggressive surgical therapy. But the nonunion rate of conservative therapy accounts for 48%. 34% of the patients may occur to femoral head necrosis[1], so doctors advocate aggressive surgical therapy in modern orthopedics. In common, surgical treatment is divided into 3 steps; they are fracture reduction, selecting internal fixation method and nail-pierced operation. Traditionally, the internal fixation method is controlled by doctors which relies solely on practical experience. This way usually leading to the rationality of structural mechanics ignored. According WOLFF principle, reasonable mechanical environment can promote the growth and healing of bone. During the process of surgical treatment, selection of internal fixation method is the only controlled variable and the only critical factor which can change the mechanical environment. Therefore, select the most reasonable internal fixation pattern is the fundamental guarantee of nail-pierced operation successfully. Aiming at selecting the optimal internal fixation method for femoral neck fracture in this paper, combining with doctors' operation and theoretical research, propose an optimal system which is available for users from biomedical engineering perspective. ANSYS Parametric Design Language (APDL) is applied to establish finite element model; the nodal displacement of fracture site under the condition of special patient's is fixed by personalized internal fixation pattern. Model is solved by FEM. Finally assisting the users find out the optimal operation plan of internal fixation based analysis results and evaluation rules.

II. Methods

A. Fractured Types and Internal Fixation Methods

The number of tightening screw, fixed angles and combinatorial ways, these 3 develop several kinds of internal fixation patterns by permutation and combination. Taking the femoral neck fractures as research object, according to the clinical treatment methods in this paper, fixed angles are set from 30°to 70°; the number of tight screw is set from 1 to 3; Combinatorial ways are set as one screw, two horizontal screws, two slant screws, two erect screws, upside-down equilateral triangles, erect equilateral triangles[2,3]. This setting way depending on special patient's fracture condition realizes the users can select the personalized best internal fixation method.

B. Assumed Conditions for Finite Element Modeling and Solving

Bone is anisotropic and is made up by a vast range of materials[4]. Because the computed result is not much contrast, the bone is assumed to be an elastic material and complies with linear elastic theory[5]. Some assumptions for finite element modeling are made in this paper:

Before any action, skeleton is in the unstressed natural state, without discontinuity inside.

doi:10.4304/jsw.9.11.2910-2917

The stress and strain of per point can be expressed by a continuous function of coordinate and has the same physical property and mechanical characteristic.

When bone is influenced under the condition of load or temperature variation, the deformation would restore if it is in elastic range. In addition, the deformation which is caused by external force is usually far less than the original size.

In view of these assumptions, the define method of element attribute and the assignment method of material prosperities are both based on the linear elastic theory.

For solving models by FEM effectively, some assumptions for solving have to be made before simulation study:

The study stage is set as the fracture reduction and internal fixation have just finished and the healing process is about to start.

The fracture site is fixed by cylindrical screws which closely integrated with bone. Even if load is applied there is no relative displacement, which is the reason that screws are simplified as cylinders.

Assume the strength of cylindrical screws is enough which cannot be broke.

According to the rough situation of bone, the frictional factor is set as 0.9.

C. Modeling

For we focus on the mechanical characters of femoral neck and overcome shortcomings of modeling by finite element software, the model is simplified in this paper. Create a precise model and calculate it by finite element software under the condition that standing on one leg. So the physiological load has to be known. Because the mechanical environment of femur is complicated, we only can get the composition of forces. According to the accurate mechanical model of femur which proposed by Scige[6] that when standing on a single leg, there are 3 composed forces, they are the joint reaction force J, iliotibial tract muscle force R, abductor force M. The mechanical model is shown in Fig.1.

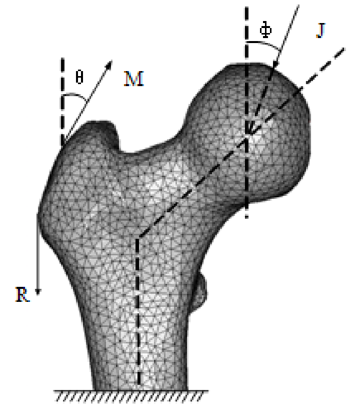


Figure 1 The mechanical model of a femur proposed by Scige. According to the accurate mechanical model of femur which proposed by Scige[6] that when standing on a single leg, there are 3 composed forces, they are the joint reaction force J, iliotibial tract muscle force R, abductor force M.

After calculating, following conclusions are got: the leading force J which transfers from femoral to femoral shaft. The upper side of femoral neck which is in the middle is subjected tension stress, and the downside is subjected compressive stress. The stress of great trochanter is little, so the finite element model is simplified at this part.

Aim to establish special patient finite element model, parameterized method is adopted in this paper. A finite element model of femur and an internal fixation model are created. Set anatomy parameters such as neck-shaft angle(b), the diameter of femoral head(d) which got from X-ray film as variables by *SET function of APDL[7]. These parameters are input into the interface of the system then form an individualized APDL file. The definition of each unit and the assignment of material property are set by EX, ET functions of APDL.

After femur neck fracture geometry model and finite element mechanical model are gotten, the number of tightening screws, the fixed angle and combinational way are set as variables which are selected by users. Due to APDL cannot distinguish string type, define the 6 kinds of combinational ways as macros combing with condition control statements for APDL to read. Create cylindrical screws and define the materials are stainless steel. Insert them into the femoral neck by Boolean operations. The radius of the screws is chosen by thread series of international standard based on constant section. Finally 6 cases of internal fixation finite element models are shown in Fig.2.

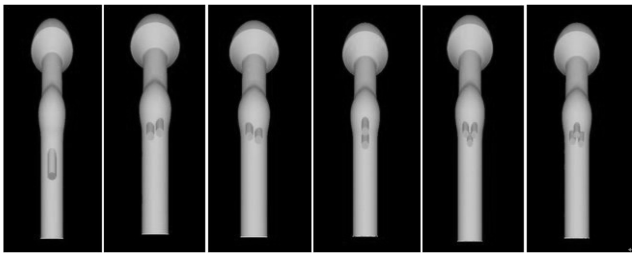


Figure 2 The schematic of 6 cases of internal fixation finite element models. Combinatorial ways are set as one screw, two horizontal screws, two slant screws, two erect screws, upside-down equilateral triangles, erect equilateral triangles.

D. Solving

FEM is adopted to solve internal fixation models. Expect the functions of APDL mentioned above, in order to apply load to the fixed node, use the function asel with the parameter loc of APDL to select nodes which should apply load J, M, and R. Meanwhile select the region to apply all constraints. Forces of J, M, and R are applied to the fixed nodes and start to solve.

Establish the FE mechanical simulation model:

$$Mx'' + C + Kx = f$$

M represents a mass matrix, C represents a damping matrix and K represents a stiffness matrix.

The tetrahedral element is adopted, so the displacement function is expressed by determinant:

$$u = \frac{1}{6V}\{(a_i + b_i x + c_i y + d_i z)u_i +$$
$$(a_j + b_j x + c_j y + d_j z)u_j +$$
$$(a_m + b_m x + c_m y + d_m z)u_m +$$
$$(a_p + b_p x + c_p y + d_p z)u_p \ \}$$

V represents the volume of a tetrahedron.

The strain matrix verified as follows:

$$\varepsilon = Ba^e = [B_i, B_j, B_m, B_p]a^e$$

$$B_i = \begin{bmatrix} \frac{\partial N_i}{\partial x} & 0 & 0 \\ 0 & \frac{\partial N_i}{\partial y} & 0 \\ 0 & 0 & \frac{\partial N_i}{\partial z} \\ \frac{\partial N_i}{\partial y} & \frac{\partial N_i}{\partial x} & 0 \\ 0 & \frac{\partial N_i}{\partial z} & \frac{\partial N_i}{\partial y} \\ \frac{\partial N_i}{\partial z} & 0 & \frac{\partial N_i}{\partial x} \end{bmatrix} = \frac{1}{6V}\begin{bmatrix} b_i & 0 & 0 \\ 0 & c_i & 0 \\ 0 & 0 & d_i \\ c_i & b_i & 0 \\ 0 & d_i & c_i \\ d_i & 0 & b_i \end{bmatrix}$$

$$N_i = (a_i + b_i x + c_i y + d_i z)/6V$$

Change the corresponding subscript, other submatrixes can be got.

E. Judgment Criteria

For the patients only can do moderate physical activity which is a static analysis, so standing on one leg is considered in this paper. According to AO principles, judging from the following parameters:

The displacement of femoral head (US): When standing on one leg, femoral head would subject load and cause displacements including lateral displacement which influence the tensile strength and immobility of fracture surfaces in x-axis; sinking displacement which influence shearing resistance in y-axis; torsion displacement which influence resisting to torsion in z-axis. The values of these 3 displacements are the smaller the better. For the sake of evaluating UX, UY, UZ, set evaluation parameters named US which represents resultant displacement of 3 directions, which is the smaller, the better.

The evaluation parameters include lateral displacement of femoral head UX, the sinking displacement UY, the torsion displacement UZ and sum displacement US.

Boundary conditions are to restrain the sagittal plane, longitudinal plane, coronal plane of distal femur.

The maximal interfragmentary movement (DS): the proximal femur is a nearly cantilever structural model. Its upside is subjected tensile stress and downside is subjected pressure stress when the external load is applied. Assume the fracture surfaces are totally ruptured in this paper. Therefore, the fracture displacement is going to generate which is represented as DS. By measuring the distance of 2 nodes between the fracture surfaces, DS can be got. The buckling resistance of the fixed structure better if DS is smaller. The schematic of DS is shown in Fig.3.

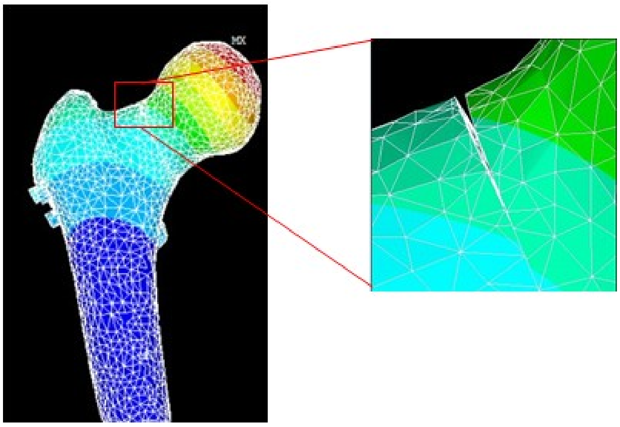


Figure 3 the schematic of DS

The evaluation parameter is the maximal interfragmentary movement DS.

Boundary conditions are to restrain the sagittal plane, longitudinal plane, coronal plane of distal femur.

Among the evaluation parameters, US and UF is the most directly evaluation parameters which can reflex the fixed-effect. However, the size relationships of US and DS are not conformity in practical situations. In addition, summarize the biomechanical experiments the other universities did, we find that the ratio of US among reasonable fixed patterns would not larger than 114% and DS would not larger than 153%. So, one third of the optimum results are selected to define US and DS. Therefore, the following criteria are proposed which compile into an evaluation algorithm by C++. The criteria are shown in Tab. 1, and the flow chat is shown in Fig.4.

TABLE I
THE JUDGMENT CRITERIA OF OPTIMAL SELECTION SYSTEM

$US_{\min} = DS_{\pi}$	Select the internal fixation pattern which minimum of US and DS correspond to.
$US_{\min} \neq DS_{\pi}$	$\frac{US}{US_{\min}} \le 1.047$
	$\frac{DS}{DS_{\min}} \le 1.177$ Select the internal fixation pattern which the minimum of US correspond to.
	$\frac{DS}{DS} > 1.177$ Do not find the most reasonable internal fixation pattern. Please reselect.

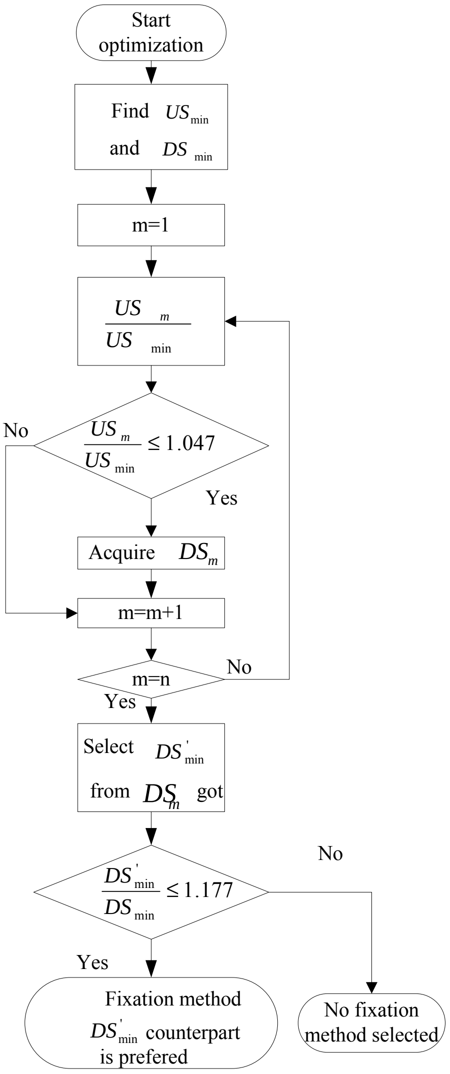

Start optimization
Find $US_{\min}$ and $DS_{\min}$
m=1
$\frac{US_m}{US_{\min}}$
$\frac{US_m}{US_{\min}} \le 1.047$
No
Yes
Acquire DS_m
m=m+1
m=n
No
Yes
Select $DS'_{\min}$ from DS_m got
$\frac{DS'_{\min}}{DS_{\min}} \le 1.177$
No
Yes
Fixation method $DS'_{\min}$ counterpart is prefered
No fixation method selected

Figure 4 The flow chat of the evaluation algorithm

The program include:

(1)Compile the communication codes between object-oriented language and APDL, to realize that collect information from interfaces and transfer to APDL to model, solve and show the results.

(2)According to the information collected, such as anteversion angle, neck-shaft angle, the diameter of femoral neck and so on, transfer them to the parameters of APDL to establish the geometry model of femur.

(3)According to the medical history of patients, considering the disease which may affect the selection of internal fixation pattern, like osteoporosis, the system will warn the users that one screw is prefered[8].

(4)Combining with patients' actual situation, initially select a method. This method does not indicate the optimal plan and can be reset frequently. Transfer the int and float type parameters of operative plan by the codes of step1 to establish the fixation finite element model which is fractured.

(5)Simulate standing on one leg, base on the calculation model Scige put forward and the assumptions above, the equilibrium equations of force and moment are list. Then calculate the 3 forces load on femoral head.

(6)The results computed by step 5 are transferred to finite element model to solve the model by FEA (finite element analysis), and the analysis results will show on the interface.

(7)The system calls the APDL codes of post-processing, show the numerical results of each internal fixation method the user chosen on the interface. Compare the results the system will give the user the optimal result by the evaluation algorithm back stage, and show it on the window.

According these steps, the flow chat of the system is shown in Fig.5.

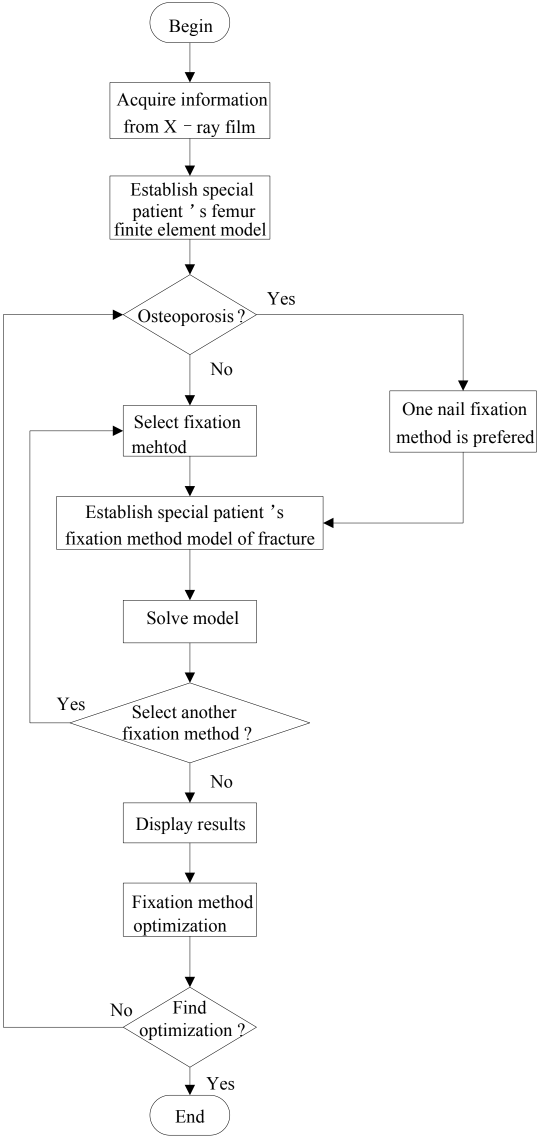

Begin
Acquire information from X－ray film
Establish special patient ' s femur finite element model
Osteoporosis ?
Yes
No
One nail fixation method is prefered
Select fixation mehtod
Establish special patient 's fixation method model of fracture
Solve model
Yes
Select another fixation method ?
No
Display results
Fixation method optimization
No
Find optimization ?
Yes
End

Figure 5 The flow chat of the optimal system

F. Optimal Program of Internal Fixation Method for Femoral Neck Fracture

Modularized programming ideas and object-oriented programming language are used in the system. Ever problem to be solved is considered as a module [9]. The whole system is divided into 3 modules which are parameterized preprocessing module, the solving module and post-processing module.

Parameterized preprocessing module. The main function of this module is summarized as build parameterized models of femur and internal fixation models. Though building the two cases of finite element models has different APDL frame codes, the way of getting parameters and the way of transferring are the same.

The method of getting the parameters and transferring is running through the whole system which is used not only in preprocessing to build finite element models but also in calculating the forces on femur. This method realizes variable parameterized design that means assigning the relevant parameters in APDL frame codes through the visualization interfaces.

MFC cannot select file as save path, so this drawback is worked out in this system first. Class CPathDialog and CPathDialogSub are friends and derived from class CWnd. Expect the constructed function, 9 functions are defined either. The classes mainly include callback function BrowseCallbackProc inside which a SWITCH statements are used to judge the situations and handle. 2 situations are considered: the first is initialization, the other is when the save path changes. The function MakeSurePathExists checks if the path exists. The function IsFileNameValid is to check if the path is valid, and the functions of Touch, ConcatPath which are able to connect the save path.

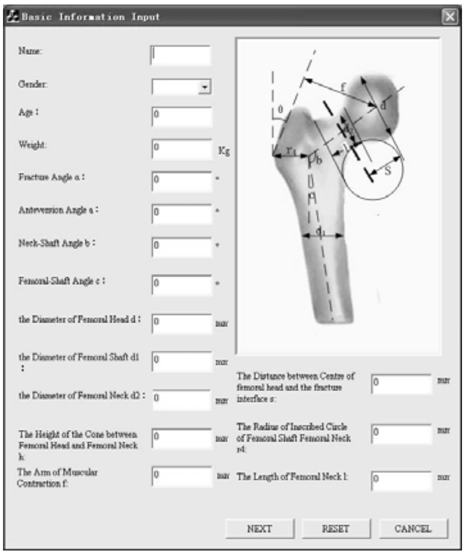
Define a function named OnApdlfile in CDialog class. Create a txt file inside OnApdlfile function by strPath+=input1.txt which is used to store the complete codes of APDL. Moreover, realize transfer the parameters. 11 pieces of information are need to model. So, 11 variables of CEdit, int and char types are defined, then call the member function GetWindowText to cast (atoi). Read and write the APDL frame codes to the txt file which already defined by the class fstream, the class wfstream. User-defined function void OnRunansys is the crucial function to solve models by finite element method inside which create a thread by AfxBeginThread (Simulation,this,THREAD_PRIORITY_NORMAL). The first parameter does not only include the definition of AnsysPath, ApdlPath and buffer but also the Macro file Bone_Ansys.mac. The function WinExec (CommandStr,SW_HIDE) is also important which realizes the batch mode is applied by finite element analysis software. The command is CommandStr ="\""+ AnsysPath +"\""+" -b nolist-i "+MacFile+" -o output.txt". This function connects the path of finite element analysis software and the path of APDL file and saves the results solved by finite element method to the output.txt.

Moreover, the parameters need in operative plan are the number of tightening screw, combinatorial ways which cannot distinguish by APDL. In order to make this happen, define the number of screw and the combinatorial ways as Macro in the system for APDL to distinguish. The interfaces of information and parameter acquisition and the selection of internal fixation pattern are shown in Fig.6. 2 vertical screws setting software implementation

```
    *IF，VAL4，EQ，20010
    CYL4，0，0，0，0，6，360，m
    Wpoff，0，0，6
CYL4，0，0，0，0，6，360，m！
*ELSEIF…
```

3 screws of positive triangle setting software implementation

```
*IF，VAL7，EQ，30010
Wpoff，0，5，0
CYL4，0，0，0，0，6，360，m
Wpoff…
```


Basic Information Input

Name:

Gender:

Age： 0

Weight: 0 Kg

Fracture Angle α： 0 °

Anteversion Angle a： 0 °

Neck-Shaft Angle b： 0 °

Femoral-Shaft Angle c： 0 °

the Diameter of Femoral Head d： 0 mm

the Diameter of Femoral Shaft d1 ： 0 mm

the Diameter of Femoral Neck d2： 0 mm

The Height of the Cone between Femoral Head and Femoral Neck h: 0 mm

The Arm of Muscular Contraction f: 0 mm

The Distance between Centre of femoral head and the fracture interface s: 0 mm

The Radius of Inscribed Circle of Femoral Shaft Femoral Neck r4: 0 mm

The Length of Femoral Neck l: 0 mm

Figure 6 The interface of information acquisition. In this figure, we can see fracture angle(α), anteversion angle(a), neck-shaft angle(b), femoral shaft angle(c), the diameter of femoral head(d), the diameter of femoral shaft(d1), the diameter of femoral neck(d2), the height of the cone between femoral head and femoral neck(h), the arm of muscular contraction(f).

The solving module. This module is applied to calculate the load on the femoral head when standing on one leg under normal circumstance. In fact, femur is subjected complicated stress. Through many researchers have simulated the stress of muscles around femur in vitro, it’s far too difficult to act load in every muscle. Taylor[10] holds the opinion that there is much nondeterminacy such as the selected quantity, the direction of load and so on, especially dynamic loading. Therefore they propose to analyze the load of leading role. The simplified mechanical model which is presented by Scige is shown in Fig.1. The primary external loads are joint reaction force J, Iliotibial tract force R, abductor force M which stress distributions are generally the same as femur interface.

J,M,R have certain relationship with body weight when standing on one leg. The gravity line of body locates the rear of public symphysis. Aim to maintain balance on one leg, the reacting force of ground is W which represents body weight, the 1/6 of W is borne by leg, and the left part bears 5/6W.

When calculating, divide the hip joint into upper separation and lower separation. 3 formulas of moment and force balances are need, and schematic of 3 forces are shown in Fig.7.

$$\frac{5}{6}W\times b-M\times f+R\times d=0 \tag{1}$$

There into:

b=2f; $d = f \times \sin\theta$

θ is 29.5°,so according to Fig.8:

$$M_X = M \times \sin\theta$$
$$M_Y = M \times \cos\theta \qquad (2)$$

According to the lower separation, the balance formula:

$$M_X - J_X = 0 \qquad (3)$$

$$M_Y - J_Y - \frac{1}{6}W + W - R = 0 \qquad (4)$$

ϕ is 22.4°,so, the force J is shown in Fig.9:

$$J_X = J \times \sin\phi$$
$$J_Y = J \times \cos\phi \qquad (5)$$

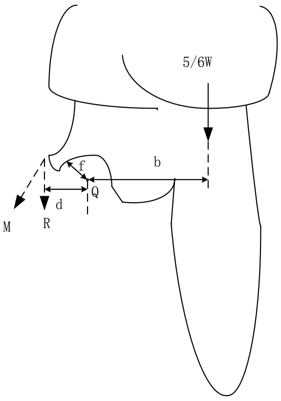


Figure 7 Force diagram of upper separation. When calculating, divide the hip joint into upper separation and lower separation. 3 formulas of moment and force balances are need. The parameters in the 3 formulas are shown in this figure. R means iliotibial tract muscle force. M means abductor force. Q means a central point of a femoral head.

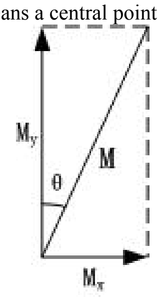


Figure 8 Block diagram of abductor force M

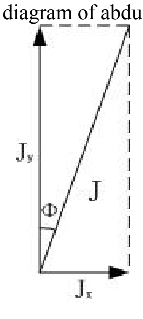


Figure 9 Block diagram of the joint reaction force J

The values of J, R, M are got by (1)~(5) which are calculated by MATLAB.

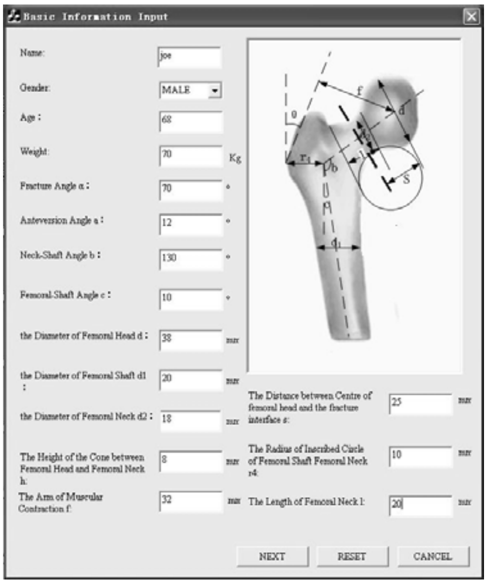
In order to show the essential information of fixed model and get W and f which are need to calculate J, R, M; the third interface should acquire parameters from the first interface which have already input. But MFC only can realize transmit the parameters between adjacent dialogs, therefore, the technique we use is set 9 invisible Static Texts to collect the parameters from the first interface and then transfer to the third interface.

Post-processing module. This module includes the submodule of showing finite element biomechanical model solving results and the submodule of result optimization. The solving results can be shown in real-time. The results output into an ouput.txt document by ANSYS batching mode in function winexec. When the content of flag.txt is 1, the ANSYS is running. The content which is saved in output.txt can be read and kept in buffer by the function Open of class CFile. Then the data in buffer are shown in edit box of which ID is IDC_STATUS through function Read. Then, according to the judgment criteria, US and DS are considered to optimize.

III. Results

Consider a female patient of intertrochanteric fractures in 70° and measure the anatomical parameters by X-ray film. Then input them into the interface (fig.10) and create a femur finite element model. Select an internal fixation method. We select the two horizontal screws with 40°and create an internal fixation model. The system acquires the information from the interface automatically and calculates the load which applied on femur. Next, the system gets the results computed and solves by finite element software. The analysis results show on the interface (fig.11) in real-time. Analyze the internal fixation methods of one screw with 40°and upside-down equilateral triangles and show all the result on the interface(fig.12 and fig.13). Finally, sort the result and find out the most reasonable internal fixation method (fig.14).

The same as the results provided by Zhang Meng[8], under the conditions of intertrochanteric fractures in 70°,we can see that without considering bone substance, materials and complications, only from the displacement changing trends of femoral head and fracture interfaces, the internal fixation method with two erect screws is the most reasonable one. Though people always hold the opinion that the number of fixed screws more, the structure is more stable. All that results is not the same. Meanwhile the result got by the system is proved to be correct and reliable.



Basic Information Input

Name: joe

Gender: MALE

Age： 68

Weight: 70 Kg

Fracture Angle α： 70 °

Anteversion Angle a： 12 °

Neck-Shaft Angle b： 130 °

Femoral-Shaft Angle c： 10 °

the Diameter of Femoral Head d： 38 mm

the Diameter of Femoral Shaft d1 ： 20 mm

the Diameter of Femoral Neck d2： 18 mm

The Height of the Cone between Femoral Head and Femoral Neck h: 8 mm

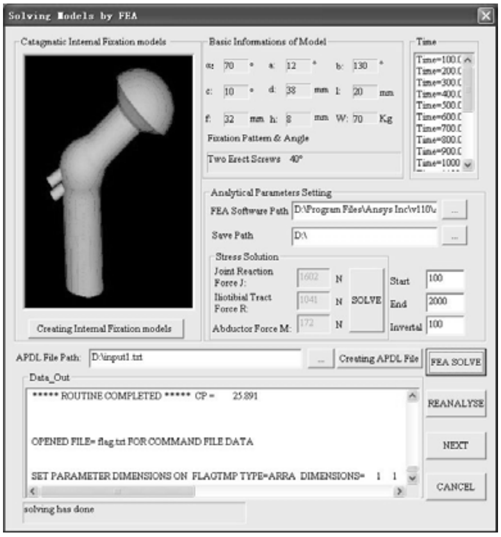
The Arm of Muscular Contraction f: 32 mm

The Distance between Centre of femoral head and the fracture interface s: 25 mm

The Radius of Inscribed Circle of Femoral Shaft Femoral Neck r4: 10 mm

The Length of Femoral Neck l: 20 mm

Figure 10 Acquire the patient’s information



Solving Models by FEA

Catagmatic Internal Fixation models

Basic Informations of Model

α: 70 ° a: 12 ° b: 130 °

c: 10 ° d: 38 mm l: 20 mm

f: 32 mm h: 8 mm W: 70 Kg

Fixation Pattern & Angle

Two Erect Screws 40°

Time

Time=100.0
Time=200.0
Time=300.0
Time=400.0
Time=500.0
Time=600.0
Time=700.0
Time=800.0
Time=900.0
Time=1000

Analytical Parameters Setting

FEA Software Path D:\Program Files\Ansys Inc\v110\

Save Path D:\

Stress Solution

Joint Reaction Force J: 1602 N

Iliotibial Tract Force R: 1041 N

Abductor Force M: 172 N

Start 100

End 2000

Inverval 100

APDL File Path: D:\input1.txt

Data_Out

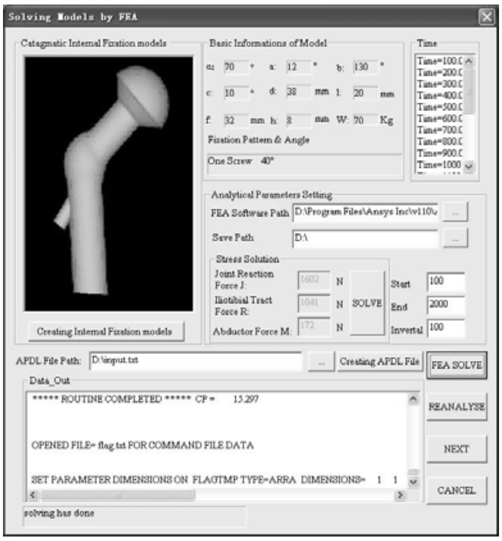
```
***** ROUTINE COMPLETED ***** CP =      25.891


OPENED FILE= flag.txt FOR COMMAND FILE DATA


SET PARAMETER DIMENSIONS ON  FLAGTMP TYPE=ARRA  DIMENSIONS=   1   1
```

solving has done

Figure 11 Calculating the load and solving the two model of two erect screws in 40°



Solving Models by FEA

Catagmatic Internal Fixation models

Basic Informations of Model

α: 70 ° a: 12 ° b: 130 °

c: 10 ° d: 38 mm l: 20 mm

f: 32 mm h: 8 mm W: 70 Kg

Fixation Pattern & Angle

One Screw 40°

Time

Time=100.0
Time=200.0
Time=300.0
Time=400.0
Time=500.0
Time=600.0
Time=700.0
Time=800.0
Time=900.0
Time=1000

Analytical Parameters Setting

FEA Software Path D:\Program Files\Ansys Inc\v110\

Save Path D:\

Stress Solution

Joint Reaction Force J: 1602 N

Iliotibial Tract Force R: 1041 N

Abductor Force M: 172 N

Start 100

End 2000

Inverval 100

APDL File Path: D:\input.txt

Data_Out

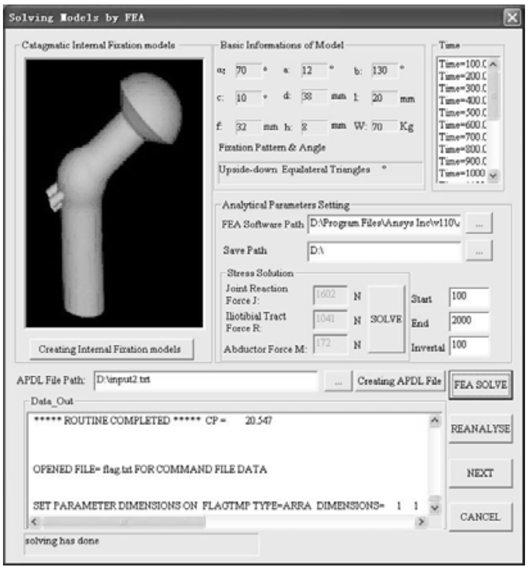
```
***** ROUTINE COMPLETED ***** CP =      15.297


OPENED FILE= flag.txt FOR COMMAND FILE DATA


SET PARAMETER DIMENSIONS ON  FLAGTMP TYPE=ARRA  DIMENSIONS=   1   1
```

solving has done

Figure 12 Calculating J、M、R and solving the FE model of one screw in 40° . J means joint reaction force. R means iliotibial tract muscle force. M means abductor force.



Solving Models by FEA

Catagmatic Internal Fixation models

Basic Informations of Model

α: 70 ° a: 12 ° b: 130 °

c: 10 ° d: 38 mm l: 20 mm

f: 32 mm h: 8 mm W: 70 Kg

Fixation Pattern & Angle

Upside-down Equilateral Triangles °

Time

Time=100.0
Time=200.0
Time=300.0
Time=400.0
Time=500.0
Time=600.0
Time=700.0
Time=800.0
Time=900.0
Time=1000

Analytical Parameters Setting

FEA Software Path D:\Program Files\Ansys Inc\v110\

Save Path D:\

Stress Solution

Joint Reaction Force J: 1602 N

Iliotibial Tract Force R: 1041 N

Abductor Force M: 172 N

Start 100

End 2000

Inverval 100

APDL File Path: D:\input2.txt

Data_Out

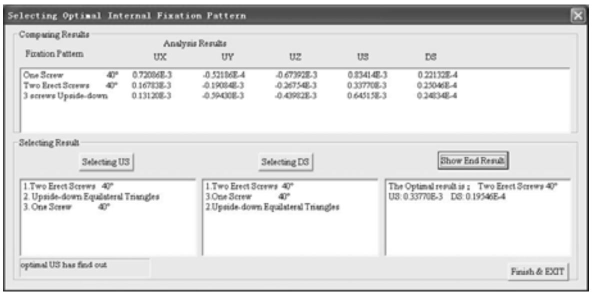
```
***** ROUTINE COMPLETED ***** CP =      20.547


OPENED FILE= flag.txt FOR COMMAND FILE DATA


SET PARAMETER DIMENSIONS ON  FLAGTMP TYPE=ARRA  DIMENSIONS=   1   1
```

solving has done

Figure 13 Calculating the load and solving the FE of model with upside-down equilateral triangles



Selecting Optimal Internal Fixation Pattern

Comparing Results

Analysis Results

Fixation Pattern		UX	UY	UZ	US	DS
One Screw	40°	0.72086E-3	-0.52186E-4	-0.67392E-3	0.83414E-3	0.22132E-4
Two Erect Screws	40°	0.16783E-3	-0.19084E-3	-0.26754E-3	0.33770E-3	0.25046E-4
3 screws Upside-down		0.13120E-3	-0.59430E-3	-0.43982E-3	0.64515E-3	0.24834E-4

Selecting Result

1.Two Erect Screws 40°
2. Upside-down Equilateral Triangles
3. One Screw 40°

1.Two Erect Screws 40°
3.One Screw 40°
2.Upside-down Equilateral Triangles

The Optimal result is： Two Erect Screws 40°
US: 0.33770E-3 DS: 0.19546E-4

optimal US has find out

Figure 14 Results optimization. The same as the results provided by Zhang Meng[8], under the conditions of intertrochanteric fractures in 70°,we can see that without considering bone substance, materials and complications, simply from the displacement changing trends of femoral head and fracture interfaces, the internal fixation method with two erect screws is the most reasonable one.

IV. DISSCUSSION

The simulation and optimal selection results show the internal fixation method with two erect screws is the most reasonable. Though people always hold the opinion that the number of fixed screws more, the structure is more stable. All that results is not the same. The result got by the system is proved to be correct and reliable.

Combing finite element method with the objected-oriented programming language, an optimal selection system of internal fixation method for femoral neck fracture is developed in this paper. According different fracture situations, the system can select the most proper method of the number of fixed screw, fixed angle and combinatorial ways intelligently and reasonably which has guiding significant for the treatment of femoral neck fracture in clinic. Meanwhile it provides an innovative and convenient method to help doctor to decide the better fixation way. From the research, the following conclusions are obtained:

The relationship between bone fracture and mechanics are studied, the parameters which are adopted to evaluate fixed effect are presented. The parameters include displacement of femoral head and the maximal interfragmentary movement. According to the biomedical experiments summarized relation, propose evaluation criteria and compile into evaluation algorithm by object-oriented programming language to optimize the internal fixation method.

A reasonable simplified femur model is presented by analyzing a precise model of femur standing on one leg which makes the modeling process can be realized by finite element parameterized language. The size of the femur model and the methods of internal fixation can change as need through interface input parameters. Then all the steps including preprocessor, solving and post-processing are realized in our optimal selection system software.

Realize the interactive communication between APDL and object-oriented programming language. In the same software, doctor can complete optimal selection based bone biomechanical analysis results.

The optimal selection system of internal fixation methods for femoral neck fracture is an executable program which compiled by C++ programming language in Visual C++6.0. Ran the optimal system software, a computer with more than memory 512MB is necessary and the assistant software ANSYS have to be installed in the computer. The running time of the optimal system is 3 minutes at least.

In addition, only testing the rationality and reliability of the system, we don't focus on the anisotropy of bone or dynamic loading. Moreover, the influence of rub, contact and bone substances also should be taken into account. In order to make our optimal selection system more perfect, we will do further research in real-time computer and accurate model in the future.

ACKNOWLEDGMENT

This work was supported by the NSFC under grant number 61272387, NCET under grant number NCET-13-0756, and Distinguished Young Scientists Funds of the Heilongjiang Province under grant number JC201302.

REFERENCES

[1] Yonghua Liao, Femoral neck fracture, First Edition, Superstar Digital Library.

[2] Walker E, Mukherjee DP, Ogden AL, Sadasivan KK, Albright JA, et al. A biomechanical study of simulated femoral neck fracture fixation by cannulated screws: Effects of placement angle and numbers. Am J orthop.2007, 36(12):680-681.

[3] Tan V, Wong KL, Born CT, Harten R, DeLong WG Jr, et al. Two-screw femoral neck fracture fixation: A biomechanical analysis of 2 different configurations. Am J orthop. 2007, 36(9): 481-512.

[4] J.Martinez-Reina, J.M.Garcia-Aznar. A bone remodeling model including the directional activity of BMUS. Biomech Model Mechanobiol, 2009, 8:111-127.

[5] Zohar Yosibash, Royi Padan, Leo Joskowicz and Charles Milgrom. A CT-based high-order finite element analysis of the human proximal femur compared to In-vitro experiments. J. Biomech. Eng., 2007, 129: 297-309.

[6] Scireg A., Arviker R J, The prediction of muscular load joint forces in the lower extremityes dwring. J. biometh, 1975, 8:123-127.

[7] APDL user guide. U.S.A:ANSYS INC, 2001.

[8] Meng Zhang. Optimazation of fixation for femoral neck fracture by finite element analysis. Harbin University of Science and Technology, 2011:30-30.

[9] Bruce Eckel. Thinking in C++. Second Edition. Prentice Hall, 2000: 210-853.

[10] Taylor ME, Tanner KE, Freeman MAR, et al. Stress and strain distribution within the intact femur: compression or bending. Med Eng Phy, 1996, 18:122-131.

[11] The MathWorks Inc.. MATLAB Compiler User's Guide. Revised for Version 4.8[EB/OL], 2008.

[12] Kevin C, Booth MD, Thomas K, et al. Femoral neck fracture fixation: a biomechanical study of two cannulated screws placement techniques. Orthopedics, 1998, 21: 1173-1176.

[13] Wirtz D. C., Schiffers N., Pandorf T., et al. Critical Evaluation of Known Bone Material Properties to Realize Anisotropic FE-Simulation of the Proximal Femur. Biomech, 2009, 33: 1325-1330.

[14] David Kruglinski, George Shepherd. Programming Micosoft Visual C++. Fifth Edition. Micosoft Press, 1998: 12-15.

[15] Sven Herrmann, Michael Kaehler, Robert Souffrant, Roman Rachholz et al. HiL simulation in biomechanics: A new approach for testing total joint replacements. Computer Methods and Programs in Biomedicine, 2012, 105(2): 109-119.

M.N. Wang received the M.S. degree in mechanical engineering from Harbin University of Science and Technology, Harbin, China, in 2000. She received the Ph.D. degree in mechatronics engineering from Harbin Engineering University, Harbin, in 2004. His current research interests include computer aided surgery and virtual surgery.

She became a full professor in the Robotics Institute, Harbin University of Science and Technology, China, from July 2007. She was a visiting professor at Scientific Computing Centre, University of ULM, ULM, Germany, between May 2010 and May 2011.

Using Game Theory to Analyze Strategic Choices of Service Providers and Service Requesters

Yi Sun
College of Computer Science and Technology, Nanjing University
of Aeronautics and Astronautics, Nanjing, 210016, China
Email: sunyiyilily@126.com

Zhiqiu Huang and Changbo Ke
College of Computer Science and Technology, Nanjing University
of Aeronautics and Astronautics, Nanjing, 210016, China
Email: zqhuang@nuaa.edu.cn

Abstract—**With the rapid development and popularization of Web services, online privacy has become a rising concern. In order to prevent service providers from abusing private information of service requesters, service providers are required to publish their privacy policies. If a service provider's privacy policy is not consistent with a service requester's privacy preference, the provider and the requester can negotiate. Whether the negotiation is successful or not depends on the strategic choices of the provider and the requester. This paper analyzes strategic choices of service providers and service requesters in negotiations by game theory. We propose a game theoretic model for negotiations between providers and requesters, discussing providers' and requesters' strategic choices in one-time negotiation and repeated negotiation. In addition, a case study is given to demonstrate how to use our game theoretic model to analyze the strategic choices of providers and requesters in negotiations.**

Index Terms—**Negotiations, Game Theory, Privacy Policies, Privacy Preferences**

I. Introduction

Nowadays, increasing people are accustomed to using Web services. When they enjoy the convenient and efficient services on the Internet, their private information is inevitably collected by service providers. As service requesters, it is difficult for them to control the way their private information is used by service providers. To avoid the misuse of service requesters' private information, service providers are asked to publish their privacy policies. Privacy policies specify how service providers will deal with the private information, which they get from service requesters. Service requesters also design their privacy preferences which state the way in which the private information will be handled. P3P privacy policy is one of the privacy policies widely used in the world [1]. P3P (The Platform for Privacy Preferences) [2] was released by World Wide Web Consortium (W3C) in April 2002. It provides a standard and machine-readable privacy policy. W3C also designs APPEL (A P3P Preference Exchange Language) [3] which allows service requesters to specify their privacy preferences.

The consistency of privacy policies and privacy preferences means that service providers will handle private information according to service requesters' requirements. In this case, requesters will use the services offered by providers. Conversely, if privacy policies are not consistent with privacy preferences, requesters will refuse to use the services so as to protect their privacy. Otherwise, providers and requesters will negotiate. The success of the negotiations relies on providers' and requesters' strategic choices. Providers have the choice to alter their privacy policies to be stricter or not alter their privacy policies. Requesters have the choice to modify their privacy preferences to be more moderate or not to modify their privacy preferences. If providers don't alter policies and requesters don't modify preferences when they are negotiating, the negotiations fail. Researchers have studied the negotiations between service providers and service requesters. Kheira Bekara and Maryline Laurent [4] proposed a negotiation mechanism, which supported automated negotiations on the inconsistencies between P3P privacy policies and privacy preferences. During the negotiations, privacy policies were changed from moderate to strict by providers, and privacy preferences were changed from strict to moderate by requesters. Salah-Eddine Tbahriti et al. [5] put forward a negotiation model to reconcile requesters' requirements with providers' policies in case of incompatibility. In their model, providers did not change their policies but offered incentives to requesters. If requesters accepted the incentives, they would modify their requirements to be compatible with providers' policies. Ke Changbo et al. [6] proposed a cloud computing oriented negotiation mechanism, in which service requesters didn't alter their privacy preferences and service providers changed their privacy policies to satisfy service requesters.

In this paper, we formulate negotiations between service providers and service requesters as a game theoretic model to analyze their strategic choices during negotiations. Both providers and requesters have two strategic choices in our game model. Providers may or may not alter their privacy policies to be stricter. Requesters may or may not modify their privacy

doi:10.4304/jsw.9.11.2918-2924

preferences to be more moderate. Based on payoffs of different strategic choices, we can infer which strategies providers and requesters will choose.

This paper is structured as follows: In Section II we introduce related works. In Section III we give an overview of game theory. In Section IV we propose a game theoretic model for negotiations between service providers and service requesters and calculate the mixed strategy Nash equilibrium of the game model. In Section V we discuss the results of one-time negotiation and repeated negotiation. In Section VI we illustrate how to analyze providers' and requesters' strategic choices by our game model with a case study. In the end, we conclude and point out the future works in Section VII.

II. Related Work

If service providers' privacy policies aren't consistent with service requesters' privacy preferences, service providers and service requesters may negotiate on the inconsistencies. To address the limitation that P3P lacks a negotiation mechanism, Kheira Bekara and Maryline Laurent [4] proposed an automated negotiation mechanism which supported negotiations on the inconsistencies between P3P privacy policies and privacy preferences. During negotiations, privacy policies were changed from moderate to strict by providers, and privacy preferences were changed from strict to moderate by requesters. Salah-Eddine Tbahriti et al. [5] put forward a negotiation model to reconcile requesters' requirements with providers' policies in case of incompatibility. In their model, providers didn't change their policies but offered incentives to requesters. Once requesters accepted the incentives, they would modify their requirements so that it could be compatible with providers' policies. Ke Changbo et al. [6] proposed a cloud computing oriented negotiation mechanism. In the mechanism, users didn't modify their privacy preferences. Privacy policies satisfying both users and service composers were gained through exchanging privacy disclosure assertion.

Although game theory is initially developed in economics, it has been used to solve problems in many fields. Lisa Rajbhandari and Einar Arthur Snekkenes [7] used a game theory based approach to perform privacy-related risk analysis. In their game model, there were two decision makers which were an online bookstore and a user. The user had the choice to provide his genuine or fake personal information to the online bookstore. The online bookstore chose to exploit the personal information of the user by selling it to third parties or didn't exploit and used it for its own internal purpose. The probabilities and outcomes to determine the level of privacy risk could be computed by obtaining the benefits from the online bookstore's and the user's strategic choices. Spyros Kokolakis et al. [8] analyzed buyers' and sellers' privacy-related strategic choices in e-commerce transactions through game theory. They explained why buyers mistrusted privacy policies and discussed possible remedies to protect privacy. Shouke Wei et al. [9] applied game theory based model to analyze and solve water conflicts concerning water allocation and nitrogen reduction in the Middle Route of the South-to-North Water Transfer Project in China. They constructed a game model including a main game and four sub-games and used statistical and econometric regression methods to formulate payoff functions of players. Justin Zhan et al. [10] proposed an approach to generate gaming strategies for the attacker and defender in a recommender system. To analyze vulnerabilities and security measures incorporated in a recommender system, they defined attack graphs, use cases, and misuse cases in their gaming framework. Wu Jiang et al. [11] analyzed gaming behaviors of graduates in scholarship competition in China by using an evolutionary game theory approach. They found that graduate individual would adopt different strategies based on different number of graduate groups, symmetry information and asymmetry information. Zhang Cheng et al. [12] focused on cheating operations of nodes happened in Opportunistic Network. In order to improve the possibility of successful message transmission and reduce the probability of cheating, they converted the phase game of nodes into repeated-game and introduced the credit mechanism and punishment mechanism into the game. Mohamed Amine M'hamdi et al. [13] proposed a scheduling algorithm to help the controller agent improve the quality of the reputation mechanism. The algorithm was based on a class of games called Bayesian Stackelberg. Sun Weifeng et al. [14] proposed a game theoretic resource allocation model in grid computing. The model guaranteed higher tasks' victorious probabilities in grid resources scheduling situations. Wu Guowei et al. [15] put forward a game theoretic energy-aware scheduling algorithm for multi-core systems. The algorithm could reduce the temperature difference between different groups of cores, which effectively avoided the local hotspot of a processor.

III. Overview Of Game Theory

Game theory was theorized by John Von Neumann and Oskar Morgenstern in 1944 [16]. Nash proved the existence of Nash equilibrium in 1950, which laid the foundation for the generalization of game theory [17][18]. Today, game theory has been widely used in many fields, such as politics, economy and biology [19]. It analyzes the behaviors of decision makers in interactions. Compared with classical analytical methods, it doesn't have to rely on subjective probabilities or accurate data. Probabilities and outcomes can be computed according to the preferences or benefits of decision makers.

A game generally comprises four parts: the players, their strategies, payoffs and the information they have [20]. According to different criteria, the game can be categorized into a static/dynamic game and a complete/incomplete game. A static game is one in which players choose simultaneously or not choose simultaneously but players who choose secondly not knowing the actions of players who choose firstly (vice versa for the dynamic game). A complete information game is one in which players know about the strategies and payoffs of others (vice versa for the incomplete information game).

If a player's chosen strategy is the best response to the strategies of others, it is the optimal strategy for the player. A Nash equilibrium is a strategy profile consisting of each player's optimal strategy. Players in the game don't have enough reasons to break the equilibrium [21]. A mixed strategy is a probability distribution of the players' pure strategies. The expected payoffs of players in a game can be calculated based on the mixed strategy Nash equilibrium.

IV. Game Theoretic Model For Negotiations

In this section, we will formulate negotiations between service providers and service requesters as a game theoretic model and calculate the mixed strategy Nash equilibrium of the model.

A. Game Formulation

We construct a game theoretic model for negotiations to analyze service providers' and service requesters' strategic choices. To formulate the game, we assume that it is a static and complete information game. The game is static as we stipulate that the provider and the requester choose their strategies simultaneously. It is of complete information as we assume that both the provider and the requester know about the strategies and outcomes of each other. In the following, we will explain the strategies of the players and how the data are collected to estimate the payoffs.

Players: It is a two-player game between the service provider and the service requester. We assume that both the provider and the requester are rational. They have the incentive to optimize their payoffs.

Strategies: Both the service provider and the service requester have two strategic choices. The service provider has the choice to alter his privacy policy to be stricter or keep his privacy policy unchanged. The strategies of the service provider are given by {Alter, NotAlter}. The service requester has the choice to modify his privacy preference to be more moderate or keep his privacy preference unchanged. The strategies of the service requester are given by {Modify, NotModify}.

Payoff:

1.Data Collection:

We need to collect data to estimate the payoffs of the provider and the requester. Table I lists the data related to the service provider.

Table I
Data Related To The Service Provider

	For Service Provider	
Strategic Choices	Alter	NotAlter
Modify	a, b	a
NotModify	a, c	d

For the service provider: In the case where the service requester modifies his preference to be more moderate, if the service provider alters his policy to be stricter, the negotiation is successful. The service provider benefits from providing services to the requester (a) and pays additional cost to implement the stricter policy (b). If the service provider keeps his policy unchanged, the negotiation is successful. The service provider benefits from providing services (a) and doesn't have to pay a cost for altering his policy. In the case where the service requester keeps his preference unchanged, if the service provider alters his policy to be stricter, the negotiation is successful. The service provider benefits from providing services to the requester (a) and pays additional cost to implement the stricter policy (c). The value of c is greater than b because the degree to which the policy is altered is bigger. If the service provider keeps his policy unchanged, the negotiation fails. The service provider suffers a loss resulting from not providing services for the requester (d).

Table II lists the data related to the service requester.

Table II.
Data Related To The Service Requester

	For Service Requester	
Strategic Choices	Modify	NotModify
Alter	e, f	e
NotAlter	e, g	h

For the service requester: In the case where the service provider alters his policy to be stricter, if the service requester modifies his preference to be more moderate, the negotiation succeeds. The service requester benefits from using services offered by the service provider (e) and suffers a loss resulting from reducing demands for privacy protection (f). If the service requester keeps his preference unchanged, the negotiation succeeds. The service requester benefits from using services offered by the provider (e) and doesn't have to suffer a loss resulting from modifying his preference. In the case where the service provider keeps his policy unchanged, if the service requester modifies his preference to be more moderate, the negotiation succeeds. The service requester benefits from using services offered by the provider (e) and suffers a loss resulting from reducing demands for privacy protection (g). The value of g is greater than f because the degree to which the preference is modified is bigger. If the service requester keeps his preference unchanged, the negotiation fails. The service requester suffers a loss resulting from refusing to use services provided by the provider (h).

2.Estimation:

The negotiation between the service provider and the service requester is shown in Figure 1. The provider may or may not alter his privacy policies to be stricter. The strategies of him are given by {Alter, NotAlter}. The requester may or may not modify his privacy preferences to be more moderate. The strategies of him are given by {Modify, NotModify}. Therefore, there are four strategy profiles which we will present in the following paragraphs. SP and SR denote respectively the provider's payoff and the requester's payoff.

{Alter, Modify} means the provider alters his policy to be stricter, while the requester modifies his preference to be more moderate. The provider benefits from providing services to the requester (a) and pays additional cost to implement the stricter policy (b). Thus, the provider's payoff is a-b. The requester benefits from using services offered by the provider (e) and suffers a loss resulting from reducing demands for privacy protection (f). His payoff is e-f.

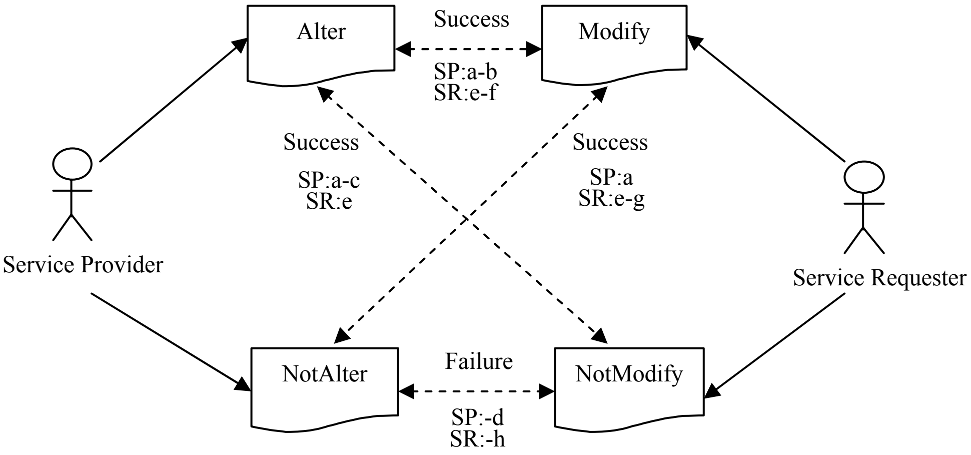


Figure 1. The negotiation between the provider and the requester

{Alter, NotModify} means the provider alters his policy while the requester keeps his preference unchanged. The provider benefits from providing services (a) and pays additional cost to implement the stricter policy (c). Thus, the provider's payoff is a-c. The requester benefits from using services (e) and doesn't have to suffer a loss resulting from modifying his preference. Thus, the requester's payoff is e.

{NotAlter, Modify} means the provider keeps his policy unchanged while the requester modifies his preference. The provider benefits from providing services (a) and doesn't have to pay a cost for altering his policy. Thus, his payoff is a. The requester benefits from using services (e) and suffers a loss resulting from reducing demands for privacy protection (g). His payoff is e-g.

{NotAlter, NotModify} means the provider doesn't alter his policy while the requester doesn't modify his preference. The provider suffers a loss resulting from not providing services (d). His payoff is -d. The requester suffers a loss resulting from refusing to use services (h). The requester's payoff is -h.

B. Game Solution

TABLE III.
PAYOFFS OF THE PROVIDER AND THE REQUESTER

		Service Provider	
		Alter	NotAlter
Service Requester	Modify	e-f, a-b	e-g, a
	NotModify	e, a-c	-h, -d

The payoffs of the provider and the requester are shown in Table III. The first value of each cell is the payoff of the service requester while the second value is the payoff of the service provider. Using the payoffs in Table III, we can find the mixed strategy Nash equilibrium in the negotiation. We assume that the provider plays the strategies Alter and NotAlter with probabilities p and 1-p respectively, and the requester plays the strategies Modify and NotModify with probabilities q and 1-q respectively. Mixed strategy Nash equilibrium makes players indifferent between their pure strategies. So, we set the expected payoffs of players' pure strategies equal to solve the mixed strategy Nash equilibrium.

When calculating p, we assume the expected payoff the requester gains by modifying his preference is E_{R1}, and the expected payoff the requester gains by keeping his preference unchanged is E_{R2}.

$$E_{R1}=(e-f)p+(e-g)(1-p)$$

$$E_{R2}=ep+(-h)(1-p)$$

In the mixed strategy Nash equilibrium, the provider's strategic choices make E_{R1} equal to E_{R2}.

$$E_{R1}=E_{R2}$$

$$(e-f)p+(e-g)(1-p)=ep+(-h)(1-p)$$

$$p=\frac{g-h-e}{g-h-e-f} \tag{1}$$

When calculating q, we assume the expected payoff the provider gains by altering his policy is E_{P1}, and the expected payoff the provider gains by keeping his policy unchanged is E_{P2}.

$$E_{P1}=(a-b)q+(a-c)(1-q)$$

$$E_{P2}=aq+(-d)(1-q)$$

In the mixed strategy Nash equilibrium, the strategic choices of the requester make E_{P1} equal to E_{P2}.

$$E_{P1}=E_{P2}$$

$$(a-b)q+(a-c)(1-q)=aq+(-d)(1-q)$$

$$q=\frac{c-a-d}{c-a-d-b} \tag{2}$$

After calculation, we get p and q. Hence, we can know the probabilities with which the provider and the requester will choose a particular strategy.

V. Results Of Negotiations

If the service provider and the service requester only negotiate once, the negotiation is a one-time game. Therefore, the provider and the requester will choose their optimal strategies based on the mixed strategy Nash equilibrium. However, if the requester asks for services offered by the provider more than once, the provider and the requester negotiate many times. We assume that changes in policies and preferences in the negotiation are not saved by the provider and the requester. These changes are only temporary compromises for the sake that the negotiation succeeds.

In the finite case, the requester asks for services several times and then stops. The provider and the requester negotiate finite times. Thus, the negotiation is a finitely repeated game. Using backward induction, we obtain the result of this game. We should first examine the last round. In the last round, the service provider and the service requester will choose their optimal strategies based on the mixed strategy Nash equilibrium. So in the penultimate round, the service provider and the service requester will also choose their optimal strategies. And so on, the service provider and the service requester will also choose their optimal strategies from the beginning. Therefore, the result of the finitely repeated game is the same as the one-time game. The provider and the requester will choose their optimal strategies based on the mixed strategy Nash equilibrium.

In the infinite case, the requester asks for services endlessly. The provider and the requester negotiate infinite times. Thus, the negotiation is an infinitely repeated game. At this point, the provider and the requester may not choose their optimal strategies. Considering their long-term interests, the provider and the requester will cooperate.

VI. Case Study

Consider the scenario between a user and an online bookstore. The user asks for services provided by the online bookstore. Because of inconsistencies between the online bookstore's privacy policy and the user's privacy preference, the online bookstore and the user negotiate with each other. The data related to the online bookstore are shown in Table IV and the data related to the user are shown in Table V.

Table IV.
Data Related To The Online Bookstore

	For online bookstore	
Strategic Choices	Alter	NotAlter
Modify	1, 1.2	1
NotModify	1, 1.7	1

Table V.
Data Related To The User

	For user	
Strategic Choices	Modify	NotModify
Alter	2, 1.5	2
NotAlter	2, 2.5	1

If the online bookstore chooses the strategy Alter and the user chooses the strategy Modify, then the online bookstore benefits from providing services to the user (1) and pays additional cost to implement the stricter policy (1.2), and the user benefits from enjoying services offered by the online bookstore (2) and suffers a loss resulting from reducing his demands for privacy protection (1.5). Thus, the online bookstore's payoff is 1-1.2=-0.2. And the user's payoff is 2-1.5=0.5.

If the online bookstore chooses the strategy Alter, while the user chooses the strategy NotModify, then the online bookstore benefits from providing services (1) and pays additional cost to implement the stricter policy (1.7), and the user benefits from using services (2) and doesn't have to suffer a loss resulting from modifying his preference. Thus, the online bookstore's payoff is 1-1.7 =-0.7. And the user's payoff is 2.

If the online bookstore chooses the strategy NotAlter, whilst the user chooses the strategy Modify, then the online bookstore benefits from providing services (1) and doesn't have to pay a cost for altering his policy, and the user benefits from using services (2) and suffers a loss resulting from reducing his demands for privacy protection (2.5). Thus, the online bookstore's payoff is 1. And the user's payoff is 2-2.5=-0.5.

If the online bookstore chooses the strategy NotAlter and the user chooses the strategy NotModify, then the online bookstore suffers a loss resulting from not providing services (1), and the user suffers a loss resulting from refusing to use services (1). Thus, the online bookstore's payoff is -1. And the user's payoff is -1.

The payoffs of the online bookstore and the user are shown in Table VI.

Table VI.
The Payoffs Of The Online Bookstore And The User

Online Bookstore User	Alter	NotAlter
Modify	0.5, -0.2	-0.5, 1
NotModify	2, -0.7	-1, -1

By substituting the data in Table VI into the expression (1) and expression (2), we get p = 25% and q = 20%, which means the online bookstore alters his policy to be stricter with a 0.25 probability and the user modifies his preference to be more moderate with a 0.2 probability. Table VII lists the probabilities with which the online bookstore and the user will choose a particular strategy.

Table VII.
The Probabilities Distributions Of Strategies

		p=25%	1-p=75%
	Online Bookstore User	Alter	NotAlter
q=20%	Modify	0.5, -0.2	-0.5, 1
1-q=80%	NotModify	2, -0.7	-1, -1

According to the data in Table VII, NotAlter is a dominant strategy for the online bookstore, since regardless of the user's choice the online bookstore gets a better payoff. For the user, NotModify is a dominant strategy which makes the user get a better payoff. Therefore, the online bookstore tends to keep his policy unchanged and the user tends to keep his preference unchanged. Probabilities 1-p and 1-q shows that the online bookstore refuses to alter his policy with a 0.75 probability and the user refuses to modify his preference with a 0.8 probability.

If the user asks for services provided by the online bookstore once, the user and the online bookstore will negotiate once. Based on the mixed strategy Nash equilibrium, the online bookstore and the user will choose the strategy profile {NotAlter, NotModify}.

If the user asks for services finite times, the user and the online bookstore will negotiate finite times. The result is the same as only negotiate once. The online bookstore and the user will choose the strategy profile {NotAlter, NotModify}. However, if the user asks for services infinite times, the user and the online bookstore will negotiate infinite times. The online bookstore and the user will not choose the strategy profile {NotAlter, NotModify} because it will make them bear the greatest losses in the long term. Considering their long-term interests, the online bookstore and the user will cooperate and choose the strategy profile {Alter, Modify}.

VII. Conclusions And Future Work

This paper analyzes service providers' and service requesters' strategic choices in negotiations through game theory. We propose a game theoretic model for negotiations between providers and requesters, discuss providers' and requesters' strategic choices in one-time negotiation and repeated negotiation, and demonstrate how to use our game theoretic model to analyze providers' and requesters' strategic choices through a case study.

However, the negotiation between the provider and the requester in this paper is for the whole privacy policy. Such negotiation is coarse grained. In fact, a privacy policy generally consists of many privacy statements. Each statement specifies the handling of a private data. So the negotiation for a privacy statement is fine grained. In future research, we plan to study the negotiation which is for a privacy statement of a particular private data. The study will help the provider and the requester to reach an agreement on the handling of a particular private data.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No. 61272083).

References

[1] L. Dong, Y. Mu, W. Susilo, P. Wang and J. Yan, "A Privacy Policy Framework for Service Aggregation with P3P", The Sixth International Conference on Internet and Web Applications and Services, pp. 171-177, March 2011.

[2] L. Cranor, M. Langheinrichand and M. Marchiori, "The platform for privacy preferences 1.0 (P3P1.0) specification", W3C recommendation, 2002, 16.

[3] L. Cranor, M. Langheinrich and M. Marchiori, "A P3P preference exchange language 1.0 (APPEL1.0)", W3C working draft, 2002, 15.

[4] K. Bekara and M. Laurent, "Privacy Policy Negotiation at User's Side based on P3P Tag Value Classification", The 2011 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, pp. 18-21, 2011.

[5] S. E. Tbahriti, B. Medjahed, Z. Malik, C. Ghedira and M. Mrissa, "How to Preserve Privacy in Services Interaction", Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on. IEEE, 2012, pp. 66-71.

[6] C. Ke, Z. Huang and M. Tang, "Supporting negotiation mechanism privacy authority method in cloud computing", Knowledge-Based Systems, vol. 51, pp. 48-59, 2013.

[7] L. Rajbhandari, E. A. Snekkenes, "Using game theory to analyze risk to privacy: An initial insight", Privacy and Identity Management for Life, Springer Berlin Heidelberg, 2011, pp. 41-51.

[8] S. Kokolakis, A. Kalliopi and M. Karyda, "An analysis of privacy-related strategic choices of buyers and sellers in e-commerce transactions", Informatics (PCI), 2012 16th Panhellenic Conference on. IEEE, 2012, pp. 123-126.

[9] S. Wei, H. Yang, K. Abbaspour, J. Mousavi and A. Gnauck, "Game theory based models to analyze water conflicts in the Middle Route of the South-to-North Water Transfer Project in China", Water research, vol. 44, no. 8, pp. 2499-2516, 2010.

[10] J. Zhan, L. Thomas and V. Pasumarthi, "Using gaming strategies for attacker and defender in recommender systems", Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE, 2011, pp. 80-87.

[11] J. Wu, H. Zhang and T. He, "Analyzing Competing Behaviors for Graduate Scholarship in China: An Evolutionary Game Theory Approach", LISS 2012. Springer Berlin Heidelberg, 2013, pp. 649-653.

[12] C. Zhang, Q. Zhu and Z. Chen, "Credit-based Repeated Game Model Applied in Transfer Decision of Opportunistic Network", Journal of Software (1796217X), vol. 6, no. 9, 2011.

[13] M. A. M'hamdi and J. Bentahar, "Scheduling Reputation Maintenance in Agent-based Communities Using Game Theory", Journal of Software (1796217X), vol. 7, no. 7, 2012.

[14] W. Sun, Q. Xia, Z. Xu, M. Li and Z. Qin, "A Game Theoretic Resource Allocation Model Based on Extended Second Price Sealed Auction in Grid Computing", Journal of Computers, vol.7, no.1, 2012.

[15] G. Wu, Z. Xu, Q. Xia and J. Ren, "An Energy-Aware Multi-Core Scheduler based on Generalized Tit-For-Tat Cooperative Game", Journal of Computers, vol 7, no.1, 2012.

[16] J. Von Neumann and O. Morgenstern, Theory of games and economic behavior. Princeton University Press, Princeton, 1944.

[17] J. F. Nash Jr, "The bargaining problem", Econometric: Journal of the Econometric Society, 1950, pp. 155-162.

[18] J. F. Nash, "Equilibrium points in n-person games", Proceedings of the national academy of sciences, vol. 36, no. 1, pp. 48-49, 1950.

[19] M. J. Osborne, An introduction to game theory. New York: Oxford University Press, 2004.

[20] R. Eric, “Games and Information: An Introduction to Game. Theory”, http://www.ebookee.net/Games-and-Information-An-Introduction-to-Game-Theory_37585.html, 2001.

[21] J. Nash, “Non-cooperative games”, Annals of mathematics, 1951, pp. 286-295.

Yi Sun was born in 1989 and received the B.S. degree in Information Security. Now she is a master candidate at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Jiangsu, China. Her research interests include service-oriented computing and privacy.
E-mail: sunyiyilily@126.com

Zhiqiu Huang was born in 1965. He received his Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics in 1999. Now he is a professor and Ph.D. supervisor at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His research interests include software engineering, formal methods, cloud computing and privacy.
E-mail: zqhuang@nuaa.edu.cn

Changbo Ke was born in 1984. He is a Ph.D candidate of Nanjing University of Aeronautics and Astronautics. His research interests include security and privacy of information system and ontology-based software engineering.
E-mail: kcb1984@163.com

Brain Tumor Segmentation Based on Structuring Element Map Modification and Marker-controlled Watershed Transform

Xiaopeng Wang, Shengyang Wan, Tao Lei
School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, 730070, China
Email: wangxp1969@sina.com

***Abstract*—Brain medical images are generally prone to noise and also fraught with intensity heterogeneity within the tumor. Fuzzy and boundary discontinuity caused by the tumor also adversely affects the accuracy of the tumor segmentation. A method based on morphological structuring element map modification and marker-controlled watershed segmentation is proposed. Firstly, a structuring element map is constructed according to the sum of the weighted variance of the specific regions within morphological gradient image, and each value of the structuring element map represents the size of structuring element (SE). Secondly, the original image is modified by morphological opening-closing, where the size of SE are determined by the structuring element map in the corresponding pixel, such an adaptive image modification can eliminate the noise and small regular details while preserve the larger object contours without less location offsets. Finally, marker-controlled watershed transform is used to complete the tumor segmentation. Experiments show that the method ensures brain tumors are more accurately segmented.**

***Index Terms*—image segmentation, brain tumor, structuring element map, marker-controlled watershed transform**

I. Introduction

Incidence of brain tumors has been on the rise in recent years. According to statistics, brain tumor accounts for about 5% of the human tumor cases and also forms about 60% of children tumor cases. About 20-30% of other malignant tumors also eventually land into intracranial categories. Because of its invasive growth, expansion in intracranial domains, once it occupies a certain space, regardless of its nature being benign or malignant, are bound to make the intracranial pressure suppress brain tissues, leading to injury to the central nervous system, thus endangering patient’s life. Brain tumor’s early detection greatly depends on the accurate diagnosis and subsequently its effective treatment. The accuracy of the result is thus a very important step to improving the disease treatment. MR and CT technologies are widely applied in the diagnosis and analysis of brain tumors. These technologies render brain tumor location information in the form of size and type, and can be used for brain tumor resection surgery and radiation therapy as important information.

Many efforts have been made to segment brain tissue and tumor from MR and CT Images [1-4]. Watershed transform [5, 6] can be used to produce single pixel width and closed contour, etc. and has been widely applied to medical image segmentation [7, 8]. However, the watershed easily leads to over-segmentation [9]. Usually, there are three kinds of schemes to eliminate over-segmentation. The first one is image pre-filtering [10], which uses filters to reduce local minima area before the watershed segmentation. The second one employs the marker-controlled method [11, 19] to limit the segment regions beforehand; the last is the post-processing after watershed, such as region merging [12]. Nowadays, several methods have been proposed for the brain tumor segmentation [1-3, 13-15], Wang [2] introduced a parametric fluid vector flow active contour model to address the issues of limited capture range and use it to implement the brain tumor segmentation; Corso [3, 13] employed multilevel segmentation and integrated Bayesian model classification to separate the brain tumor. Kowar [14] presented a method for the detection of brain tumor using histogram thresholding; Christ [15] applied K-Means clustering integrated with marker-controlled watershed algorithm to segment MR brain images; In fact, the brain tissue structure is more complex, and the boundary of tumor region and normal tissue is not obvious. The CT or MR image itself may contain noise and low-contrast regions, such a case likely cause discontinuities and fuzzy boundaries, and maybe lead to resultant inaccurate segmentation of contours. We thus present a novel method by applying morphological structuring element map to modify the original image, then use the marker-controlled watershed transform to segment the modified image. During the segmentation process, the selection of the appropriate size of SE is the forte of morphological modification. For each pixel of the original image, we apply opening-closing with different SE to modify each pixel, where the size of the SE is determined by the structuring element map.

Manuscript received March 16, 2014; revised May 15, 2014; accepted July 1, 2014.
Corresponding author: Xiaopeng Wang, wangxp1969@sina.com

doi:10.4304/jsw.9.11.2925-2932

II. Morphological Modification

When applying the marker-controlled watershed transformation to segment the image of brain tumors, it can accurately mark the tumor area and produce closed contours, yet usually the tumor can be present in normal brain tissue. In a variety of medical imaging of tumor shape, the gray value of the interior is not uniform; sometimes the tumor and surrounding tissues are very close resulting in a less accurate segmentation. If the image does not undergo pre-filtering or smoothing, a direct use of marker-controlled watershed transform may result in inaccurate target contour positioning.

Morphological opening and closing can eliminate the bright and dark regions less than the SEs. But the problem is that when apply opening and closing with invariant smaller SE to modify the tumor image, some bright and dark regular details can not be eliminated completely. In reverse, lager SE may lead to the tumor contour occur location offset. To this end, we present a method based on morphological opening and closing to modify image, where the SE's size is variant for the different pixel. The modification mainly based on morphology theory [6,17] is such that, every gray image can be seen as a three-dimensional topographical map, and then using different viscous fluids that are in a flooded landscape. When the viscosity is at a low temperature, the fluid can reach more irregular detailed regions, conversely, when the viscosity is at a high temperature, the fluid can only reach the wider region. Closing operation can eliminate dark regions smaller than SE, this would be equivalent to use different size of the SEs to effect modification of the image by a morphological closing. On the other hand, opening can eliminate bright regions smaller than the SE. The combination of the two operations will release the bright and dark small details and noise within the image, and largely reduce the factors that result in over-segmentation. The gradient can reflect the degree of the image's gray change, but sometimes the pixel's gradient does not accurately reflect the topography image information, therefore, we employ the sum of pixel's gradient weighted gradient variance to determine the size of SE that corresponding to viscosity.

A. Structuring Element Map

For ease and exactly of calculation, the normalization processing is carried out. In the following sections, we will use the normalized images everywhere. The morphological gradient is given by

$$g = (f \oplus s) - (f \Theta s) \tag{1}$$

where s is a circular SE with 1 as the radius, g is the gradient image, and f is the original image, $\oplus$ and Θ respectively denote morphological dilation and erosion.

For the aim to modify each pixel using variant SE, we construct a circular structuring element map $M(x,y)$, its size is equal to original image and each pixel value represents the size of the SE which will be used to modify original image in corresponding pixel. Each SE's size in $M(x,y)$ is determined by the morphological gradient image. If the gradient image is seen as a topography map, gradient value would represent the altitude of each point, in general the target contour points corresponds to higher elevations. It is well known that the variance is a measure of how far a set of numbers is spread out. For example, a pixel $A(x,y)$ is shown in Fig. 1. The sum of the square difference of each pixels gradient with A can reflect the difference between the regions of the A 's 3×3 neighborhood. The sum of the variance is defined as following.

$(x-1,y-1)$	$(x-1,y)$	$(x-1,y+1)$
$(x,y-1)$	(x,y)	$(x,y+1)$
$(x+1,y-1)$	$(x+1,y)$	$(x+1,y+1)$

Figure1. 3×3 neighborhood.

$$\begin{aligned} V(x,y) = {} & (g(x-1,y-1)-g(x,y))^2 + (g(x-1,y) \\ & -g(x,y))^2 + (g(x-1,y+1)-g(x,y))^2 + (g(x,y-1) \\ & -g(x,y))^2 + (g(x,y+1)-g(x,y))^2 + (g(x+1,y-1) \\ & -g(x,y))^2 + (g(x+1,y)-g(x,y))^2 + (g(x+1,y+1) \\ & -g(x,y))^2 \end{aligned} \tag{2}$$

where $V(x,y)$ denotes the sum of variance, $g(x,y)$ is the corresponding gradient value. The distance between pixel A to each neighborhood is different, therefore its impact on the target point varies. If the distance is shorter, the impact is greater and vice versa. Therefore the weighted variance is thus defined as

$$\begin{aligned} V(x,y) = {} & w_1(g(x-1,y-1)-g(x,y))^2 + w_2(g(x-1,y)- \\ & g(x,y))^2 + w_3(g(x-1,y+1)-g(x,y))^2 + w_4(g(x,y-1) \\ & -g(x,y))^2 + w_5(g(x,y+1)-g(x,y))^2 + w_6(g(x+1,y-1) \\ & -g(x,y))^2 + w_7(g(x+1,y)-g(x,y))^2 + w_8(g(x+1,y+1) \\ & -g(x,y))^2 \end{aligned} \tag{3}$$

where w_i denotes the weighting coefficient and is defined as following.

$$w_i = \frac{1}{r} \tag{4}$$

where r is the distance between the current point (x,y) to the neighborhood point (x',y'), and the distance is

$$r = \sqrt{(x-x)^2 + (y-y')^2} \tag{5}$$

Equation (6) reflects the relationship between the weighted variance and structuring elements map.

$$M(x,y) = \left|-\log(\alpha \times V(x,y))\right|, \quad (0 \le M(x,y) \le R_{\max}) \tag{6}$$

where $|\bullet|$ indicates rounding, α is a factor to adjust the value of $M(x,y)$, $R_{\max}$ is the maximal size of SE.

B. Image Modification

Morphological opening-closing operation employs different SE to modify each pixel of the image, and this is different from the traditional opening-closing by fixed SE. Such an adaptive opening-closing operation will eliminate the small bright and dark details and maintain the accuracy of the larger object contours. Our morphological modification is defined as

$$f_d(x,y) = g(x,y) \circ M(x,y) \bullet M(x,y) \tag{7}$$

where $f_d(x,y)$ is the modification image, $\circ$ and $\bullet$ are respectively denote morphological opening and closing operation.

III. Controlled Watershed

Small regular details are largely eliminated after the image modification. In order to segment the brain tumors and limit the allowable divided regions, marker-controlled watershed is employed to segment the modification image by the following steps.

Step 1, Tumor Marker Extraction Extraction: The purpose of this step is to locate the inner tumor regions. Since markers are picked from original modified image and the brain tumor regions usually have higher gray values than other brain tissue [20], the tumor regions can be extracted by thresholding(*T*) processing, where pixels value larger than *T* are labeled as tumor markers M .

Step 2, Background Marker Extraction: In order to determine the inside and outside catchments basins, background markers are also needed. This can be achieved by calculating the watershed transform of the Euclidian distance of the inner tumor regions. The Euclidian distance [19] is defined as following.

$$D_{(i,j)} = \min_{(x,y \in M)} D[(i,j),(x,y)] \tag{8}$$

$$D[(i,j),(x,y)] = \sqrt{(i-x)^2 + (j-y)^2} \tag{9}$$

where $D_{(i,j)}$ denotes the minimal distance between tumor marker pixel (x,y) and other pixel (i,j) , $D[(i,j),(x,y)]$ is the Euclidian distance between pixel (x,y) and (i,j) .

Step3, Watershed segmentation: After the foreground and background markers respectively corresponding to inside and outside of tumor have both marked out, minima imposition is applied to modify the gradient image so that the regional minimum occur at the markers location. Finally, watershed transform is performed on the modified gradient image to implement the tumor segmentation.

IV. Implementation

Fig.2 shows the proposed segmentation process, firstly, the morphological gradient image is calculated from the original image, and then the sum of variance is computed according to the pixel value of the gradient image. After constructing a structuring elements map with the size equal to the original image, its value of each pixel can be determined by the sum of variance. Modify each pixel of the original image by the different SE that size corresponding to the structuring element map at same location. Then mark the modified gradient image by the foreground and background markers. Finally watershed transform is used to implement the tumor segmentation.

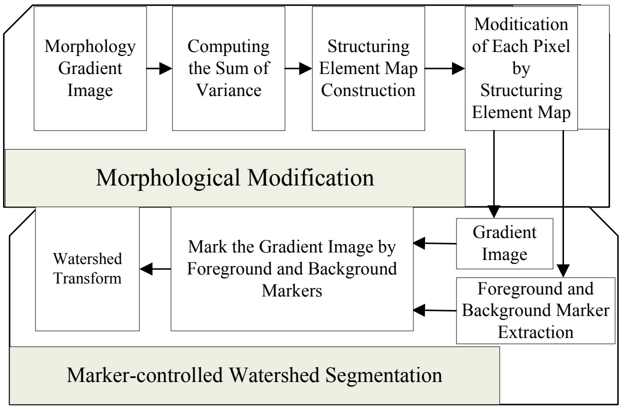

Morphology Gradient Image
Computing the Sum of Variance
Structuring Element Map Construction
Moditication of Each Pixel by Structuring Element Map
Morphological Modification
Gradient Image
Foreground and Background Marker Extraction
Mark the Gradient Image by Foreground and Background Markers
Watershed Transform
Marker-controlled Watershed Segmentation

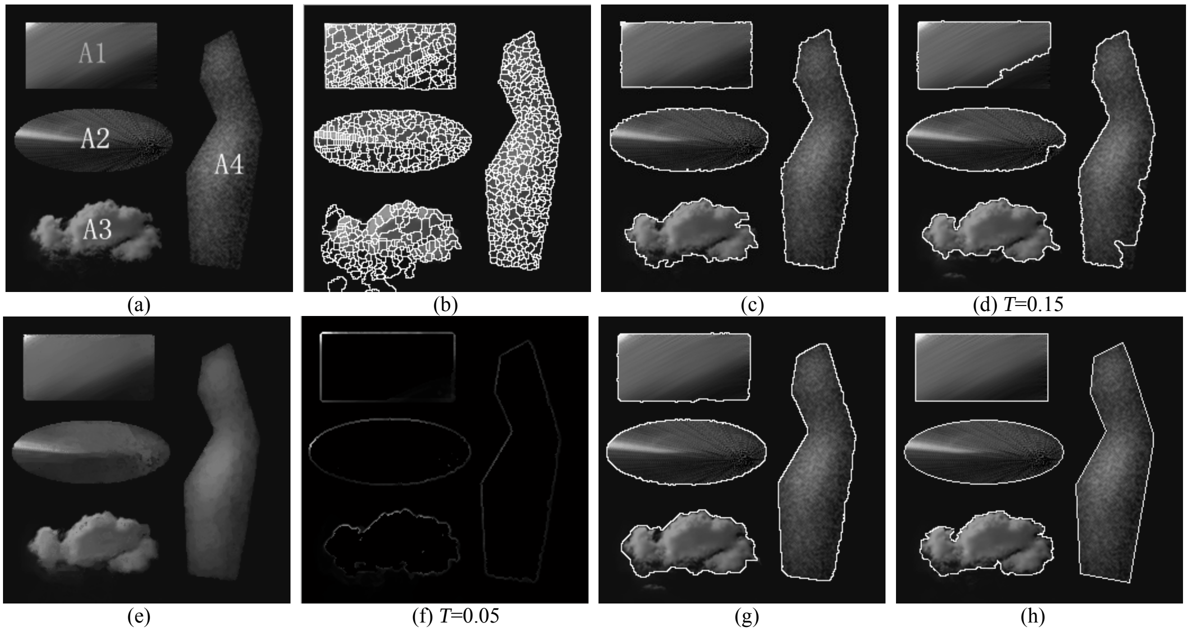
Figure2. The flow chart of the proposed segmentation

V. Experiments And Discussion

In order to verify the validity and performance of the proposed method, we choose a synthetic image and several clinical brain tumor CT images, and implement the simulation on MATLAB7 platform. The synthetic image as shown of Fig. 3(a), it contains four regions labeled as A1, A2, A3 and A4.

Fig. 3(b) is the result of watershed transform on Fig. 3(a), it produces a serious over-segmentation. Fig. 3(c) shows the segmentation of watershed transform followed by the maximal similarity based region merging [16], it can be seen that over- segmentation is largely released, but object contours occur offset. Fig.3 (d) gives the marker-controlled watershed segmentation, where $T = 0.15$. It is obvious that the bottom right corner of A1 is missing, and the other object shape contour is not accurate. Fig. 3(e) shows the result of the proposed modification, where $R_{\max} = 10$ and $\alpha = 6$. The marker of gradient image of the modified image by the foreground and background with $T = 0.05$ is given as Fig. 3(f); the marker-controlled watershed was performed on Fig. 3(f) producing the final segmentation. Compared with manual segmentation as Fig. 3(h), the proposed method result as shown in Fig. 3 (g) has accurately segmented the four desired object contours.

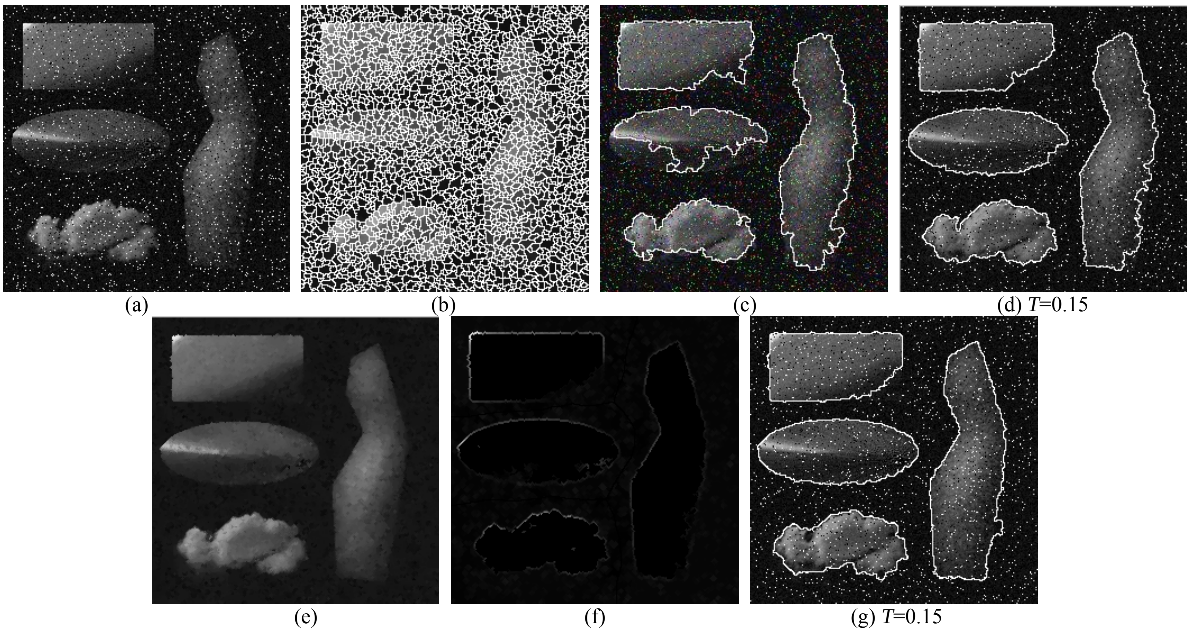
For the purpose to test the performance of the proposed method under noisy condition, we add Gaussian noise (0.1%) and salt-and-pepper noise (5%) to the Fig. 3(a). We can see from Fig. 4(b) when watershed transform is directly applied on such a noisy image, a serious over-segmentation appears. The maximal similarity based region merging is sensitive to noise and produce under-segmentation (Fig. 4(c)). Sole marker-controlled watershed shown in Fig. 4(d) led to inaccurate shape contours, especially, in low-contrast regions. Our method shown from Fig. 4(e) to (f) indicates that it is more robust to noise. TABLE I shows the time-costing of the different segmentation methods, where watershed transform is fast, watershed with region merging is more time-costing, and our method is slower than watershed transform but faster than region merging.



(a) (b) (c) (d) *T*=0.15

(e) (f) *T*=0.05 (g) (h)

Figure3. The flow chart of the proposed segmentation: The method for segmenting different shapes. (a) The original images; (b) Watershed segmentation; (c) The result of region merging of Fig. 3(b); (d) The result of maker-controlled watershed transform; (e) The result of modification; (f) Maker the gradient image; (g) The proposed segmentation result. (h) Manual segmentation.



(a) (b) (c) (d) *T*=0.15

(e) (f) (g) *T*=0.15

Figure4. The method for segmenting different shapes with noise. (a) Noisy image of Fig. 3(a); (b) Watershed transform for noisy image; (c) The result of region merging of Fig. 4(b); (d) The result of maker-controlled watershed transform; (e) The result of modification, (f) Maker the gradient image; (g) The proposed segmentation result.

TABLE I
THE TIME-COSTING OF DIFFERENT SEGMENTATION

Image	Segmentation Time(s)		
	Watershed Transform	Region Merging	Proposed method
Figure3 (a)	0.6523	1.5482	9.8764

In order to quantitatively analyze the segmentation accuracy of the different methods, we introduce the *TM* (Tanimoto Metric) [2] to evaluate the results of the segmentation, it is defined as following.

$$TM = \frac{\left\| R_x \cap R_g \right\|}{\left\| R_x \cup R_g \right\|}, \quad (0 \le TM \le 1) \tag{10}$$

where R_x denotes the amount of the segmented regions, R_g is the amount of region by hand-sketched, $\|\bullet\|$ denotes the total number of pixels within the collection. Typically, if *TM* is more close to 1, it indicates that the result of region is more close to the real contour. TABLE II shows the *TM* of the different segmentation methods for four objects in Figure.3 and Figure 4.

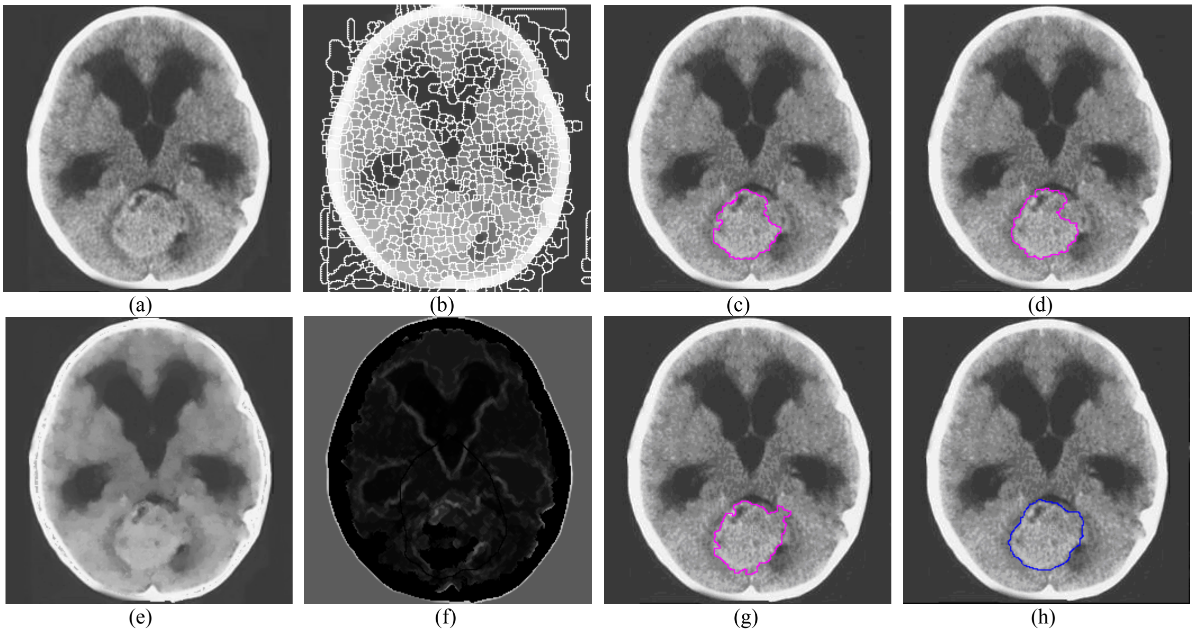


Figure5. Segmentation of CT-1 by different methods. (a)The original image; (b) Watershed transform; (c) Region merging of Fig.5(b); (d) Marker-controlled watershed transform; (e) Modification; (f) Marker the gradient image; (g) The proposed segmentation; (h) Manual segmentation.

TABLE.II
TM FOR DIFFERENT SEGMENTATION METHODS

Image		*TM*		
		Region merging	Marker-controlled watershed	Proposed method
Original	A1	0.9376	0.8104	0.9745
	A2	0.9785	0.9109	0.9787
	A3	0.9201	0.8778	0.9859
	A4	0.9683	0.9290	0.9821
Noisy	A1	0.7343	0.8361	0.8974
	A2	0.6289	0.9126	0.9774
	A3	0.8176	0.8726	0.8996
	A4	0.6688	0.8856	0.9486

It indicates that the accuracy of the proposed method is superior to the others, especially in noisy condition. To validate the performance of our method to segment the brain tumor, we firstly choose a clinical CT image named CT-1 as Fig.5(a). Fig.5(b) is the direct segmentation by watershed transform; Fig.5(c) shows the result after the maximal similarity based region merging, where most part of the tumor region is separated from the brain tissue, but the contour location occur bias. The marker - controlled watershed transform is almost the same as shown in Figure 5(d). Our method from Fig.5(e) to (g) indicates that it is closer to the manual segmentation (Fig.5(h)) than the others, where $R_{\max} = 10$, $T = 0.5$ and $\alpha = 12$. TABLE III shows the proposed method has a higher accuracy than the other methods.

TABLE.III
TM VALUE FOR DIFFERENT IMAGES

Tumor image	TM
a1	0.9164
a2	0.9273
a3	0.9418
a4	0.8978
a5	0.9102

To verify the capability of positioning tumor edge of the proposed method, we choose another five clinical tumor CT images (Fig.6 (a1-a5)). The parameters are identical with the CT-1 except that $T_{a1} = T_{a2} = 0.46$ and $T_{a3} = T_{a4} = T_{a5} = 0.36$. The second column of Fig. 6 shows the proposed segmentation results and the third column is the manual tumor segmentation. It can be seen that the proposed method is close to the desired manual segmentation. TABLE IV shows the *TM* of the proposed method for different images, and the average *TM* value is 0.9187, which indicates the proposed method has higher segmentation accuracy.

TABLE.IV
TM OF DIFFERENT METHODS

Image	Watershed with region merging	Marker-controlled watershed	Proposed method
CT-1	0.8403	0.8149	0.8883

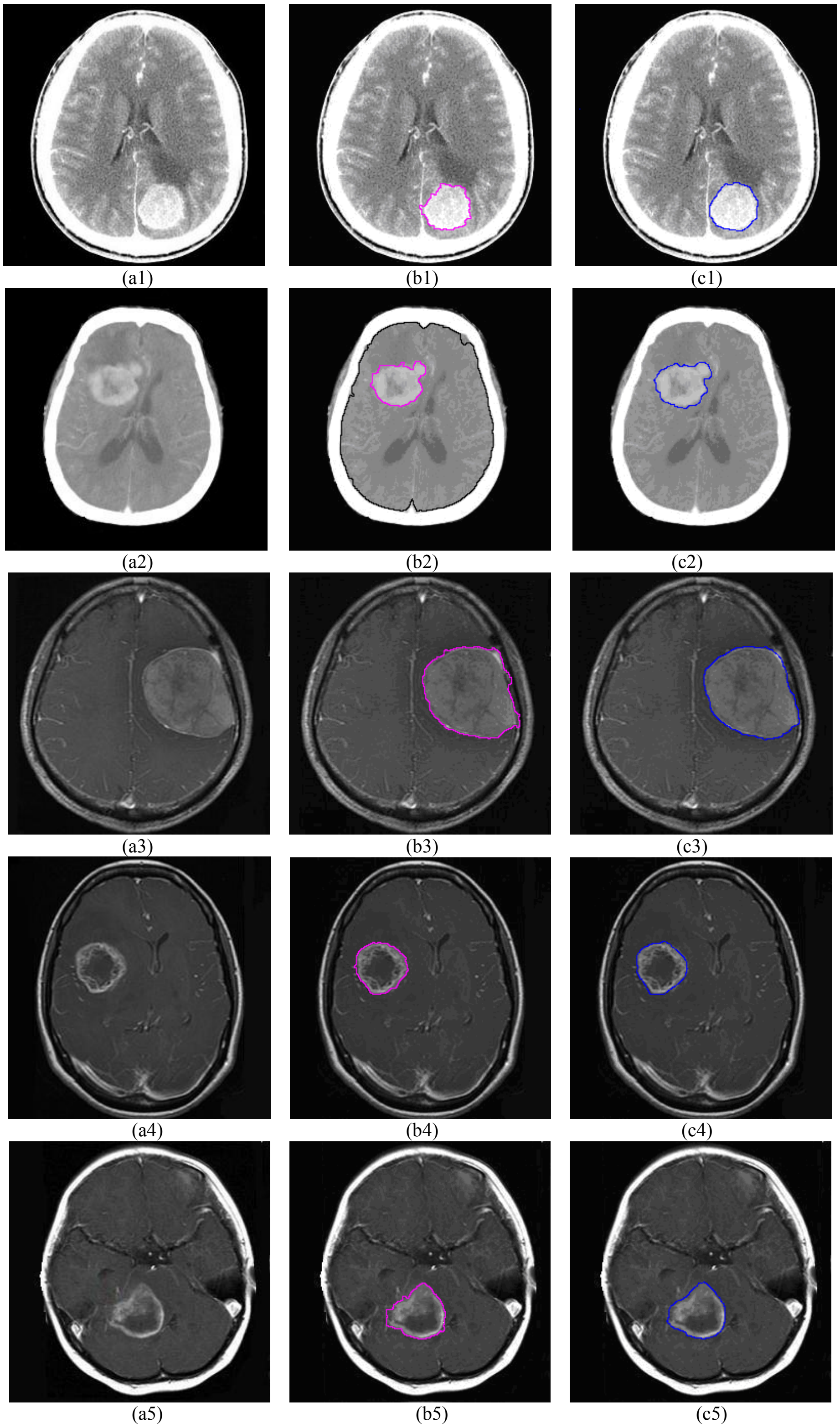


Figure6. Segmentation results of different images. (a) The original images; (b) The proposed method segmentation results; (c) Manual segmentation.

VI. Conclusion

Brain tumor segmentation plays an important role in the treatment. We propose a hybrid method which combines morphological image modification and marker-controlled watershed transform to segment the brain tumors. The original image is modified by opening-closing with the constructed structuring element map to release the bright and the dark regular details while preserve the objects contour with less offset. Marker-controlled watershed transform is used to localize and segment the tumors. Synthetic and several clinical images experimental results show that the proposed method can release the over-segmentation of the traditional watershed, and allows reliable and precise segmentation of the brain tumors.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61261029, 61202314), The Science and Technology Support Plan of Gansu Province (1204GKCA051), and Jinchuan Company Research Foundation (JCYY2013009).

References

[1] S.D. Salman and A.A. Bahrani, "Segmentation of tumor tissue in gray medical images using watershed transformation method," *Intl. Journal of Advancements in Computing Technology,* Vol. 2, No. 4, 2010, pp.123-127.

[2] T. Wang, I. Cheng, A. Basu. "Fluid vector flow and applications in brain tumor segmentation," *IEEE Transactions on medical Engineering*, Vol. 53, No. 3, 2009, pp.781-789.

[3] J.J Corso, E Sharon, A. Yuille, "Multilevel segmentation and integrated Bayesian model classification with an application to brain tumor segmentation," *Med. Image Comput. Comput. Assisted Intervention*, vol. 2, 2006, pp. 790–798.

[4] J. H Liu, J. W Wang, "Research on Contour Correction in Medical CT", *Journal of Computers*, Vol. 7, No. 3, 2012, pp. 762-767.

[5] KARANTZALOS K, ARGIALAS D. "Improving edge detection and watershed segmentation with anisotropic diffusion and morphological levellings". *International Journal of Remote Sensing*, Vol. 27, No.24, 2006, pp. 5427-5434.

[6] J.H Li, "Morphological Segmentation of 2-D Barcode Gray Scale Image", *Journal of Computers*, Vol. 8, No. 10, 2013, 2461-2468.

[7] V. Grau, A. U. J. Mewes, M. Alcañiz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging*. Vol. 23, No.4, 2004, pp. 447-458.

[8] J. M. Sharif, M. F. Miswan, M. A. Ngadi, Md Sah HjSalam. "Red blood cell segmentation using masking and watershed algorithm: A preliminary study," *Proceedings of 2012 International Conference on Biomedical Engineering* (ICOBE), 2012, pp. 258-262.

[9] J. Cousty, G. Bertrand, L. Najman, Michel Couprie. "Watershed Cuts: Thinnings, Shortest Path Forests, and Topological Watersheds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No.5, pp. 925-939, 2010.

[10] S. V Kumar, M. N Lazarus, C. Nagaraju, "A novel method for the detection of microcalcifications based on Multi-scale morphological gradient watershed segmentation algorithm," *International Journal of Engineering Science and Technology*, Vol. 2, No.7, 2010, pp. 2616-2622.

[11] S. Xu, H. Liu, E Song, "Marker-controlled watershed for lesion segmentation in mammograms," *Journal of Digital Imaging*, Vol. 24, No.5, 2011, pp. 754-763.

[12] B. Peng, L. Zhang, D. Zhang. "Automatic Image Segmentation by Dynamic Region Merging." *IEEE Transactions on Image Processing*, Vol. 20, No.12, 2011, pp. 3592-3605.

[13] J. J. Corso, E. Sharon, S. Dube. "Efficient Multilevel Brain Tumor Segmentation with Integrated Bayesian Model Classification," *IEEE Transactions on Medical Imaging,* Vol. 27, No.5, 2008, pp. 629-640.

[14] M. K. Kowar, S. Yadav, "Brain tumor detection and segmentation using histogram thresholding," *International Journal of Engineering and Advanced Technology*, Vol. 1, No.4, 2012, pp. 16-20.

[15] M. C. Jobin Christ, R. M. S. Parvathi, "Segmentation of medical image using K-Means clustering and marker controlled watershed algorithm," *European Journal of Scientific Research*. Vol. 71, No.2, 2012, pp. 190-194.

[16] N. Bouaynaya, D. Schonfeld, "Theoretical foundations of Spatially-Variant mathematical morphology part II: Gray-Level images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 30, No.5, 2008, pp. 837–850.

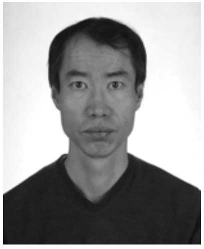
[17] C.Vachier, F. MEYER, "The viscous watershed transform," *Journal of Mathematical Imaging and Vision*, Vol. 22, No.2, 2005, pp. 251-267.

[18] J. Ning, L.Zhang, D. Zhang and C. Wu, "Interactive Image Segmentation using Maximal Similarity based Region Merging," *Pattern Recognition*, Vol. 43, No.2, 2009, pp. 445-456.

[19] L.L Xu and H.X Lu, "Automatic Morphological Measurement of the Quantum Dots Based on Marker-Controlled Watershed Algorithm", *IEEE Transactions on Nanotechnology*, Vol.12, No. 1, 2013, pp. 51-56.

[20] S. H. Lewis, A J Dong, "Detection of Breast Tumor Candidates Using Marker-controlled Watershed Segmentation and Morphological Analysis", *SSIAI2012*, 2012, pp.1-4.

[21] C.Y Lui, "Gaussian Kernelized Fuzzy c-means with Spatial Information Algorithm for Image Segmentation", *Journal of Computers*, Vol. 7, No. 6, 2012, pp. 1511-1518.



Xiaopeng Wang received his Ph.D degree in signal and information processing from Northwestern Polytechnical University, China, in 2005. His interested research fields are image analysis and recognition.

Shengyang Wan received his BA degree in electronic and Information engineering from Nanchang HangKong University, China, in 2010. He is currently pursuing an MS. degree in Signal and Information Processing at Lanzhou Jiaotong University, China. His research interests include image processing and pattern recognition.

Lei Tao received his Ph.D degree in information and communication engineering from the Northwestern Polytechnical University, Xian, in 2011. Currently, he is an Associate Professor at Lanzhou Jiaotong University. His research interests include image processing, pattern recognition and computer vision.

A Cloud Computing-based Television Program Opinion Monitoring and Analysis System

Dingguo Yu
School of New Media, Zhejiang University of Media and Communications, Hangzhou, China
Email:zjydg@163.com

Huxiong Li, Zhiwen Hu
School of New Media, Zhejiang University of Media and Communications, Hangzhou, China
Email: jsj_lhx@126.com, sunneyhu@gmail.com

Abstract—**Today, many television viewers are also experienced internet users. People often post reviews of television programs on social networks and have created influential and powerful opinions. To strengthen the monitoring and analysis of these opinions, this paper developed a cloud computing-based internet public opinion monitoring and analysis system for television programs. The purpose of this novel system is to assist the government and television stations with program monitoring in order to provide a decision-making basis for improvement. This paper introduced the new cloud computing-based monitoring and analysis system, including the necessary cloud computing technology, system framework, functional structure, and system workflow.**

Index Terms—**Television Program, Internet Public Opinion, Cloud Computing**

I. Introduction

This internet public opinion monitoring and analysis service for television programs is monitors and analyzes the comments and opinions of internet users to relevant television programs through the monitoring and tracking of websites (including portal websites, industry websites, forums, blogs, and microblogs, etc.). It further investigates and analyzes the viewer base and the social impact of certain television programs, and relays this information to the regulatory and decision-making departments of government institutions and television stations for references.

Due to the development and popularity of the internet, especially the wide spread of social networks, and the fact that most television viewers are also experienced internet users, the discussions and comments regarding television programs through social media such as blogs and microblogs, have developed into a strong influence on television program opinion, which has also further impacted the public opinion. Currently, there are numerous television programs with uneven qualities. A certain number of programs, especially some comments have large influences on the ideology of the viewers. Therefore, it is essential to develop an internet public opinion monitoring and analysis system for television programs in order to launch the monitoring and analysis of public opinion on television programs, assist the government and television stations with the regulations of television programs, understand the viewer base and social impact of television programs, find problem, divert public opinions, and provide evidences for the decision-making of government and the regulatory and improvement of television programs.

II. Related Work

The key of the television program opinion monitoring technology is topic detection and tracking (TDT), which is an algorithm that finds relevant information from vast amounts of data streams[1,2]. The technology has been well-developed and primarily includes two categories: clustering algorithms [3-7] and repeating string matching [8-10].

Currently, there are no monitoring and analysis systems designed for domestic television programs opinions. There are only some commonly available internet public opinion monitoring and analysis systems which are difficult to use for monitoring and analysis of the opinions specifically directed at television programs. A few commonly used systems include: Goonie internet public opinion monitoring system [11], Junquan internet public opinion monitoring system [12], TRS internet public opinion management system [13], and Founder internet public opinion and information monitoring and analysis system [14].

III. System Design

A. Cloud Computing Technology

The cloud computing technology for this television program public opinion monitoring and analysis system primarily involves the following three aspects:

(1) Uses cloud storage technology for program opinion resource data management in order to meet the need for large amounts of data.

(2) Adopts cloud computing technology architecture with a centralized television program opinion monitoring center to uniformly perform information collection and public opinion monitoring and analysis services.

doi:10.4304/jsw.9.11.2933-2937

Monitoring and analysis application terminals can be PCs, tablets, or smart phones, which shows the flexibility of the system.

(3) The television program opinion monitoring and analysis is task-driven. Each television program is a single monitoring task, and the system allocates different monitored keywords information to different programs. Targeted monitoring and analysis can be carried out for various programs and different monitoring and analysis reports will be given.

B. The Design of System Framework and Functional Structure

The framework and functional structure for the cloud computing based television program opinion monitoring and analysis system is shown in Figure 1.

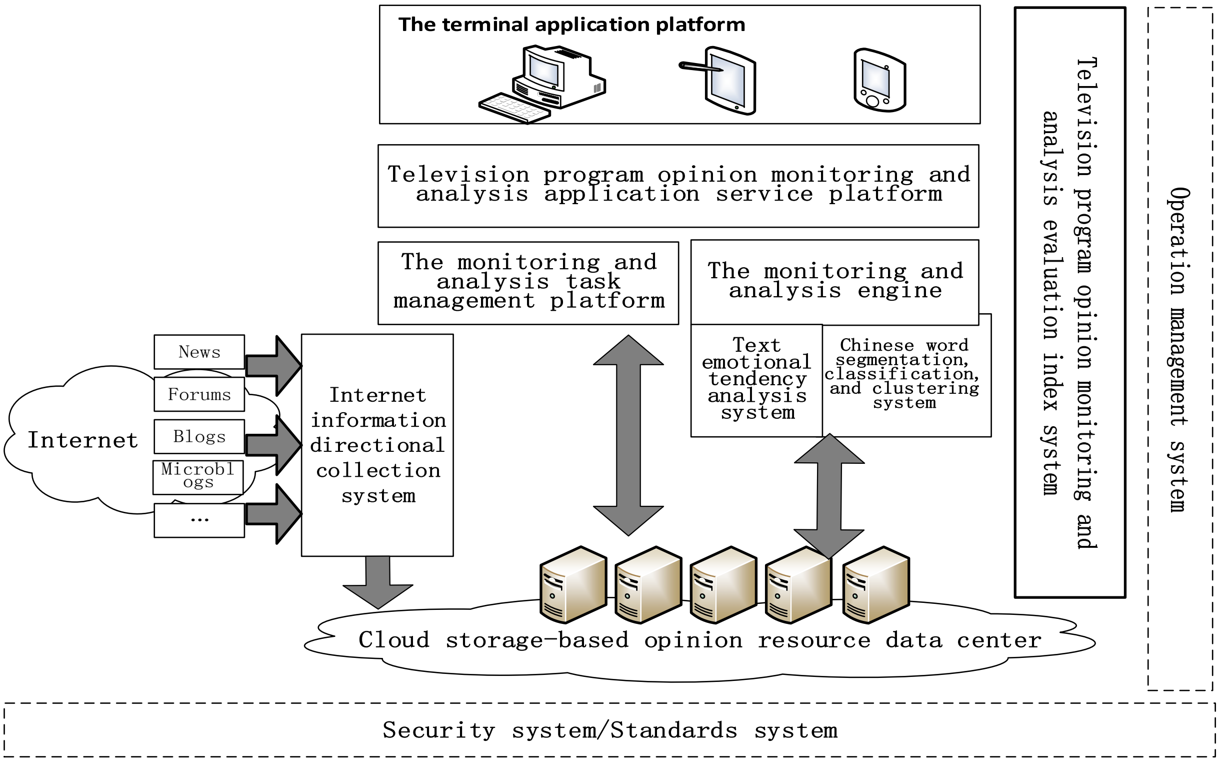

The terminal application platform
Television program opinion monitoring and analysis application service platform
The monitoring and analysis task management platform
The monitoring and analysis engine
Text emotional tendency analysis system
Chinese word segmentation, classification, and clustering system
Internet
News
Forums
Blogs
Microblogs
...
Internet information directional collection system
Cloud storage-based opinion resource data center
Television program opinion monitoring and analysis evaluation index system
Operation management system
Security system/Standards system

Figure 1. The design of system framework and functional structure

The main functional modules of the system include:

(1) Cloud-storage-based opinion resource data center

The data center primarily stores and manages the television program opinions corpus data. The system adopts a distributed cluster data storage management, which realizes the management technology of big data.

(2) Internet information directional collection system

The internet information directional collection system is the front and foundation of the cloud computing-based television program opinion monitoring and analysis system. This system automatically collects and classifies the information of designated websites and interprets the collected information. Specifically, it performs smart partitioning, and duplicate and noise removal for web pages. This system then identifies the page structure, transforms information into structured data, and stores them to the data center for the follow-up function modules. Additionally, the system will regularly update the collected information in order to perform information tracking.

Directional collection: Limits information collection objects through system configuration. If the collection sites are set to be “Sina.com” and “Sohu.com”, the system will only collect information from web pages of the two portal websites.

Smart interpretation: Automatically partitions the collected web page files, removes noise, identifies web page structure, extracts the main body of the web pages and other key information (such as titles, article sources, time of publication, and authors.), and converts it into structured data.

Regular updates: Performs regular scans (for example, every 24 hours) on the collected data and examines whether the contents of the collecting sites have been updated. If updated, the original web page data will be substituted with the new content.

Classified collection: In order to increase the efficiency of the collection and the accuracy of the interpretation, the collected objects are classified. The collection and interpretation program can also be adjusted for targeted collection and interpretation. Currently, there are three resources from which to collect: news and blogs, forums, and microblogs.

(3) Television program opinion monitoring and analysis engine

This engine primarily realizes the analysis of opinion corpus data that are generated by an internet information directional collection system and stored in the data center. It's main operations include Chinese word segmentation, classification, and clustering of corpus information in the data center. It also analyzes the emotional tendencies of the data. The engine is a Lucene-based full-text search system that can be used by television opinion monitoring and analysis application service platform for searching, statistics, and invoking.

(4) The opinion monitoring and analysis task management system

The opinion monitoring and analysis task management system primarily manages and maintains the monitor tasks submitted by users, and configures and modifies the corresponding monitor indices. Its main functions include:

User management: Primarily manages users who manage the monitoring and analysis tasks, and maintains user information. Its major operations include user information maintenance and password management.

Monitoring and analysis task management: Maintains and manages relevant information of all television program opinion monitoring and analysis tasks. Its major operations include maintenance of task information and monitor keywords which are some key field information need to be monitored and tracked for some monitoring and analysis task.

(5) Television program opinion monitoring and analysis application service platform

The application service platform performs statistical analysis of the data generated by the television program opinion monitoring and analysis engine module and offers statistical analysis data for the use of various television opinion monitoring and analysis application terminals. The statistical analysis links to the opinion corpus data files that have already been indexed by the monitored keywords. The analysis results are provided to all kinds of monitoring and analysis application terminals (such as web sites and mobile applications) in the form of service.

(6) Television programs opinion monitoring and analysis terminal application platform

The terminal application platform is the collection of various television program opinion monitoring and analysis terminal application programs, including web sites and IOS or Android-based mobile applications.

C. *System Workflow Design*

The workflow of all function modules are shown in Figure 2.

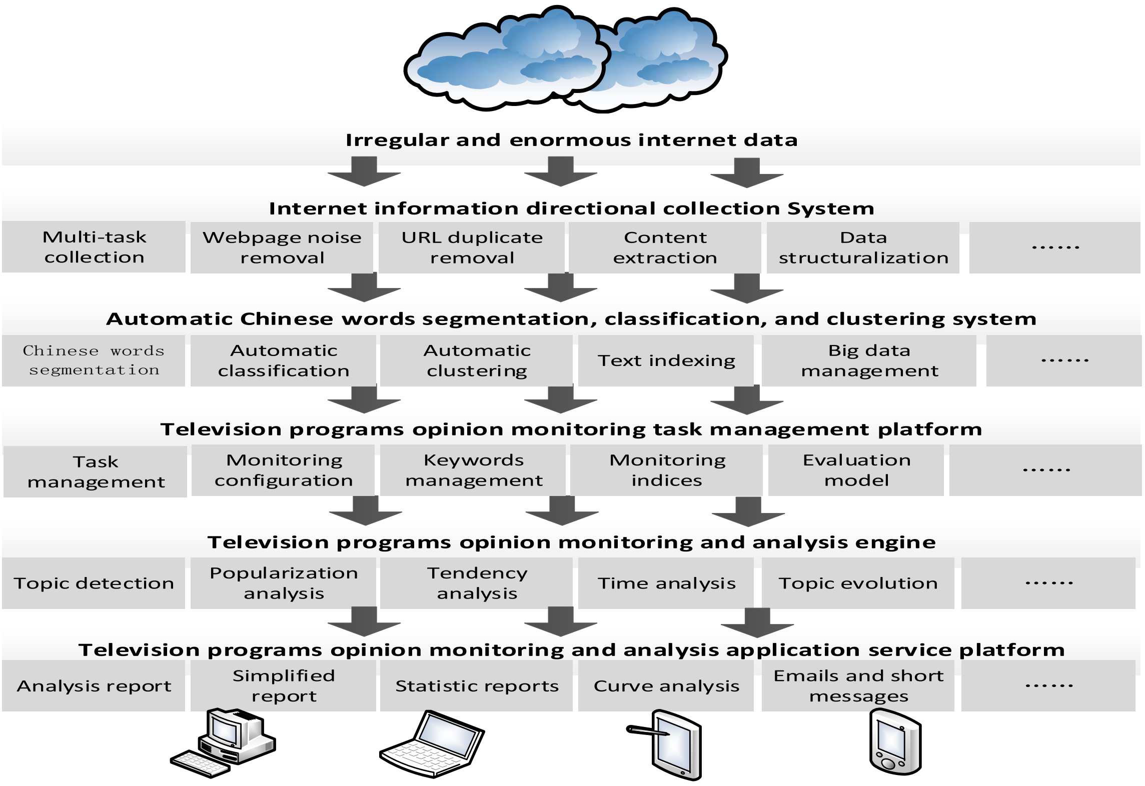


Figure 2. System workflow chart

IV. System Operational Deployment

The operational deployment of the cloud computing-based television program opinion monitoring and analysis system is shown in Figure 3.

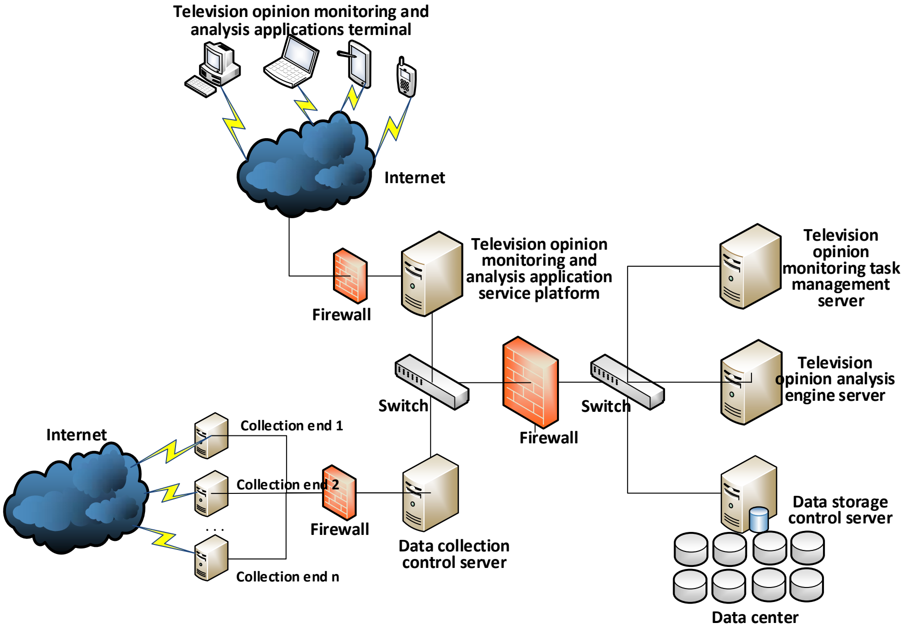

Television opinion monitoring and analysis applications terminal
Internet
Television opinion monitoring and analysis application service platform
Firewall
Television opinion monitoring task management server
Switch
Firewall
Switch
Television opinion analysis engine server
Internet
Collection end 1
Collection end 2
Firewall
Collection end n
Data collection control server
Data storage control server
Data center

Figure 3. System operational deployment chart

The entire system deployment includes the following key parts:

(1) Internet information collection server (group)

The server (group) operates the internet information directional collection system. It is composed of a data collection control server and several data collection ends. The data collection control server uniformly manages and monitors the collection ends. The system realizes classified multi-task webpage information collection and content interpretation. Each collection end undertakes a collection task where the multi-thread webpage information collection and content interpretation is performed for each designated collection site to convert irregular webpage information into structured data and store them to the specified data center.

(2) Data storage server (group)

The server (group) is a data center composed by a data storage control server and several data storages locations. It stores the corpus data of television opinions. The data storage capacity of the data center is 20T at the time of the preliminary design.

(3) Television opinion analysis engine server

The server operates the television program opinion monitoring and analysis engine program. It analyzes the opinion corpus data that are generated by the internet information directional collection system and stored in the data center, its main function including Chinese word segmentation, classification, and clustering of corpus information in the data center. It also analyzes the emotional tendencies of the data. The server also constructs index files according to task-specified monitored keywords for retrieval, census, and invoking by the television opinion monitoring and analysis application service platform.

(4) Television opinion monitoring task management server

The server operates the television opinion monitoring task management program which manages and maintains the monitoring task information submitted by users.

(5) Television opinion monitoring and analysis application service providing server

The server operates the television opinion monitoring and analysis application server platform which performs statistical analysis on the data generated by the television program opinion monitoring and analysis engine module and produces statistical analysis data in the form of service for the use of various television opinions monitoring and analysis application terminals.

V. Conclusions

In order to enhance the monitoring and analysis of relevant television program opinions on the internet, a cloud computing-based television programs monitoring and analysis system was developed and a 20T data center was established for the regular collection and update of relevant data from approximately 30 portal websites and news websites, 30 forum sites, and Sina microblogs. It can be seen from the testing that the intended result was achieved. Using a 100M campus network, the average speed of information collection and interpretation for news websites on a single PC (one collection terminal) is

47 pages per second. The interpretation accuracy reached 94.8%.

ACKNOWLEDGMENT

This work was supported by the Program for Zhejiang Leading Team of Science and Technology Innovation under Grant No. 2011R50019, the science and technology project of state administration of press，publication，Radio，film and television of China under Grant No. 201309, and the international cooperation project of Science Technology Department of Zhejiang Province under Grant No. 2012C24019.

REFERENCES

[1] James Allan, “Introduction to Topic Detection and Tracking,” Topic Detection and Tracking, The Information Retrieval Series, Vol.12, pp.1-16, 2002.
[2] Yu Hong, Yu Zhang, and Ting Liu, et al., “Topic Detection and Tracking Review,” Journal of Chinese Information Processing, Vol.21(6), pp.71-87, 2007.
[3] Chunshan Li, Yunming Ye, and Xiaofeng Zhang, et al., “Clustering Based Topic Events Detection on Text Stream,” Intelligent Information and Database Systems Lecture Notes in Computer Science, Vol.8397, pp. 42-52, 2014.
[4] Shu-Wei Liu,and Hsien-Tsung Chang, “A Topic Detection and Tracking System with TF-Density,” Recent Progress in Data Engineering and Internet Technology Lecture Notes in Electrical Engineering, Vol.156, pp.115-120, 2013.
[5] Maximilian Walther,Michael Kaisser, “Geo-spatial Event Detection in the Twitter Stream,” Advances in Information Retrieval Lecture Notes in Computer Science, Vol.7814, pp.356-367, 2013.
[6] Jianfang Wang, Xiao Jia, Longbo Zhang, “Identifying and Evaluating the Internet Opinion Leader Community Through k-clique Clustering,”. Journal of Computers, Vol.8(9), pp.2284-2289, 2013.
[7] Yongping Du, Changqing Yao, “Performance Evaluation of the Cyberspace Public Opinion Detection and Tracking,” Journal of Computers, Vol.7(5), pp. 1284-1288, 2012.
[8] ZENG Yi-ling, XU Hong-bo, “Research on Internet Hotspot Information Detection,” Journal on Communications, Vol.28(12), pp.141-146, 2007.
[9] Xiaoming Zhang, Zhoujun Li, “Automatic Topic Detection with an Incremental Clustering Algorithm,” Web Information Systems and Mining Lecture Notes in Computer Science, Vol.6318, pp.344-351, 2010.
[10] He Qi, Chang Kuiyu, and Lim Ee-Peng, et al., “Keep It Simple with Time: A Reexamination of Probabilistic Topic Detection Models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.32(10), pp.1795-1808, 2010.
[11] Goonie Internet public opinion monitoring system, http://www.goonie.cn/products, May 13, 2013.
[12] Junquan internet public opinion monitoring system, http://www.54yuqing.com, March 17, 2014.
[13] TRS internet public opinion management system, http://www.trs.com.cn/product, March 15, 2014
[14] The solution of Founder internet public opinion and information monitoring and analysis system, http://wenku.baidu.com, June 14, 2010.

Dingguo Yu is currently an associate professor in the Zhejiang University of Media and Communications, China. He received his M.S. and Ph.D. degrees from Tongji University, China in 2005 and 2011 respectively. His research interests are in the fields of cyber security, mobile computing and internet public opinion, etc.

Huxiong Li was born in Hubei, China. He received his Ph.D. degrees in 2009 respectively in Northwestern Polytechnical University. He is currently an associate professor at Zhejiang University of Media and Communications in Hangzhou, China. His current research Software engineering, Web application development, etc.

Zhiwen Hu was born in Hubei, China. He received his B.S., M.S., and Ph.D. degrees in 1996, 2003, and 2006 respectively in China University of Geosciences, Hefei University of Technology and Chinese Academy of Sciences. He is currently an associate professor at Zhejiang University of Media and Communications in Hangzhou, China. His current research interests reside in information fusion, intelligent control, etc.

A Spatial Skyline Query for a Group of Users

Mohammad Shamsul Arefin†, Geng Ma‡, and Yasuhiko Morimoto‡
†Department of CSE, Chittagong University of Engineering and Technology (CUET), Bangladesh
‡Graduate School of Engineering, Hiroshima University, Japan
Email: sarefincse@gmail.com, [M115245, morimo]@hiroshima-u.ac.jp

Abstract— **A skyline query finds objects that are not dominated by another object from a given set of objects. Skyline queries can filter unnecessary information efficiently and provide us important clues for various decision making tasks. Now a days, GPS devices and location based services are very popular and they can easily connect users and make groups. Conventional skyline queries are not sufficient to obtain valuable knowledge to fulfil the needs of such groups. Considering this fact, in this paper, we proposed a spatial skyline query for groups of users located at different positions. Our proposed skyline query algorithm selects a set of spatial objects to fulfil the groups' needs. For example, if a group wants to find a restaurant to hold a meeting, our method can select a convenient place for all users of the group. We performed several extensive experiments to show the effectiveness of our approach.**

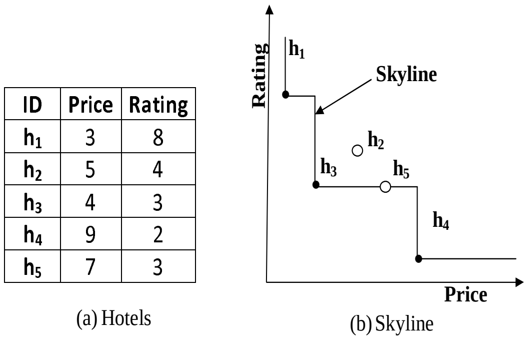
Index Terms— **Spatial skyline, Skyline for a group, Voronoi diagram.**

I. Introduction

Given a k-dimensional database DB, a skyline query retrieves a set of skyline objects, each of which is not dominated by another object. An object p is said to dominate another object q if p is not worse than q in any of the k dimensions and p is better than q in at least one of the k dimensions. Figure 1 shows a typical example of skyline. The table in Figure 1 is a list of five hotels, each of which contains two numerical attributes "Price" and "Rating". In the list, h_2 and h_5 are dominated by h_3, while others are not dominated by any other hotel. Therefore, the skyline of the list is h_1, h_3, h_4. Such skyline results are important for users to take effective decisions over complex data having many conflicting criteria. In database literature, there are many recent studies for efficient computation of skyline queries from databases [1]–[9]. All of these works just consider non-spatial information like price and rating.

Recently, GPS devices and location based services become popular. As a result, we have large databases containing spatial information. Therefore, we often have to select spatial objects from a spatial database. Conventional skyline queries are not sufficient to handle spatial objects. To solve the problem, spatial skyline queries have been proposed [14]–[21]. Most of those spatial skyline queries select a set of objects based on proximity from a given query point.

Different from other works, we consider a spatial skyline query for a group of users located at different positions. This is because there are situations where a group of users at different locations may want to choose a particular object that can fulfil the group's needs. For example, assume that members of a multidisciplinary task force team located at different offices want to put together in a restaurant to hold a lunch-on meeting. Conventional spatial skyline query cannot take into account the group's convenience.



ID	Price	Rating
h_1	3	8
h_2	5	4
h_3	4	3
h_4	9	2
h_5	7	3

(a) Hotels

(b) Skyline

Figure 1. Skyline example

The problem of spatial skyline queries can be defined as follows. Given the two sets P of data points and Q of query points, the spatial skyline of P with respect to Q is the set of those points in P, which are not *spatially dominated* by any other point of P. A data point p_1 is said to spatially dominate another point p_2 with respect to Q iff we have $d(p_1, q_i) \leq d(p_2, q_i)$ for all $q_i \in Q$ and $d(p_1, q_j) < d(p_2, q_j)$ for some $q_j \in Q$, where $d(p, q)$ is the Euclidean distance between p and q. Figure 2 shows a set of nine points and two query points q_1 and q_2 in a plane. The point p_1 spatially dominates the point p_2 since both q_1 and q_2 are closer to p_1 than to p_2.

Social network services can connect users of different positions and make groups easily. Therefore, we often have to solve this spatial problem. Some of the existing spatial skyline queries consider the same spatial problem. However, most of those works only consider spatial information such as locations of the users and objects and do not take into account non-spatial features of objects, such as price and rating. Since both spatial and non-spatial features of objects are very important for efficient knowledge discovery tasks, we consider a method that can select objects based on both spatial and non-spatial features.

A. Motivating Example

Assume there is a database of restaurants as in Table I. The database has two non-spatial attributes: "Rating"

doi:10.4304/jsw.9.11.2938-2947

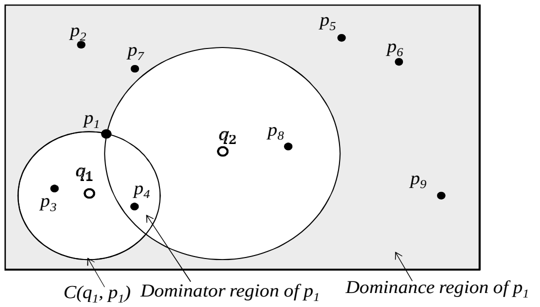


Figure 2. Spatial skyline example

TABLE I.
RESTAURANT DATABASE

ID	Location	Rating	Price
$\mathbf{r_1}$	(3, 9)	3	2
$\mathbf{r_2}$	(7, 5)	2	2
$\mathbf{r_3}$	(7, 7)	3	4
$\mathbf{r_4}$	(5, 1)	3	2
$\mathbf{r_5}$	(4, 4)	2	3
$\mathbf{r_6}$	(4, 8)	3	3
$\mathbf{r_7}$	(5, 6)	3	1
$\mathbf{r_8}$	(1, 3)	3	2
$\mathbf{r_9}$	(5, 3)	2	2
$\mathbf{r_{10}}$	(9, 3)	1	1

and "Price", in addition to the "Location" attribute. We assume that lower value is better in each of the non-spatial attributes. We also assume there are four users u_1, u_2, u_3, and u_4, whose current locations are at (4.5, 5.5), (5, 6.8), (6, 5), and (5, 3.8), respectively, as in Table II.

To select a good restaurant for the four users, at first, we calculate the Euclidean distance of each restaurant from each of the four users (query points) and construct the table as shown in Table III. In the table, the attribute r-u_1 represents Euclidean distances of the restaurants from user u_1. Similarly, r-u_2, r-u_3, and r-u_4 are the Euclidean distances of restaurants from u_2, u_3, and u_4, respectively. Sum-$Distance$ attribute in Table III contains the sum of Euclidean distances of each data point (restaurant) from the users u_1, u_2, u_3, and u_4.

Note that a restaurant that is the closest from one user can be an attractive candidate. In addition, a restaurant whose sum of Euclidean distances from the four users is smallest must be an attractive candidate. Therefore, we use those five spatial attributes for the four users problem.

Next, we join the non-spatial attributes of Table I and spatial information of Table III and obtain the information of Table IV. After computing Table IV, we can get the skyline for the four users by using conventional skyline query, which are r_2, r_5, r_7, r_9, and r_{10}. However, we have to compute spatial features like Table III for each of different query, which are time-consuming and not affordable.

TABLE II.
USERS' LOCATION DATABASE

ID	Location
$\mathbf{u_1}$	(4.5, 5.5)
$\mathbf{u_2}$	(5, 6.8)
$\mathbf{u_3}$	(6, 5)
$\mathbf{u_4}$	(5, 3.8)

TABLE III.
SPATIAL ATTRIBUTES OF RESTAURANTS

ID	r-u_1	r-u_2	r-u_3	r-u_4	Sum-$Distance$
$\mathbf{r_1}$	3.81	2.97	5.12	5.57	17.47
$\mathbf{r_2}$	2.55	2.69	1	2.33	8.57
$\mathbf{r_3}$	2.92	2.01	2.24	3.77	10.94
$\mathbf{r_4}$	4.53	5.8	4.12	2.8	17.25
$\mathbf{r_5}$	1.58	2.97	2.24	1.02	7.81
$\mathbf{r_6}$	2.55	1.56	3.61	4.32	12.04
$\mathbf{r_7}$	0.71	0.89	1.41	1.48	4.49
$\mathbf{r_8}$	4.30	5.52	5.39	4.08	19.29
$\mathbf{r_9}$	2.54	3.8	2.24	0.8	9.38
$\mathbf{r_{10}}$	5.15	5.18	3.61	4.08	18.38

In this paper, we consider an efficient method for computing such a spatial skyline query without constructing all the information of Table IV for a group of users of different locations. Instead, we only compute necessary spatial information for each of different query (group) efficiently. For simplicity, we consider the above examples as running examples throughout the paper.

The proposed method can be summarized as follows:

- First, we compute skyline objects based on "spatial sub-space" of the data points. In this step, we do not compute all values in Table III but compute only necessary distances to find dominated objects on the spatial sub-space.
- Next, we compute dominated objects on "non-spatial sub-space" of the data points.
- Then, we integrate those information to compute final skyline result.

We intensively evaluate our framework using both synthetic and real data and validate the effectiveness of our method.

The remainder of this paper is organized as follows. In Section II, we provide a brief survey of related works. In Section III, we describe some preliminary concepts related to our work. Section IV briefly explains over all procedure of the proposed skyline computation method. In Section V, we report our evaluation results, and finally this paper is concluded in Section VI.

II. RELATED WORKS

A. Skyline Computation

Skyline queries were originally considered for maximal vectors computation [1]. Borzsonyi et al. [2] first introduced skyline queries in database applications and proposed Block Nested Loop (BNL), Divide-and-Conquer, and B-tree based algorithms. Later, a number of different algorithms such as progressive skyline computation algorithm [3], nearest neighbor algorithm [4], branch and bound skyline (BBS) algorithm [5], and sort-filter-skyline (SFS) algorithm [6] were proposed for efficient skyline computation.

Due to the increase in data dimensionality, there have been many research efforts to address the dimensionality

TABLE IV.
NON-SPATIAL AND SPATIAL ATTRIBUTES OF RESTAURANTS

ID	Rating	Price	r-u_1	r-u_2	r-u_3	r-u_4	*Sum-Distance*	
$\mathbf{r_1}$	3	2	3.81	2.97	5.12	5.57	17.47	dominated by r_2, r_7
$\mathbf{r_2}$	2	2	2.55	2.69	1	2.33	8.57	not dominated
$\mathbf{r_3}$	3	4	2.92	2.01	2.24	3.77	10.4	dominated by r_7
$\mathbf{r_4}$	3	2	4.53	5.8	4.12	2.8	17.25	dominated by r_7
$\mathbf{r_5}$	2	3	1.58	2.97	2.24	1.02	7.81	not dominated
$\mathbf{r_6}$	3	3	2.55	1.56	3.61	4.32	12.04	dominated by r_7
$\mathbf{r_7}$	3	1	0.71	0.89	1.41	1.48	4.49	not dominated
$\mathbf{r_8}$	3	2	4.30	5.52	5.39	4.08	19.29	dominated by r_2, r_7, r_9
$\mathbf{r_9}$	2	2	2.54	3.8	2.24	0.8	9.38	not dominated
$\mathbf{r_{10}}$	1	1	5.15	5.18	3.61	4.08	18.38	not dominated

problem of skyline queries such as skyline frequency [7], k-dominant skylines [8], and k-representative skylines [9].

All these efforts, however, do not consider spatial relationships between data points.

B. Spatial Skyline Query

Spatial query processing was first studied for ranking neighboring objects. Several works [10]–[12] considered spatial query mechanism for ranking neighboring objects using the distance to a single query point. Papadias et al. [13] considered ranking of objects using aggregate distance of multiple query points.

Sharifzadeh et al. [14] first addressed the problem of spatial skyline queries. They proposed two algorithms, B^2S^2 and VS^2, for static query points and one algorithm, VCS^2, for the query points whose locations change over time. VCS^2 exploits the pattern of change in query points to avoid unnecessary re-computation of the skyline. The main limitation of VS^2 algorithm is that it can not deliver correct results in every situation. To overcome the limitation of VS^2 algorithm, Son et al. [15] presented a simple and efficient algorithm that can compute the correct results. Guo et al. [16] introduced the framework for direction-based spatial skyline computation that can retrieve nearest objects around the user from different directions. They also developed an algorithm to support continuous queries. However, their algorithm for direction-based spatial skyline can not handle more than one query point. Kodama et al. [17] proposed efficient algorithms to compute spatial objects based on a single query point and some non-spatial attributes of the objects.

There are some considerations about spatial skyline computation in road networks. Deng et al. [18] first proposed multi-source skyline query processing in road network and proposed three different spatial skyline query processing algorithms for the computation of skyline points in road networks. In [19], Safar et al. considered nearest neighbour based approach for calculating skylines over road networks. They claimed that their approach performs better than the approach presented in [18]. Huang et al. [20] proposed two distance-based skyline query techniques those can efficiently compute skyline queries over road networks. Zheng et al. [21] proposed a query processing method to produce spatial skylines for location-based services. They focus on location-dependent spatial queries (LDSQ) and consider a continually changing user location (query point). In their approach, it is not easy to decide how often the skyline result needs to be updated.

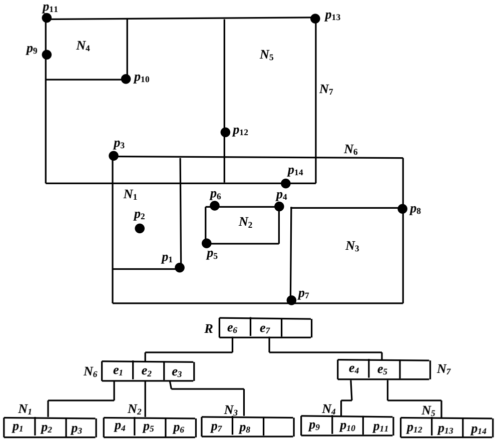


Figure 3. Example of an R-tree

None of the above works considered the computation of spatial skyline objects for a group of users based on both spatial and non-spatial information. In this paper, we consider the issue and propose an efficient method for computing such spatial skyline objects.

III. PRELIMINARIES

A. Skyline Queries

Let p and q be objects in a database DB. Let $p.a_l$ and $q.a_l$ be the l-th attribute values of p and q, respectively, where $1 \leq l \leq k$. An object p is said to dominate another object q, if $p.a_l \leq q.a_l$ for all the k attributes a_l, $(1 \leq l \leq k)$ and $p.a_j < q.a_j$ on at least one attribute a_j, $(1 \leq j \leq k)$. The skyline is a set of objects which are not dominated by any other object in DB.

B. Spatial Skyline Queries

Assume that there are two point sets. One is a set of data points, say P, and the other is a set of query points, say Q. We also assume that each point in P and Q has spatial attributes, which are 2-dimensional coordinate attributes. Let us also consider that the distance function $d(p, q)$ returns the Euclidean distance between a pair of points p and q.

Definition 1: We say that p_1 "spatially dominates" p_2 if and only if $d(p_1, q) \leq d(p_2, q)$ for every $q \in Q$, and $d(p_1, q) < d(p_2, q)$ for some $q \in Q$.

The spatial skyline of P with respect to Q is the set of those points in P, which are not *spatially dominated* by any other point of P.

C. R-Tree

R-tree is the most prominent index structure widely used for spatial query processing. Figure 3 shows an R-tree containing $P = \{p_1,\cdots,p_{14}\}$. We set the capacity of each node to three. The leaf nodes N_1, ..., N_5 store the coordinates of the grouped points together with optional pointers to their corresponding records. Each intermediate node contains the Minimum Bounding Rectangle (MBR) of the sub-tree of the nodes. For example, node e_1 corresponds to MBR N_1, which covers the points, p_1, p_2, and p_3. Similarly, node e_6 and node e_7 correspond to MBR N_6 and MBR N_7, respectively.

D. Voronoi Diagram

Let P is the set of n distinct data points on the plane. The Voronoi diagram of P is the subdivision of the plane into n cells. Each cell contains only one point of P, which is called the Voronoi point of the cell. In this paper, we denote $V(p_j)$ as a cell of a Voronoi point p_j, $p_j \in P$, and $VN(p_j)$ as a set of cells that are adjacent to $V(p_j)$. Assume that P contains fourteen data points $\{p_1, p_2, \cdots, p_{14}\}$ and two query points q_1 and q_2. Figure 4 shows the Voronoi diagram of the points in P. We can say that a query point is nearest to a data point if the query point is within Voronoi cell of the data point. As for example, from the Voronoi diagram of Figure 4, we can find that the nearest Voronoi point of the query point q_1 is p_8, since q_1 is within the Voronoi cell of p_8. Similarly, the nearest Voronoi point of query point q_2 is p_1.

Voronoi diagram provides an efficient data structure to compute the nearest Voronoi point for a given query point q. We use Fortune's algorithm [22] to construct Voronoi diagram for a set of points. Fortune's algorithm is a sweep line algorithm for generating a Voronoi diagram from a set of points in a plane. Though the worst time complexity for constructing Voronoi diagram for a set of n points using Fortune's algorithm is $O(n^2)$, the expected time complexity is $O(n \log n)$.

E. VoR-Tree

A VoR-tree [23] is a variation of R-tree that index the data points using the concepts of Voronoi diagram and R-tree. Each leaf node stores a subset of data points. Each leaf node also includes the data records containing extra information about the corresponding points. In the record of a data point p_j in a VoR-tree, we store the pointer to the location of Voronoi neighbors $VN(p_j)$ and the vertices of $V(p_j)$, i.e., vertices of the Voronoi cell of p_j. Here, a vertex represents a common endpoint of two edges of a Voronoi cell.

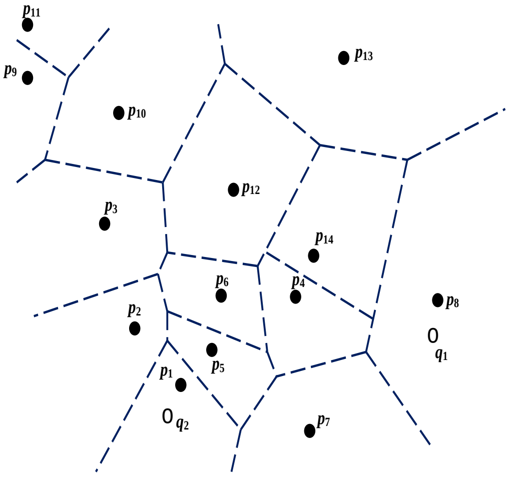


Figure 4. Example of a Voronoi diagram

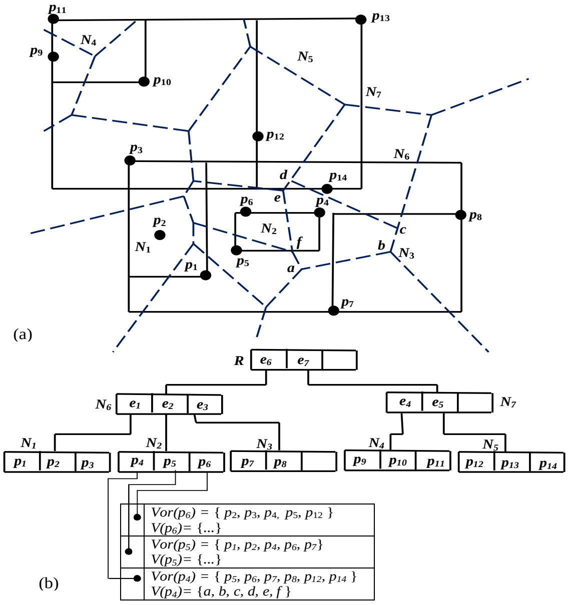


Figure 5. (a) Voronoi diagram (b) VoR-tree (adapted from [22])

For constructing VoR-tree, at first, we index the data points using an R-tree. Then, we use the Voronoi diagram of the data points to find the Voronoi neighbors and vertices of a Voronoi cell for each data point p_j. Next, we store both information as a record associated with each data point p_j. Each Voronoi neighbor of p_j in this record is a pointer to the disk block storing the information of that Voronoi neighbor. A disk block also known as a sector is a sequence of bytes for storing and retrieving data.

Figure 5(b) shows an example of VoR-tree for the data points of Figure 3. Each rectangular in Figure 5 is a node of the VoR-tree. In Figure 5, rectangular N_2 contains three points, i.e., p_4, p_5, and p_6. N_2 and two other rectangular boxes N_1 and N_3 are contained by the parent, which is the rectangular N_6. For simplicity, we show only the contents of the records of the data points of node N_2. From Figure 5 (b), we can see that data point p_5, p_6, p_7, p_8, p_{12}, and p_{14} are Voronoi neighbors of p_4 and its Voronoi cell has vertices a, b, c, d, e, and f.

Since the expected time complexity for constructing a Voronoi diagram using Fortune's algorithm is $O(n \log n)$, we can expect to construct the VoR-tree with a time-complexity very close to $O(n \log n)$. Since the locations of spatial objects, such as restaurants, are static, we can construct VoR-tree before processing the groups' skyline query.

VoR-tree provides us an efficient way to search non-dominated objects in spatial sub-space, since we can find the nearest spatial object in VoR-tree from a given query point in $O(\log n)$ time. Using VoR-tree, we can significantly reduce the search space that dramatically improves the performance of our query. We give detail explanation of how VoR-tree improves our query performance in subsection IV-A.

IV. Query Processing

It is possible to calculate skyline query after constructing a table like Table IV by conventional skyline queries. However, the number of data points such as restaurants is too large that the construction of a table like Table IV and computation of skyline result from such a table using any conventional skyline query algorithm are not affordable.

Considering this fact, in this paper, we compute the skyline results in two phases.

In the first phase, we compute skyline results in the spatial sub-space like ($r - u_1$, $r - u_2$, $r - u_3$, $r - u_4$, $Sum\text{-}Distance$) of Table IV. We utilize the concept of $Sum\text{-}Distance$ for spatial processing which can easily eliminate a large number of objects during the computation of skyline objects in the spatial sub-space.

Based on the skyline result of the spatial sub-space, the second phase efficiently computes whether some other objects can be in the skyline in the non-spatial sub-space like ($Rating$, $Price$) of Table IV. In this phase, we check the dominance of non-skyline objects of spatial sub-space against the skyline objects of spatial sub-space. Such an approach can easily eliminate many objects from domination check.

A. Spatial Processing

We say that an object is "spatially dominated" if the object is dominated in the spatial sub-space. For example, we can say that a restaurant in Table III is "spatially dominated", if the restaurant is dominated in its sub-space $\{r\text{-}u_1, r\text{-}u_2, r\text{-}u_3, r\text{-}u_4, Sum\text{-}Distance\}$.

For selecting non-dominated objects in spatial sub-space, at first, we select the Voronoi point (restaurant) that is nearest to the centroid of the query points (user locations). For example, if we consider the users (query points) of Table II, we can find that the centroid of r-u_1, r-u_2, r-u_3, and r-u_4 is (5.13, 5. 28). From Table I, we can find that r_j is nearest to (5.13, 5. 28). So, we select r_7. Next, for each of the user, we draw a circle. The radius of each circle is the Euclidean distance from the user and r_j. Let $C(u_i, r_j)$ be a circle whose center is the position of user u_i. The radius of $C(u_i, r_j)$ is the Euclidean distance from u_i to data point r_j. We denote this distance by $D(u_i, r_j)$. We call the region within the union of the circles of r_j as the "search region" of r_j.

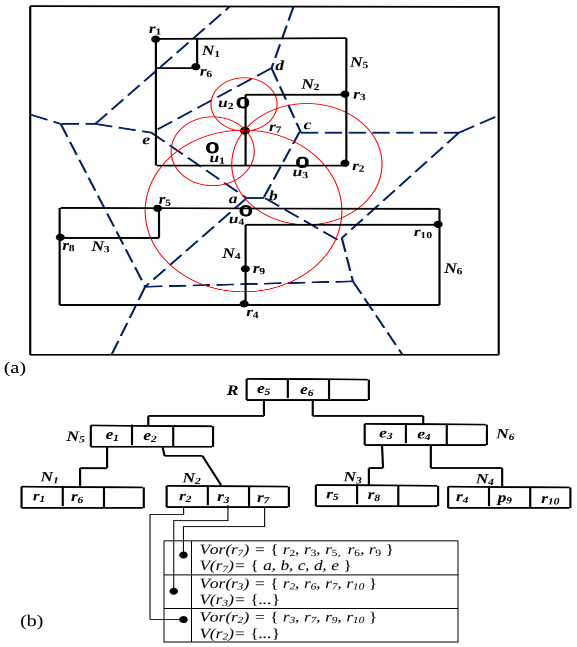


Figure 6. (a) Location of users and restaurants (b) VoR-tree

We, then, search for the data points within the "search region". To obtain the data points within the "search region", we just consider the Voronoi cells those are either completely inside the "search region" or those have some intersections with any of the circles. If a Voronoi cell is completely inside the search region, we can say that corresponding data point is within the "search region". If a Voronoi cell intersects with any of the circles, we need to check the distance of the corresponding data point from the center of the circles. If we find that the Euclidean distance is less than or equal to the radius of any of the circles, we can decide that the data point is inside the "search region". Otherwise, it is outside the "search region".

Later, we compute the sum of Euclidean distances of a data point (restaurant) from the query points (users). We call this distance "Sum Distance".

We can efficiently compute the set of objects those are not spatially dominated using "search region", "Sum Distance" and VoR-tree that incrementally returns the skyline points as explain below.

First, we compute the sum of Euclidean distances for each data point within the "search region". Then, we pick the data point, say r_k that has minimum "Sum Distance" and add r_k along with its "Sum Distance" to a heap. Next, we examine the Voronoi neighbours of r_k, $VN(r_k)$ and add the Voronoi neighbors within the search region in the heap in increasing order of their "Sum Distance". When

TABLE V.
SPATIAL INFORMATION OF THE DATA POINTS WITHIN SEARCH REGION

ID	r-u_1	r-u_2	r-u_3	r-u_4	Sum-$Distance$
$\mathbf{r_2}$	2.55	2.69	1	2.33	8.57
$\mathbf{r_5}$	1.58	2.97	2.24	1.02	7.81
$\mathbf{r_7}$	0.71	0.89	1.41	1.48	4.49
$\mathbf{r_9}$	2.54	3.8	2.24	0.8	9.38

TABLE VI.
HEAP FOR TRAVERSING VoR-TREE

Step	Heap content	Skyline S
1	(r_7, 4.49)	$\oslash$
2	(r_7, 4.49), (r_5, 7.81), (r_2, 8.57), (r_9, 9.38)	$\oslash$
3	(r_5, 7.81), (r_2, 8.57), (r_9, 9.38)	$\{r_7\}$
4	(r_2, 8.57), (r_9, 9.38)	$\{r_7, r_5\}$
5	(r_9, 9.38)	$\{r_7, r_5, r_2\}$
6	$\oslash$	$\{r_7, r_5, r_2, r_9\}$

a data point r_k is explored, we pop it from the heap and add it to the skyline list if it is not dominated in spatial sub-space by some other objects already in the skyline. We continue the process until the heap becomes empty.

Now, consider the computation process of skyline objects in spatial sub-space from the example as shown in Figure 6. In the Figure 6(a), white dots are locations of four users and black dots are locations of restaurants. We first pick up r_7 and compute $C(u_i, r_7)$ for each user u_i ($i = 1, ..., 4$) to get the "search region". We, then, find that restaurants r_2, r_5, r_7, and r_9 are within the "search region" of r_7. Next, we compute the "Sum Distance" for each of these restaurants and construct the table as shown in Table V. In the process, we keep the heap data structure like Table VI.

Looking at the information of Table V, we can find that returant r_7 has minimum "Sum Distance". So, we add $(r_7, dist(r_7, U))$ to the heap and marks r_7 as "checked". Next, we collect the Voronoi neighbors of r_7 and find that its Voronoi neighbors r_2, r_5, and r_9 are inside the "search region" (union of $C(u_i, r_7)$ for user u_i ($i = 1, ..., 4$)). Then, we add $(r_2, dist(r_2, U))$, $(r_5, dist(r_5, U))$ and $(r_9, dist(r_9, U))$, to the heap in ascending order of their "Sum Distance".

After the steps, restaurant r_7 is added to the skyline list S as shown in step-3 of Table VI. Next, we pick the top element r_5 from the heap and find that its Voronoi neighbours are r_1, r_6, r_7, r_8 and r_9. Among them r_1, r_6 and r_8 are outside the search region and r_7 and r_9 are already checked. Therefore, no new entry is added in the heap by r_5. After that, we examine the spatial dominance of r_5 against r_7. Since r_5 is not spatially dominated by r_7, we add r_5 in S as in step-4. Similarly, we continue the process and add r_2 and r_9 to the skyline. After the process of r_9, the heap becomes empty. Finally, we get $S = \{r_2, r_5, r_7, r_9\}$ as skyline result based on spatial sub-space.

TABLE VII.
NON-SPATIAL INFORMATION OF DOMINATED OBJECTS IN SPATIAL SUB-SPACE

ID	Rating	Price
$\mathbf{r_1}$	3	2
$\mathbf{r_3}$	3	4
$\mathbf{r_4}$	3	2
$\mathbf{r_6}$	3	3
$\mathbf{r_8}$	3	2
$\mathbf{r_{10}}$	1	1

TABLE VIII.
NON-SPATIAL INFORMATION OF THE SKYLINE OBJECTS IN SPATIAL SUB-SPACE

ID	Rating	Price
$\mathbf{r_2}$	2	2
$\mathbf{r_5}$	2	3
$\mathbf{r_7}$	3	1
$\mathbf{r_9}$	2	2

Since the "search region" is relatively very small compared with the whole space, such computation is very much efficient with respect to space and time.

B. Non-spatial Processing

In non-spatial processing, at first, we collect all dominated data points at spatial sub-space. Table VII shows such data points with non-spatial information. From Table VII, we can see that data points r_1, r_3, r_4, r_6, r_8, and r_{10} are spatially dominated. So, we need to check their dominance in the non-spatial sub-space.

To obtain non-dominated objects at non-spatial sub-space, we check their dominance against the skyline objects r_2, r_5, r_7, and r_9 of spatial sub-space. Table VIII shows non-spatial information of these skyline objects in spatial sub-space. Note that objects of Table VIII are in the final skyline as well.

If we check the objects of Table VII against the objects of Table VIII, we can find that r_7 also dominates r_1, r_3, r_4, r_6, r_8 in non-spatial sub-space. So, they are not in the skyline. However, object r_{10} is not dominated in its non-spatial sub-subspace by any object of Table VIII and there is no other non-dominated object in Table VII. So, r_{10} is also in the skyline. Finally, we find r_2, r_5, r_7, r_9 and r_{10} as final skyline result.

Algorithm 1 shows the proposed computation procedure of the spatial skyline queries. It first computes "spatially dominated" objects based on spatial sub-space (line 3-20). Then, **Algorithm 1** computes whether there are skyline objects among the "spatially dominated" objects by examining non-spatial sub-space (line 21-29). Finally, the algorithm returns the spatial skyline objects (line 30).

C. Correctness of Algorithm

The correctness of **Algorithm 1** follows some basic properties of geometry and skyline query. From **Algorithm 1**, we can see that for a set of query points Q, it first adds the data point r_j with minimum "Sum Distance" to the skyline S. All the Voronoi neighbors of r_j are

Algorithm 1 Computation

```
Input: Set of query points $U = \{u_1, u_2, \cdots, u_i\}$ and data points $R = \{r_1, r_2, \cdots, r_j\}$
Output: Spatial skyline objects Set $S$, $S \subseteq R$
 1: begin
 2: set $D, (D \subseteq R)$ = the set of dominated objects in spatial sub-space
 3: select a data point $r_j$ that is closest to the centroid of the query points $U = \{u_1, u_2, \cdots, u_i\}$
 4: compute the search region of $r_j$
 5: obtain the data points set, say $T$ within the “search region“, $T \subseteq R$
 6: compute the “Sum Distance” $dist_k$ of each data point $r_k$, $r_k \in T$
 7: select the data point $r_k$ that has minimum “Sum Distance”
 8: add ($r_k$, $dist_k$ ) to the heap $H$
 9: select the Voronoi neighbors of $r_k$ those are within the “search region” and add
    them to $H$ in increasing order of their “Sum Distance”
10: remove ($r_k$, $dist_k$ ) from $H$ and add $r_k$ to $S$
11: repeat
12:    choose the top element, say $r_l$ from $H$
13:    select the Voronoi neighbors of $r_l$ those are within the “search region” and add
       them to $H$ in increasing order of their “Sum Distance”
14:    pop $(r_l, dist_l)$ from $H$
15:    if $r_l$ is not dominated by some other objects in $S$ in spatial sub-space then
16:       add $r_l$ to $S$
17:    else
18:       add $r_l$ to $D$
19:    end if
20: until $H$ becomes empty
21: for each data point $r_m \in D$ do
22:    if $r_m$ is dominated by some other objects of $S$ in non-spatial sub-space then
23:       $r_m \notin S$
24:    else if $r_m$ is dominated by some other objects of $D$ in non-spatial sub-space then
25:       $r_m \notin S$
26:    else
27:       add $r_m$ to $S$
28:    end if
29: end for
30: return $S$ as the spatial skyline result
31: end
```

TABLE IX.
DATASETS FOR EXPERIMENTS

Datasets	Total Objects	Density
r	50,747	–
$\mathbf{s_1}$	80,000	0.08
$\mathbf{s_2}$	50,000	0.05
$\mathbf{s_3}$	20,000	0.02

then checked and added to the heap in increasing order of their their “Sum Distance” if they are within the “search region”.

The traversal started from the data point with minimum “Sum Distance” towards the Voronoi neighbors in increasing order of “Sum Distance” and we can find that the data point r_j with minimum “Sum Distance” is in the skyline S. The reason is that “Sum Distance” is considered as an attribute in the spatial sub-space. During the consideration of Voronoi neighbors of a data point, we just consider the Voronoi neighbors within the “search region”. We can easily ignore the Voronoi neighbors of a data point those are outside the “search region”. This is because, the Euclidean distances between a Voronoi neighbor that is outside the “search region” and query points must be larger than the Eucledian distances between r_j and query points. Hence, any Voronoi neighbor that is outside the “search region” will never be in the skyline in the spatial sub-space. However, the Voronoi neighbors those are within the “search region” can be in the skyline of spatial sub-space. So, **Algorithm 1** further checks such Voronoi neighbors against the data points in S to determine whether they are in the skyline of the spatial sub-space or not.

Line 21-29 of **Algorithm 1** shows the computation of skyline objects in non-spatial sub-space. The correctness of **Algorithm 1** for computing skyline objects in non-spatial sub-space comes from the basic idea of skyline. If an object is in the skyline of $d - i$ (i = 1 to d -1) dimensions, it will also be in the skyline of d dimensions.

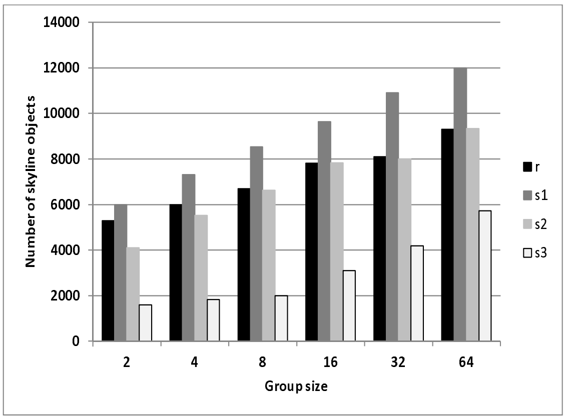


Figure 7. Number of skyline objects

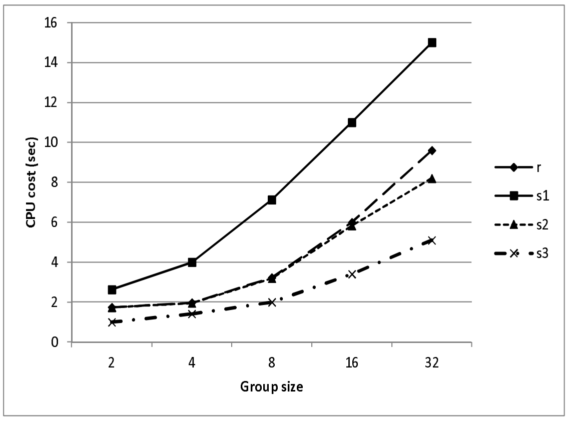


Figure 8. Running time varying group size

V. Performance Evaluation

To evaluate the efficiency and effectiveness of the proposed skyline queries algorithm, we conducted extensive experiments. We implemented all algorithms using Microsoft Visual C++ V6.0, and conducted the experiments on a PC with Intel core i5 processor, 2.3 GHz CPU, 4G main memory and 200G hard disk, running Microsoft Windows 7 Professional Edition. Our developed system is able to handle large volume of data containing both spatial and non-spatial information.

A. Experimental Setup

We implemented the experiments by deploying both real and synthetic datasets. The real datasets came from line segment data of Long Beach from the TIGER database [24]. We made this point set by extracting the midpoint for each road line segment. The set consists of 50,747 points normalized in [0,1000] × [0, 1000] space. There are three synthetic datasets $s1$, $s2$, and $s3$ with different densities normalized in [0,1000] × [0,1000] space as in Table IX. In Table IX, r stands for real dataset of TIGER database and density means how many points fall into one square unit in average. The points in each synthetic dataset are distributed randomly. We indexed all datasets by using a VoR-tree. By default, we consider a location attribute and two category attributes for each data set.

B. Experimental Results

The first experiment studies the numbers of skyline objects under different densities and different group size. Figure 7 shows the total numbers of skyline objects from datasets r, s_1, s_2, and s_3. From Figure 7, we can see that total number of skyline objects increases with the increase in density and group size.

The second experiment explores the performance of the algorithm under different group size and different densities. From Figure 8, we can observe that the running time increases with the increase in group size. Also, it is observed that running time increases if the density of data points increases.

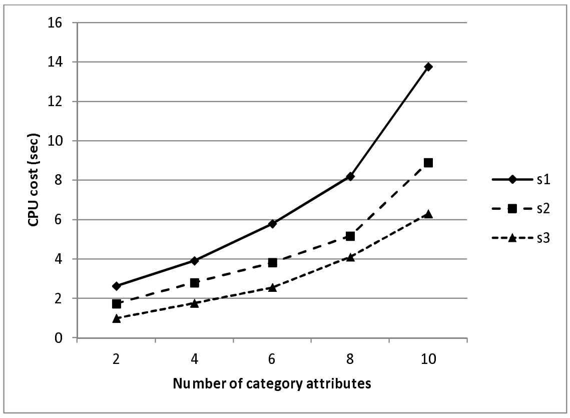


Figure 9. Running time varying number of category attributes

Next experiment shows the effect of the increase in the number of category attributes while keeping the group size to 32. In this experiment, we considered three synthetic datasets. Figure 9 shows result. From the result, we can see that there is an increase in computation time with the increase in the number of category attributes.

In the fourth experiment, we compared our algorithm with BBS approach using the dataset r. Although there are some other spatial skyline query algorithms, we considered BBS algorithm for comparison due to its effectiveness in handling both spatial and non-spatial attributes. From the result of Figure 10, we can see that our algorithm (VR) significantly outperforms BBS algorithm.

Next experimental results are shown in Figure 11. It shows the relative dominance check between our algorithm and BBS algorithm. From Figure 11, we can see that our algorithm constantly performs less number of dominance check compared with BBS algorithm.

Figure 12 shows the results of our sixth experiment. It shows the effectiveness of our algorithm while there is an increase in the number of category attributes. In this experiment, we considered the synthetic dataset s_1 and group size 2. From the result of Figure 12, we can see that in case of fixed number of users and more category attributes, the performance of our algorithm is still better

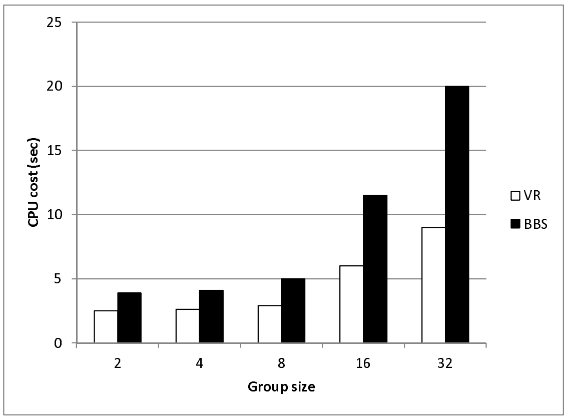


Figure 10. Comparative performance in running time varying group size

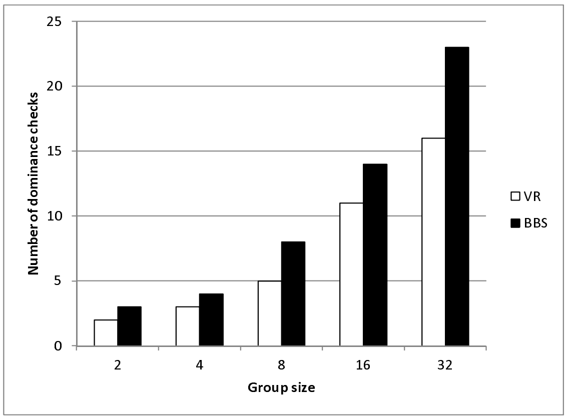


Figure 11. Comparative performance in dominance check varying group size

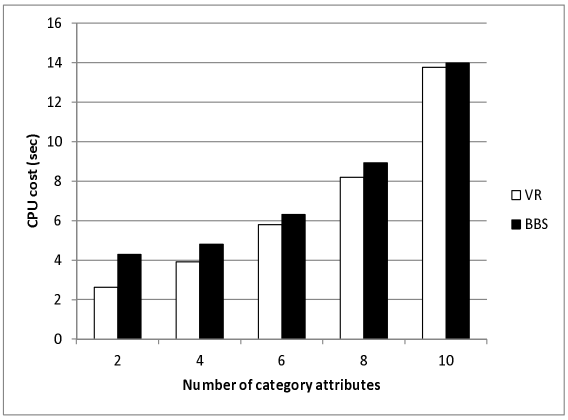


Figure 12. Comparative performance in running time varying number of category attributes

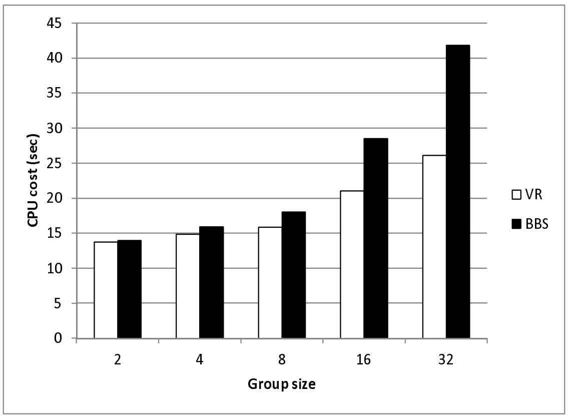


Figure 13. Comparison in running time varying group size for large number of category attributes

than BBS algorithm.

The final experiment shows the effectiveness of our algorithm in case of large number of category attributes while there is an increase in group size. In this experiment, we considered ten category attributes. From the result of Figure 13, we can find that our algorithm becomes comparatively better than BBS algorithm with an increase in group size.

VI. Conclusion

In this paper, we proposed a framework for computing skyline of spatial objects for a group of users located at different locations. In the proposed framework, different from existing works, we took into account not only spatial features, but also non-spatial features of the objects.

Recently, many social network services create groups considering users located in different places. Spatial skyline queries for a group can be able to play an important role in such environments.

In our computation framework, we utilized VoR-tree and "Sum Distance" to calculate spatial skyline objects for a group of users of different locations efficiently. Experimental results demonstrate that the proposed algorithm is scalable enough to handle large and high dimensional datasets.

In this paper, we have considered static query points, which mean all query points do not move. However, in general, query points are not static. Therefore, we have to develop an efficient algorithm that can handle the change in the locations of query points in our future works.

References

[1] H. T. Kung, F. Luccio, and F. Preparata, "On finding the maxima of a set of vectors", *Journal of the Association for Computing Machinery*, vol. 22 no. 4, pp. 469-476, 1975.

[2] S. Borzonyi, D. Kossmann, and K. Stocker, "The skyline operator", *In Proc. of the 17th International Conference on Data Engineering*, pp. 421-430, 2001.

[3] K. L. Tan, P. K. Eng, and B. C. Ooi, "Efficient progressive skyline computation", *In Proc. of the 27th International Conference on Very Large Data Bases*, pp. 301-310, 2001.

[4] D. Kossmann, F. Ramsak, and S. Rost, "Shooting stars in the sky: An online algorithm for skyline queries", *In Proc. of the 28th International Conference on Very Large Data Bases*, pp. 275-286, 2002.

[5] D.Papadias, Y. Tao, G. Fu, and B. Seeger, "An optimal and progressive algorithm for skyline queries", *In Proc. of ACM SIGMOD International Conference on Management of Data*, pp. 467-478, 2003.

[6] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang, "Skyline with presorting", *In Proc. of the 19th International Conference on Data Engineering*, pp. 717-719, 2003.

[7] C. Y. Chan, H. Jagadish, K. Tan, and A. K. Tung, Z. Zhang, "On high dimensional skylines", *In Proc. of LNCS, Springer, Heidelberg*, pp. 478-495, 2006.

[8] C. Y. Chan, H. Jagadish, K. L. Tan, and A. K. Tung, Z. Zhang, "Finding k-dominant skylines in high dimensional space", *In Proc. of ACM SIGMOD International Conference on Management of Data*, pp. 444-457, 2006.

[9] X. Lin, Y. Yuan, Q. Zhang, and Y. Zhang, "Selecting stars: The k most representative skyline operator", *In Proc. of the 23rd International Conference on Data Engineering*, pp. 86-95, 2007.

[10] N. Roussopoulos, S. Kelley, and F. Vincent, "Nearest neighbor queries", *In Proc. of the ACM SIGMOD International Conference on Management of Data* , pp. 71-79, 1995.

[11] S. Berchtold, C. Bohm, D. A. Keim, and H. P. Kriegel, "A cost model for nearest neighbor search in high-dimensional data space", *In Proc. of the 16th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* , pp. 78-86, 1997.

[12] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?", *In LNCS, Springer, Heidelberg*, vol. 1540, pp. 217-235, 1999.

[13] D. Papadias, Y. Tao, K. Mouratidis, and C. K. Hui, "Aggregate nearest neighbor queries in spatial databases", *ACM Transactions on Database Systems*, vol. 30 no. 2, pp. 529-576, 2005.

[14] M. Sharifzadeh, and C.Shahabi, "The spatial skyline queries", *In Proc. of the 32nd International Conference on Very Large Data Bases*, pp. 751-762, 2006.

[15] W. Son, M. Lee, H. Ahn, and S. Hwang, "Spatial skyline queries: an efficient geometric algorithm", *In Proc. of 11th International Symposium on Spatial and Temporal Databases*, pp. 247-264, 2009.

[16] X. Guo, Y. Ishikawa, and Y. Gao, "Direction-based spatial skylines", *In Proc. of the ACM SIGMOD International Conference on Management of Data*, pp. 73-80, 2010.

[17] K. Kodama, Y. Iijima, X. Guo, and Y Ishikawa, "Skyline queries based on user locations and preferences for making location-based recommendations" *Proc. of International Workshop on Location Based Social Networks*, pp. 9-16, 2009.

[18] K. Deng, X. Zhou, and H. T. Shen, "Multi-source skyline query processing in road networks", *In Proc. of 23rd International Conference on Data Engineering*, pp. 796-805, 2007.

[19] M. Safar, D. E. Amin, and D. Taniar, "Optimized skyline queries on road networks using nearest neighbors", *Journal of Personal and Ubiquitous Computing*, vol. 15 no. 8, pp.845-856, 2011.

[20] Y. K. Huang, C. H. Chang, and C. Lee, "Continuous distance-based skyline queries in road networks" *Journal of Information Systems*, vol. 37, pp. 611-633, 2006.

[21] B. Zhang, K. C. K. Lee, and W. C. Lee, "Location-dependent skyline query", *In Proc. of 9th International Conference on Mobile Data Management*, pp. 3-8, 2008.

[22] S. Fortune, "A sweep line algorithm for Voronoi diagrams", *In Proc. of the Second Annual Symposium on Computational geometry*, pp. 313-322, 1986.

[23] M. Sharifzadeh, and C. Shahabi, "VoR-Tree: R-trees with Voronoi diagrams for efficient processing of spatial nearest neighbor queries", *In Proc. of the 36th International Conference on Very Large Data Bases*, pp. 1231-1242, 2010.

[24] Tiger. Available at: http://tiger.census.gov/

Mohammad Shamsul Arefin received his B.Sc. Engineering in Computer Science and Engineering from Khulna University, Khulna, Bangladesh in 2002, and completed his M.Sc. Engineering in Computer Science and Engineering in 2008 from Bangladesh University of Engineering and Technology (BUET), Bangladesh. He received his Doctor of Engineering Degree from Hiroshima University with support of the scholarship of MEXT, Japan in 2013. He is a member of Institution of Engineers Bangladesh (IEB) and currently working as an Associate Professor in the Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong, Bangladesh. His research interest includes privacy preserving data mining, cloud privacy, multilingual data management, semantic web, and object oriented system development.

Geng Ma received his M.Sc in Information Engineering from Hiroshima University, Japan in 2013. His current research interest includes spatial databases, preference-based queries, and recommendation systems.

Yasuhiko Morimoto is an Associate Professor at Hiroshima University. He received B.E., M.E., and Ph.D. from Hiroshima University in 1989, 1991, and 2002, respectively. From 1991 to 2002, he had been with IBM Tokyo Research Laboratory where he worked for data mining project and multimedia database project. Since 2002, he has been with Hiroshima University. His current research interests include data mining, machine learning, geographic information system, and privacy preserving information retrieval.

An Effective Method to Solve Flexible Job-shop Scheduling Based on Cloud Model

Xiaobing Liu and Xuan Jiao
School of Management, Dalian University of Technology, Dalian 116024,China
jxis9ms@163.com

Tao Ning and Ming Huang
Institute of Software, Dalian Jiaotong University, Dalian 116045,China
daliannt@126.com

Abstract—**In order to solve the problem of flexible job-shop scheduling, this paper proposed a novel quantum genetic algorithm based on cloud model. Firstly, a simulation model was established aiming at minimizing the completion time, the penalty and the total cost. Secondly, the method of double chains structure coding including machine allocation chain and process chain was proposed. The crossover operator and mutation operator were obtained by the cloud model X condition generator because of its randomness and stable tendency. The non dominated sorting strategy was introduced to obtain more optimal solution. Finally, the novel method was applied to the Kacem example and a mechanical mould scheduling, the simulation results demonstrated that the proposed method can reduce the precocious probability and obtain more non dominated solutions comparing with the existing algorithms.**

Index Terms—**Flexible job-shop scheduling, Cloud model, Quantum genetic algorithm, Double chains structure coding**

I. Introduction

Flexible Job-shop Scheduling Problem (FJSS) is the extension of Job-shop Scheduling Problem (JSS). The domestic and foreign scholars have studied FJSS with various methods and achieved corresponding results[1].Bucker P.and Schlie R. [2] proposed FJSS in 1990, then the research hotspot about FJSS focused on the application of genetic algorithm and other intelligent algorithms. Chen H.[3] used genetic algorithm (GA) to solve FJSS aiming at minimizing the completion time, and simulated the chromosome with graph theory whose coding constituted with the routing and the process. Ho N.B.[4] proposed an optimization algorithm of three layer structure to solve FJSS. Najid N.M.[5] used simulated annealing algorithm(SAA) integrating neighbor function to minimize the maximum completion time of FJSS. Kacem I.[6] solved single objective and multi-objective FJSS respectively. He solved the machine allocation problem with the local search method and constructed the initial population firstly, and then improved the quality of solution with the optimization. D.Y.Sha[7] solved FJSS through updating the speed of particle swarm optimization(PSO) and combined tuba search algorithm(TS). B.Liu[8] used PSO based on genetic algorithm for permutation flow shop scheduling on the basis of the combination of PSO operator and local search operator. K.Fan[9] designed a novel algorithm to improve the binary PSO and obtained the approximate optimal solution. XIA W.J.[10] used the integration of PSO and SAA to solve FJSS. He solved the machine allocation with PSO and process scheduling with SAA. YU X.Y.[11] proposed multi workshop planning and scheduling based on the parallel cooperative evolutionary genetic algorithm. LIU A.J.[12] proposed multi-objective FJSS algorithm based on a multi population genetic algorithm by introducing fuzzy number to describe the completion time and delivery. ZHANG J[13] proposed the particle position update algorithm directly in the discrete domain on the basis of sequence and machine allocation. SHI J.F.[14] used continuous space ant colony algorithm to optimize the multi constraints of FJSS through establishing the simulation model of flexible routing.

There will be various shortcomings when the above methods are used such as low search efficiency, the weak ability of local search and premature convergence because of the loss of population diversity in later period. Considering the randomness and stable tendency of cloud droplets in the cloud model may improve the crossover operator and mutation operator of the adaptive genetic algorithm, this paper proposed a novel quantum genetic algorithm based on the cloud model to improve the convergence and robustness. The coding method of double chains was used on the basis of initializing the machine distribution chain with quasi level uniform design and heuristic initializing the process chain. The crossover operator and mutation operator were generated by the cloud model X condition generator, and the new population was obtained through rotation angle of quantum gates. The non dominated sorting strategy was introduced based on the fuzzy set theory. Finally, the proposed method is verified to be effective through the application to Kacem instances and the comparison with the existing algorithms.

doi:10.4304/jsw.9.11.2948-2954

II. Model Of Multi-Objective FJSS

A. Problem Description

FJSS is described as follows: there are N workpieces to be processed and M machines in workshop, each workpiece $i(i\in\{1,2,\ldots,N\})$ includes $n_i(n_i\geq 1)$processes, and the process should be processed with the specified route. R_{ij} means the j^{th} $(j\in\{1,2,\ldots, n_i\})$ process of workpiece i, M_{ij} $(M_{ij}\subseteq\{1,2,\ldots, M\})$ means the machine set, each R_{ij} may be processed by any machine m $(m\in\{1, 2,\ldots,M_{ij}\})$ with processing capacity, and m can process different workpieces [13]. The performance of different machines m makes the completion time different for R_{ij}.

B. Objective Function

The objective of FJSS is to select the suitable machine for each process and determine the optimum processing sequence, the objective function is established as follows:

1) To minimize the maximum completion time:

$$f1=\min(F)=\min[\max(\sum_{m=1}^{M}F_m)] \qquad (1)$$

$$F_m=\sum_{i=1}^{N}\sum_{j=1}^{n_i}(S_{ijm}b_{ijm}+S_{ijm}t_{ijm}) \qquad (2)$$

In formula(1), F means the total completion time of all the machines, which acts as an important index to measure the machine load. In formula(2),F_m means the total completion time of machine m, b_{ijm} means the start time of R_{ij} in m, t_{ijm} means the processing time of R_{ij} in m, S_{ijm} takes the value of either 1(processed in machine m) or 0(not).

2) To minimize total cost:

$$f2=\min(C)=\min[\sum_{i=1}^{N}(M_i+\sum_{j=1}^{n_i}\sum_{m=1}^{M}C_{ijm}S_{ijm})] \qquad (3)$$

$$C_{ijm}=(\mu_{ijm}+\nu_{ijm}) \qquad (4)$$

In formula(3), C means the total costs of workpiece i, M_i means the commodity cost of workpiece i, C_{ijm} means the processing cost of R_{ij} in m. In formula(4), μ_{ijm} and ν_{ijm} mean the labor cost and machine cost of R_{ij} in m respectively.

3) To minimize penalty:

$$f3=\min(P)=\min\{\sum_{i=1}^{N}[pe_i\max((d_i-t_i),0)+pl_i\max((t_i-d_i),0)]\} \qquad (5)$$

In formula(5), pe_i and pl_i mean the earliness penalty and tardiness penalty respectively, t_i and d_i mean the completion time and delivery for workpiece i.

4) To maximize the satisfaction:

$$f4=\max[\frac{1}{N}\sum_{i=1}^{N}gI_i(\tilde{t}_i)] \qquad (6)$$

In formula(6),$gI_i()$ means the satisfaction function for the customer to the completion time of workpiece i. Satisfaction is one of the important indexes to evaluate the fuzzy scheduling, its value depends on the fuzzy completion time $\tilde{t}_i$,$gI_i(\tilde{t}_i)=(\tilde{t}_i\wedge\tilde{D}_i)/\tilde{t}_i$, $\tilde{D}_i$ means the fuzzy delivery of workpiece i. In actual production, it is uncertain for the processing time and completion moment, and these factors will change in a certain interval to delivery.

C. Constraint Conditions

1) Process constraint

The sequence constraint of different processes in the same workpiece, where $S_{ijm}= S_{i(j-1)m}=1$:

$$\sum_{m=1}^{M}b_{ijm}S_{ijm}\geq\sum_{m=1}^{M}[(b_{i(j-1)m}t_{i(j-1)m})]S_{i(j-1)m} \qquad (7)$$

2) Machine constraint

The same machine can only do one process at the same time, that is to say, if $\exists S_{ijm}=1$at the moment of t, then there mustn't be S_{xym}=1.

3) Continuity constraint

R_{ij} can't be interrupted in the processing, in formula(8), c_{ijm} means the completion time of R_{ij}.

$$c_{ijm}=\begin{cases}\max\{c_{i(j-1)m},b_{ijm}\}+t_{ijm}, & j>1;\\ b_{ijm}+t_{ijm}, & j=1.\end{cases} \qquad (8)$$

III. Improved Algorithm Based On Cloud Model

A. Concept of Cloud Model

The concept of cloud model was proposed by professor Li Deyi[15] on the basis of probability theory and fuzzy mathematics in 1995.It is a conversion model between qualitative and quantitative concept through a specific algorithm, and reveals the inherent relationship between randomness and fuzziness. During less than 20 years since the cloud model was proposed, it has been applied to many fields such as intelligent control, data mining and decision analysis successfully.

The definition of the cloud and the cloud droplet[15]: Assuming that T is a fuzzy subset on domain U, the mapping $C_T(x)$ *which is from U to* [0,1] is a random number with stable tendency, that is $\forall$ $x\in U$, $x\rightarrow C_T(x)$, and the distribution of $C_T(x)$ in U is called the membership cloud of T, or the cloud model of T, and then each variable x is called a cloud droplet. $\forall$ $x\in U$, $C_T(x)$ is not a clear membership curve but consists of a large number of cloud droplets. When the mapping $C_T(x)$ follows normal distribution, it is called the cloud model of normal distribution.

The cloud model describes some qualitative concept with three digital features including expected value E_x, entropy E_n and hyper entropy H_e. E_x is the expectation of distribution for the cloud droplet in domain. E_n is the uncertainty measure of the qualitative concept, H_e which can be also call entropy's entropy is the uncertainty measure of E_n.

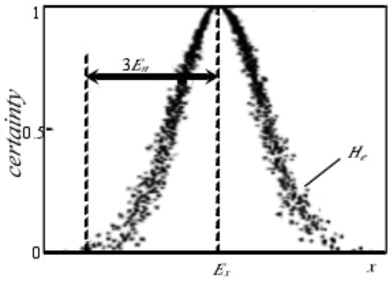


Figure 1. Digital feature of normal cloud model

The algorithm of normal cloud generator[16] is as follows:

Step 1: Generate a normal random number E_n' taking E_n as expectation, H_e as standard deviation according to the three digital features (E_x,E_n,H_e);

Step 2: Generate a normal random number x taking E_x as expectation, |E_n| as standard deviation, x is a cloud droplet in domain U, $Drop(x_i,u_i)$;

Step 3: Calculate the certainty $\mu = e^{-\frac{(x-E_x)^2}{2(E_n')^2}}$ according to Step 1 and Step 2;

Step 4: Repeat Step 1 to 3 until generate N cloud droplets.

B. Improved Algorithm

1) Double chains quantum coding

This paper introduced a novel compensation factor $\gamma(\gamma \geq 1)$ based on probability coding. Assuming that p_i means a quantum chromosome, the encoding scheme of the i^{th} chromosome is as follows:

$$p_i = \left[\begin{vmatrix}\alpha_{i1}\\ \beta_{i1}\end{vmatrix}\begin{vmatrix}\alpha_{i2}\\ \beta_{i2}\end{vmatrix}\cdots\begin{vmatrix}\alpha_{im}\\ \beta_{im}\end{vmatrix}\right] = \left[\begin{vmatrix}\cos(\gamma t_{i1})\\ \sin(\gamma t_{i1})\end{vmatrix}\begin{vmatrix}\cos(\gamma t_{i2})\\ \sin(\gamma t_{i2})\end{vmatrix}\cdots\begin{vmatrix}\cos(\gamma t_{im})\\ \sin(\gamma t_{im})\end{vmatrix}\right] \tag{9}$$

Where $t_{ij} = 2\pi \times rad$,rad means a random number between (0,1), $i = 1, 2, \cdots, n$; j=1,2,…,m; n means the population size, and m means the number of qubits. γ extends the cycle from 2π to a multi-cycle, which can improve the convergence probability of the algorithm. Each chromosome consists of two parallel gene chains, which means the machine allocation chain and process chain of FJSS respectively. If each gene chain means an optimal solution, then each chromosome has two optimal solutions in the search space, that is:

$$p_{i\cos} = (\cos(t_{i1}), \cos(t_{i2}), \cdots, \cos(t_{in}))$$

$$p_{i\sin} = (\sin(t_{i1}), \sin(t_{i2}), \cdots, \sin(t_{in}))$$

$p_{i\cos}$ and $p_{i\sin}$ are called the cosine and sine solution. The two solutions can be updated synchronously in each chromosome iteration, which can expand the search space and increase the number of global optimal when the population size is the same.

2) Non dominated sorting

It is difficult to obtain the optimal solution which will meet all the objectives for FJSS. A sorting method of non dominated is proposed based on the fuzzy theory, and the method realizes the classification depending on the parameters N_p and n_p of individual p in population S, the specific steps are as follows:

Step 1: Initialize the parameter set N_p which includes all the individuals dominated by p and make $Np = \varnothing$;

Step 2: Initialize the variable n_p. n_p means the number of individuals which can dominate p;

Step 3: Calculate dominance relationship, $p,q \in$ S, if p can dominate q, then N_p=N_p U{q}, else if q can dominate p, then n_p=n_p+1; if n_p=0, then p is a non dominated individual, denoted as p_r=1,and p joins R_1, that is to say R_1=R_1 U{p};

Step 4: Q means the set of residual individuals, making i=1,when $R_i \neq \varnothing$,Q$= \varnothing$. If $q \in N_p$, then n_q= n_q-1, else if n_q=0, then q_r=i+1; make $Q = Q \cup \{q\}$,i=i+1,R_i=Q;

Step 5: Judge whether R_i is empty or not, if it is empty, then stop, otherwise turn to Step 4.

In order to maintain the diversity of the population, the selection is made according to the crowding distance of the chromosomes based on non dominated sorting.

3) Cloud quantum genetic algorithm

Cloud quantum genetic algorithm(CQGA) is a novel optimization method of GA combining of cloud model theory and quantum theory. The specific steps are as follows:

Step 1: Initialize the population and execute the double chains quantum coding to the chromosome;

Step 2: Design the fitness function as $fit(x)$= $1/Z(x)$, $Z(x)$ is the individual objective function value, the smaller the function value is, more excellent the individual is;

Step 3: Select the excellent individuals from the population into the next generation using the best one preservation strategy and fitness proportional;

Step 4: Generate the crossover operator p_{cr} using the cloud model X condition generator. Put the gene region between inter-section in the first of the sub generation and remove the same code in the father generation, then copy the rest code to the sub generation according to the order. If the individual in the sub generation is beyond the constraints, adjust the position of 0;

Step 5: Generate the mutation operator p_{mt} using the cloud model X condition generator. Select two codes r_1 and r_2 from the gene coding of mutational individuals randomly and exchange the selected codes to generate the new one;

Step 6: Construct new population which consists of m excellent individuals from the father generation and m individuals in the sub generation, and then to extend the size of the population and the search space;

Step 7: Get the next m generation population through the operation of GA;

Step 8: Update the quantum gates according to Schrodinger equation;

Step 9: Judge whether the stopping condition is met, if not, turn to Step 3, else stop running the algorithm.

C. Analysis of Convergence

CQGA is a novel hybrid method which consists of the diversity of the quantum population and generating new individual through rotation angle of the quantum gates. It won't influent the convergence of the algorithm because of using the cloud model X condition generator. The state transformation of CQGA is as follows:

Assuming the length of the chromosome is L, the population size is P, the size of the state space about GA is 2^{LP}, and the size of the state space about CQGA is u^{LP}(u is the dimension of the state space), then the state transfer process of the population described by Markov chain is as follows:

$$Q_t \xrightarrow{observe} P_k (\xrightarrow{cross} p_t{}' \xrightarrow{mutate} p_t{}'') \xrightarrow{keep\ \ the\ \ optimal\ solution\ and\ update\ Q_k} Q_{t+1}$$

The upgrading operation of CQGA is influenced not only by the evolutionary constraints of GA but also by that of quantum rotation gates. Its convergence is not affected after the transformation.

IV. Analysis And Verification

A. Analysis of Simulation Experiment

In order to verify the performance of the proposed method, this paper takes the minimization of completion time, minimization of penalty and minimization of cost as targets to test. The testing data based on the classic Kacem [17] example is as follows: the size of the population is 200, the maximum number of iterations is 100, the crossover probability is 0.45 and the mutation probability is 0.02.

Five standard Kacem examples are solved by CQGA and they were compared with the existing AL+CGA[17], PSO+TS[18] and HBCA [19] at the same time. The results are shown in Table Ⅰ. n is the number of the workpieces, m is the number of the machines, Sol_n(n=1,2,3,4) are the solutions obtained by different algorithms, T_x is the maximum completion time of the machine, M_t is the total load to the machines, M_x is the maximum load in balancing the load of the machines. It can be seen from Table 1 that the proposed method can obtain more non dominated solutions and has got the current optimal solution. For example, for case 10×7 although both HBCA and CQGA obtain three non dominated solutions, the solution (12,61,11) obtained by HBCA is dominated by (11,60,11),(12,61,10) and (12,60,11) obtained by CQGA, and (11,61,11) obtained by HBCA is dominated by (11,60,11) obtained by CQGA, and (12,60,12) obtained by HBCA is dominated by (11,60,11) and (12,60,11) obtained by CQGA.

TABLE I.
COMPARISON WITH DIFFERENT ALGORITHMS ON KACEM

n×m	Obj	AL+CGA		PSO+TS		HBCA			CQGA			
		Sol$_1$	Sol$_2$	Sol$_1$	Sol$_2$	Sol$_1$	Sol$_2$	Sol$_3$	Sol$_1$	Sol$_2$	Sol$_3$	Sol$_4$
4×5	*Tx*	16		11		11	12	13	11	11	12	11
	Mt	34		32		31	31	33	30	30	31	31
	Mx	10		10		10	8	7	9	8	7	8
8×8	*Tx*	15	16	15	15	14	15	16	14	14	15	14
	Mt	79	75	77	75	76	75	73	73	74	73	73
	Mx	13	13	12	12	12	12	13	12	12	12	11
10×7	*Tx*					12	11	12	11	12	12	
	Mt					61	61	60	60	61	60	
	Mx					11	11	12	11	10	11	
10×10	*Tx*	7		7		8	6	7	7	6	6	6
	Mt	45		43		41	42	41	41	40	41	40
	Mx	5		6		6	5	5	4	5	5	6
15×10	*Tx*	23		11		11	11		10	11		
	Mt	95		93		90	91		90	89		
	Mx	11		11		11	11		10	11		

B. Analysis of Scheduling Experiment

Further more tests were conducted in the mould workshop of a machinery company to verify the performance of the novel method for multi-objective FJSS. The data in Table II (6workpieces×8machines) are the results after the raw data of the mould have been processed.

These data consist of machines, processing time and costs, in which the unit of time is minute and the unit of cost is yuan. The set of machines are as follows: rough turning lathe (M1), fine turning lathe(M2),rough milling machine(M3), fine milling machine(M4), boring machine(M5), planer(M6), grinder(M7) and machining center (M8).

TABLE II
.DATA OF 6×8 EXAMPLE

workpiece	process 1	process 2	process 3	process 4	process 5	process 6
1	M6,48,75.0	M2,45,69.8	M2,43,68.0	M4,48,62.0	M5,22,20.8	M8,32,27.0
2	M6,47,74.6	M5,25,24.5	M6,46,68.6	M3,28,34.8	M2,40,66.0	M8,30,25.0
3	M1,12,9.8	M4,46,57.0	M4,46,54.8	M7,24,40.9	M6,48,76.8	M4,48,63.5
4	M3,28,35.0	M7,24,40.2	M8,31,25.5	M6,49,77.8	M1,13,10.2	M3,28,34.6
5	M3,30,37.2	M8,32,26.2	M6,46,69.9	M7,26,48.2	M2,38,60.4	M4,44,55.8
6	M5,22,21.6	M3,32,38.0	M5,26,26.8	M1,14,10.1	M1,12,9.6	M4,45,57.6

The proposed CQGA is compared with several existing algorithms after being conducted 50 times. The result is shown in Table 3. It can be seen that the optimal solution and the average solution with CQGA are both better than those of other algorithms, besides, it can get less penalty than the other two.

TABLE III
.COMPARISON OF THREE ALGORITHMS

Objective	PSO[20]		IPSO[21]		CQGA	
	Opt Solution	Avg solution	Opt Solution	Avg solution	Opt Solution	Avg solution
Process time	1198	1246	1105	1148	998	1065
Process cost	1380	1428	1296	1378	1009	1135
Penalty	231	246	216	230	170	188

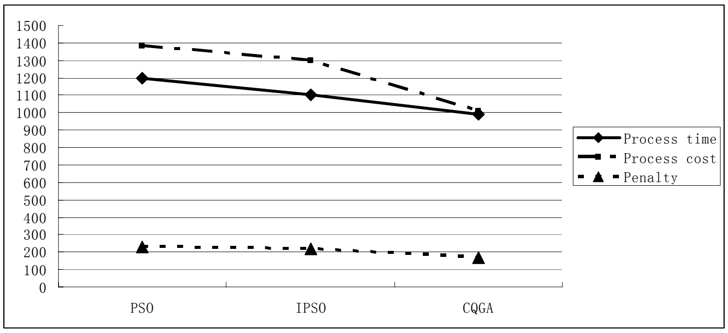


Figure 2. Comparison of optimal solutions

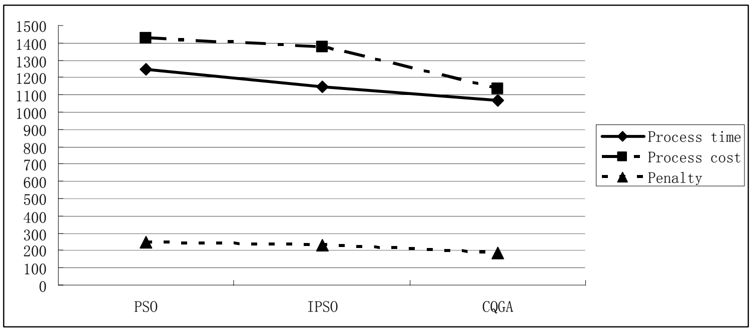


Figure 3. Comparison of average solutions

Ⅴ. CONCLUSION

The mathematics model of multi-objective FJSS was established in this paper, and the coding method of double chains was used including the initialization of the machine distribution chain with quasi level uniform design and the heuristic initialization of the process chain. On the basis of the theory of cloud and quantum, the crossover operator and mutation operator were generated by the cloud model X condition generator, and the new population was obtained through rotation angle of quantum gates. The non dominated sorting strategy was introduced based on the fuzzy set theory. Finally, the proposed method was applied to Kacem instances and the scheduling of a mould workshop, and then compared the data with that of the existing algorithms. The comparison of the results verified the proposed method could not only reduce the maximum completion time and process cost but decrease the penalty effectively.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation, China (No.61034003), the National Key Technology R&D Program, China (No.2012BAF12B08), the National 863 Plan Project, China (No.2012AA 041402-4), the Talented Young Scholars Growth Plan of Liaoning Province Education Department, China (No.LJQ2013048). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

CONFLICT OF INTERESTS

The author declares no conflict of interests.

REFERENCES

[1] LIANG Xu,LIU Pengfei, HUANG Ming, “Genetic algorithm for multi-order Job Shop scheduling under mixed production patterns”, Computer Integrated Manufacturing Systems, vol.18,no.10, pp.2217-2223,2012.

[2] Bucker P. and Schlie R, “Job-shop scheduling with multi-purpose machines”, Computing, vol.45, no. 4,pp.369-375,1990.

[3] Chen H.,Ihlow J. and Lehmann C, “A genetic algorithm for flexible job-shop scheduling”, In: Proceedings of the 1999 IEEE International Conference on Robotics &Automation, Detroit Michigan:IEEE,vol.2,pp.1120-1125,1999.

[4] Ho N.B. and Tay J.C, “Genance:an efficient culture algorithm solving the flexible job shop problem”, In:Proceedings of Congress on Evolutionary Computation,pp.1759-1766,2004.

[5] Najid N.M.,Dauzere-Peres S.and Zaidat A, “A modified simulated annealing method for flexible job shop scheduling problem”, In:Proceedings of the IEEE International Conference on Systems Man and Cybernetics.NJ,USA,IEEE,pp.89-94,2002.

[6] Kacem I. “Genetic algorithm for the flexible job shop scheduling problem”, IEEE International Conference on Systems, Man and Cybernetics, vol.4,pp.3464-3469,2003.

[7] D.Y.Sha, C.Y.Hsu, “A hybrid particle swarm optimization for job-shop scheduling problem”, Computers and Industrial Engineering, pp.791-808, 2006.

[8] B.Liu,L.Wang,Y.H.Jin, “An Effective PSO-Based Memetic Algorithm for Flow Shop Scheduling”, IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics,

vol.37, no.1, pp.18-27,2007.

[9] Kun Fan,Ren-qian Zhang, Guoping Xia, “An Improved Particle Swarm Optimization Algorithm and Its Application to a Class of JSS Problem”, Proceedings of IEEE International Conference on Grey Systems and Intelligent Services, Nanjing, 2007.

[10] Xia W J,Wu Z M,“An effective hybrid optimization approach for multi-objective flexible job shop scheduling problems”, Computers & Industrial Engineering,vol.48,no.2, pp.409-425, 2005.

[11] YU Xiaoyi,SUN Shudong, CHU Wei, “Parallel collaborative evolutionary genetic algorithm for multi-workshop planning and scheduling problems”, Computer Integrated Manufacturing Systems, vol.14,no.5,pp.991-1000,2008.

[12] LIU Aijun,YANG Yu,XING Qingsong,et al, “Multi-population genetic algorithm in multi- objective fuzzy and flexible Job Shop scheduling”, Computer Integrated Manufacturing Systems, vol.17,no. 9,pp.1954-1961,2011.

[13] ZHANG Jing,WANG Wanliang,XU Xinli, et al, “Improved particle swarm algorithm for bath splitting flexible job shop scheduling”, Control and Decision,vol.27,no.4,pp.35-40,2012.

[14] SHI Jinfa, JIAO Hejun, CHEN Tao, “Multi- objective Pareto Optimization on Flexible Job-Shop Scheduling Problem about Due Punishment”, Journal of Mechanical Engineering, vol.48, no.12,pp. 188-196,2012.

[15] Li D.Y,Du Y, “Artificial Intelligence with Uncertainty”, Beijing: National Defense Industry Press,2005.

[16] Dai C.H,Zhu Y.F.,Chen W.R, “Cloud model based genetic algorithm and its applications”, acta electronic sinicaA,vol.35,no.7, pp.1419- 1424, 2007.

[17] KACEM I,Hammani S Borne P, “Approach by localization and multi-objective evolutionary optimization for flexible job-shop scheduling problems”, IEEE Trans Syst Man Cyb C,vol.32, no.1, pp.1-13,2002.

[18] Partha Pratim Das, Sriyankar Acharyya, Hybrid Local Search Methods in Solving Resource Constrained Project Scheduling Problem, Journal of Computers,vol.8, no.5, pp.1157- 1166, 2013.

[19] Tran Quang Tuan,Phan Xuan Minh. Adaptive Fuzzy Model Predictive Control for non-minimum phase and uncertain dynamical nonlinear systems, Journal of Computers, vol.7, no.4, pp. 1014-1024,2012.

[20] Fadi A. Aloul, Syed Z. H. Zahidi, et al. Solving the Employee Timetabling Problem Using Advanced SAT & ILP Techniques. Journal of Computers, vol.8,no.4,pp.851-858, 2013.

[21] YANG Hongan,SUN Qifeng, SUN Shudong, et al, “Job Shop earliness/tardiness scheduling problem based on genetic algorithm”, Transactions of the Chinese Society for Agricultural Machinery, vol.17,no.8,pp.1798-1805,2011.

Xiaobing Liu was born in Changchun city, Jilin Province of the P. R. China in 1956. He received his M.S.degree from Dalian University of Technology in 1984, and his Ph.D.degree from the Germany Dortmund University in 1992. He is now working in the Dalian University of Technology. His research is mainly in the area of artificial intelligence and computation integrated manufacturing system . He has involved with 8 projects supported by the national Natural Science Foundation of China, and 2 projects supported be the national High Technique Developing Program. He has published over 150 research papers.

Xuan Jiao was born in Baishan City, Jilin Province of China in1986. She received the M.S.degree from Liaoning University, China in 2012. She is pursuing Ph.D.degree from school of management of Dalian University of Technology under the supervision of Prof. Xiaobing Liu. Her current research interests are in the computer integrated manufacturing system.

Tao Ning was born in Penglai City, Shandong Province of China in 1979. He received his M.S.degree in Dalian Maritime University in 2006 and his Ph.D.degree in Dalian Maritime University in 2013. He majors in computer information management and computer integrated manufacture and is working as a professor in.Dalian Jiaotong University.

Ming Huang was born in Changchun City, Jilin Province of China in 1961. He received his B.S. degree in Jilin University, China in 1982. He is currently working as a professor in Dalian Jiaotong University, China. His research interest fields include computational geometry, large scale software development, design and analysis of algorithms.

A Software Platform Design for Objective Video QoE Assessment

Yitong Liu, Hao Liu, Yuchen Li, Yun Shen, Jianwei Wu, Dacheng Yang
Beijing University of Posts and Telecommunications, Beijing, China
Email: liuyitong@bupt.edu.cn

***Abstract*—Along with the rapid development of 3G and 4G technologies, mobile video services have gained its popularity among users around the world. Consequently, Content Providers (CPs), Service Providers (SPs), and especially, the operators are paying increasing attentions to the quality of experience (QoE) of the video services which could be easily affected by the quality of network. In this paper, a novel real-time objective video QoE assessment method is proposed and a software assessment system is built to test the video service quality in the real network. Firstly, in the test terminals, the QoE measurement of the entire video services is conducted by collecting all of the customers' experience in full-reference method, and then the QoE scores are evaluated through an accurate mathematic model. Secondly, the artifact of compression caused by video encoding should also be taken into account. Model in this part adopts no-reference method in consideration of the varied screen sizes in different terminals. What's more, the platform also evaluates the error of network in the part of video transmission by associating no-reference PSNR with network delay, jitter, and packet loss ratio. The results of Mean Opinion Score (MOS) tests show that the proposed models estimate QoE with high quality estimation accuracy respectively. We develop a software toolkit using the test methodologies above, which can help the operators to make measurements for its network. This software toolkit is useful as a QoE monitoring tool on video streaming services and can be deployed on real network conveniently.**

***Index Terms*—QoE, video service, objective assessment, software toolkit**

I. Introduction

The advances in video encoder technologies and broad IP networks lead to the popularity of video streaming services. Furthermore, with the development of 3G/4G technologies, the number of customers attracted by mobile video streaming services is growing rapidly. In recent years, mobile network operators in China have launched a variety of video services, including VOD, video telephone, and etc. Besides, the increase in the amount of Internet services is making the video services much more bustling.

As shown in Fig.1 [1], Web service, video, IPTV and P2P contribute to the major part of the total Internet traffic since 2011. Among these services, Video, IPTV and P2P are relative to video distribution.

In China, a statistic analysis report from CNNIC shows that the mobile terminals became the NO.1 internet access device in China by the year of 2013 [2] , used by 75% of users.

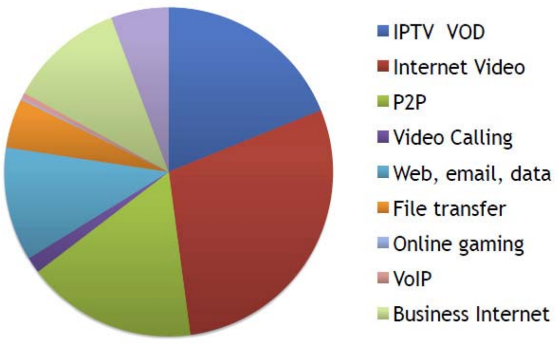


Figure 1. 2011 Internet Traffic distribution

A survey from iReseach revealed that the video traffic is turning to mobile market. They took 3 most popular videos appeared in one year as examples. The mobile share of traffic doubles nearly every 6 months [3], shown in Fig.2.

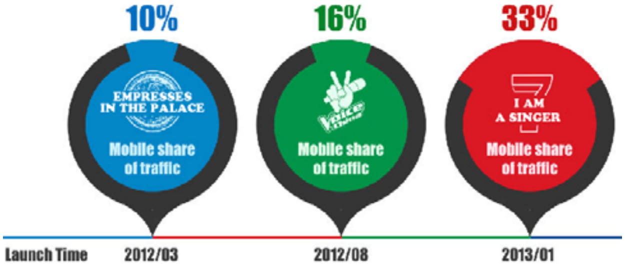


Figure 2. The mobile traffic of 3 most popular videos in 2012

The quality of video service is a reflection to the quality of network. The services' perceived quality draws the most attention form operators since it is closely related to customers' personal feelings. The better the quality is, the more customers could be attracted. Nevertheless, traditional indicators to evaluate the performance of services, in terms of Quality of Service (QoS) for the network, are not accurate enough to reflect the customers' degree of satisfaction.

Then came the question. In our paper, we discuss how to evaluate the customers' feeling, experience, or

doi:10.4304/jsw.9.11.2955-2965

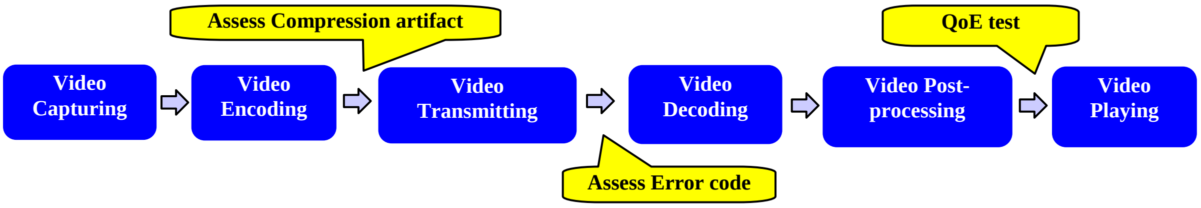


Figure 3. The test purpose of proposed QoE-based evaluation toolkit in a video service

reception on the video service objectively, and how to use this kind of evaluation result as the baseline for the provision of network quality for this kind of service.

In[4],the Quality of Experience (QoE) is proposed. In recent years, many proposals are made in order to evaluate, measure, and improve QoE of the video services.

Nicolas Staelens proposed a novel subjective quality assessment methodology based on full-length movies [5]. Their subjects took DVD together with a questionnaire enclosed in a sealed envelope home to watch it in real-life environments. Ozgur Oyman reviewed the recently standardized QoE metrics and reporting framework in 3GPP and presented an end-to-end QoE evaluation study conducted over 3GPP LTE networks [6].Ricky K.P. Mok investigated the relationship between network QoS, application QoS, and QoE and then proposed their QoE measurement [7]. They evaluated QoE of Flash video perceived by users and quantified how the QoE is influenced by the application QoS.Hyun Jong Kim developed a QoS/QoE correlation model, which could evaluate the QoE using QoS parameters in offered network environment [8].Karan Mitra proposed a novel approach [9] to estimate and predict QoE in Heterogeneous Access Networks (HAN). This system is based on Hidden Markov Models (HMM) and Multi-homed Mobility Management Protocol (M-MIP), which improved the accuracy of QoE estimation in many network conditions. All of these researches about the QoE of video services are quite constructive. However, the limits of QoE assessment can be not ignored. Above all, the subjective test consumes lots of manpower and material resources, which is unacceptable for both of the operators and the customers. Therefore, the objective and quantitative QoE test is introduced in this paper.[10][11]

Besides, QoE assessment is posteriori so that the evaluation result could only show the QoE level of the whole service no matter how many factors and parts there are. In consequence, it is very difficult to monitor what caused the QoE down.

So our video QoE assessment system is divided into 3 platforms to describe the performance of the entire service, impairment from the video encoding, and the impairment in delivering.

The paper is organized as follows. In section II, the methodology of QoE-based assessment is introduced. The structure and deployment of our software toolkit is described. Section III describes the method and the software platform we proposed to assess the quality of entire video service. A reference sequence is chosen, and the service is recorded while playing. Full-reference method is adopted. KPIs in the video playing process, which are directly felt by customers, are collected to map to QoE score. In Section IV, we introduce the software platform which evaluates the compression artifact. Because there are various encoded videos in different sizes and the source video cannot be acquired in most of time, the no-reference test method is adopted. Through this platform, the behaviors of different providers and encoders, as well as the quality of videos with different contents and sizes are assessed and presented for the operators to monitor and control. Then in section V, assessing error code software is introduced to test and evaluate the impacts of network environment on the video quality. The structure and modules are introduced in this part. Finally, paper is ended with conclusion and some future works in section VI.

II. QoE-Based Evaluation Software Platform Of Video Quality

Fig.3 shows a complete processing procedure for a video service, including video capturing, video encoding, video transmitting, video decoding, video post-processing, and video playing. Obviously, the QoE that worked out from the terminal should be the final result of the entire video service. And there are only 2 processes, video encoding and video transmission, that could be monitored and controlled by the operators. It is significant for the operators to find out which part causes the degradation of service quality, and then to quantize and compare the QoE loss.

Video encoding is a kind of loss compression coding due to the limitation of storage and bandwidth. Thus, video encoding is one of the main sections causing quality degradation. Take H.264 encoding for example, the quantization of conversion coefficients, which is controlled by a quantization parameter (QP), degrades the image quality via increasing the QP value.[12] This degradation can be reflected by blocking artifact, blur, and Peak Signal to Noise Ratio (PSNR).

The quality of encoded video will degrade again while transmitted in the network. As a result of the network delay, jitter, and packet loss, frame skipping and frame frozen will appear in the received video services.

In this paper, we build a video QoE evaluation platform, which could provide QoE information to the operators, in order to help them in optimizing their services and to define the responsibilities clearly. The designed QoE evaluation framework functions in three parts.

The first part is placed in the test terminals. The target of this part of assessment is to give the QoE result of the whole service. This platform will be described in detail in the section III. The online video service is involved and the full-reference method is adopted. To begin with, we establish the standard original videos and upload them to the web servers. In the terminals, the service is requested and at the same time the course of the service is recorded. Three groups of KPIs is extracted from the comparison between the recorded video and the original video, including KPIs in the connecting period, KPIs of the image and quality of the voice. We use an accurate mapping model to obtain the QoE score from network response delay, KPIs of the images and the sound QoE scores from the Perceptual Evaluation of Speech Quality (PESQ) [13], an international standard to evaluate the QoE on the voice service and also a prevailing algorithm inbuilt by many network optimizing instruments. The QoE score is calculated out and presented in the assessment windows and the detail parameters are illustrated as graphs and tables to compare with each other. This part of this assessment system is named the online video QoE assessment and this assessment can be performed in anywhere we aim to acquire the quality of the network and service.

In the first part, the operators have obtained the value of the service quality. However, it is difficult to judge whether the damage of the service comes from the quality of the network provided by Themself. The second part is to assess compression artifact of video services. This platform is installed in the video center of the operators, and is implemented to test the encoded videos in various contents and sizes uploaded by the CPs. It is of great importance to distinguish the quality degradation owing to the coding compression from the loss because of the bandwidth limitation. The video content can also be an important factor for video coding.[14] Generally, videos in the same content would be encoded into different sizes and different qualities because of the demand of various mobile terminals. The performance of the different CPs and different commercial video encoders are monitored in 3 video types, the rapid movement, the slow movement and the colorful scenario.

The compression artifact assessment is performed after the CPs upload all of the encoded video chips. The encoded videos are in different sizes and the source videos cannot be acquired in most of the time. Therefore, we use the no-reference method to estimate the KPIs of the image, including PSNR, blur, block and motion. It is the most significant and difficult issue in the assessment. Our toolkit records these KPIs and the final QoE results in case of subsequent statistic analysis.

The third part mainly focuses on assessing error code. This part is distributed to the net nodes of individual province companies through Content Delivery Network (CDN) as the video of demand for customers. This part of assessment is to test the quality of transmission. The operators can monitor the behaviors of the individual province companies. We extracts the KPIs of the networks and the qualities of the video service. Packet loss, delay, jitter and so on are associated with the no-reference PSNR and other parameters. The assessment toolkit shows and records the real-time parameters in every transmission node.

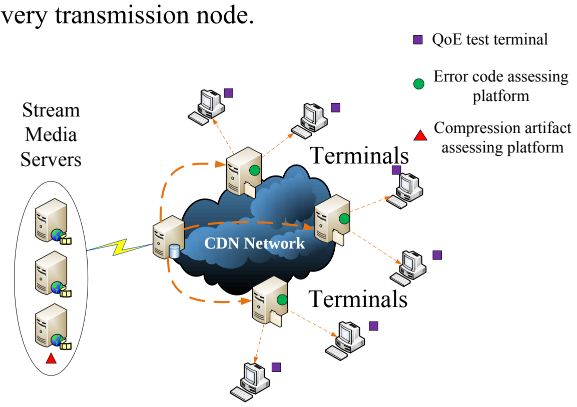

■ QoE test terminal
● Error code assessing platform
▲ Compression artifact assessing platform
Stream Media Servers
Terminals
CDN Network
Terminals

Figure 4. The network framework of the QoE evaluation platform

As previously mentioned, Fig.4 illustrates the network framework of the QoE-based evaluation platform. The online QoE test terminal is used to obtain the QoE performance of the video service in any point. The compression artifact assessment is deployed to administrate the encoded videos required by a variety of screens. Besides, the error code assessment is monitor of the transmission quality in the CDN and even in the wireless connections.

III. Online Video QoE Evaluation Platform In Terminals

A. *Procedures of the Video QoE Test*

We propose an online test module structure in this section. The progressive streaming scenario is considered. The full-reference method, which compares the degraded video against the original video to get results, is adopted. The standard source video samples are divided into 3 groups of different durations: 30s, 1mins and 3mins. Besides, different contents are involved. The procedures of the assessment are as follows:

1. Build a network service server and upload all of the original standard videos.

2. Choose a video sequence in the test terminals to test.

3. Request a video service, and record the playing process automatically. The KPIs during the service request, such as *successful access ratio* and *response delay*, are counted.

After recording the test videos online,the degraded sequences are obtained.

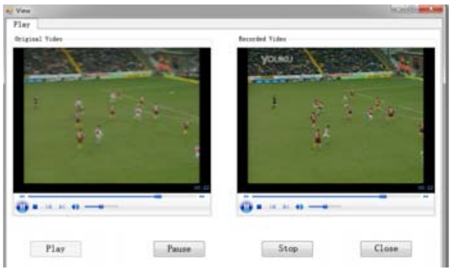
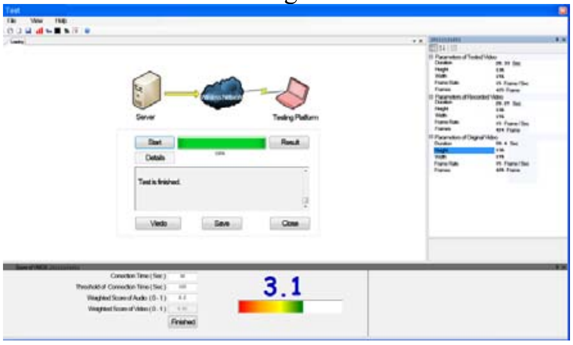


Figure 5. The comparison between the source videos and the recorded videos

4. Build a new project to test this video service. Input the original video and the degraded video for comparison and start the QoE testing automatically. The QoE score of the whole service is calculated. This score is based on image quality, audio quality and service access delay.

5. Detailed parameters and KPIs can also be viewed. Besides, the original video and replay of the recorded video service can be watched, as shown in Fig.5.

Some Graphical User Interfaces (GUIs) of this part of QoE test platform are demonstrated in Fig.6. The main interface gives the assessment progress and the basic parameters of the video in Fig.6.


Test
File
View
Help
Server
Testing Platform
Start
Result
Details
Video
Save
Close
Connection Time (Sec)
Threshold of Connection Time (Sec)
Weighted Score of Audio (0-1)
Weighted Score of Video (0-1)
Finished
3.1
Parameters of Tested Video
Parameters of Recorded Video
Parameters of Original Video
Duration
Height
Width
Frame Rate
Frames

Test is finished.

Figure 6. The main interface of the online video QoE evaluation platform

The final QoE score is presented in the end of the assessment. After the test finished, the detail KPIs, such as delay, PSNR, frame skipping and frame frozen, can be analyzed in graphs and can be output to a file named after the test time. The analysis window is demonstrated in Fig.7.

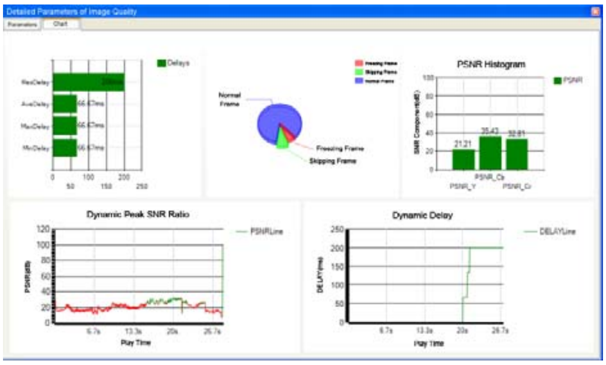


Figure 7. The detailed KPIs analysis interface

B. *KPIs Extraction Fliter in Video QoE Test*

We defined several KPIs that influence the customers' experience. These KPIs is obtained in the toolkit and presented for the operators.

1.KPIs during the period of connecting server, including *successful access ratio* and *service access delay.*

Concretely, *service access delay* is defined as T :

$$T = t_1 + t_2 \tag{1}$$

Where t_1 is *network access time* and the t_2 is *response delay*, described in the Fig.8. The *network access time* equals to the time length from the time when customer demands the video to the time when video starts to buffer. It is used to estimate the network response time, related to the net environment. And *response delay*, depending on the predetermined strategy, records the time lag from the time that video starts to buffer to the time that video starts to play. Therefore, *service access delay* equals to the waiting time after the costumers request the service.

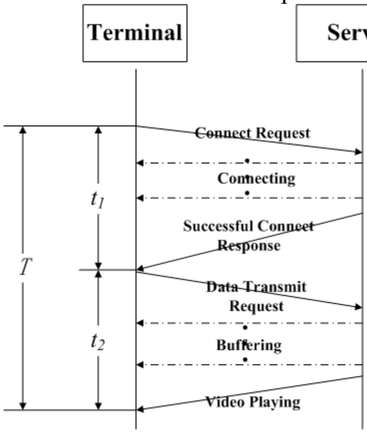


Figure 8. The definitions of the service access delay, the network access time and the response delay

2.KPIs of video images influence the fluency and resolution, including:

--*Image activity ratio*

--*Activity in time domain* which means image variation degree in time domain

--*Spatial complexity* which depends on video types

--*Luminance*

--*PSNR*

--*Frame skipping* caused by frame loss

--*Frame frozen* caused by frame repetition

--*Block*

--*Blur*

--*Delay distribution while playing*

PSNR is the most frequently used indicator for video quality. It can be calculated from the luminance values of the source signal *p(x, y)* and the degraded signal *q(x, y)* as follows.

$$MSE = \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} (q(x,y) - p(x,y))^2 \tag{2}$$

$$PSNR = 10\lg\left(\frac{(2^Q - 1)^2}{MSE}\right) \tag{3}$$

MES is the mean squared error and *Q* is the bit of an intensity value. *X,Y* is the frame width and height separately. The classic *PSNR* algorithm provides three types of values: $PSNR_Y$, $PSNR_{Cr}$ *and* $PSNR_{Cb}$.

The blur parameter is calculated based on an extreme luminance value in a frame. We use the zero-crossing rate to extract blur. The image definition reflects the degree of changes in image details. Namely, the higher the definition is, the better the image presents. In consequence, gray scale can be more sensitive to the changes in location and variations in image details could also be high, resulting in a good degree of recognition

[15][16]. The value range of blur is (0,1], where higher value means higher ambiguity.

Generally speaking, blocking artifacts are caused by low bit rate encoding in the flat area of image and border of moving objects. The key of Blockiness algorithm is to extract block border areas and non-block border areas. The block border area consists of pixels that are adjacent to a block border or pixels that include a block border, which can be detected by canny-edge detector. And non-block border area consists of the rest pixels. The block is calculated by reference to the computing method proposed by [17].

Delay parameters are divided into *maximum delay, minimum delay* and *average delay*, which come from the full reference algorithm. And these detailed parameters are listed on the interface separately.

Some of the parameters mentioned above work as intermediate variables to calculate parameters afterwards. Others are used to build the eventual QoE model.

3.Quality of sound in the video

For the sound in the video,we choose the PESQ test, which is introduced in the ITU P.862. PESQ is a most prevailing voice service QoE assessment method and has been inbuilt by many network optimizing instruments. However, our former research have drawn a conclusion that it is not convinced enough to use PESQ to assess Chinese voice service directly [18]. So the English voice is preferred in the test video sequences.

C. Structure of Image Quality Assessment

As it shows in Fig.9, the image quality assessment module is composed of pre-process and analysis module, parameter extraction module and quality estimation module. First, the pre-process and analysis module takes source video signal and degraded video signal as inputs, and extract the Region of Interest (ROI) and some other information from spatial domain and temporal domain. Then, the parameter-extraction module derives delay parameters, PSNR, block and blur parameters. Finally, the quality-estimation module estimates video quality using these parameters.

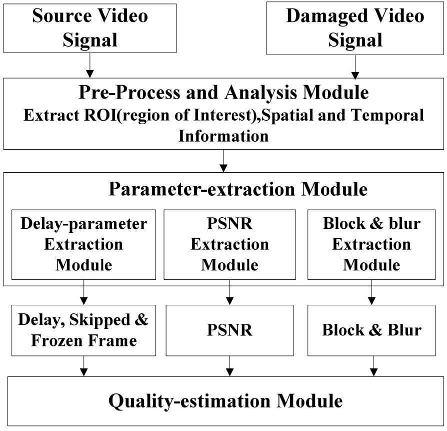

Source Video Signal
Damaged Video Signal
Pre-Process and Analysis Module
Extract ROI(region of Interest),Spatial and Temporal Information
Parameter-extraction Module
Delay-parameter Extraction Module
PSNR Extraction Module
Block & blur Extraction Module
Delay, Skipped & Frozen Frame
PSNR
Block & Blur
Quality-estimation Module

Figure 9. Full-reference QoE online test module structure

D. QoE Mapping Modeling

The QoE assessment model is established to map *PSNR, delay, frameskipped, framefrozen, blockiness, blur* and frame rate *Rate_{fps}* to QoE score. The modeling method and result have been introduced in our previous research paper [19].

$$Score_{PSNR} = a*PSNR_Y + b*PSNR_{Cb} + c*PSNR_{Cb} \quad (4)$$

$$\begin{aligned} Score_{frame} &= \ln(framefrozen) \\ &\quad + d*frameskipped \end{aligned} \quad (5)$$

Here, $\begin{matrix} PSNR, \\ framefrozen, \\ frameskipped \end{matrix} \in [1,5]$·

$$\begin{aligned} Score_{Block} &= (e*\ln(Block+1)+f) \\ &\quad *(Score_{frame}+Score_{psnr})+g \end{aligned} \quad (6)$$

$$Score_{Blur} = \begin{cases} Score_{Block} + 10/Blur - h & 10/Blur <= 5 \\ (\frac{i}{Blur} + j)*Score_{Block} & 10/Blur > 5 \end{cases} \quad (7)$$

$$S^{'} = Rate_{fps}*(k*Score_{Blur}+l) \quad (8)$$

$$Rate_{fps} = \min(1, m*fps^n + o) \quad (9)$$

The final score is

$$S_{QoE} = \begin{cases} 1 & \text{if } S^{'} < 1 \\ 5 & \text{else if } S^{'} > 5 \\ S^{'} & \text{else} \end{cases} \quad (10)$$

Where *a, b, c, d, e, f, g, h, j, k, l, m,* and *o* are coefficients. These coefficients are optimized using least-square method to minimize the difference between subjective video quality and estimated video quality.

E. Subjective Test and Data Acquisiton

According to [20], 30 seconds standard videos are accepted in the subjective test. In this section, 29 sequences with different damages are produced as counterparts. No less than 10 viewers vote every counterpart.

Before beginning the subjective test, every viewer accepts a simple training using 5 examples. And each viewer should rate the counterpart using the integral MOS scale of 1, very bad, to 5, excellent.

After cancelling the invalid test samples, according to [21], 315 votes are accepted and the coefficients in (4)-(10) are determined, listed in Table I.

TABLE I.
COEFFICIENTS OF QOE MODEL FOR ONLINE VIDEO TEST

Coeff.	Value	Coeff.	Value	Coeff.	Value
a	0.3	***f***	1	***k***	0.9654
b	0.1	***g***	1.3	***l***	-0.0421
c	0.1	***h***	5	***m***	-10.12
d	0.2	***i***	1.348	***n***	-0.03116
e	-0.279	***j***	-0.069	***o***	10.45

The Pearson Correlation Coefficient of above QoE model is calculated as high as 0.925, as shown in Fig 10.

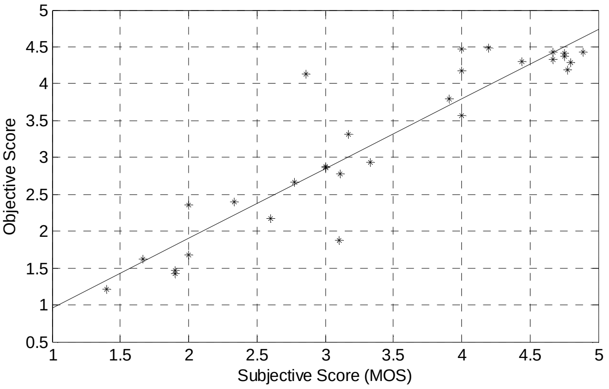


Figure 10. Scatter diagram between MOS and QoE score

IV. COMPRESSION ARTIFACT ASSESSING PLATFORM

The mobile Internet users has been dramatically increasing and the usage scenarios are diverse. Internet access is common to various mobile devices, such as smart phones, tablet PCs, laptops, TVs and so on. All kinds of video content are provided by a sight of CPs, whereas the videos are aimed to be applied to different terminals. Therefore, even videos of the same content should have different sizes and qualities.

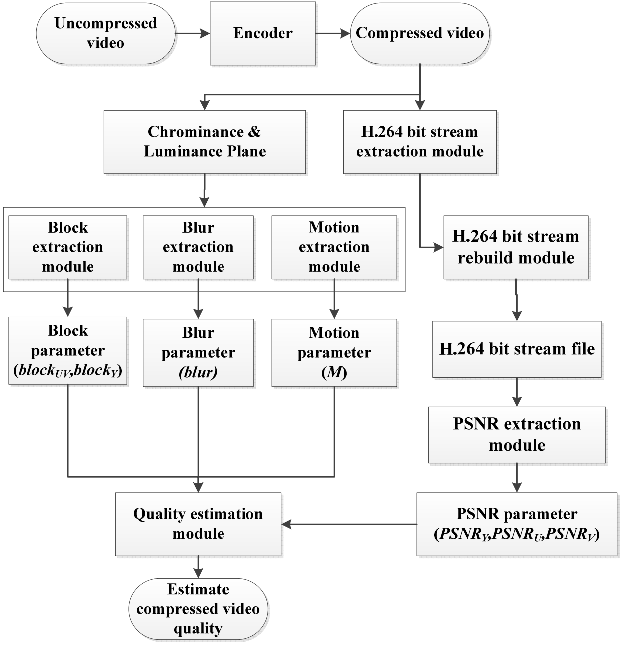
As known, different encoding types lead to different compression artifacts, which cause varying degrees damage of quality. Even the same encoding type can make different degrees of artifacts by using different encoders.

According to the different content, we divide videos into 3 groups: slight movement, rapid movement and colorful scenario. All of the 3 video content types are considered. We choose 6 original video samples with 1.5Mbps and D1 (720*480) resolution. Rapid movement group contains *Car Racing* and *Tennis*. Slight movement group contains *News* and *Football*. The Colorful group has *Movie* and *Natural scenario*.

5 CPs are invited to participate in the assessment. 5 different commercial encoders are involved to produce compression artifact counterparts. In each encoder, every original video is encoded into 3 versions with different bitrates and resolutions: D1 for high bitrate, CIF for middle bitrate and QCIF for low bitrate. Finally, 90 degraded samples are collected.

Same as that in section III, no less than 10 viewers are recruited to rate each counterpart. 2700 votes are accepted in the end, half of which is used to establish the QoE model and some other is used to verify the performance of the model. In addition, the invalid votes are filtered out as specification in [21].

The objective of assessing compression artifact is to get optimal encoders for different type of videos. In the video encoding section, the full-reference method is infeasible. It is difficult to compare the degraded video and the original video because the encoded video chips have different sizes from the original samples or even the original video samples may not be acquired. Therefore, the proposed model adopts no-reference method to assess compression artifact by using h.264 encoded video instead. The model structure can be seen in Fig.11.



Uncompressed video
Encoder
Compressed video
Chrominance & Luminance Plane
H.264 bit stream extraction module
Block extraction module
Blur extraction module
Motion extraction module
H.264 bit stream rebuild module
Block parameter ($block_{UV}$,$block_Y$)
Blur parameter (blur)
Motion parameter (M)
H.264 bit stream file
PSNR extraction module
Quality estimation module
PSNR parameter ($PSNR_Y$,$PSNR_U$,$PSNR_V$)
Estimate compressed video quality

Figure 11. Structure of assess compression artifact platform

The complete algorithm is composed of two modules, Chrominance and Luminance Plane and H.264 Bit Stream Rebuilt Module. There are totally 4 parameters that can be extracted by the model.

In the first part, the signal is sent to Block Module, Blur Module and Motion Module respectively. In these three modules, video is processed to obtain its block, blur and motion value. In our platform, the algorithm of calculating these three parameters are a no-reference method, which means that there is no need to acquire the source video. Many studies provided solution of this aspect, here we adopted to the algorithm from [22].

In the second pare, the PSNR can only be extracted from bit stream of the whole video signal. Usually, the traditional PSNR algorithm is a full-reference method which needs the source signal in (2) and (3). However, it's hard to calculate PSNR when lacking of source signal, i.e., uncompressed video. Thus, we adopt a no-reference method to estimate the value of PSNR, according to [23]. This algorithm uses the coded transform coefficients to estimate the PSNR in a statistical manner. And the Pearson Correlation Coefficient between full-reference PSNR and this no-reference PSNR is as high as 0.96. By the simulation results in [23][24], this algorithm is confirmed to suit different content types and resolutions.

Based on the above 4 KPIs, in our previous research paper [24], we have given the QoE evaluation formula and verified a good performance of the model in estimation the subjective experience. In the Quality Estimation Module of the software platform, the QoE model is adopted to acquire the QoE score.

For the slow movement and rapid movement videos, the QoE evaluation formula is:

$$Score_{PSNR} = a_1 * PSNR_Y + a_2 * PSNR_U + a_3 * PSNR_V + a_4 \quad (11)$$

$$Score_{Block} = (b_1 * block_Y + b_2 * block_{UV} + b_3) * Score_{PSNR} \quad (12)$$

$$S_{QoE} = \begin{cases} c_1 * Score_{Block} + c_2 * Blur + c_3 & 176 \le W < 352 \\ c_4 * Score_{Block} + c_5 * Blur - c_6 & 352 \le W < 720 \\ c_7 * Score_{Block} + c_8 * Blur + c_9 & W \ge 720 \end{cases} \quad (13)$$

Where $PSNR_Y$, $PSNR_U$, and $PSNR_V$ are the no-reference PSNR values on chrominance and luminance plane. W is the width in pixels.

The formula reflects the map relationship between PSNR value and QoE score. The coefficients *a1-c9* depend on the specific content type of video, which are shown in TABLE II. The Pearson Correlation Coefficient is approaching to 0.97 for the Rapid movement video and 0.84 for the slow movement.

TABLE II.
COEFFICIENTS OF QOE MODEL FOR RAPID&SLOW MOVEMENT VIDEO

Coeff.	Rapid movement	Slow Movement	Coeff.	Rapid movement	Slow movement
a_1	0.9104	1.1590	c_2	0.1394	-0.7480
a_2	-2.2971	-3.0432	c_3	0.3378	0.5694
a_3	1.6015	1.9305	c_4	0.2137	-9.9208
a_4	-0.9782	2.9416	c_5	0.322	8.8242
b_1	0.7421	0.8251	c_6	-0.1591	0.2134
b_2	-0.5819	-0.0822	c_7	0.316	-1.7434
b_3	0.8908	1.2075	c_8	0.5762	1.6814
c_1	0.0688	1.1194	c_9	0.0954	0.6651

For the colorful scenario, another QoE evaluation model is more suitable:

$$Score_{PSNR} = a_1 * PSNR_Y + a_2 * PSNR_U + a_3 * PSNR_V + a_4 \quad (14)$$

$$Score_{Block} = b_1 \cdot Score_{PSNR} + b_2 \cdot \ln(block_Y + 1) + b_3 \cdot \ln(block_{UV} + 1) + b_4 \quad (15)$$

$$S_{QoE} = c_1 \cdot Score_{block} + c_2 \cdot blur + c_3 \quad (16)$$

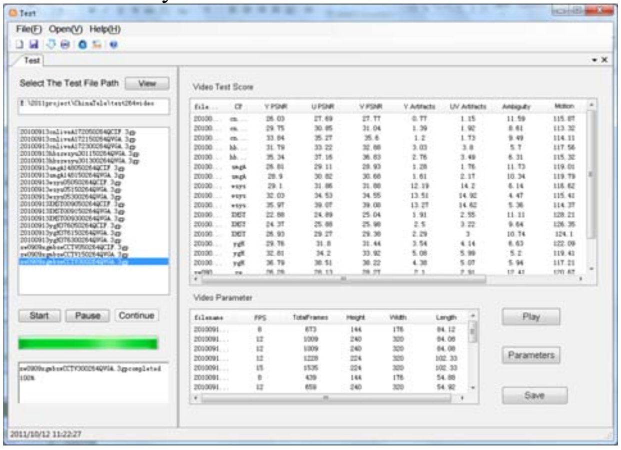
With the huge subjective test results, the coefficients in (14) (15) (16) are listed in TABLE III. The Pearson Correlation Coefficient for the colorful scenario is as high as 0.95.

The compression artifact assessing platform can be located at the video center servers in which the CPs upload the encoded videos. This software tool tests the encoded video quality provided by content providers, and furthermore, evaluates the behaviors of different CPs and different encoders, as well as the quality of videos with different contents and different sizes.

TABLE III.
COEFFICIENTS OF QOE MODEL FOR COLORFUL SCENARIO

Coeff.	Resolution(pixel*pixel)		
	QCIF (176x144)	CIF (352x288)	D1 (720x480)
b_1	0.1943	-0.4637	-0.1549
b_2	-7.842	39.1546	2.1186
b_3	7.5542	-38.5645	-1.7643
b_4	1.315	7.7655	4.7507
a_1	0.6903		
a_2	0.0041		
a_3	-0.6606		
a_4	3.6908		
c1			1.0218
c2			0.0214
c3			-0.3951

Fig.12 shows the GUI of compression artifact assessing toolkit. After CPs uploaded the encoded videos to the center server of the operators, the software tool loads all the video sequences automatically and tests them successively. The user chooses the file folder and the videos are loaded in the left bar. Through the QoE test, the quality score, PSNR, Blockiness, Blur, Motion, Luminance, Skipped Frame, Frozen Parameter and other KPIs are calculated out in the right side of the window. The video information is listed below the QoE results. The results can be compared in different KPIs, CPs, contents, or other dimensions. All the results are stored for further analysis and research.



E:\2011project\ChinaTele\test264video

20100913cnliveA1720502264QCIF.3gp
20100913cnliveA172150264QVGA.3gp
20100913cnliveA172300264QVGA.3gp
20100913hbsrwsyu301150264QVGA.3gp
20100913hbsrwsyu301300264QVGA.3gp
20100913smgkl40050264QCIF.3gp
20100913smgkl40150264QVGA.3gp
20100913wxys050050264QCIF.3gp
20100913wxys05150264QVGA.3gp
20100913wxys05300264QVGA.3gp
20100913DMST009050264QCIF.3gp
20100913DMST009150264QVGA.3gp
20100913DMST009300264QVGA.3gp
20100913ygR076050264QCIF.3gp
20100913ygR076150264QVGA.3gp
20100913ygR076300264QVGA.3gp
sw0909sgmbzsCCTV050264QCIF.3gp
sw0909sgmbzsCCTV150264QVGA.3gp
sw0909sgmbzsCCTV300264QVGA.3gp

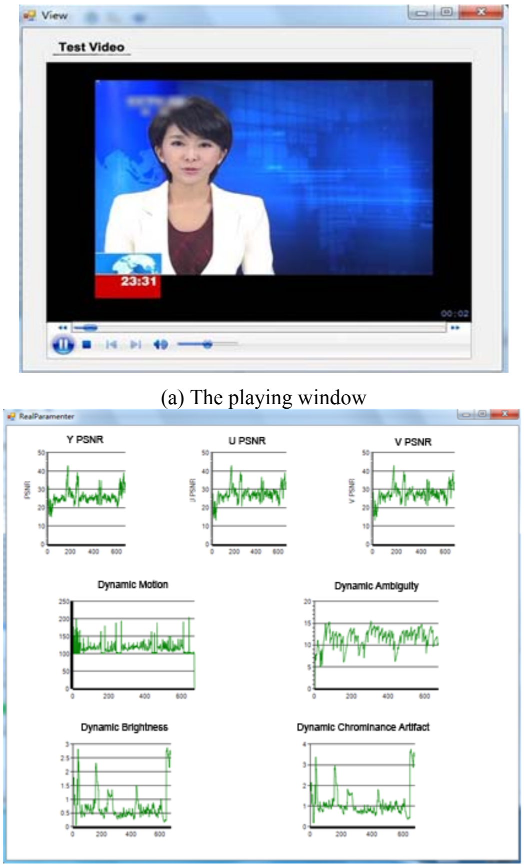
sw0909sgmbzsCCTV300264QVGA.3gpcompleted 100%

File...	CP	Y PSNR	U PSNR	V PSNR	Y Artifacts	UV Artifacts	Ambiguity	Motion
20100...	cn...	26.03	27.69	27.77	0.77	1.15	11.59	115.07
20100...	cn...	29.75	30.85	31.04	1.39	1.92	8.61	113.32
20100...	cn...	33.84	35.27	35.6	1.2	1.73	9.49	114.11
20100...	hb...	31.79	33.22	32.88	3.03	3.8	5.7	117.56
20100...	hb...	35.34	37.16	36.83	2.76	3.49	6.31	115.32
20100...	smgk	26.81	29.11	28.93	1.28	1.76	11.73	119.01
20100...	smgk	28.9	30.82	30.68	1.61	2.17	10.34	119.79
20100...	wxys	29.1	31.86	31.88	12.19	14.2	6.14	118.62
20100...	wxys	32.03	34.53	34.55	13.51	14.92	4.47	115.41
20100...	wxys	35.97	39.07	39.08	13.27	14.62	5.36	114.37
20100...	DMST	22.88	24.89	25.04	1.91	2.55	11.11	128.21
20100...	DMST	24.37	25.88	25.98	2.5	3.22	9.64	126.35
20100...	DMST	26.93	29.27	29.38	2.29	3	10.74	124.1
20100...	ygR	29.76	31.8	31.44	3.54	4.14	8.63	122.09
20100...	ygR	32.81	34.2	33.92	5.08	5.99	5.2	119.41
20100...	ygR	36.79	38.51	38.22	4.38	5.07	5.94	117.21
sw090...	zz...	26.28	28.13	28.27	2.1	2.91	12.41	120.67

filename	FPS	TotalFrames	Height	Width	Length
2010091...	8	673	144	176	84.12
2010091...	12	1009	240	320	84.08
2010091...	12	1009	240	320	84.08
2010091...	12	1228	224	320	102.33
2010091...	15	1535	224	320	102.33
2010091...	8	439	144	176	54.88
2010091...	12	658	240	320	54.92

Figure 12. Compression artifact assessing platform main GUI—the QoE and KPIs results

In the Fig.13 (a), different versions of the test videos can also be played on this platform. The users could decide which video to play on the basis of their test results. In addition, the results can be further analyzed on the platform. Real time parameters are provided for detailed observation. For example, brightness, chrominance and other parameters of every frame are calculated in real time and shown in line chart, which are demonstrated in Fig.13 (b).



(a) The playing window

(b) Real-time results analysis GUI

Figure 13. Assess compression artifact GUIs—the playing window and real-time parameters analysis in graphs

V. Error Code Assessing Platform

When the compressed video signal transmits through the network to the receiver following the Real Time Streaming Protocol (RTSP), different kinds of error code would appear under the influence of real network defects, such as delay, jitter, and packet loss. The target of assessing error code is to analyze the influence to video quality caused by network error code.

A. *Procedure of the Error Code Assessment*

The structures of error code assessing platform are as follows:

1. A RTSP video service connecting should be built in the first place. After opening the platform, the customer should input the server IP, home IP and URL address, and choose the recording equipment in the computer.

2. Press the Play button, and begin to view the RTSP video on demand. Meanwhile, the platform records the screencast of the video automatically. The blur, blockness, movement and PSNR are calculated in the recorded video.

3. Capture all the RTP packets to rebuild a bit stream file to extract the PSNR parameter when the video are playing. By analyzing the RTP packets, the network KPIs are calculated, such as delay, jitter and packet loss ratio.

4. Evaluate the video quality based on the KPIs from the RTP packet capture and the KPIs of the images. The final QoE score and the KPIs are presented in the GUIs.

5. After the QoE test, the KPIs and real time parameters can be reviewed in another windows.

B. *KPIs in the Error Code assessment*

KPIs that influence the customers’ experience are chosen. Both the average result and the real-time parameters are presented on the toolkit.

The network level KPIs: *delay, jitter* and packet loss ratio. The definition of this KPIs is specified in [25].

The video information: *video size, duration of the video, frame rate, file name, provider’s name* and so on.

The KPIs of the image quality: *luminance, chrominance, PSNR, the frame frozen, blur, blockness* and so on.

C. *Modules of the KPI extractions*

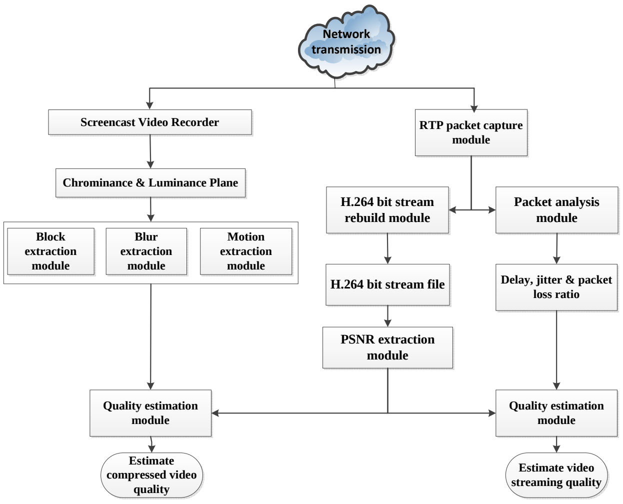

Network transmission
Screencast Video Recorder
RTP packet capture module
Chrominance & Luminance Plane
H.264 bit stream rebuild module
Packet analysis module
Block extraction module
Blur extraction module
Motion extraction module
H.264 bit stream file
Delay, jitter & packet loss ratio
PSNR extraction module
Quality estimation module
Quality estimation module
Estimate compressed video quality
Estimate video streaming quality

Figure 14. Structure of assessing error code

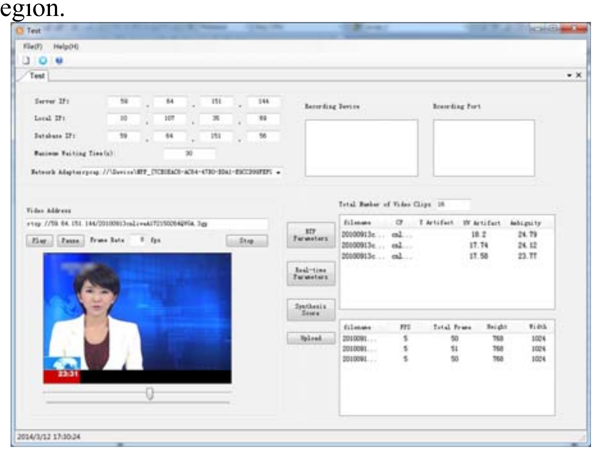
As Fig.14 shows, the assessment is deployed in the network transmission. This module can also be divided into two sections. One is RTP packet capture module. In this part, the H.264 bit stream file is rebuilt from the RTP packets on the one hand. The model calculating no-reference PSNR is the same with the one in section IV. On the other hand, the packets are analyzed to calculate the delay, jitter and packet loss ratio[25].

Another sections bases on the recorded videos. The no-reference method is also accepted. It is important to acquire the frozen frames caused by the delay or packet loss during the transmission. Besides, blur, blockness and movement are obtained with the same algorithms as former sections.

D. *GUIs in the Error Code Assessment*

In the Fig.15, Fig.16 and Fig.17, the GUIs of the error code assessing platform are revealed and the assessment process is specified.

Fig. 15 gives the main interface of the error code assessing platform. To start with, users should fill the server’s IP and the maximum waiting time on the left blanks followed by choosing voice record devices on the right. Demanded videos will be played in the bottom while its assessment results is shown in the lower right region.



Server IP: 59 . 64 . 151 . 144
Local IP: 10 . 107 . 35 . 69
Database IP: 59 . 64 . 151 . 56
Maximum Waiting Time(s): 30
Network Adapter: rpcap://\Device\NPF_{7CE0EAC6-AC84-47B0-BDA1-E9CC390FEF\

Video Address
rtsp://59.84.151.144/20100913cnliveAI7215026492AGA.3gp
Frame Rate 0 fps

Total Number of Video Clips 16

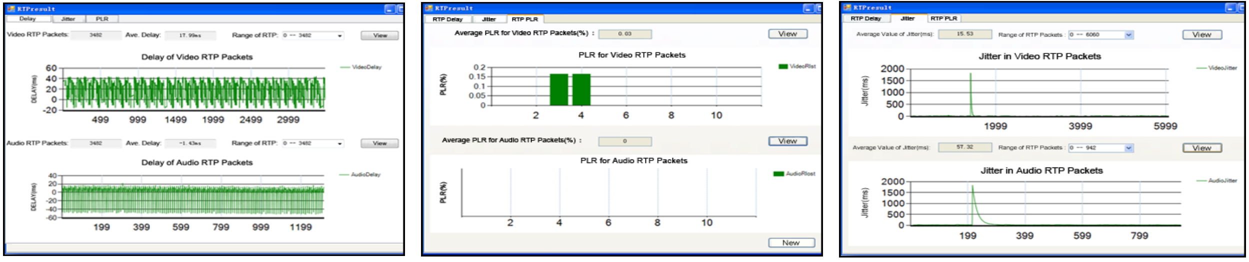
Filename	CP	Y Artifact	UV Artifact	Ambiguity
20100913c...	cnl...		18.2	24.79
20100913c...	cnl...		17.74	24.12
20100913c...	cnl...		17.58	23.77

Filename	FPS	Total Frame	Height	Width
2010091...	5	50	768	1024
2010091...	5	51	768	1024
2010091...	5	50	768	1024

2014/3/12 17:30:24

Figure 15. Main GUI of error code assessing platform

The users can also review the real-time parameters after pressing the "Real-time parameters" button as shown in Fig.16. The RTP packet arrived delay and the jitter can be recorded in chronological order.



Video RTP Packets: 3482 Ave. Delay: 17.99ms Range of RTP: 0 -- 3482
Audio RTP Packets: 3482 Ave. Delay: -1.43ms Range of RTP: 0 -- 3482

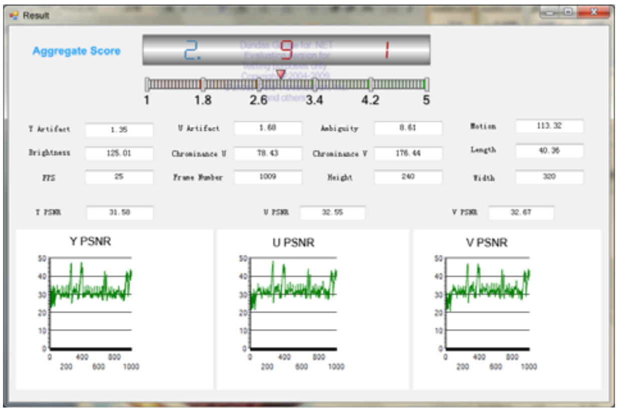
Average PLR for Video RTP Packets(%) : 0.03
Average PLR for Audio RTP Packets(%) : 0

Average Value of Jitter(ms): 15.53 Range of RTP Packets : 0 -- 6060
Average Value of Jitter(ms): 57.32 Range of RTP Packets : 0 -- 942

(a) Network delay (b) Packet loss ratio (c) Jitter

Figure 16. The real-time KPI analysis GUIs in the error code assess software platform

The QoE evaluation result and information of current video service is given in the QoE score windows, illustrated in Fig.17.



Y Artifact 1.35 U Artifact 1.60 Ambiguity 0.61 Motion 113.32
Brightness 125.01 Chrominance U 78.43 Chrominance V 176.44 Length 40.36
FPS 25 Frame Number 1009 Height 240 Width 320
Y PSNR 31.50 U PSNR 32.55 V PSNR 32.67

Figure 17. QoE score and video information GUI

VI. Conclusion

In this paper, an objective QoE assessment methodology for the video service is proposed. This evaluation can be used as the baseline for provision of network quality for video streaming services.

The KPIs that can influence the feelings of the customers are extracted, which are based on our former and others' constructive research achievements. The QoE mapping models are built and huge subjective tests are implemented. The result demonstrates our objective QoE assessment models are accurate.

Based on this assessment methodology, a software platform is built for the Operators to evaluate the quality of video streaming service and the performance of 3G/4G network. The software platform has 3 parts. The first part, video QoE Evaluation platform in terminals, can assess the QoE of the entire video service. Moreover, the second part, compression artifact assessing platform can quantify the QoE loss in the video encoding process. This part of the software platform can be deployed in the video service's center server to test the behavior of the CPs. Through evaluating the quality of videos with different content and different sizes, the performances of different video encoders are ascertained. It is helpful for the operators to distinguish the QoE loss in the video encoding process from the loss in the transmission process. The third part is error code assessing platform, which monitors the network parameters and map them to the QoE score. The output of this part is the QoE loss between any two points in the network. The error code assessing platform is deployed in the CDN.

The GUIs of these 3 software platforms are shown. The assess procedure is automatic and the real-time parameters are recorded and can be remote reviewed.

The proposed video service assessment methodology solves the problem of the objective and automated QoE testing. And the QoE-based assessment software is useful as a QoE monitoring tool on video streaming services and can be flexibly deployed on real network.

REFERENCES

[1] Hui Zhang. “Internet Video: The 2011 Perspective,” unpublished.

[2] CNNIC. 32nd Statistical Report on Internet Development in China. July, 2013.

[3] iResearch Inc. http://www.iresearch.com.cn/, 2013.3.

[4] GB923 TMF.v3.0-2004[J]. *Wireless service measurements handbook*, 2004.

[5] Staelens N, Moens S, Van den Broeck W, Marien IIse, et al. "Assessing quality of experience of IPTV and video on demand services in real-life environments," *Broadcasting, IEEE Transactions on* 56.4 (2010): 458-466.

[6] Ozgur Oyman, Sarabjot Singh. "Quality of experience for HTTP adaptive streaming services," *Communications Magazine, IEEE* 50.4 (2012): 20-27.

[7] Mok R K P, Chan E W W, Chang R K C. "Measuring the quality of experience of HTTP video streaming," *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*. IEEE, 2011.

[8] Hyun long Kim, t Seong Gon Choi. "A study on a QoS/QoE correlation model for QoE evaluation on IPTV service," *Advanced Communication Technology (ICACT), 2010 The 12th International Conference on*. Vol. 2. IEEE, 2010.

[9] Karan Mitra, Christer Ahlund, Arkady Zaslavsky. "QoE estimation and prediction using hidden Markov models in heterogeneous access networks," *Telecommunication Networks and Applications Conference (ATNAC), 2012 Australasian*. IEEE, 2012.

[10] Markus Fiedler, Tobias Hossfeld, Phuoc Tran-Gia. "A generic quantitative relationship between quality of experience and quality of service," *Network*, IEEE 24.2 (2010): 36-41.

[11] Brooks, Peter, Bjørn Hestnes. "User measures of quality of experience: why being objective and quantitative is important," *Network, IEEE* 24.2 (2010): 8-13.

[12] Jinbao Cai, Yalin Wu. An Adaptive Rate Control Initialization Method for H. 264 at Low Bit Rate[J]. Journal of Software (1796217X), 2013, 8(10).

[13] ITU-T Rec. P.862 “Perceptual Evaluation of Speech Quality,” *International Telecommunication Union, Geneva, 2001*.

[14] Daxing Qian, Hongyu Wang, Wenzhu Sun and Kaiyan Zhu. Bit Stream Extraction Based on Video Content Method in the Scalable Extension of H. 264/AVC[J]. Journal of Software (1796217X), 2011, 6(10).

[15] Yan Tan, Chao Wang, “Fuzzy degree of image based on fuzzy mathematics,” *Journal of Computer Aided Design and Computer Graphics*, vol. 14, No.8, August 2002.

[16] Jiegu Li, *Image Processing Technology*. Shanghai Jiaotong University Press, 1988.

[17] Xu Zheng, Bo Yang, Yawen Liu, Guangjian Shi. “Blockiness evaluation for reducing blocking artifacts in compressed images,” *2009 Digest of Technical Papers International Conference on Consumer Electronics*. 2009: 1-2.

[18] Yitong Liu, Yun Shen, Qianhong Liu, Shi Ran, Yang Hongwen, Yang Dacheng, et al. “Performance Evaluation and Accuracy Upgrading of PESQ in Chinese Environment,” *Vehicular Technology Conference (VTC Spring)*, 2013 IEEE 77th. IEEE, 2013: 1-5.

[19] Kui Tian, Yitong Liu, Yongyu Chang, Dacheng Yang, “A novel full-reference video quality assessment method.” *Modern Science and Technology of Telecommunications*, vol.6, June 2011.

[20] ITU-R BT.500-11,“Methodology for the subjective assessment of the quality of television pictures,” 2002.6

[21] ITU R BT.1788, “Methodology for the subjective assessment of video quality in multimedia applications,” 2007.

[22] Taichi Kawano, Kazuhisa Yamagishi. "No reference video-quality-assessment model for video streaming services," *Packet Video Workshop (PV),* 2010 18th International. IEEE, 2010.

[23] Arnd Eden. "No-reference estimation of the coding PSNR for H. 264-coded sequences," *Consumer Electronics*, IEEE Transactions on 53.2 (2007): 667-674.

[24] Yun Shen, Yitong Liu, Nan Qiao, Lin Sang, Dacheng Yang. "QoE-based evaluation model on video streaming service quality," *Globecom Workshops (GC Wkshps),* 2012 IEEE**.**

[25] Network Working Group Request for Comments:3550, “RTP: A Transport Protocol for Real-Time Applications,” *IETF RFC3550 (2003)*.

Yitong Liu received M.S degree in communication engineering from BUPT, China, in 2007. She is a Ph.D. student in School of Information and Communication Engineering in BUPT. Her research interests include QoE of streaming service, quality and performance for wireless network, and wireless applications.

Hao Liu received his B.E. degree in Communication Engineering from BUPT in 2013. Currengtly, he is a Master in WT&T Lab in BUPT. His research interests mainly focus on video coding, nature language processing(NLP) and Dynamic Adaptive Streaming of HTTP (DASH).

Yuchen Li is an M.S. candidate in Wireless Theories and Technologies Lab, Beijing University of Posts and Telecommunications (BUPT), China. He received her B.E. degree of Communication Engineering in BUPT in 2012. His research interest now is Quality of Experience.

Yun Shen received B.E. degree in communication engineering from BUPT, China, in 2010. He is currently a Ph.D. student in WT&T Lab. His research interests include source/channel coding, multimedia communication, and QoE study on application services.

Jianwei Wu is a senior engineer in BUPT. He received his M.S. degree in 2007. His research interests focus on the computer networks, switching and routing technologies, new generation networks theory and technologies.

Dcheng Yang received Ph.D. degree in communication engineering from BUPT in 1988, and is currently a professor in BUPT. He is also the director of the WT&T Lab at BUPT. Since 1988, he has engaged in studies on communication systems. His recent interests are in wireless transmission techniques and systems.

Grid Resource Discovery Algorithm Based on Distance

Zhongping Zhang[1,2], Long He[1], Chao Zhang [1]
[1]The School of Information Science and Engineering, Yanshan University,
Qinhuangdao, Hebei, 066004, China
[2]The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province,
Qinhuangdao, Hebei, 066004, China
Email: zpzhang@ysu.edu.cn, hl_helong1988@126.com, zcmysky310@163.com

***Abstract*—Resource discovery in a Grid environment is a critical problem and also a fundamental task which provides searching and locating necessary resources for given processes. Based on the domain and resource routing nodes, a grid resource discovery model of multilayer overlay network is given in this paper. And on this basis, using a linear combination of the block distance and chessboard distance instead of Euclidean distance, grid resource discovery algorithm based on distance is proposed. Experiment shows that the algorithm has low cost, fast response and can obtain better success rate of lookup, as well as the effectiveness of the resource discovery algorithm.**

***Index Terms*—Grid; model; distance; resource discovery; multi-layer overlay network**

I. Introduction

Now, the grid is the international frontier of research topics. With the development of Grid technology, the Grid begins to be used in various fields. The most important one is data grid which is widely used in data-intensive industries. Data Grid can access, storage, move and manage the data which is heterogeneous, distributed and mass. It has a very broad application prospects.

Grid computing connects a large amount of geographically distributed heterogeneous resources to form a virtual network environment, enabling the users to share resources dynamically, which can effectively improve the utilization rate of resources and system performance. In the dynamic grid environment, resources are huge and have strong heterogeneity. The problems of resource management and lookup under ensuring high efficiency and use as possible as low cost become very complicated. Grid resource discovery mechanism in grid system is the key to realize resources sharing. Grid resource discovery mechanism performance directly determines the performance of the grid system. So it is necessary to design grid resource discovery algorithm with a low cost, good scalability and high efficiency.

In order to develop an efficient and scalable solution for Grid resource discovery, a series of challenges must be faced, due to the following reasons [1]:

1. dynamic property of Grid resources;
2. absence of central authority;
3. heterogeneous and large scale character of Grid environments;
4. unpredictability of faults;
5. difficulty in handling complex multi-attribute range queries.

To deal with these challenges, a resource discovery mechanism must have the following important characteristics [1]:

1. support for intermittent resource availability;
2. independence from any centralized/global control and global knowledge;
3. support for attribute-based discovery;
4. support for multi-attribute range queries;
5. scalability in terms of number of users and resources, and type of resources;
6. provide excellent query performance.

The existing fully centralized or fully distributed resource discovery method is not very ideal. They have some flaws respectively. However, the grid resource discovery method which is based on hierarchical model focuses on the advantages of centralized and distributed discovery method. So, using the combination of the layered structure and tree structure, the model based on Multi-Layer Overlay Network to discover resources is given in this paper. And on this basis, using a linear combination of the block distance and chessboard distance instead of Euclidean distance, grid resource discovery algorithm based on distance is proposed.

The rest of the paper is organized as follows. Related work on grid resource discovery is given in section II. A Multi-Layer Overlay Network(MLON) Model is proposed for resource discovery in data grid in section III. Grid Resource Discovery Algorithm Based on Distance(BC-DIS) is proposed in section IV. The performance evaluation is presented in Section V. Finally, the conclusion and proposed future work are discussed.

II. Related Work

At present, many organizations and researchers both at home and abroad study grid resource discovery model

Manuscript received March 2014, revised April 1, 2014, accepted April 15, 2014.

doi:10.4304/jsw.9.11.2966-2973

from different perspectives and different levels. The unified strategy to manage resources in the grid is adopted in the Globus centralized model proposed in [2]. It has a good global control and high efficiency of resource discovery, but it lack adaptability and extensibility in grid. Resources are managed through the interaction of different resource management systems in the P2P distributed model proposed in [3-5]. And just the opposite of the centralized model, it has a good scalability and lacks in the overall control. Besides, the communication costs of whole model are far higher than these of the centralized model. The flooding technology to find the resources is adopted in [6]. In the face of the grid dynamic environment, it has good fault tolerance and availability, but poor extensibility. The distributed resource discovery based on the virtual structure is adopted in the resource discovery mechanism based on Overlay Network (ON) proposed in [7] and Tree-Structured Overlay Network (TSON) in dynamic grid environment proposed in [8]. The ON and TSON can shield dynamic characteristics of grid resources and heterogeneity effectively. But they also have some deficiencies: because they partition Organizational virtual structured domain based on the physical area and use attribute matching method to find the resources, they cause large network consumption and poor efficiency.

A peer-to-peer architecture for resource discovery in a large and dynamic collection of resources has been proposed in [9]， it is similar to Gnutella combined with more sophisticated query forwarding strategies taken from the Free-net overlay network. Requests are forwarded to one neighbor only based on experiences obtained from previous requests, thus trying to reduce network traffic and the number of requests per peer compared with simple query flooding as used by Gnutella. Because a suitable peer was not reached simply, the approach suffers from higher numbers of required hops to resolve a query compared to our approach and provides no lookup guarantees.

A model called the Hierarchical Resource Organizational Model has been proposed in [10]. The model consists of three layers to process the information in a grid. These three layers include: Physical Network, Resource Information and Index Information. The Physical Network Layer is at the lowest level containing the physical resources linked with each other on the Internet. For each resource, a resource node is placed in the Resource Information Layer. Therefore, the Resource Information layer contains virtual organizations (VO), which is a group of resource nodes in a star topology with a super node in the center. The super node stores all the information regarding the resources of a VO as adjacent lists. The Index Information Layer stores information pertaining to all super nodes of the middle layer and composition of the multiple layers form the basis for hierarchical resource discovery.

At present, the resource discovery mechanism is mainly divided into two types: one is centralized style; one is distributed style. Traditional centralized resource discovery mechanism has certain advantages [11].

1. the topology structure of the system is relatively simple and it is easy to build and maintain, besides the consumption is small;
2. services focus, easy to resources sharing and system security is better;
3. there is no problem of inconsistent of information resources and the efficiency of resource discovery is higher in a small area.

Meanwhile, there are same disadvantages [11] as follows:

1. Reliability is Relatively poor. When the central server fails, the system will not work properly and the system has not fault-tolerant;
2. Scalability of the system is Relatively poor.

However, the characteristics of fully distributed resource discovery mechanism and centralized resource discovery mechanism is opposite, the main drawback is that:

1. the disorder and structurelessness of resource information space make resource discovery have a certain blindness;
2. Nodes can join or leave at any time, which makes the security of system difficult to control.

III. Resource Discovery Model

Overlay network is built on a physical network virtual network connected by virtual or logical links. It can provide more reliable and better fault-tolerant application services without changing existing large-scale network architecture. Each layer of overlay network uses a structured P2P technology, which is advantageous to dynamically join, leave and forward service of the virtual node. Multi-Layer Overlay Network(MLON) resources organization mechanism proposed in this paper is make full use of the overlay network technology and resources which are large, widely distributed, heterogeneous and dynamically changed in the grid of P2P technology organizations to better meet the dynamic environment of the grid system.

A. Model Organization Structure

MLON uses the combination of the hierarchical structure and tree structure to organize overlay network. As the service nodes at the bottom, all the resources in the grid constitute the physical resource layer of the model framework. A large number of grid resources are classified according to the type of resources. The virtual organization with an internal structure is managed by the corresponding overlay network node. Multi-layer overlay network structure is shown in Figure 1.

Overall, MLON structure will be divided into two levels: the upper layer is the overlay network layer, or MLON layer; the lower layer is the physical resource layer, or the Internet layer. MLON consists of multiple virtual layers. All nodes in each layer are classified by the resource type and belong to different domains respectively. Each domain is independent, and each node in the domain uses the graph structure to connect each other in the form of structured peer-to-peer. Each layer adopts the method of the upper managing the lower. A virtual node of the upper manages a domain of the layer. Looking down from the top of a node, virtual nodes adopt tree structure to organize them.

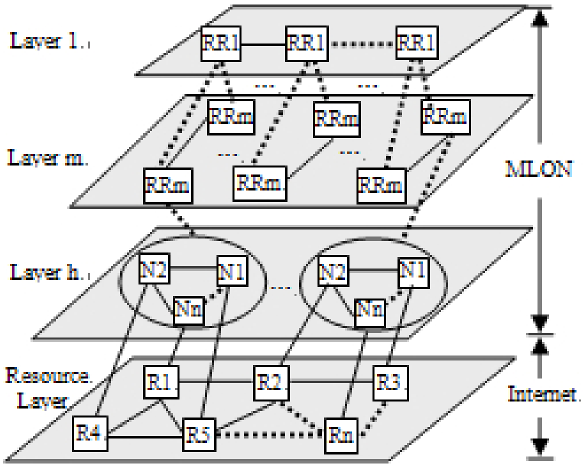


Figure 1. The frame of MLON

Domain and layer as the basic logic unit is used in the MLON. A vast amount of resources are logically divided into several domains through the classification. Each domain is managed by the Resource Router. Resource router holds a large number of service information of the domain nodes. Resources belong to different domains by type. When virtual nodes of a domain increases to a certain number, the domain was divided into two or more smaller domains, which leads to the granularity of resource type small; on the contrary, when the nodes of a domain is reduced to a certain number, the domain nodes is merged, which leads to the granularity of resource type large. Thus a multilayer overlay network is formed.

B. *The Basic Definition*

Definition 1: Grid Resource $R =< ID, A, op() >$, where *ID* represents the resource *R* identifier; *A* represents the attribute collection of the resource *R*; *op()* indicates the set of operations of the resource *R*. When $r \in R$: $a_i(r) = \{a_{i1}, a_{i2}, \cdots, a_{ij}\}$, $op_i(r) = \{op_{i1}(r), op_{i2}(r), \cdots, op_{ik}(r)\}$.

Definition 2: Resource Management Model $G =< N_G, E, R, T >$, where N_G represents the set of all the nodes in the grid, that is to say $N_G = \bigcup_{i=1}^{n} n_i$; *E* denotes the set of edges in the network; *R* denotes the set of all resources in the grid, that is to say $R = \bigcup_{i=1}^{n} r_i$; *T* indicates the collection of the type of resource, that is to say $T = \bigcup_{k=1}^{m} t_k$. As $r_i \in R$, for $\forall$ i , $j \in [1,n] : r_i \cap r_j = \varnothing$ .When the type of resource r_i is represented by $r_i(a)$,i.e. $r_i(a) = t_i$, $t_i \in T \Rightarrow R(A) \in T$,for $\forall$ i , $j \in [1,m] : t_i \cap t_j = \varnothing$.

Definition 3: Domain For each type of resource, the connected and structured P2P, called domain, is constructed by lots of information nodes which register with the type of resources.

Definition 4: Overlay Network Node $n_V =< VID, des, indexlist >$, where *VID* represents the virtual node n_V identifier in the overlay network; *des* indicates the description of the node n_V ; *indexlist* indicates the pointer list of the node n_V , it contains pointers of linking to the parent nodes, the child nodes and the brother nodes, namely $indexlist = \{index\_f, index\_s, index\_b\}$ .Each resource is assigned a globally unique *VID* address by overlay network, which is bundled with resource node *ID*. *VID* includes two parts of domain number (VID_r) and intra-domain node number (VID_S), then two parts are divided and each segment represents a certain type of resources.

Definition 5: MLON Model $G_V =< N_V, rf, IndexList >$, where N_V indicates the set of all the nodes in MLON, namely $N_V = \bigcup_{i=1}^{n} n_{vi}$; *rf* is the function of resources registration; *IndexList* means the set of the pointer list in MLON.

Definition 6: Resource Router Define the non-leaf nodes in the MLON model as *RR*(Resource Router)

Definition 7: the function of resources registration *rf* The domain of the MLON is divided by resource type which is determined by resource properties. For each resource, whether it is static attribute or dynamic attribute, it is only active in a particular area[12]. If $a_j(r) \in [c,d]$, *m* control points are inserted in the interval $[c,d]$, when $c = c_0 < c_1 < \cdots < c_{m-1} = d$ and $b = (b_0, b_1, \cdots, b_{m-1}) = (0, 0, \cdots, 0)$, for any $a_j(r_i)$:

$$b_k = \begin{cases} 1, a_j(r_i) = [c_k, c_{k+1}) \\ 0, \quad else \end{cases} \tag{1}$$

On the basis of formula (1): $rf_j(a_j(r_i)) = b = (0, \cdots, 1, \cdots, 0)$, then $rf(a(r_i)) = \bigcup_{j=1}^{n} rf(a_j(r_i))$

According to the definition 1 and 2, the following assumptions are made:

Assumption 1: A resource belongs to only one type of resource, i.e. $|r_i(a)|=1$. A resource can be registered on multiple resource routers, one is the resource registration node, others are nodes of the copy.

In order to better describe the structure of the MLON model, according to the definition, the following assumptions can be made:

Assumption 2: The neighbor of the node n_V is only its brother. i.e. $index\_f(n_V) = index\_f(index\_b(n_V))$.

According to the definition and description, node n_{vi} in MLON has the following characteristics:

(1) $\forall n_{vi} \in N_V$,if $indexlist_i(n_{vi}) \neq \varnothing$, then $\exists indexlist_i \in IndexList$.

(2) $\forall n_{vi} \in N_V$, if $index\_s(n_{vi}) \neq \varnothing$, then $n_{vi} \in RR$.

(3) $\forall n_{vi} \in N_V$, if $index\_f(n_{vi}) = \varnothing$, then n_{vi} is the top-level resource router node in MLON.

(4) $\forall n_{vi} \in N_V$, if $index\_s(n_{vi}) = \varnothing$, then n_{vi} is the leaf node in MLON.

Non-leaf node (i.e. resource router) in MLON is only responsible for managing resource sub-tree which uses it as the root and provides the sub-tree information for the upper node. The nodes on the bottom are called leaf nodes and the leaf nodes correspond to resource nodes of the physical network, which contain all the information of resources. In MLON, leaf nodes and non-leaf nodes only distinguish logically.

IV. Resource Discovery Algorithm

Resource discovery algorithm is focused on resource searching. If the grid dynamic changes like the joining or exiting of the nodes are mastered, good resource discovery algorithm is able to resource search. Considered with the dynamic change of grid environment from the resource routing nodes service deployment and the resource nodes register, the resource discovery algorithm of the Multi-Layer Overlay Network model based on distance is obtained.

Because of the dynamic of the resources in the grid environment, MLON must be able to self organized and maintained. It mainly includes the registration of resources, the cancellation of resources, the update of resources information and the failure and exception handling of resource nodes. These issues have been discussed in other articles, so they are not described in detail here

A. *The Problem Description*

Definition 8: Grid resource discovery model for request processing $K = \{G, request(R(A)), l\}$, where $request(R(A))$ signifies the search request, it defines the various attribute constraints which the resources of finding should be met; l denotes the request-forward strategy in the process of resource searching; K indicates the collection of resources found in resource discovery process, which matches the search request.

It has a lot of similarity between resource nodes in each domain of virtual network. When looking for resources, users usually not concerned about a resource or a certain type of resources, but only pay attention to the most relevant k results with the needs. The proposed resource discovery algorithm in this paper combines Source Searching Technique Based on Style Matching Routing(SMRT) with Source Searching Technique Based on Topk Algorithm(TopkT) to finally find the collection of resources K which match the search request of user.

Definition 9: Euclidean distance The $P_1 = (x_1, x_2 \cdots, x_n)$ and $P_2 = (y_1, y_2 \cdots, y_n)$ are regarded as the point in V_n , so we can define the formula of Euclidean distance $d(P_1, P_2)$ as follows:

$$d(P_1, P_2) = \left\|P_1 - P_2\right\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

Definition 10: Blocks distance For any $P_1 = (x_1, x_2 \cdots, x_n)$ and $P_2 = (y_1, y_2 \cdots, y_n)$, $P_1, P_2 \in R^n$, so we can define the formula of Blocks distance $d_s(P_1, P_2)$ as follows:

$$d_s(P_1, P_2) = \sum_{i=1}^{n} \left|x_i - y_i\right| \tag{4}$$

Definition 11: Chessboard distance For any $P_1 = (x_1, x_2 \cdots, x_n)$ and $P_2 = (y_1, y_2 \cdots, y_n)$, $P_1, P_2 \in R^n$, so we can define the formula of Chessboard distance $d_c(P_1, P_2)$ as follows:

$$d_c(P_1, P_2) = \max_{1 \le i \le n} (\left|x_i - y_i\right|) \tag{5}$$

Theorem 1: For any $P_1 = (x_1, x_2 \cdots, x_n)$ and $P_2 = (y_1, y_2 \cdots, y_n)$, $P_1, P_2 \in R^n$,then $d_c(P_1, P_2) \le d(P_1, P_2) \le d_s(P_1, P_2)$.

After the user requests resource, grid system can quickly find the resource domain which matches the request of user, because the *RR* in the resource domain is responsible for assigning the most appropriate resources. Resources in the resource domain can be uniformly represented with a point of the m-dimensional coordinates. Searching a resource can be seen seeking a point that satisfies the query conditions in the m-dimensional coordinates. So, $request(R(A))$ can be converted into seeking the k resources which satisfy the query conditions and whose distance is nearest from enquiry point in the m-dimensional coordinates.

By the Euclidean distance formula, m-dimensional resource attribute query can be converted into a one-dimensional range query to meet the grid resource discovery mechanism principles of multiple attribute queries. According to the formula (3), (4), (5), it can be seen that the calculation of d_S and d_C is simpler than the calculation of d. According to theorem 1, if we use d_S instead of d, the calculated value will be larger; if we use d_C instead of d, the calculated value will be smaller; deviation value of the two methods are often large. So we can consider using an appropriate linear combination of the d_S and d_C instead of d. This will not only improve the efficiency of calculation, but also make the deviation value smaller. According to the proposed method in [13], we can calculate $\alpha(d_S + d_C)$ instead of d.

When searching for resource, users often have a preference for a certain attribute, so the resource attributes are weighted according to users' need when they query. Let $W(w_{m-1}, \cdots, w_1, w_0)$ be the attribute weight of resource r_i, where $\sum_{i=0}^{m=1} w_i = 1$. So the distance between resource r_i and enquiry point r is given below.

$$d_i = \alpha\left(\sum_{j=0}^{m-1}\left(\left|a_j^i - a_j\right| \times w_j\right) + \max_{0 \le j \le m-1}\left\{\left|a_j^i - a_j\right| \times w_j\right\}\right) \tag{6}$$

The method that uses an appropriate linear combination of Blocks distance and Chessboard distance instead of Euclidean distance reduces multiplication operation and has a high operation speed, thereby the efficiency of resource discovery has been improved.

B. Algorithm Description

Resource discovery algorithm based on MLON can be divided into two parts: the user's request is passed between nodes *RR*; the request is transferred and searching for resource in matching types of resources domain.

After $request(R_i(A))$ is submitted, the Resource Router in the network generates VID_i by the function rf, then it makes and operation with bitwise between VID_i and VID of node *RR*. If the result is equal to VID, it shows that the resources of query are in resource domain which is managed by node *RR*, i.e. if $\exists(VID_i \wedge VID) = VID$,then $r_i \in R(VID)$. Otherwise, the request is passed, until the resource domain that accords with the request is found or Time To Live (TTL) of request message is zero.

When the bottom resource router RR_j receives a request, $request(R_i(A))$ will be transponded towards each node of its management domain by the router. When receiving the request, the leaf node first tests itself whether meet the requirements of the query. If it meets, the distance d_i between local resource and query point is calculated. Then d_i is returned to the resource router. Based on the optimal matching principle, the nearest *k* resources with $request(R_i(A))$ will be returned to the user by the resource router according to d_i. The resource discovery algorithm of the Multi-layer Overlay Network model based on distance(BC-DIS) is given below.

Algorithm: The resource discovery algorithm of the Multi-layer Overlay Network model based on distance(BC-DIS)

```
Inputs: W, $request(R_i(A))$, k, ttl;
Outputs: The optimal k resources.
BC-DIS (W, $request(R_i(A))$, k, ttl)
Begin
(1) user sends $request(R_i(A))$ to MLON;
(2) $r_i(a) = request(R_i(A))$;
(3) $VID_i = rf(r_i(a))$;
(4) While ( $(VID_i \wedge VID) \neq VID$ ) {   /*search the domain of matching resources*/
    (5) ttl-1, if ttl is less than 0, then the identifier which represents the end of the life cycle of registration message is returned.
    (6) If ( $index\_f(VID) \neq Null$ )
        (7) Node RR delivers $request(R_i(A))$ and $VID_i$ towards father node;
    (8) Else
        (9) Node RR delivers $request(R_i(A))$ and $VID_i$ towards brother node by the best neighbor strategy; }
(10)While( $index\_s(index\_s(VID)) \neq Null$ )
{ /*search father node $RR_j$ of matching resources */
    (11) ttl-1, if ttl is less than 0, then the identifier which represents the end of the life cycle of registration message is returned.
    (12) Node RR delivers $request(R_i(A))$ and $VID_i$ towards child node, the request is passed by the best neighbor strategy in the domain; }
(13) $RR_j$ transmits the request towards child node;
(14) The distance $d_j$ between leaf node $r_j$ and $r_i(a)$ is calculated by leaf node, then returns $d_j$;
(15)Return TopK = $SelectMinD(d_0, d_1, \cdots, d_n)$;
End
```

C. The Algorithm Time Complexity Analysis

Assuming the number of grid resource type is n_t, namely $|T|=n_t$, the number of nodes with the lowest layer resource domain that requests to be matched is n_v, the height of MLON is *h*, the number of resource return is *k*.

From the above algorithm, we can get to know that the resource discovery process can be divided into two stages: (1) the first stage is that user requests to find resource domain. The worst case is that the requests of user are transferred from the leaf nodes to the top of MLON. Then it traverses the top of the resource routers. Finally, it sent a request to (*h*-1) layer resource routers of the MLON. The number of comparison between requests and nodes is ($2h+n_t$). As the number of resources managed by the top-level resource routers has a threshold, the depth *h* is a constant. (2) the second stage is that the request is propagated in the matching resource domain. When all the resources in the domain satisfy the request, the number of comparisons is ($n_v+k(n_v-k)$). Therefore, the time complexity of the algorithm is $O(n_t+n_v)$.

V. The Experimental Evaluation

Experiment employs GridSim[14] simulation toolkit to simulate experiment about the grid resource discovery algorithm BC-DIS and grid resource discovery algorithm based on flooding[15]. Use the JDK and JCreator as a programming environment. The program is running on Windows XP platform.

A. Experimental Environment Settings

The experiment only simulates computing resources in the grid, not considering other grid resources. In the process of simulation, we need to ignore change of network topology in resource discovery at a time, and all the resource information should be kept stable and effective in the network. According to the macroscopic statistical properties showed by the system, we will simulate. When we ignore the dynamic changes of the various details of the system, we can simplify the problem under the condition of grasping the essence of the system.

Supposing the number of resource nodes of the model is 50000, namely $|N_G|=50000$, $|R|=50000$; the number of resource type is 10, $|T|=10$, Various types of resources are evenly distributed and $Max(N_{vi}(R(A)))=5000, Min(N_{vj}(R(A)))=50$;

According to the experimental environment, the node n_{vi} of the MLON must meet the following two conditions:

(1) $\exists f(n_{vi})\in\varnothing, then 50\le\left|N_{vi}\right|\le 5000, de(n_{vi})=4$

(2) $\forall S(n_{vi})\in\varnothing, then 50\le\left|N(f(n_{vi})\right|\le 500, de(n_{vi})=4$

where $de(n_i)$ represents the degree between nodes n_i and its sibling nodes in a certain layer of the MLON. $f(n_i)$ represents the parent node of node n_i, $S(n_i)$ represents the child node of node n_i.

In flooding resource discovery mechanism, an average degree between nodes is 4 with life cycle in the way of TTL controlling request information in the network.

In order to make the comparability between the two kinds of resource discovery mechanism, the concept of Resource Density [16] is introduced:

Definition 12: Resource Density is the proportion between the resources that meet the search conditions and the total number of all the resources, that is to say $d=\frac{|N_R|}{|N|}$, where *d* indicates the density of the resource *r* which is corresponded with *request (R (A))*, its unit is 1/1000. N_R indicates the resource node set that matches with *request(R(A))*. *N* represents the set of all resource nodes.

B. Simulation Analysis

In the process of simulation experiment, experiment runs many times with different user requests, and simulation results are the average of the experiment results. There are three important performance indexes of the resource discovery algorithm: the number of nodes involved in the process of resource discovery, response time and the success rate of lookup. This section mainly compares BC-DIS resource discovery algorithm with flooding resource discovery algorithm through the three indexes.

B1. The Comparison of the Number of Nodes Involved in the Process of Resource Discovery

The comparison of the number of nodes in the resource searching for the two kinds of resource discovery algorithm is shown in Figure 2. Seen from the Figure 2, with the increase of density of resources, the number of nodes involved in BC-DIS resource discovery algorithm is far less than flooding algorithm. Especially when the density is small, BC-DIS algorithm has more obvious advantages.

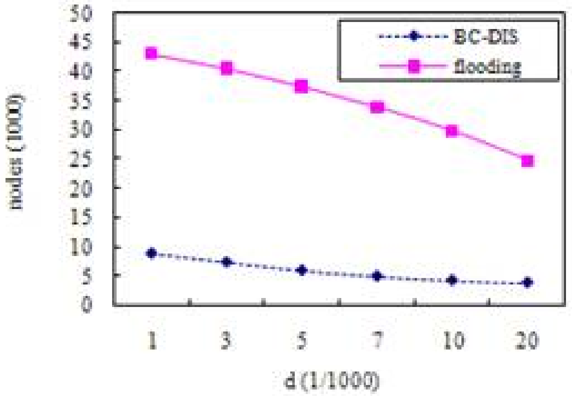


Figure 2 Referenced nodes

B2. The Comparison of Response Time

Response time is the average time from a resource request to receiving a successful response by user. Judge the response time according to hops in the path in which service request is forwarded. The comparison of the number of hops in the resource searching for the two kinds of resource discovery algorithm is shown in Figure 3. Seen from the Figure 3, the hops of the BC-DIS resource discovery algorithm are less than the hops of flooding algorithm. But with the increase of density of the network resources, resources satisfying the requests in the system increase, and the hops of two kinds of resource discovery algorithm become smooth and consistent.

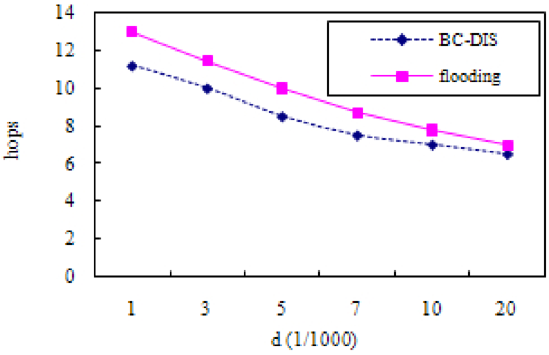


Figure 3 Hops

B3. The Comparison of the Success Rate of Lookup

The comparison of the success rate of lookup for the two kinds of resource discovery algorithm is shown in Figure 4. Seen from the Figure 4, two algorithms can always find the resources with the increasing of lifecycle of information. The BC-DIS algorithm discovers resources more effectively when TTL value is small.

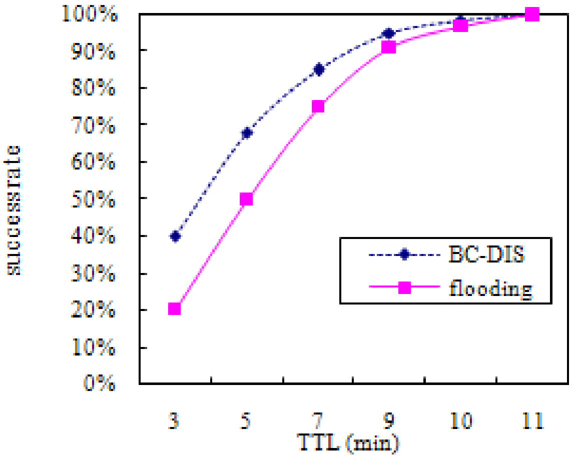


Figure 4 Research success rate

It can be seen from the above simulation results that the use of resource discovery mechanism proposed in this paper can achieve better query efficiency with less traffic and node hops in large quantities of certain resources.

In addition, resource search results of two algorithms are found: grid resource discovery algorithm based on flooding that adopts the resource search method of comparison of the attributes often returns to results with deviation of attribute too much; grid resource discovery algorithm based on BC-DIS supports multiple attributes and range search, and it can return to the k most appropriate resources according to customer request, so it has the highest customer satisfaction.

VI. The Conclusion And Future Work

Because of disadvantage that layered resource discovery mechanism is strongly dependent on resource routing nodes in the grid, a grid resource discovery model based on MLON is given. And on the basis of MLON, the grid resource discovery algorithm based on distance is proposed by using a linear combination of the block distance and chessboard distance instead of Euclidean distance. Through performance analysis and simulation results, the model has fault tolerance and scalability. And it also meets the requirements of the grid dynamics, distribution and scalability. Then it can shield heterogeneity among the resources. The next step is to deploy the model to experiment in real grid environment.

Acknowledgements

This work was financially supported by Hebei Provincial Natural Science Foundation of China (F2012203087), National Natural Science Foundation of China (61272124),and National Natural Science Foundation of China (61073060).

References

[1] Anand Padmanabhan, Sukumar Ghosh, Shaowen Wang. A Self-Organized Grouping (SOG) Framework for Efficient Grid Resource Discovery[J]. Journal of Grid Computing, 2010, vol.8, no.3, pp.365–389.

[2] Karl Czajkowski, Steven Fitzgerald, Ian Foster. Grid information services for distributed resource sharing[C]. //Proceedings of the 10th IEEE HPDC. Washington, DC: IEEE Computer Society, 2001, vol.9, no.14, pp.181-194.

[3] Yingjie Xia, Mingzhe Zhu, Yang Li. Towards Topology-and-Trust-Aware P2P Grid[J]. Journal of Computers, 2010, vol.5, no.9, pp.1315-1321.

[4] Zhengzhen Zhou, Yonglong Luo, Liangmin Guo, Meijing Ji. A Trust Evaluation Model based on Fuzzy Theory in P2P Networks[J]. Journal of Computers, 2011, vol. 6, no.8, pp.1634-1638.

[5] Javad Akbari Torkestani. A distributed resource discovery algorithm for P2P grids[J]. Journal of Network and Computer Applications, 2012, vol.35, no.6, pp.2028-2036.

[6] Shikha Goyal. Flooding Algorithm Based Resource Discovery on Virtual Organization[C]. Proceedings of National Conference on Trends in Signal Processing & Communication(TSPC'13), 2013, pp.12-14.

[7] Francesco Palmieri. Introducing Virtual Private Overlay Network services in large scale Grid infrastructures[J]. Journal of Computers, 2007,vol.2,no2, pp.61-72.

[8] M Marzolla, M Mordacchini, S Orlando. Resource Discovery in a Dynamic Grid Environment. Proc. DEXA Workshop 2005, pp.356-360.

[9] A. Iamnitchi and I. T. Foster. On fully decentralized resource discovery in grid environments. In GRID '01: Proceedings of the Second International Workshop on Grid Computing, pages 51–62, London, UK, 2001. Springer-Verlag.

[10] Mingyong Li, Yan Ma, Yuanyuan Liang. Study the Model of Information Resource Classified Register and Discovery based on Hierarchy in Grid[J]. Journal of Software, vol.7, no.7,July 2012.

[11] Xing Liu, Weidong Xiao, Lei Xu; Grid Resource Discovery Mechanism Based on Complex Topology [J]; Computer Engineering and Applications; vol.9, pp. 132-136, 2005.

[12] Moreno Marzolla, Matteo Mordacchini, Salvatore Orlando. Resource discovery in a dynamic grid environment[C].

//Database and Expert Systems Applications(DEXA), 2005, pp.356-360.

[13] Wang Yuxuan, Li Haijun, Zhou Chunguang. Using a linear combination of the Blocks and Chessboard distance instead of computing Euclidean distance. The small microcomputer system, 2004, vol.25, no.12, pp.2121-2125.

[14] Weifeng Sun, Qiufen Xia, Zichuan Xu, Mingchu Li. A Game Theoretic Resource Allocation Model Based on Extended Second Price Sealed Auction in Grid Computing[J]. Journal of Computers, 2012, vol.7, no.1, pp.65-75.

[15] Noghabi, Hossein Boroumand; Ismail, Abdul Samad; Ahmed, Aboamama Atahar; Khodaei, Masoumeh. An Optimized Search Algorithm for Resource Discovery in Peer to Peer grid[C]. Proceedings - 1st International Conference on Informatics and Computational Intelligence, ICI 2011, pp.21-24, 2011.

[16] Cheng Zhu, A Decentralized Grid Resource Discovery Scheme based on Resource Classification, National University of Defense Technology, 2004.

Zhongping Zhang, Male, Born in 1972, professor, Ph.D., post-doctoral, CCF Senior Member (E20-0006458S). His main research interests are the grid computing, data mining and semi-structured data etc. He has undertaken 1 project of provincial level and participated in 2 projects funded by national natural science foundation of China. He rewarded the provincial Scientific and Technological Progress second-class Award. On the domestic and international academic conferences and journals, He published more than 80 papers, 15 of them are cited by EI.

Long He, Male, Born in 1988, Current Master Student, the main research interest is the grid computing.

Chao Zhang, Male, Born in 1988, Current Master Student, the main research interest is the grid computing.

Single-channel Speech Separation Using Orthogonal Matching Pursuit

Haiyan Guo
School of Information Science and Engineering, Southeast University, Nanjing, China;
College of Engineering, Nanjing Agricultural University, Nanjing, Jiangsu, China
Email: haiyan.guo@seu.edu.cn

Xiaoxiong Li, Lin Zhou and Zhenyang Wu
School of Information Science and Engineering, Southeast University, Nanjing, China;
Email: {220120726, linzhou, zhenyang}@ seu.edu.cn

Abstract—**In this paper, we propose a new sparse decomposition based single-channel speech separation method using orthogonal matching pursuit (OMP). The separation is performed using source-individual dictionaries consisting of time-domain training frames as atoms. OMP is used to compute sparse coefficients to estimate sources. We report the separation results of our proposed method and compare them with a separation method based on sparse non-negative matrix factorization (SNMF) which is a classical sparse decomposition based separation method. Experiments show that our proposed method results in higher signal-to-noise ratio (SNR) and signal-to-interference ratio (SIR).**

Index Terms—**Single-channel speech separation (SCSS), sparse decomposition, orthogonal matching pursuit (OMP), dictionary**

I. Introduction

In a natural environment, several speech signals are usually mixed. Speech separation aims to estimate such individual speech sources from their mixture. It has several obvious applications, e.g., in hearing aids or as a preprocessor to offer robustness in speech recognition, speaker recognition, and speech coding [1-2]. Single-channel speech separation (SCSS) discussed in this paper is an extreme case, where only one mixture is known. It is considered as the most difficult case since no information of mixing matrix can be used. However, the human auditory system has impressive ability to solve this problem, that is, even using an ear, we can still isolate each individual speech when multi talkers speak at the same time.

SCSS aims to recover underlying speech sources from a mixture. It is an ill-conditioned problem since the number of mixture is less than the number of sources.

Previous state-of-the-art SCSS approaches can be divided into two groups: source-driven method or method based on computational auditory scene analysis (CASA) [3], and model-driven method [4-6]. CASA-based method tries to achieve human performance in auditory scene analysis (ASA) based on the perceptual organization of sound. Ideal binary time-frequency (T-F) mask has been proposed as the main computational goal of CASA [7]. CASA generally consists of two major stages: segmentation and grouping. In the segmentation stage, the input mixture is decomposed into time-frequency cells dominated by one individual source. In the grouping stage, multiple segments are grouped into simultaneous streams, and subsequently streams are organized into whole streams corresponding to individual sources. The grouping principles in speech organization prominently used are harmonicity of voiced speech, temporal continuity, onset and offset synchrony, common amplitude modulation, etc. CASA-based method does not rely heavily on priori knowledge of sources. It seeks discriminative features in the mixed signal for separation. However, in general, its separation performance is not as good as that of model-based method.

Model-driven method relies heavily on a priori knowledge about the speakers, hence generally outperforms CASA-based method. From a separation viewpoint, model-driven method can be divided into two classes: statistical model-driven method and decomposition based method. Statistical model-driven method is based on statistical models (e.g. vector quantization (VQ) [8], Gaussian mixture model (GMM) [4] [8] [9-11], hidden Markov model (HMM) [6] and sinusoidal model [12]) or codebooks (e.g. independent component analysis (ICA) basis [5], VQ codebook [10] [12]) trained for individual speakers. It tries to solve out model parameters or find codebook atoms which can generate mixture optimally to estimate sources by statistical methods, e.g. minimum mean square error (MMSE) estimation [9], maximum likelihood (ML) estimation [5] [10], maximum a posterior (MAP) estimation [5] [9], etc. Though statistical model-driven method has been reported to be effective, its training is rather time consuming and estimation is significantly complex. In [12], every possible combination needs to be

Manuscript received January 27, 2014; revised February 22, 2014; accepted March 7, 2014.
Corresponding author: Haiyan Guo, haiyan.guo@seu.edu.cn.

doi:10.4304/jsw.9.11.2974-2980

considered during distortion function minimization to find the optimal codebook atoms.

In model-driven SCSS method, sparsity has been proven to be useful for SCSS. In [5], a priori sets of ICA basis filters are learned for SCSS. The associated coefficients of ICA basis functions are made use of as a function of learning algorithm. Based on the observation that only a small number of coefficients of ICA basis functions differ significantly from zero, generalized Gaussian distribution is used. In [13], a sparse-distribute code of spectro-temproral basis functions where basis functions are more than the dimensionality of the space is generated , leading to better separation results than a compact code of basis functions extracted in [14]. On the other hand, due to sparsity, there is less overlap between the sparse decomposition coefficients of different sources, which means different sources are less likely to be simultaneously active in the sparse domain. Obviously, this feature is helpful for separation.

Decomposition based method is a more intuitive way to use sparsity to perform separation [15-18]. It generally works by two steps. First, personalized dictionaries are learned to give sparse representation of training signals. Second, sparse coefficients are obtained by computing sparse decomposition of the mixture on the union of learned dictionaries, and used to perform separation by combining dictionary atoms assigned to each speaker. In decomposition based method, sparsity is exploited to shrink the feasible solution region of the corresponding underdetermined problems, thus further simplifying the search for the optimal solution. Sparse non-negative matrix factorization (SNMF) is a classical sparse decomposition based SCSS method, and has achieved comparable performance [17-18]. Moreover, it has been proved in [19] that the unique sparse representation of a signal in a union of bases can be found by l_0 optimization if the union of bases and the unique sparse representation satisfy certain conditions. Motivated by this result, we expect to perform high quality single-channel speech separation based on sparse decomposition.

Recently, various methods have been proposed for sparse decomposition. The most typical methods are Basis Pursuit (BP) [20] and orthogonal matching pursuit (OMP) [21-22]. The principle of BP is to find a representation of a signal whose l_1 norm is minimal. A BP problem can be equally reformulated as a linear program and solved by linear programming (LP) [20]. OMP proposed in [21-22] is a recursive algorithm to compute sparse decomposition. It is a modification to the MP algorithm and leads to improved convergence. OMP achieves similar performance as BP, but more quickly. It is a greedy algorithm and selects atoms iteratively for signal recovery. At each iteration, an atom most correlated with the residual is chosen and then residual is updated by subtracting off the contribution of the chosen atom.

In this paper, we propose a sparse decomposition based separation method using OMP. A source-specific dictionary is generated as a matrix consisting of time-domain training frames of each speaker as columns termed as atoms. OMP is used to compute sparse coefficients in the union of source-specific dictionaries for the estimation of sources. Experiments show that the proposed separation method is effective since it results in higher signal-to-noise ratio (SNR) and higher signal-to-interference ratio (SIR) than the separation method using SNMF.

The remainder of this paper is structured as follows. In Section II, we introduce the general model of sparse decomposition based SCSS. In Section III, we introduce SCSS algorithm using OMP. The experiment results are reported in section IV. Finally, we conclude and give future perspectives in Section V.

II. Sparse Decomposition Based Scss

Consider the SCSS problem where the mixed signal $y(t)$ is the sum of two individual speech signals: $s_1(t)$ and $s_2(t)$. $y(t)$ is given as

$$y(t) = a_1 s_1(t) + a_2 s_2(t) \tag{1}$$

where $a_i (i = 1,2)$ is the gain of each source fixed over time. Note that the same mixture can be obtained by adjusting the value of a_i or the power of sources. Therefore, the problem in (1) can be simplified as the following equivalent mixing model

$$y(t) = s_1(t) + s_2(t) \tag{2}$$

It can be written in vector as

$$\vec{y} = \vec{s}_1 + \vec{s}_2 \tag{3}$$

where $\vec{y}$ and $\vec{s}_i (i = 1,2)$ denote the mixed signal and the i^{th} individual speech source in time-domain respectively.

Suppose that $\vec{y}$ can be sparsely represented in a known overcomplete dictionary $\mathbf{D}$ which is the concatenation of the source-individual dictionaries $\mathbf{D} = [\mathbf{D}_1 \quad \mathbf{D}_2]$. That is, [17]

$$\vec{y} = \mathbf{D}\vec{\theta} \tag{4}$$

where $\vec{\theta}$ is the sparse code which is the concatenation of the source-individual codes $\vec{\theta}_i (i = 1,2)$, that is, $\vec{\theta} = \left[\vec{\theta}_1^{\mathrm{T}} \quad \vec{\theta}_2^{\mathrm{T}}\right]^{\mathrm{T}}$. The sparsest representation of $\vec{y}$ in $\mathbf{D}$ can be found by solving the following problem,

$$\min \left\|\vec{\theta}\right\|_0, s.t. \quad \vec{y} = \mathbf{D}\vec{\theta} \tag{5}$$

where $\|\cdot\|_0$ denotes the l_0 norm of a vector. The problem (5) is equivalent to the following problem,

$$\min f(\vec{\theta}_1, \vec{\theta}_2) = \min \sum_{i=1}^{2} \left\|\vec{\theta}_i\right\|_0, s.t. \quad \vec{y} = \sum_{i=1}^{2} \mathbf{D}_i \vec{\theta}_i \tag{6}$$

The solution of problem (5) is denoted as $\hat{\vec{\theta}}$, which is the concatenation of estimated source-individual codes $\hat{\vec{\theta}}_i (i=1,2)$, the solution of problem (6), that is, $\hat{\vec{\theta}} = \left[\hat{\vec{\theta}}_1^{\mathrm{T}} \quad \hat{\vec{\theta}}_2^{\mathrm{T}} \right]^{\mathrm{T}}$. If the source-individual dictionaries are diverse enough, we can separate $\vec{y}$ into its individual sources $\hat{\vec{s}}_i$ as [17]

$$\hat{\vec{s}}_i = \mathbf{D}_i \hat{\vec{\theta}}_i \tag{7}$$

As a consequence of above, there are two connected tasks to be solved in sparse decomposition based SCSS: learning source-individual dictionaries and computing sparse decomposition in (5). In this paper, we do not focus on dictionary learning and generate $\mathbf{D}_i (i=1,2)$ as the matrix consisting of the i^{th} speaker's training frames as columns called atoms. The proposed separation method using that dictionary is proved effective since it leads to a higher SNR and SIR in our experiments than the separation method based on SNMF which is a classical sparse decomposition based method. (The performance of separation using $\mathbf{D}_i$ generated as unsupervised clustering of training frames has also been tested, and it is much lower in SNR and SIR.)

For the computation of sparse decomposition, two classical approaches are used extensively: BP and OMP. The principle of BP is to find the optimal decomposition coefficients having the smallest l_1 norm. That is, one solves the problem [20]

$$\min \left\| \vec{\theta} \right\|_1 , s.t. \quad \vec{y} = \mathbf{D}\vec{\theta} \tag{8}$$

where $\|.\|_1$ denotes the l_1 norm of a vector. In [23], it has been proven that, the solution to the problem (5) can be approximated by the solution of the problem (8). Since problem (5) is NP-hard and difficult to solve, one turns to solve the problem (6) to obtain an approximate solution to (5). The solution of BP problem (6) can be obtained by solving an equivalent linear program as [20]

$$\min \vec{c}^{\mathrm{T}} \vec{z}, s.t. \quad \mathbf{A}\vec{z} = \vec{b} \tag{9}$$

where

$\mathbf{A} \Leftrightarrow (\mathbf{D}, \ -\mathbf{D})$; $\vec{b} \Leftrightarrow \vec{y}$; $\vec{c} \Leftrightarrow (1; \ 1)$; $\vec{z} \Leftrightarrow (\vec{u}; \ \vec{v})$;

$\vec{\theta} \Leftrightarrow \vec{u} - \vec{v}$.

OMP is a sparse approximation algorithm which is not based on optimization. It is a recursive algorithm to compute coefficients of atoms in a dictionary which is nonorthogonal and possibly overcomplete. It is a modification to MP by maintaining orthogonality of residual at each iteration, thus leads to improved convergence. The major advantage of OMP is its speed and ease of implementation. It achieves comparable performance as BP. The recovery of $\vec{y}$ based on sparse decomposition with OMP is given as follows [21-22].

Algorithm (Signal recovery with OMP)

INPUT:
Dictionary $\mathbf{D}$
Mixed signal $\vec{y}$
The sparsity level m of $\vec{y}$
OUTPUT:
An estimate $\hat{\vec{y}}$ of mixed signal
A set $\mathbf{D}^m$ containing chosen atoms to approximate $\vec{y}$

A sparse approximation $\hat{\vec{\theta}}^m$ of $\vec{y}$

A residual $\vec{r}^m = \vec{y} - \mathbf{D}^m \hat{\vec{\theta}}^m$

PROCEDURE:

1) Initialize the residual $\vec{r}^0 = \vec{y}$, the matrices of chosen atoms $\mathbf{D}^0 = \varnothing$, the iteration counter $t = 0$.

2) Find the index that solves the optimization problem [21-22]

$$\lambda^t = \arg \quad \max_{j=1,\dots,K} \left| \left\langle \vec{r}^t, \vec{d}_k \right\rangle \right| \tag{10}$$

where $\vec{d}_k$ is the k^{th} atom in $\mathbf{D}$, and K is the number of atoms in $\mathbf{D}$. Select $\vec{d}_{\lambda_t}$ as the newly selected atom.

3) Argument the matrix of chosen atoms $\mathbf{D}^t = \left[\mathbf{D}^{t-1} \quad \vec{d}_{\lambda^t} \right]$.

4) Solve a least squares problem to obtain a new approximation of $\vec{y}$ supported in $\mathbf{D}^t$ [21-22],

$$\vec{\theta}^t = \arg\min_{\vec{\theta}} \left\| \vec{y} - \mathbf{D}^t \vec{\theta} \right\|_2 \tag{11}$$

5) Calculate residual as [21-22],

$$\vec{r}^{t+1} = \vec{y} - \hat{\vec{\theta}}^t \mathbf{D}^t \tag{12}$$

6) Increment t, and return to step 2) if $t < m$.

The mixed signal is estimated by $\hat{\vec{y}} = \mathbf{D}^m \hat{\vec{\theta}}^m$.

As stated above, OMP picks atoms to recover the original signal in a greedy fashion. At each iteration, an atom that is most strongly correlated with the remaining part of the original signal called residual is chosen from the dictionary, and its contribution to the residual is subtracted off for the next iteration.

III. Proposed Separation Method

We will now proceed to describe the proposed new sparse decomposition based separation method using OMP. Fig.1 shows the block diagram of the proposed separation algorithm.

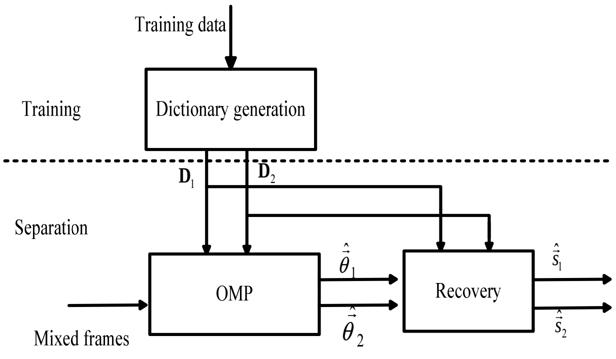


Fig.1. Block algorithm of proposed speech separation method using OMP.

As stated in Fig.1, the proposed algorithm works in two stages: training and separation. In the training stage, two source-individual dictionaries consisting of training frames of individual speakers as atoms are generated. In the separation stage, separation is performed frame after frame and then speech is synthesized by overlap-adding. Each mixed frame is separated based on sparse decomposition using OMP algorithm proposed.

In the following, we give the separation algorithm using OMP.

Algorithm (Separation using OMP)

INPUT:

Source-individual dictionaries $\mathbf{D}_1, \mathbf{D}_2$

Mixed frame $\vec{y}$

Stopping criterion

OUTPUT: (suppose procedure stops after m iterations):

Two estimates $\hat{\vec{s}}_1, \hat{\vec{s}}_2$ for original individual source frames $\vec{s}_1, \vec{s}_2$

Two sets $\mathbf{D}_1^m, \mathbf{D}_2^m$ containing chosen atoms to approximate $\vec{s}_1, \vec{s}_2$

Two decomposition coefficients $\hat{\vec{\theta}}_1^m, \hat{\vec{\theta}}_2^m$ supported in $\mathbf{D}_1^m, \mathbf{D}_2^m$ to approximate $\vec{s}_1, \vec{s}_2$

A residual $\vec{r}^{\,m} = \vec{y} - \left[\mathbf{D}_1^m \quad \mathbf{D}_2^m\right]\left[\hat{\vec{\theta}}_1^{m\mathrm{T}} \quad \hat{\vec{\theta}}_2^{m\mathrm{T}}\right]^{\mathrm{T}}$

PROCEDURE:

1) Initialize the residual $\vec{r}^{\,0} = \vec{y}$, the matrices of chosen atoms $\mathbf{D}_1^0 = \varnothing, \mathbf{D}_2^0 = \varnothing$, the iteration counter $t = 0$.

2) Find the index that solves the optimization problem (10).

3) Merge the newly selected atom with the previous matrices of chosen atoms,

$$\mathbf{D}_1^t = \begin{cases} \left[\mathbf{D}_1^{t-1} \quad \vec{d}_{\lambda^t}\right], & \text{if} \quad \lambda^t \le K_1 \\ \mathbf{D}_1^{t-1}, & \text{others} \end{cases} \tag{13}$$

$$\mathbf{D}_2^t = \begin{cases} \left[\mathbf{D}_2^{t-1} \quad \vec{d}_{\lambda_t}\right], & \text{if} \quad K_1 \le \lambda^t \le K_1 + K_2 \\ \mathbf{D}_{t-1}^2, & \text{others} \end{cases} \tag{14}$$

where K_i is the number of atoms in $\mathbf{D}_i$, satisfying $K = \sum_{i=1}^{2} K_i$.

4) Solve the least squares problem (11) to obtain a new decomposition coefficient to approximate $\vec{y}$ supported in $\mathbf{D}^t$ [21-22], the concentration of $\mathbf{D}_1^t$ and $\mathbf{D}_2^t$, $\mathbf{D}^t = \left[\mathbf{D}_1^t \quad \mathbf{D}_2^t\right]$ .The solution of (11) is given by $\hat{\vec{\theta}}^t = \left(\mathbf{D}^t\left(\mathbf{D}^t\right)^{\mathrm{T}}\right)^{-1}\mathbf{D}^t\vec{y}$ [21-22]. We denote $\hat{\vec{\theta}}_1^t$ and $\hat{\vec{\theta}}_2^t$ as the parts of $\hat{\vec{\theta}}^t$ belonging to the support $\mathbf{D}_1^t$ and $\mathbf{D}_2^t$ respectively.

5) Update residual as (12).

6) Increment t, and return to step 2) until satisfying $\left\|\vec{r}^{\,t}\right\|_2 \le \delta_e$ or $\max_{j=1,\ldots,K}\left|\left\langle \vec{r}^{\,t}, \vec{d}_k \right\rangle\right| \le \delta_c$ where δ_e and δ_c are chosen thresholds.

The separated speech source is estimated by $\hat{\vec{s}}_i = \hat{\vec{\theta}}_i^m \mathbf{D}_i^m (i = 1,2)$.

IV. Experiments

As a proof of concept, we evaluate the proposed separation algorithm using the Grid corpus provided for SCSS by Cooke et. al [24] and compare its performance with SNMF based SCSS method which is a classical sparse decomposition based method [17-18]. We selected four speakers including two female (speakers 18 and 20) and two male speakers (speakers 1 and 2) from the database and denoted them as F1, F2, M1 and M2 in sequel. For each speaker, half of the sentences in the database were used for training and ten other sentences are selected randomly for testing. Speech sources are added directly at 0 dB SNR for each speech pair to have 400 female-male mixtures, 100 female-female mixtures and 100 male-male mixtures. The original sampling frequency was decreased from 25 kHz to 8 kHz and a hamming window of duration 32 ms with a frame-shift of 16 ms was used.

To evaluate the separation performance, average of signal-to-noise ratio (SNR) and average of source-to-interferences ratio (SIR) are used. SNR is defined as

$$\mathrm{SNR}_i = 10\times \lg(\frac{\vec{x}_i\vec{x}_i^{\mathrm{T}}}{(\vec{x}_i-\hat{\vec{x}}_i)(\vec{x}_i-\hat{\vec{x}}_i)^{\mathrm{T}}}) \qquad (15)$$

where $\vec{x}_i$ and $\hat{\vec{x}}_i$ are original and estimated source speech signals respectively, and SNR_i is the SNR of estimated $\hat{\vec{x}}_i$. SIR is defined as in [25].

In our experiments we set $\delta_e = 10^{-10}$ and $\delta_c = 10^{-5}$.

We compare the separation results of the proposed method using OMP with that of the separation method using SNMF in SNR and SIR respectively. The average results of 600 mixtures tested are shown in TABLE I and II. The same training sentences are used to generate each SNMF source dictionary with the sparsity $\lambda = 0.1$ and the size of 560 as in [36].

TABLE I
PERFORMANCE OF SEPARATION USING SNMF AND PROPOSED SEPARATION METHOD USING OMP IN SNR (DB) ON 600 MIXTURES

	SNMF	Proposed method
F/F	4.7/4.5	4.8/4.7
F/M	5.5/5.3	5.7/5.8
M/M	3.3/3.9	3.7/4.4

TABLE II
PERFORMANCE OF SEPARATION USING SNMF, TRADITIONAL OMP AND OMP IN SIR (DB) ON 600 MIXTURES

	SNMF	Proposed method
F/F	5.0/6.3	8.7/11.4
F/M	9.3/8.7	12.7/11.7
M/M	3.3/4.9	8.9/8.9

As shown in TABLE I and II, the proposed separation method outperforms the method using SNMF in both SNR and SDR. The average SNR result of the proposed method is 0.15dB, 0.35dB and 0.45dB higher than that of the method using SNMF for female/female, female/male and male/male mixture respectively. The average SIR result of the proposed method is 4.4dB, 3.2dB and 4.8dB higher than that of the method using SNMF for female/female, female/male and male/male mixture respectively.

We also report the separation results of the sentences which are shown in TABLE III.

TABLE III
LABELS OF SPEAKERS AND FILE NAMES USED FOR TESTING

F1	speaker 18	"lwix2s" "sbil4a" "prah4s"
F2	speaker 20	"lwwy2a" "sbil2a" "prbu5p"
M1	speaker 1	"pbbv6n" "sbwozn" "prwkzp"
M2	speaker 2	"lwwm2a" "sgai7p" "priv3n"

For each speech pair, the speech sources are added directly to form a mixture, resulting 54 mixtures including 36 male-female mixtures, 9 male-male mixtures and 9 female-female mixtures.

Fig.2. shows the first 16 atoms chosen using OMP to separate a mixed frame. In this example, 70 and 51 atoms are selected to estimate two sources based on OMP.

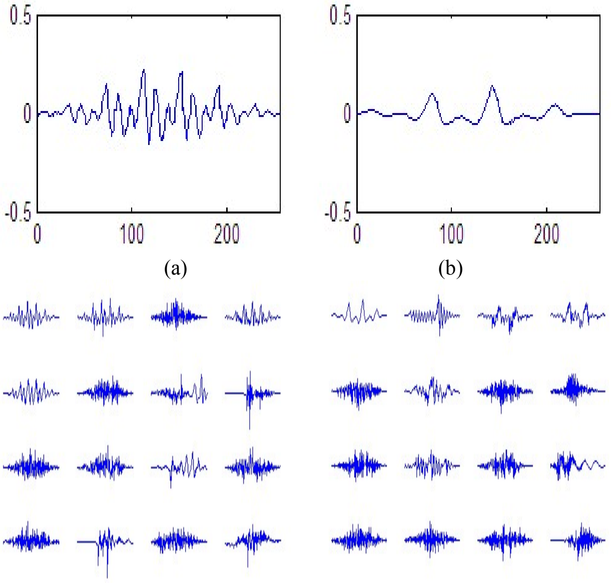


Fig.2. Waveforms of sources and selected atoms.(a-b) source frames 1,2; (c-d) the first 16 atoms selected to estimate source frames 1,2 using traditional OMP;

We compare the average results of separation using OMP with the separation method using SNMF in SNR and SIR respectively. The results are shown in TABLE IV and V.

From TABLE IV and V, it is also observed that the proposed method outperforms the separation method using SNMF in both SNR and SIR. The proposed method achieves 0.2dB higher SNR and 2.55dB higher SIR results respectively than the method using SNMF for female/female mixtures. It achieves o.15 dB higher SNR and 1.35dB higher SIR result than the method using SNMF for female/male mixtures. It achieves o.3 dB higher SNR and 3.9dB higher SIR result than the method using SNMF for male/male mixtures.

TABLE IV
PERFORMANCE OF SEPARATION USING SNMF AND PROPOSED SEPARATION METHOD USING OMP IN SNR (DB) ON 54 MIXTURES

	SNMF	Proposed method
F/F	4.4/4.1	4.2/4.3
F/M	5.5/5.7	5.7/5.8
M/M	3.2/3.3	3.4/3.7

TABLE V
PERFORMANCE OF SEPARATION USING SNMF, TRADITIONAL OMP AND OMP IN SIR (DB) ON 54 MIXTURES

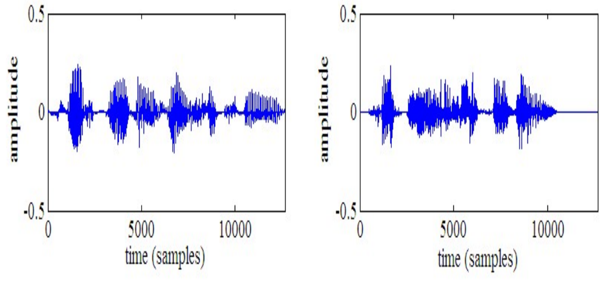
	SNMF	traditional OMP
F/F	6.8/3.5	6.0/7.8
F/M	8.3/9.6	9.3/11.3
M/M	1.3/3.6	5.1/7.6

Finally, Fig.2. illustrates the waveforms of the original sources and the separated results for the mixture of a female and male speech.

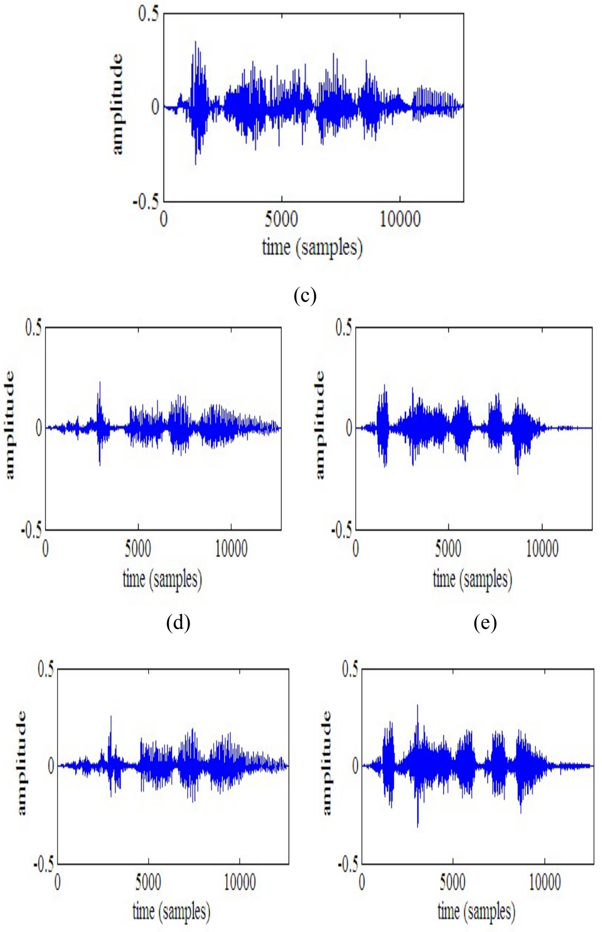
V. CONCLUSION AND FUTURE WORK

In this paper, we presented a new sparse decomposition based SCSS method using OMP. In the proposed method, we generate source dictionaries as matrices consisting of time-domain training frames as atoms and use OMP for the computation of sparse coefficients in the union of source dictionaries. We compared our separation results to the results of separation using SNMF, and showed that our proposed method achieved higher SNR and SIR.

In the proposed separation method, matrices consisting of training frames as atoms are used as source-individual dictionaries directly. In the future, we plan to unite dictionary learning and the presented separation work to improve separation speed. Moreover, we plan to discuss whether we can improve the OMP algorithm for separation by considering mutual independence between sources.



(a) (b)



(c)

(d) (e)

(f) (g)

Fig.2. Separated speech waveforms of a female and a male speech. (a-b)original sources; (c) mixed signal ; (d-e) separated sources estimated using SNMF based method; (f-g) separated sources estimated using OMP.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (No. 61302152, No. 61201345 and No. 61271240), the Beijing Key Laboratory of Advanced Information Science and Network Technology. (No. XDXX1308) and the Engineering Colledge of Nanjing Agricultural University Introduced Talent Start-up Fund Project (No. rcqd11-02).

REFERENCES

[1] J. Wen 1, S. Zhang, and J. Yang, “A fast algorithm for undedetermined mixing matrix identification based on mixture of guassian (MoG) sources model,” *Journal of Software*, vol. 9, no. 1, pp. 184-189, 2014.

[2] B. Yu, H.F Li, C.Y. Fang, “Speech Emotion Recognition based on Optimized Support Vector Machine,” *Journal of Software*, vol. 7, no. 12, pp. 2726-2733, 2012.

[3] Y. Shao, S. Srinivasan, Z.Z Jin and D.L. Wang, “A computational auditory scene analysis system for speech segregation and robust speech recognition,” Computer Speech and Language, vol. 24, pp. 77-93, 2010.

[4] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” *J. Acoust.Soc. Amer.*, vol. 126, pp. 1486–1494, 2009.

[5] G.J. Jang and T.W. Lee, “A maximum likelihood approach to single-channel source separation”, *Journal of Machine Learning Research*, 4, pp. 1365-1392, 2003.

[6] M. Stark, M. Wohlmayr and F. Pernkopf, “Source–Filter-Based Single-Channel Speech Separation Using Pitch Information,” *IEEE Transcations on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, Feb. 2011.

[7] D.Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” *Speech Separation by Humans and Machines*, Kluwer Academic, Norwell MA, pp. 181-197, 2005.

[8] M.G. Christensen, “Metrics for vector quantization-based parametric speech enhancement and separation,” *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp.3062-3065, 2013.

[9] M.H. Radfar and R.M. Dansereau, “Single-channel speech separation using soft mask filtering,” *IEEE Transcations on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[10] M.H. Radfar, R.M. Dansereau and A. Sayadiyan, “A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation”, *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(1).

[11] R.J. Weiss andD.P.W. Ellis, “Speech separation using speaker-adapted eigenvoice speech models,” Computer Speech and Language, vol. 24, pp. 16-29, 2010.

[12] P. Mowlaee, M.G. Christensen, S.H. Jensen, “New results on single-channel speech separation using sinusoidal modeling,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no. 5, pp. 1265-1277, 2011.

[13] M.V.S Shashanka, B. Raj and P. Smaragdis, “Sparse overcomplete decomposition for single channel speaker separation,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. II-641-II-644, 2007.

[14] B. Raj and P. Smaragdis, “Latent variable decomposition of spectrograms for single channel speaker separation,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 17-20, 2005.

[15] M. Moussallam, G. Richard and L. Daudet, “Audio source separation informed by redundancy with greedy multiscale decompositions,” *IEEE European Signal Processing Conference (EUSIPCO)*, pp. 2644-2648, 2012.

[16] B.A. Pearlmutter and R.K. Olsson, “Linear program differentiation for single-channel speech separation,” *IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pp. 421-426, 2006.

[17] M.N. Schmidt and R.K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” *International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

[18] B.L. Zhu, W.L., R.J. Li and X.Y. Xue, “Multi-Stage Non-Negative Matrix Factorization for Monaural Singing Voice separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no.10, pp. 2096-2107, 2013.

[19] Rémi Gribonval and Morten Nielsen, “Sparse Representations in Unions of Bases,” IEEE Transactions on Information Theory, vol.49, no.12, pp: 3320-3325, 2003.

[20] Y. Li, A. Cichocki, and S.-I. Amari, “Analysis of sparse representation and blind source separation,” *Neural Comput.*, vol. 16, no. 6, pp. 1193–1234, 2004.

[21] Y.C. Pati, R. Rezaiifar and P.S. Krishnaprasad, “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” *1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pp. 40-44, 1993.

[22] J.A. Tropp and A.C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Transactions on Information Theory*, vol. 53, no.12, pp. 4655-4666, 2007.

[23] Y.Q. Li, A. Cichocki, S. Amari, S.L. Xie and C. Guan, “Equivalence Probability and Sparsity of Two Sparse Solutions in Sparse Representation,” IEEE *Transactions on Neural Networks*, vol. 19, no. 12, pp. 2009-2021, Dec 2008.

[24] M.P. Cooke, J. Barker, S.P. Cunningham, and X. Shao, “An audiovisual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[25] E. Vincent, C. Fevotte and R. Gribonval, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

Haiyan Guo received her Ph.D. degree in signal and information processing from Nanjing University of Posts & Telecommunication (NUPT), Nanjing, China in 2011.

She is a lecturer in the college of engineering, Nanjing Agricultural University. She is currently working as a postdoctoral researcher in the school of information science and engineering, Southeast University. Her research interests mainly include speech signal processing, sparse decomposition and compressed sensing.

Xiaoxiong Li received his B.S. degree in signal and information processing from Hohai University, Nanjing, China in 2012. He is currently pursuing the M.S. degree in signal and information processing in the school of information science and engineering, Southeast University.

His research interests mainly include speech signal processing.

Lin Zhou received her Ph.D. degree in signal and information processing from Southeast University, Nanjing, China in 2005.

She is currently an associate professor of signal and information processing in the school of information science and engineering, Southeast University. Her research interests mainly include speech signal processing and spatial hearing.

Zhenyang Wu received his M.S. degree in electrical engineering from Southeast University, Nanjing, China in 1982.

He is currently a professor of signal and information processing in the school of information science and engineering, Southeast University. During 2000–2008, he was the vice dean of the School of Information Science and Engineering, Southeast University. His research interests include speech and audio processing, image and multimedia processing, and sensor array signal processing. He has published more than 100 international journal and conference paper. From 1990s, he has been an invited reviewer of several famous scientific journals. He is a number of IEEE.

Prof. Wu served as a member of technical program committee for the 13th IEEE International Symposium on Consumer Electronics (ISCE2009). In 2008, he received the National Award for Distinguished Teacher.

A Partner Selection Method for Forming Innovation Alliance

Yang Yang
School of Management and Economics, Beijing Institute of Technology, Beijing, China
Email: dexter@bit.edu.cn

Wendai Lv
School of Business, Renmin university of China, Beijing, China

Guangming Hou
School of Management and Economics, Beijing Institute of Technology, Beijing, China

Junpeng Wang
School of Management and Economics, Beijing Institute of Technology, Beijing, China

Abstract—**Establishing innovation alliance has now become a very important and indispensable way for achieving the mission of research and development (R&D) of national mega projects in China. As an important component for constructing the collaborative partnership system, an appropriate approach of partner selection should be able to evaluate the performance of candidates objectively. Innovation alliance has members with different backgrounds, but the most suitable partner should be involved in the alliance by launcher before establishing alliance. For this reason, the aim of this paper is to take the different characteristics of organization as the influence factors into the account of weight setting, and propose a fuzzy analytic hierarchy approach (fuzzy AHP) to effectively evaluate candidates. According to the extension principle of fuzzy set theory, linguistic variables defined as fuzzy numbers are applied to pair-wise comparisons to avoid the vague situation, and this study proposes an approximate method for calculating the multiplication products of fuzzy number. This method can handle the vagueness and incompletion during the process of evaluation. A case study is also given to demonstrate the potential of the methodology.**

Index Terms—**Innovation Alliance, National Mega Project, Partner Selection, Fuzzy AHP**

I. Introduction

With the development of hi-tech economies, a series of national mega projects has been launched by the Chinese government in recent years. For example, research and development (R&D) about aircraft engines has been listed as a national mega project in 2012. A hundred billion RMB is going to be invested in this project in order to conduct R&D over the next five years. National mega projects need huge investment, advanced technology, products innovation, and the acquisition of sufficient resources and fundamental research. The most important feature of these projects is innovation range from product to technique, and the success of them is also considered the symbol of innovation ability of China [1-2]. All kinds of organizations are eager to be involved in the R&D or manufacturing activities of these projects due to the high profits and bright perspectives. No organization, however, is able to fulfill the whole process of the project because of lack of resources and capabilities. Therefore, the importance of co-operation within different organizations has been emerging. Establishing a form of innovation alliance through different organizations, such as government sectors, enterprises, universities and academic institutions, may be an attainable way to acquire necessary resources and techniques for innovation. Agility and innovation are becoming increasingly important for creating value from products. Innovation alliance has currently become an important and indispensable form of co-operation between different organizations.

Although the concept about strategy alliance or virtual enterprise has been widely applied in the theoretical and practical field in the previous literature, some researchers [3-5] still found that incompatibility of partners is one of the most common reasons for failure, which means organizations in alliance cannot be satisfied with each other, or they were unable to achieve their assigned responsibility and finally resulted in collapse. Hence, selecting the appropriate partners must be carefully considered before such a partnerships system can be built.

It can be seen from previous literature that the research methods for partner selection have been studied from simple weighted scoring models to complex mathematical programming approaches [6]. Analytic hierarchy process (AHP) is one of the most common methods, introduced by Saaty [7], which is for solving unstructured problems. Although this method has been widely applied for evaluating the relative importance of a set of activities in a multi-criteria decision problem, it cannot handle the uncertainty and vagueness associated

doi:10.4304/jsw.9.11.2981-2988

with mapping of one's decision to a number. Some researchers employed the fuzzy set theory to deal with the imprecision and vagueness, which from the subjective perception and the experience of humans in the decision-making process. These studies have extended the method of AHP to deal with the pair-wise comparison process using fuzzy utilities represented by fuzzy numbers. Chang [8] proposed an extent analysis approach for the synthetic extent values of the pair-wise comparison for handling fuzzy AHP. In the study of Bevilacqua, he proposed an approach of fuzzy quality function deployment to conduct supply partner selection. His approach addresses both internal and external variables to rank the potential partners, and transform the decision makers' verbal assessments to linguistic variables, which are more accurate than other non-fuzzy methods [9]. Yucel et al developed a weighted additive fuzzy programming approach for multi-criteria partner selection [10]. Fuzzy set theory allows the decision makers to incorporate unquantifiable information, incomplete information and non-obtainable information into decision model [11]. Their model has diminished the computational procedure, so it can deal with the rating of factors effectively.

However, the weighting process of criteria is affected by influence factor has not been considered in most prior research. In our case, the innovation alliance is a dynamic structure formation. Due to the particular merits, the partner in an innovation alliance might be a member in another alliance as well. It means that the biggest, richest or the most powerful organization may be not the most suitable partner for this alliance, so what type of organization they needed most for innovation has to be known before establishing alliance. Since the candidates needed for collaboration from different backgrounds, we address the characteristics of organization as influence factors for the weight setting of criteria because a general set of criteria cannot consider the priority of each organization. Additionally, the relative weights for each criterion with respect to each characteristic are calculated by the composite relative important weights.

The remainder of this paper is organized as follows. Section 2 describes the details of the proposed evaluation framework and the criteria. The process of weight setting and candidate evaluation is given in section 3. Section 4 illustrates an example with proposed method. In section 5, conclusions for this study are given.

II. Framework Of Evaluation

We structure the AHP model hierarchically based on the organization characteristics and criteria. The objective of partner selection is the first level of evaluation model, organization characteristics as influence factors in the second level, the criteria and sub-criteria are on the third and the forth level respectively.

A. *Organization Characteristics*

For achieving the mission of R&D, the launcher of alliance intends to build cooperative partnership with other organizations for relieving financial pressure, reducing R&D risk, shortening the research time, exchanging information, increasing the market share and so on [12-13]. Different patterns of innovation usually have different efforts and needs. When these partners are involved in R&D activities, launcher must decide whether and how to cooperate with other organizations. In practical situation, the organizations within national mega project usually are possibly constituted of governmental sector, enterprise, university and academic institution. All of them own their particular merits, which may play a special role during innovation process. Hence, the function of each partner in alliance also has to be known before establishing. Based on this, the types of organizations will be defined to strategy-based, capital-based, resource-based and learning-based in this study.

Strategy-based: To collaborate with the type of this organization like governmental organizations could obtain benefits of tax policy, the classified information or political privilege, which can accelerate the speed of innovation.

Capital-based: Capital investment has a positive impact on technological innovation, because more money invested in innovation activities can accelerate the speed of innovation. Alliance stands poised to benefit from the investments of cash-rich organization. These organizations can provide the financial support to develop the research quality of product, and also to share the cost of R&D.

Resource-based: Innovation is a complex process and requires many significant resources. The launcher of alliance could get these critical resources from other organizations. These resources include equipment, techniques, marketing channels, experts and other key resources for innovation.

Learning-based: The newest knowledge and technology is the key element for innovation. These organizations could also help launcher to solve the problem about human resources. Researchers in alliance can learn from the partners by conducting joint technology development.

B. *Evaluation Criteria*

As different organizations have different purposes and motivations for establishing innovation alliance, the identification of universal criteria weights for use in any situation will not be appropriate. The purpose of establishing innovation alliance in this paper is to make the breakthrough of new products or techniques, and finally to achieve the objective of the project. Innovation alliance needs to select partners that have common goals, burning desire, sophisticated skills, and complementary resources. Thus, every member should have the idea of sharing investment, management, risks and responsibility for profits and losses. Although meeting the requirement of innovation is the primary purpose of alliance, it can also bring huge profits and opportunities to the members in alliance. There is no doubt that it is a win-win situation.

The criteria for evaluating the performance of candidates have been discussed in the literature both theoretical and practical field. Geringer [14] was one the first man to conduct the study of partner selection criteria, he found even though there is no optional criteria for partner selection procedure, partners' culture, past experience, size, and structure were as important as the traditional criteria, such as financial assets, access to markets, and technical know-how. In the study of [15], it emphasized that complementary resources, symmetrical position and extension of social resources are necessary conditions for becoming a partner in an alliance.

Based on a detail literature survey, we employ the following four criteria for innovation partner selection mechanism, which are cooperative willingness, financial ability, complementary resources and technological ability. For each criterion, a cluster of sub-criteria for evaluating the suitability of candidate partners are also addressed.

The framework for partner selection is established as described in Figure.1.

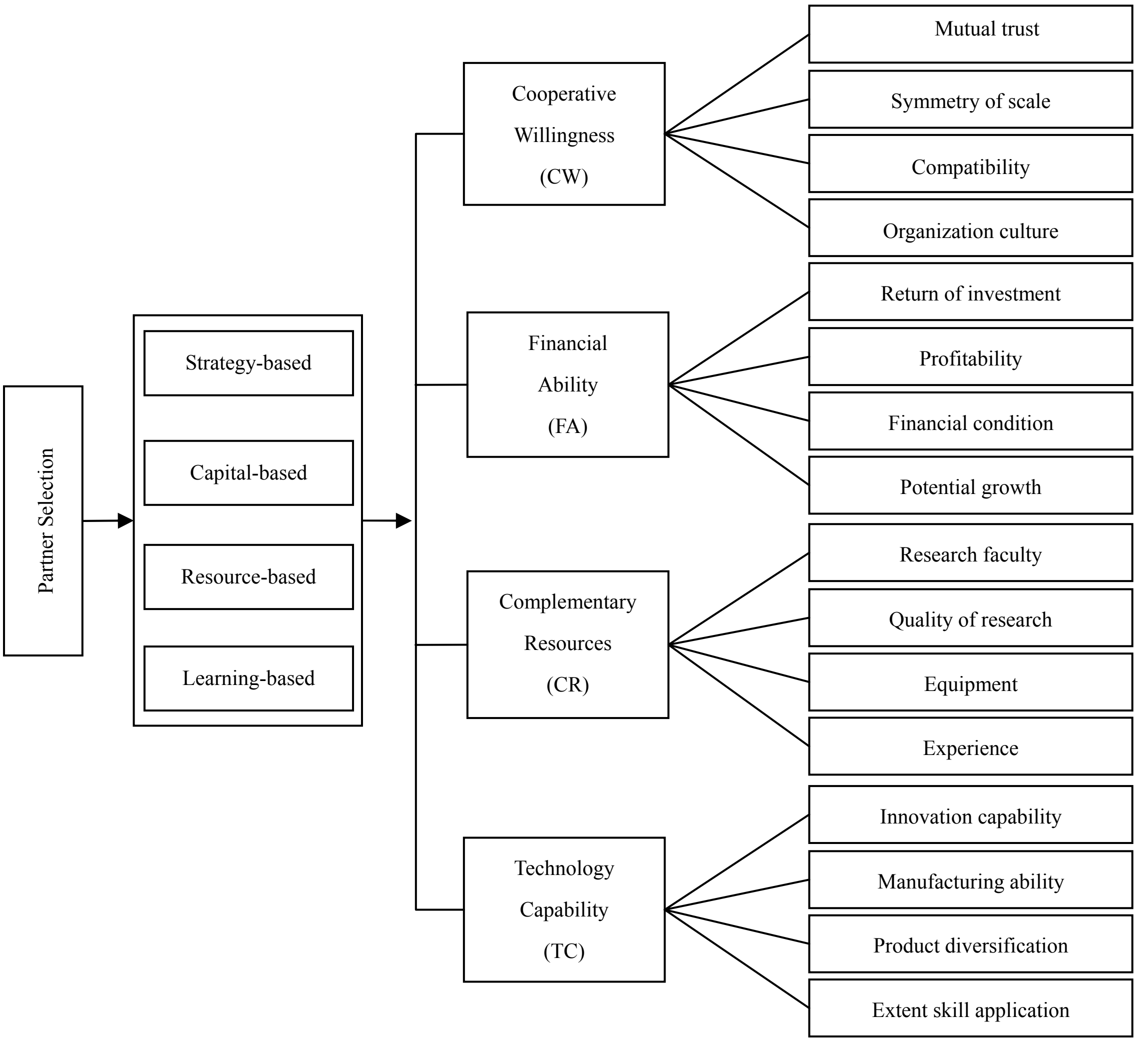


Figure 1. AHP model for partner selection

III. The Proposed Fuzzy AHP Modol

This paper proposes fuzzy AHP to solve the problem of multiple criteria decision-making. The framework of evaluation in this study is designed by AHP, which must be settled before these methods are effectively employed to assess. However, weight setting and partners are selected by pair-wise comparison, which is composed of the linguistic variables defined as fuzzy numbers. The process of weight setting and evaluation is composed of the following steps:

Step 1: Determine the intensities of each characteristic by combining the scores from all decision makers.

Step 2: Identify the relative importance of the criteria with respect to characteristics and calculate the composite weights of relative importance of criteria.

Step 3: Calculate the composite weights of sub-criteria as last step.

Step 4: Use the pair-comparison matrices to evaluate the performance of candidate with each other according to the measurable sub-criteria. Linguistic variables are used in this step.

Step 5: Synthesize the suitability index of each candidate by summing up the results of multiplying the normalized weight score of each criterion with the normalized relative importance of this criterion.

During this process, we apply an approximate method of Yager's to deal with the multiplication of fuzzy number [16]. This method can rank the fuzzy number and also ensure to properly reflect the evaluations from decision maker. It diminishes the load of calculation compare to the conventional method.

A. Method for Weight Determination

Innovation alliance may have members with different kinds of characteristics according to the needs of innovation in term of the particularity and complexity of national mega project, each of whom might play a special role during the process of innovation. Thus, the launcher of alliance has to realize what type of partner is the most needed for conducting the collaboration innovation activities. And different intensities of organizational characteristics will affect the weight set for the criteria importance. For the convenience of illustration, we assume there are k experts who are involved in a partner selection issue. A unit scales is employed to express the degrees of intensity ranging from very unimportant, unimportant, moderate and important, to very important and denoted by consecutive decimal numbers from 0 to 1. For example, suppose the number x_{ip} is the evaluation from the i th expert for the degree of intensity of p th organizational characteristic. By combining the scores of all experts, the composite fuzzy weight of the p th organizational characteristic could be expressed by the following triangular fuzzy number[17]:

$$\widetilde{X}_p = (a_p, b_p, c_p) \tag{1}$$

In which

$$a_p = \min_i(x_{ip}),\ c_p = \max_i(x_{ip}),$$

$$b_p = \left(\frac{\Pi_{i=1}^k x_{ip}}{a_p \times c_p}\right)^{\frac{1}{k-2}},\ p = 1,...,p$$

The following step is to determine the relative importance of the four criteria relating to each organizational characteristic. Adjusting the weight for importance of the criteria is emphasized as a particular organizational characteristic, which can also make sure that the most satisfied candidate for the particular organizational characteristic preference of alliance should be considered as a partner firstly. Similar as the previous step, suppose q represents the criteria and y_{ipq} means the evaluation from the i th expert for the q th criteria with regard to the p th organizational characteristic. Therefore, the composite relative importance is obtained as following:

$$\widetilde{Y}_{pq} = (d_{pq}, e_{pq}, f_{pq}) \tag{2}$$

In which

$$d_p = \min(y_{ipq}),\ e_p = \max(y_{ipq}),$$

$$f_p = \left(\frac{\Pi_{i=1}^k y_{ipq}}{d_{pq} \times f_{pq}}\right)^{\frac{1}{k-2}},\ p = 1,...,p;\ q = 1,...,Q$$

Accordingly, multiplying the Eq. (1) with Eq. (2) can obtain the composite fuzzy relative importance for the q th criterion as follow:

$$\widetilde{Z}_q = \sum_{p=1}^{p} \widetilde{X}_p \otimes \widetilde{Y}_{pq},\ q = 1,...,Q \tag{3}$$

With reference to the extension principle of fuzzy sets and the definition of the triangular fuzzy number, $\widetilde{Z}_q$ is still a fuzzy number. For simplicity, the relationship function of fuzzy number can be expressed as a non-fuzzy number by using the approximation formula, as follow:

$$\widetilde{Z}_q = \sum_{p=1}^{p} \widetilde{X}_p \otimes \widetilde{Y}_{pq} \cong (c_{q1} = c_1, c_{q2} = c_{2,}\ c_{q3} = c_3) \quad （4）$$

For diminish the load of calculation, we applied the centroid ranking method, which was proposed by Yager, for rank the fuzzy number. After some mathematical rearrangement, the centroid rank value of the approximated triangular fuzzy number is:

$$R(\widetilde{Z}_q) = \frac{1}{4}(c_{q1} + 2c_{q2} + c_{q3}),\ q = 1,...,Q \tag{5}$$

B. Fuzzy evaluation for candidate partners

Each criterion must assess each potential partner via a set of measurable sub-criteria. The relative importance of four sub-criteria related with its upper criterion must be determined before conducting the evaluation. Similarly, the composite relative importance for the s th sub-criterion with respect to its upper q th criterion for the expert members could be expressed as the following triangular fuzzy number:

$$\widetilde{L}_{qs} = (a_{qs}, b_{qs}, c_{qs}) \tag{6}$$

In which,

$$a_{qs} = \min(l_{iqs}),\ c_{ps} = \max(l_{iqs}),\ b_{ps} = \left(\frac{\Pi_{i=1}^k l_{iqs}}{a_{qs} \times c_{qs}}\right)^{\frac{1}{k-2}}$$

$$q = 1,...,Q;\ s = 1,...,S$$

According to the definition of each sub-criterion, each expert conducts a series of pair-wise comparisons to evaluate the performance of these candidates in the next step. A seven-point linguistic scale is employed to express their relative performance. The relationship functions of the linguistic values are shown in Figure 2, and defined as follows:

Extremely poor (EP): (0, 0, 0.1)
Very poor (VP): (0.05, 0.2, 0.35)
Poor (P): (0.2, 0.35, 0.5)
Mediate (M): (0.35, 0.5, 0.65)
Good (G): (0.5, 0.65, 0.8)
Very good (VG): (0.65, 0.8, 0.95)
Extremely good (EG): (0.9, 1, 1)

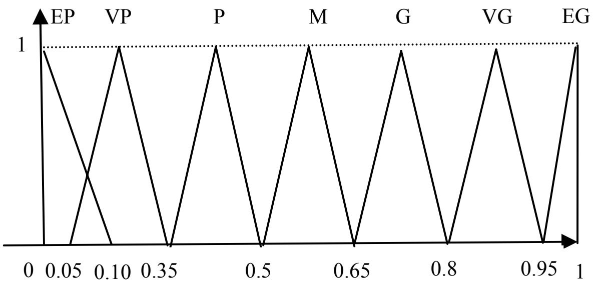


Figure 2. Linguistic variables for criteria rating

The results about the linguistic variable of pair-wise comparison are described as a matrix form. The decision makers just gives their own opinion in the right part of the matrix according to each criterion. The left part then is obtained automatically as the "reciprocal" of the right part of the matrix. For the convenience of calculation let $\widetilde{A} = [\widetilde{m}]_{T\times T}$ be the comparison matrix. The arithmetic average of each row can be calculated as the performance evaluation of the t th candidate partner on the s th sub-criterion of its upper-level criterion, which evaluated by the i th expert. This average can be denoted by the triangular fuzzy number

$$\widetilde{P}_{iqst} = \frac{1}{T}\sum_{v=1}^{T} \widetilde{m}_{tv} = (m_{at},\ m_{bt},\ m_{ct}),\ \ t = 1,...,T$$

Similarly, the composite performance evaluation of one candidate for a sub-criterion from experts could be expressed as:

$$\widetilde{M}_{qst} = \frac{1}{K}\sum_{i=1}^{k} \widetilde{P}_{iqst} = (d_{pq}, e_{pq}, f_{pq}) \qquad (7)$$
$$t = 1,...,T$$

Then, the composite weighted performance evaluation of the t th candidate on the q th criterion can be calculated:

$$\widetilde{N}_{qt} = \frac{1}{K}\sum_{s=1}^{qs} \widetilde{L}_{qs} \otimes \widetilde{M}_{qst} \qquad (8)$$
$$q = 1,...,Q;\ t = 1,...,T$$

The centroid ranking method can be used again, as foll

$$R(\widetilde{N}_{qt}) = \frac{1}{4}(c_{qt1} + 2c_{qt2} + c_{qt3}) \qquad (9)$$
$$q = 1,...,Q;\ t = 1,...,T$$

Finally, the suitability index is employed to compare the performance of these candidates and indicates which one is the most suitable partner, each of which could be synthesized the product of multiplying the composite performance on each criterion with its relevant composite important weight as following equation:

$$S_t = \sum_{q=1}^{Q} R(\widetilde{Z}_q) \otimes R(\widetilde{N}_{qt}) \qquad t = 1,...,T \qquad (10)$$

IV. Illustrative Example

The proposed model for the innovation alliance partner selection is used in one project, which provides the theoretical support for the decision maker. This project is one of the important parts about electric vehicle. The car industry currently is an important part of Chinese economic system. Many organizations are throwing themselves into the innovation activities of new energy vehicles. Like some developed counties, Lots of Chinese institutions and organizations are also eager to be involved in car energy programs as the bright market prospects. Along with the rapid development of Chinese car industry, low-cost, high-quality, and customer-oriented new products are needed to satisfy the requirement of customers agilely. The organization for innovation can effectively employ the product development and innovation capacity of the members within alliance. Therefore, the case company set up an expert group for select appropriate partner for innovation, which consists of different fields, such as technique, product, financial, innovation and strategy. We suppose there are five experts in the group, who held meeting to discuss the issue of partner selection following the steps detailed in the previous sections, and trying to gain a consensus by giving their own opinion about each criterion.

As seen in the data of Table 1, the left part depicts the weight distribution of each characteristic from five experts' opinion respectively. The fuzzy intensity index for four characteristics was calculated by Eq. (1) in the right three columns. The fuzzy index illustrates that the most important characteristic of organization is resource-based, followed by learning-based, which means that the objective of launcher for establishing alliance is mainly looking for a partner with complementary resources for innovation.

TABLE I.
The Intensity Index For Each Characteristic

	E.1	*E.2*	*E.3*	*E.4*	*E.5*	***Fuzzy intensity***		
						a_p	b_p	c_p
Strategy-based	0.350	0.205	0.122	0.200	0.215	0.122	0.207	0.350
Capital-based	0.180	0.340	0.136	0.160	0.225	0.136	0.186	0.340
Resourced-based	0.300	0.230	0.422	0.310	0.335	0.230	0.315	0.422
Learning-based	0.170	0.225	0.320	0.330	0.225	0.170	0.253	0.330

Same as the last step, after collecting and normalized the data from experts, the relative importance of four criteria relating to each characteristic are obtained by Eq. (2). We summarize the fuzzy weight for these criteria with different organizational characteristics in Table 2. Following this step, the composite importance weights of criteria are available by Eq. (3) with data in Table 1 and Table 2. It can be seen in Table 3 that defuzzified weight of each criterion is established by the mathematical rearrangement by means of Eq. (5). Apparently, complementary resources and technology capability are emphasized, which might be affected by the fuzzy intensity of characteristic. Additionally, the normalized weight in Table 3 is employed in the final step of evaluation.

In the next phase, the sub-criteria of each criterion are used to evaluate the performance of the candidate partners. The relative weights of importance of the sub-criteria about the upper criterion from which they develop must be determined before they can be applied to the evaluation process. Eq. (7) is employed to set up the fuzzy weights of these sub-criteria with respect to their upper-level criterion. Table 4 describes the composite

TABLE II.
THE APPROXIMATED FUZZY RELATIVE IMPORTANCE WEIGHT OF THE CRITERIA FOR CHARACTERISTICS

	Strategy-based			***Capital-based***			***Resourced-based***			***Learning-based***		
	a_p	b_p	c_p	a_p	b_p	c_p	a_p	b_p	c_p	a_p	b_p	c_p
Cooperative Willingness	0.190	0.239	0.320	0.220	0.303	0.330	0.220	0.237	0.260	0.150	0.192	0.260
Financial Ability	0.240	0.293	0.360	0.125	0.168	0.220	0.160	0.196	0.215	0.150	0.165	0.200
Complementary Resource	0.100	0.189	0.310	0.240	0.260	0.330	0.250	0.280	0.340	0.280	0.344	0.450
Technology Capability	0.220	0.243	0.350	0.240	0.272	0.280	0.230	0.288	0.320	0.250	0.270	0.340

TABLE III.
THE COMPOSITE FUZZY WEIGHTS OF THE RELATIVE IMPORTANCE OF CRITERIA

	Composite fuzzy weight			***Defuzzified weight***	***Normalized weight***
	C_{q1}	C_{q2}	C_{q3}		
Cooperative Willingness	0.129	0.229	0.420	0.252	0.240
Financial Ability	0.109	0.195	0.358	0.214	0.204
Complementary Resources	0.150	0.263	0.513	0.297	0.284
Technology Capability	0.155	0.260	0.465	0.285	0.272

TABLE IV.
COMPOSITE FUZZY WEIGHTS OF EACH SUB-CRITERION

	Cooperative Willingness				***Financial Ability***				***Complementary Resource***				***Technology Capability***		
	a_p	b_p	c_p		a_p	b_p	c_p		a_p	b_p	c_p		a_p	b_p	c_p
CW1	0.160	0.250	0.400	***FA1***	0.185	0.258	0.310	***CR1***	0.180	0.233	0.325	***TC1***	0.150	0.219	0.320
CW2	0.175	0.263	0.360	***FA2***	0.170	0.300	0.360	***CR2***	0.210	0.256	0.300	***TC2***	0.210	0.282	0.330
CW3	0.170	0.193	0.340	***FA3***	0.185	0.235	0.340	***CR3***	0.140	0.239	0.320	***TC3***	0.165	0.248	0.375
CW4	0.150	0.253	0.350	***FA4***	0.150	0.196	0.320	***CR4***	0.185	0.259	0.360	***TC4***	0.160	0.207	0.400

TABLE V.
PAIR-WISE COMPARISON OF A SUB-CRITERION

	C1	***C2***	***C3***	***C4***	***Fuzzy Performance***		
					m_{at}	m_{bt}	m_{ct}
Candidate 1	I	VP	P	VG	0.475	0.588	0.700
Candidate 2	VG	I	G	EG	0.763	0.863	0.938
Candidate 3	G	P	I	VG	0.588	0.700	0.813
Candidate 4	VP	EP	VP	I	0.275	0.350	0.450

TABLE VI.
COMPOSITE PERFORMANCE EVALUATION FIE THE CANDIDATE PARTNERS

	Candidate 1			***Candidate 2***			***Candidate 3***			***Candidate 4***		
	d_{qs1}	e_{qs1}	f_{qs1}	d_{qs1}	e_{qs1}	f_{qs1}	d_{qs1}	e_{qs1}	f_{qs1}	d_{qs1}	e_{qs1}	f_{qs1}
CW1	0.605	0.713	0.810	0.483	0.608	0.708	0.623	0.733	0.838	0.350	0.448	0.555
CW2	0.568	0.675	0.773	0.495	0.600	0.710	0.628	0.735	0.833	0.403	0.500	0.608
CW3	0.558	0.658	0.753	0.513	0.625	0.738	0.615	0.725	0.830	0.395	0.493	0.600
CW4	0.565	0.672	0.790	0.465	0.575	0.680	0.668	0.778	0.883	0.385	0.460	0.555
FA1	0.683	0.793	0.898	0.615	0.725	0.830	0.423	0.535	0.648	0.328	0.425	0.533
FA2	0.680	0.788	0.885	0.483	0.595	0.708	0.468	0.580	0.693	0.443	0.548	0.658
FA3	0.653	0.763	0.868	0.513	0.625	0.738	0.493	0.595	0.698	0.413	0.518	0.628
FA4	0.713	0.815	0.898	0.543	0.655	0.768	0.513	0.625	0.738	0.323	0.405	0.508
CR1	0.333	0.423	0.528	0.758	0.860	0.943	0.528	0.655	0.768	0.460	0.573	0.685
CR2	0.305	0.380	0.480	0.750	0.853	0.935	0.600	0.710	0.815	0.445	0.558	0.670
CR3	0.343	0.440	0.548	0.695	0.803	0.900	0.573	0.685	0.798	0.438	0.550	0.663
CR4	0.300	0.383	0.485	0.693	0.798	0.888	0.623	0.733	0.838	0.475	0.588	0.700
TC1	0.470	0.573	0.675	0.645	0.755	0.860	0.388	0.485	0.593	0.578	0.688	0.793
TC2	0.505	0.580	0.693	0.650	0.758	0.855	0.340	0.430	0.535	0.588	0.700	0.813
TC3	0.460	0.573	0.685	0.733	0.840	0.938	0.325	0.415	0.520	0.550	0.663	0.775
TC4	0.468	0.563	0.663	0.660	0.770	0.875	0.368	0.535	0.648	0.518	0.623	0.733

TABLE VII.
COMPOSITE WEIGHTED PERFORMANCE EVALUATION FOR THE CANDIDATE PARTNERS

Candidate	***C.1***			***C.2***			***C.3***			***C.4***		
Criterion	C_{q11}	C_{q12}	C_{q13}	C_{q21}	C_{q22}	C_{q23}	C_{q31}	C_{q32}	C_{q33}	C_{q41}	C_{q42}	C_{q43}
CW	0.376	0.653	1.135	0.321	0.576	1.028	0.414	0.713	1.226	0.251	0.455	0.839
FA	0.470	0.780	1.179	0.372	0.641	1.009	0.326	0.574	0.924	0.261	0.475	0.778
CR	0.228	0.400	0.666	0.519	0.817	1.195	0.417	0.688	1.051	0.325	0.560	0.888
TC	0.327	0.548	0.967	0.460	0.747	1.256	0.242	0.441	0.821	0.384	0.641	1.106

TABLE VIII
NORMALIZED SCORE AND SUITABILITY INDEXATION OF ERCH CANDIDATE

Candidate ***Criterion***	***C.1***	***C.2***	***C.3***	***C.4***
CW(0.240)	0.278	0.213	0.298	0.220
FA(0.204)	0.317	0.228	0.234	0.219
CR(0.284)	0.168	0.285	0.277	0.256
TC(0.272)	0.237	0.274	0.190	0.305
Suitability	**0.244**	**0.253**	**0.249**	**0.253**

fuzzy weights for these sub-criteria from the experts group.According to the 16 sub-criteria, the composite fuzzy performance of the candidate partners was calculated by Eq. (8) in Table 6. In Table 7, the defuzzified score of each candidate about each criterion is obtained by Eq. (10).

After the calculation of the composite fuzzy weights for these sub-criteria, we use the method of pair-wise comparison as above mentioned to evaluate the performance for comparing each candidate with others. Table 5 shows the result according to one sub-criterion.

The normalized score for each candidate of each criterion is also calculated and given. The suitability indexation of each candidate is calculated by summing up the product of normalized score of each criterion with the weight of importance of this criterion, which is shown in the last row of Table 8. We can see that candidate 2 and candidate 4 have the same suitability score, which is higher than the other 2 candidates and prove both of them are qualified partners. Furthermore, we can also select the most satisfied partner between these two candidates, which has the most needed organizational characteristic stronger than the other one if the launcher only needs one partner.

To select an appropriate partner for conducting the activities of innovation is not a simply work in some complicated projects due to the vary needs and purposes. But the launcher of alliance has to have a clear and definite goal of what organizational characteristic is the key one for further collaborative innovation. The suitable index could be considered as an important reference for partner selection, but to conduct a comprehensive and comparative analysis of the candidates is much more necessary.

V. CONCLUSIONS

An effective and efficient partner selection approach is one of the most fundamental steps before building partnerships system. What characteristic of partner is the most needed for innovation has to be identified before partner selection, in spite of the innovation alliance is constituted of members with different backgrounds, each of whom may play a special role during the collaborative innovation process. In this study, we proposed a fuzzy AHP method to solve the problem of partner selection, which has considered the priority of organizational characteristics as the factors of the weighting process of criteria and integrate the merits of each candidate into the evaluation criteria. Linguistic variables then are applied to pair-wise comparisons for weighting criteria and evaluating the performance of candidates. This approach is able to avoid the vagueness and effectively solve multi-criteria decision making issues. Finally, the suitability index for each candidate is acquired by an approximate method, which diminish the load of calculation compared to other fuzzy AHP methods. Although the model is developed and tested for serving in one particular innovation alliance, it can also be used with slight modification in any alliance.

ACKNOWLEDGEMENT

This work is supported by the National Nature Science Foundation of China (No: 71173016), and the Fund of Beijing Institute of Technology Fundamental Research (No: 20122142020).

REFERENCES

[1] Hou Guangming. “Research on the development of civil-military integration”, Science Press, 2009.

[2] [2] Maria Kapsali. “How to implement innovation policies through projects successfully”, Technovation, Vol.31, No.12, pp. 615-626, 2011.

[3] Giuseppe Bruno, Emilio Esposito, Andrea Genovese, Renato Pssaro. “AHP-based approaches for supplier evaluation: Problems and perspectives”, Journal of Purchasing & Supply Management, Vol.18, No.03, pp. 159-172, 2012.

[4] Mohinder Chand, Anastasia A. Katou. “Strategic determinants for the selection of partner alliances in the Indian tour operator industry: A cross-national study”, Journal of World Business, Vol.47, No.02, pp.167-177, 2012.

[5] Jing Yang, Hua Jiang. “Fuzzy evaluation on supply chains’ overall performance based on AHM and M(1,2,3)”, Journal of Software, Vol.7, No.12, pp. 2779-2786, 2012.

[6] Chong Wu, David Barnes. “A literature review of decision-making models and approaches for partner selection in agile supply chain”, Journal of purchasing & supply management. Vol.17, No.04, pp.256-274, 2011.

[7] Saaty, Y.L. “The analytic hierarchy process”, McGraw-Hill, 1980.

[8] D.Y. Chang. “Application of the extent analysis method on fuzzy AHP”, European of Journal of Operational Research, Vol.95, No.03, pp.649-655, 1996.

[9] M. Bevilacqua, F.E. Ciarapica, G. Giacchetta. “A fuzzy-QFD approach to supplier selection”, Journal of Purchasing and Supply Management, Vol.12, No.01, pp.14-27, 2006.

[10] Atakan Yucel, Ali Fuat Guneri. “A weighted additive fuzzy programming approach for multi-criteria supplier selection”, Expert Systems with Applications, Vol.38, No.05, pp.6281-6286, 2011.

[11] Tzu-An Chiang. “Multi-objective decision-making methodology to create an optimal design chain partner combination”, Computers & Industrial Engineering, Vol.63, No.04, pp.875-889, 2012.

[12] Sheu Hua Chen, Pei Wen Wang, Chien Min Chen, Hong Tau Lee. “An analytic hierarchy process approach with linguistic variables for selection of an R&D strategic alliance partner”, Computers and industrial engineering, Vol.58, No.02, pp.278-287, 2010.

[13] Bierly III P E, Gallagher S. “Explaining alliance partner selection: fit, trust and strategic expediency”, Long Range Planning, Vol.40, No.02, pp.134-153, 2007.

[14] Geringer, J.M. “Joint venture partner selection: strategies for developed countries”, Quorum Books, 1988.

[15] Bo Feng, Zhi-Ping Fan, Jian Ma. “A method for partner selection of codevelopment alliances using individual and collaborative utilities”, International Journal of Production Economics, Vol.124, No.01, pp.159-170, 2010.

[16] Guojing Fan, Erik D. Goodman, Zhijun Liu. “AHP (Analytic Hierarchy Process) and computer analysis software used in tourism safety”, Journal of Software, Vol.8, No.12, pp.3114-3119, 2013.

[17] Zeydan M, Çolpan C, Çobanoğlu C. “A combined methodology for supplier selection and performance evaluation”, Expert Systems with Applications, Vol.38, No.03, pp.2741-2751, 2011.

Yang Yang is currently a PhD candidate at School of Management and Economics of Beijing Institute of Technology. His research interests include organisational innovation, organisational evaluation, and decision support system.

Wendai Lv is currently a PhD candidate at School of Business of Renmin university of China. Her research interests include supply chain innovation, quantitative analysis, and management decision.

Guangming Hou is a professor of business management at Beijing Institute of Technology, who used to be the vice chancellor of this university. He is the member in a lot of important academic committees in China, and he is also the expert of the National Nature Science Foundation of China. His research fields include systematic science, organisational science, decision support system, and innovative methodology.

Junpeng Wang is a lecturer of business management at Beijing Institute of Technology, who got his PhD degree in 2013. His research interests include decision support system, innovative methodology and quantitative analysis.

Call for Papers and Special Issues

Aims and Scope.

Journal of Software (JSW, ISSN 1796-217X) is a scholarly peer-reviewed international scientific journal focusing on theories, methods, and applications in software. It provide a high profile, leading edge forum for academic researchers, industrial professionals, engineers, consultants, managers, educators and policy makers working in the field to contribute and disseminate innovative new work on software.

We are interested in well-defined theoretical results and empirical studies that have potential impact on the construction, analysis, or management of software. The scope of this Journal ranges from the mechanisms through the development of principles to the application of those principles to specific environments. JSW invites original, previously unpublished, research, survey and tutorial papers, plus case studies and short research notes, on both applied and theoretical aspects of software. Topics of interest include, but are not restricted to:

- Software Requirements Engineering, Architectures and Design, Development and Maintenance, Project Management,
- Software Testing, Diagnosis, and Validation, Software Analysis, Assessment, and Evaluation, Theory and Formal Methods
- Design and Analysis of Algorithms, Human-Computer Interaction, Software Processes and Workflows
- Reverse Engineering and Software Maintenance, Aspect-Orientation and Feature Interaction, Object-Oriented Technology
- Component-Based Software Engineering, Computer-Supported Cooperative Work, Agent-Based Software Systems, Middleware Techniques
- AI and Knowledge Based Software Engineering, Empirical Software Engineering and Metrics
- Software Security, Safety and Reliability, Distribution and Parallelism, Databases
- Software Economics, Policy and Ethics, Tools and Development Environments, Programming Languages and Software Engineering
- Mobile and Ubiquitous Computing, Embedded and Real-time Software, Database, Data Mining, and Data Warehousing
- Internet and Information Systems Development, Web-Based Tools, Systems, and Environments, State-Of-The-Art Survey

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the “Call for Papers” to be included on the Journal’s Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. “Special Issue: Selected Best Papers of XYZ Conference”.
- Sending us a formal “Letter of Intent” for the Special Issue.
- Creating a “Call for Papers” for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jsw/.

(Contents Continued from Back Cover)