Journal of Emerging Technologies in Web Intelligence

ISSN 1798-0461

Volume 5, Number 3, August 2013

Contents

REGULAR PAPERS

DBSoft: A Toolkit for Testing Database Transactions 205
Zuhoor A. Al-Khanjari, Youcef Baghdadi, Abdullah Al-Hamdani, and Sara Al-Kindi

Development of University Ontology for aSPOCMS 213
Sanjay K. Dwivedi and Anand Kumar

Priority Recommendation System in an Affiliate Network 222
Zeeshan Khawar Malik, Colin Fyfe, and Malcolm Crowe

Adaptive Search and Selection of Domain Ontologies for Reuse on the Semantic Web 230
Jean Vincent Fonou-Dombeu and Magda Huisma

Web Usage Mining: An Analysis 240
Mehak Jain, Mukesh Kumar, and Naveen Aggarwal

Prevention of Losing User Account by Enhancing Security Module: A Facebook Case 247
M. Milton Joe, B. Ramakrishnan, and R.S. Shaji

Automatic Text Summarization System for Punjabi Language 257
Vishal Gupta and Gurpreet Singh Lehal

JAPL: the JADE Agent Programming Language 272
Mohamed Bahaj and Abdellatif Soklabi

A Direction-Based Vertical Handoff Scheme 278
Abdelnasser Banihani, Mahmoud Al-Ayyoub, and Ismail Ababneh

Proposing a New Structure for Web Mining and Personalizing Web Pages 287
Hamid Alinejad-Rokny, Mostafa Keikhay Farzaneh, Amir Goran Orimi, Mir Mohsen Pedram, and Hojjat Ahangari Kiasari

Applying Clustering Approach in Blog Recommendation 296
Zeinab Borhani-Fard, Behrouz Minaei, and Hamid Alinejad-Rokny

Automatic Extraction of Place Entities and Sentences Containing the Date and Number of Victims of Tropical Disease Incidence from the Web 302
Taufik Fuadi Abidin, Ridha Ferdhiana, and Hajjul Kamil

Widespread Mobile Devices in Applications for Real-time Drafting Detection in Triathlons 310
Iztok Fister, Dušan Fister, Simon Fong, and Iztok Fister Jr.

RISING SCHOLAR PAPERS

Solving Problems of Imperfect Data Streams by Incremental Decision Trees 322
Hang Yang

DBSoft: A Toolkit for Testing Database Transactions

Zuhoor A. Al-Khanjari
Sultan Qaboos University/Department of Computer Science, Muscat, Oman
Email: zuhoor@squ.edu.om

Youcef Baghdadi, Abdullah Al-Hamdani, and Sara Al-Kindi
Sultan Qaboos University/Department of Computer Science, Muscat, Oman
Email: {ybaghdadi, abd}@squ.edu.om; sara.al.kindy@gmail.com

***Abstract*—Databases (DBs) are used in all enterprise transactions, which require attention not only to the consistency of DB, but also to existence, accuracy and correctness of data required by the transactions. While the Atomicity, Consistency, Isolation, and Durability (ACID) properties of a transaction ensure that DB is consistent after the execution of each transaction, it is not sure that the transactions retrieve the correct data. Indeed, the testing phase of the transactions, in the development process, is often ignored. Therefore, there is a need for testing techniques and tools. This paper proposes an architecture, a design, and an implementation of a tester, we refer to as DBSoft, to test transactions, in terms of required data they need to access. The architecture of DBSoft is a layered one. It is made of five components having separate concerns and serving each other: (C1) a parser to collect information, specifically for the metadata, (C2) an input generator to generate test cases, (C3) an output generator to implement the test cases, (C4) an output validator to validate test cases, and (C5) a report generator to generate test reports. DBSoft aims at avoiding cost effective transaction run-time errors.**

***Index Terms*—Databases, Transactions, Testing Tools, Metadata, XML**

I. INTRODUCTION

Database Management Systems (DBMSs) play a crucial role in storing, accessing, and managing data. Most organizations deal with a certain form of DBMS, as these systems provide, through a uniform interface that is SQL, an easy, efficient access to large amount of data by hiding the low-level details of how the data is physically structured and accessed. All enterprise transactions perform a kind of Create/Retrieve/Update/Delete (CRUD) operations against DBs through SQL.

The handled data, through transactions, is used in day-to-day activities or decision-making, which requires a certain attention not only to the consistency of DB, but also to existence, accuracy and correctness of data required by the transactions.

Yet, when transactions running against the DBs abort due to lack or inaccuracy of data can prove to be costly in terms of time and money to organizations. While the Atomicity, Consistency, Isolation, and Durability (ACID) properties of the transaction ensure that the DB is consistent after execution of each transaction, it is not sure that the transactions match the correct data description or value at run-time. When DBs and transactions are not tested before implementation, errors or bugs may appear at any time during the implementation phase or the exploitation (e.g., data population).

In the development process, guided by the three-level architecture, the step that deals with the mapping external schema/conceptual schema is often ignored. This would ensure that the required data by all transactions map into the conceptual schema and vice-versa. This critical testing-kind step ensures that transactions retrieve the correct data description or values, which avoids any costly run-time errors.

One would think that the amount of research being invested in the field would be vast given the importance of DB systems. Unfortunately, the opposite is true. Testing DBs is not easy [1]. For instance, relational DBs have hundreds of interrelated tables structured in a specific way, and described in a metadata (aka catalog), which makes it complex. In addition to this complexity, the metadata is not static as the DB administrator always alters it when business requirements change. That is, the schema and the states of these tables need to be consistently validated to avoid transactions run-time errors.

Therefore, there is a crucial need for techniques and tools to test the correctness of data against the transaction requirements in terms of data. These techniques and tools should be integrated within the development process, preferably before the implementation, i.e., at the mapping interface between transactions and DB.

doi:10.4304/jetwi.5.3.205-212

We propose an architecture, a design, and an implementation of tester, DBSoft, to test transactions, in terms of required data, against DBs they access. The architecture of DBSoft is a layered one. It is made of five components having separate concerns, and serving each other:

C1. A parser: this component generates XML files from the metadata about the input DB to test.

C2. An input generator: this component creates test cases by using the information generated by the parser.

C3. An output generator: this component runs the DB tests by using the populated test cases and saving the results.

C4. An output validator: this component validates the results produced by the output generator.

C5. A report generator: this component produces reports, graphs, and other information about the executions of the test.

DBSoft aims at avoiding cost effective transaction run-time errors.

In this paper, we present the architecture and design of the DBSoft, but we limit the implementation to the crucial component of the toolkit that is the parser. This component parses the metadata, generated by the DBMSs during the creation phase of the DB schema, in order to extract the required information in a standard XML format. Then the XML documents will be used in generating test cases and their expected outcome as well as generating a DB test.

The remainder of the paper is organized as follows: Section 2 provides an overview of related work and existing tools. Section 3 discusses DBMS metadata required by the toolkit. Section 4 presents the architecture of DBSoft. Section 5 details the DBSoft parser. Section 6 presents an implementation of the parser component. Section 7 includes concluding remarks and future direction of the work and research.

II. RELATED WORKS AND TOOLS

The closely related work to what is being proposed in this paper is done by Chays *et al.* [1][2][3] with the AGENDA toolkit. It is composed of five components: a parser, a state generator, an input generator, a state validator and an output validator. AGENDA parses a database transaction, generates test cases, and validates the result. The AGENDA parser is the focal step of interest for now. Based on a modified form of PostgreSQL's [4], the internal parser collects relevant information about a DB supplied to it and stores the information in an internal DB called the AGENDA DB.

In [5], a framework is presented to perform efficient regression tests on DB transactions. In [6], issues with the automatic running of DB regression tests are listed. A framework for creating a test database is presented in [7] and [8], while in [9], the framework tests the features of DBMSs and also includes an automatic DB generator called QAGen.

Due to security and confidentially issues tied with "live" DB data, an automatic data generating tool is proposed in [10]. It is called ADG. Automatic Data Generation is also highly useful in terms of its ability to create a desired problem situation within a DB for testing.

A number of different methods in creating the test cases exist, and these are discussed in [11], [12], [13] and [14].

There is no evident related work in the academic community to collecting information from DBMS metadata to be used in DB testing, and it is safe to say that DBSoft toolkit is the first to apply this concept.

The main two differences between our approach and AGENDA are:

1. Instead of developing a new parser, we extract relevant information from DBMS metadata.
2. DBSoft toolkit does not store the information in tables, but represents it as standard XML documents. Indeed, storing the information in an internal DB as done by AGENDA is not efficient in terms of simplicity and specifically extensibility. DBSoft uses XML document to store information due to its nature of being self-descriptive, easily processed and human-readable. However, XML tags and attributes are based on AGENDA DB tables.

One of the aims of the DBSoft toolkit is to enable performing regression tests on DB while they are updated. DBSoft will generate test cases and a test DB state by means of the information stored in XML documents that are produced in the data extraction step. The test cases will be used to run the test DB.

III. DBMS METADATA

There exist many definitions of the concept metadata. The most general is “data about data”. In DB systems, a metadata describes and provides information about all the objects of a DB such as tables, views, indexes, sequences, stored procedures, functions, etc. For example, in Oracle, a table is described by more than fifty data such as name, number of columns, number of rows, etc. In PostgreSQL, objects are described in Information_schema.

The SQL standard defines a uniform interface to access this data, but not all DBMSs implement this feature. Hence, a number of different mechanisms have evolved with accessing metadata over different systems: For instance:

- Oracle has data dictionary and metadata registry.
- PostgreSQL provides system catalogs and Information_Schema.

- SQL Server and MySQL contain system catalogs and the Information_Schema, but the method for querying it differs from that of PostgreSQL.

Additionally, due to the nature of where the information is being collected from, it is safe to say that the integrity of the information collected is increased and hence more trustworthy.

Metadata is the most relevant input to the testing step, as it should describe all the data required by the transactions running against the DB.

IV. DBSOFT ARCHITECTURE

The architecture of DBSoft is a layered one. It is made up of components having separate concerns, but serving each other as shown in Figure 1. This ensures:

- A smooth and independent design and implementation of the components.
- A straightforward wrapping of the components into services for an easy adoption of Service-Oriented Architecture (SOA).
- A replacement of any of the components by other components when some non-functional requirements such as reliability or performance.
- Security of the components.

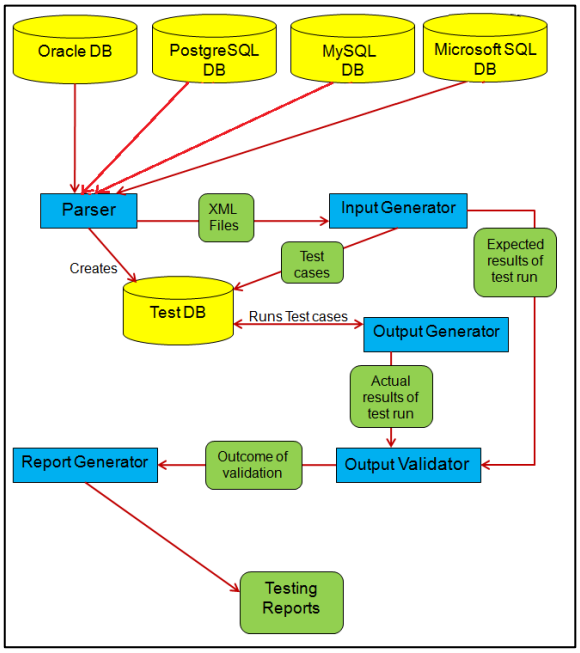


Figure 1. DBSoft Architecture

The components are specified as follows:

C1. A parser: this component is involved in collecting information from the metadata about the DB to test. It generates XML files consisting of the collected information for the DB test.

C2. An input generator: this component creates test cases by using the information in the XML files generated by the parser component. It populates the DB test with test cases and creates an XML file of the expected results of each test case.

C3. An output generator: this component runs the DB tests by using the populated test cases and saving the results.

C4. An output validator: this component validates the results produced by the output generator by comparing them with the expected results from the input generator.

C5. A report generator: this component produces reports, graphs, and other information about the executions of the test.

The aforementioned architecture guides us towards a testing process that consists of five steps. Each component translates into a step in the testing process.

1. *Step 1*: information collection
2. *Step 2*: test case generation
3. *Step 3*: test case implementation
4. *Step 4*: test case validation
5. *Step 5*: report generation

The stepping-stone into the DBSoft testing tool is the parser. In this work, much of attention is given to the parser, as it constitutes the corner stone of the system. Through the correct design and implementation of the parser, we would ensure that the rest of the components will work properly, thus meeting the requirements of DBSoft as a powerful tool for testing DBs.

V. DBSOFT PARSER

Using the parser, a DB could be parsed; and relevant information would be extracted into XML documents.

In order to be able to create an efficient and real-world usable tool for DB testing, the first consideration is that DBs need to be transformed into a uniform object, as there is a number of DBMS in existence today with different metadata, catalogs, and physical storage mechanisms.

The creation of standard XML documents, outlining the details of the physical organization and structure of DB schema, will also aid in:

(i) the creation of efficient test cases, and a test DB for testing.

(ii) the generation of test data for DB population.

(iii) the validation of the results of the test runs.

A. DBSoft parser functioning

The following are the sub-steps taken by the DBSoft parser:

1. User inputs location of DB

The DBSoft parser collects the relevant information from the metadata in the DBMS. The program needs to be pointed to the location of the DB, to be tested, in order to access the metadata. The user will provide the correct information on the location of the DB. This includes the host, the port number, the DB name, the username, and the password.

2. DBSoft parser creates connection

Once the location of the database is input correctly, the DBSoft parser opens a connection to the DB.

3. Information collected from the metadata

The DBSoft parser will begin collecting the relevant information from the metadata. This is done by querying the metadata by means of relevant commands through the JDBC API as shown in Figure 2. Since the mechanism of accessing metadata differs from one DBMS to another, the commands issued will be according to the DBMS in question.

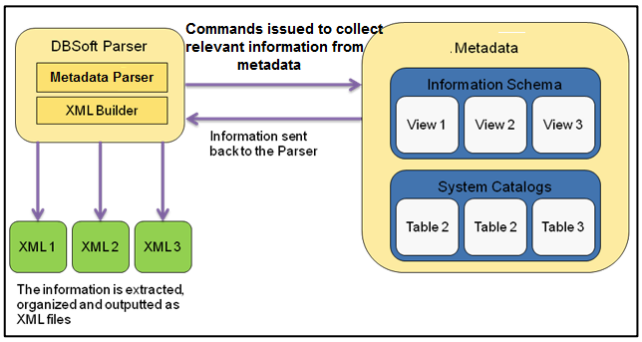


Figure 2. DBSoft Parser Component: Collecting Information from Metadata

4. Information extracted into XML document

The information that has been collected in the parsing step will be organized and extracted into XML documents, such as:

- Tables (containing information on tables, number of attributes, type, etc).
- Attributes (containing information on attributes, tables they are contained in, type, etc).
- Boundary values (containing information on boundary values, range, type, etc).
- Table relationships (relationships between tables in the database and their associated attributes).

Figure 3 illustrates a standard structure for XML documents to represent relational databases maintained in different DBMS. For each database system, an XML tree is generated by extracting information from its DBMS metadata. The database is composed of a set of table elements. Each table is composed of three main elements: *name* (the table name), *attributelist* (list of attributes in the table with their types and constraints), and *constraints* (other constraints in the table such as the primary key, foreign keys, unique attributes). Information about each table can be extracted from the tables in the DBMS metadata catalog.

The *attributelist* element for each table is composed of a list of attribute elements to represent information about each attribute such as name, type, default value, maximum and/or minimum values. The information about each attribute element in the XML tree can be extracted from the attributes and the boundary values in the metadata catalog. The constraint element is composed of a list of constraints, including primary, foreign keys, and other constraints on the database. In the implementation section, we illustrate how an XML document is generated using the proposed XML tree structure for an existing database system as shown in Figure 7.

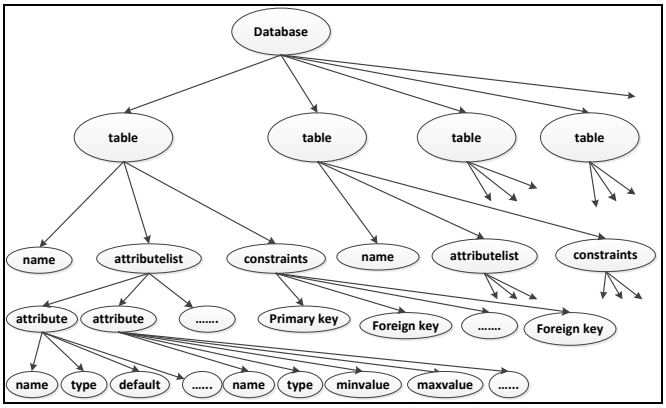


Figure 3. XML Tree Structure for a Relational Database

B. Transactions of DBSoft Parser

The XML files produced by the DBSoft parser can be employed in several testing transactions. The aim for the DBSoft toolkit is to cover most if not all DB testing methods.

Analysis can be made in regards to whether the transaction program is behaving as specified, i.e., checking correctness, in addition to reflecting upon whether the DB schema correctly models the organization of real-world data.

Test cases can be produced by using the information in the XML files, such as creating specific queries that will make a DB transaction 'break'. Data tailored for specific problem areas found within the DB can be generated, and in turn will be populated within a test DB that has been built using the information in the XML documents.

Regressions tests can be performed on databases when they are updated, ensuring that everything still work as specified. Validity of the results will be made using the information extracted in the parsing step (constraints, types, etc) in addition to looking into other automatic means.

VI. IMPLEMENTATION

To implement DBSoft, we have used PostgreSQL [15] DBMS and the same approach can be used to extract metadata from other DBMSs. It is an object-relational DBMS, originally developed at UC Berkeley. It is open-source. In this work, we will be dealing with the

PostgreSQL's (version 8.4.5) metadata, namely its system catalogs and Information_Schema.

System catalogs in PostgreSQL are regular tables that store the schema or metadata on a parallel DB. There are about 60 system catalog tables, each catalog deals with a specific type of metadata.

The Information_Schema is a set of views that provides information about objects defined in the current DB. It is portable and more stable, as it is defined with SQL standard, contrary to the system catalogs mentioned above which are specific to PostgreSQL. There are about 51 views.

The information stored within each form of metadata is nearly the same. However, there exist some differences between the two with the amount of information stored. Hence, they complement each other in terms of the information they contain. Thus, we can extract information from both metadata rather than concentrating on one only.

PostgreSQL metadata storages can be accessed using an Application Programming Interface (API) such as Java Database Connectivity API (JDBC) [15]. JDBC is a middleware that consists of a set of classes that enables transactions developed with Java to interact with DBs, whereas the relevant information can be extracted from the metadata sources using SQL commands. This makes both the system catalogs and the Information_Schema, a trove of information fit to be employed in parsing a PostgreSQL DB. Figure 4 shows an example of query that gets the names of all the tables in a DB by means of the Information_Schema. The last part of the command ensures that the names of the tables contained within the system catalogs and the Information_Schema will be retrieved as well.

```
SELECT table_name
  FROM information_schema.tables
 WHERE table_type = 'BASE TABLE'
       AND table_schema NOT IN
              ('pg_catalog',
    'information_schema');
```

Figure 4. Querying the Schema to get Table Information

A. Parser front-end

The UI of the DBSoft toolkit gives users the ability to parse an input DB, output the DBSoft schema, and to recreate the information collected from the parsing stage as the SUT DB as shown in Figure 5 and Snapshot 1.

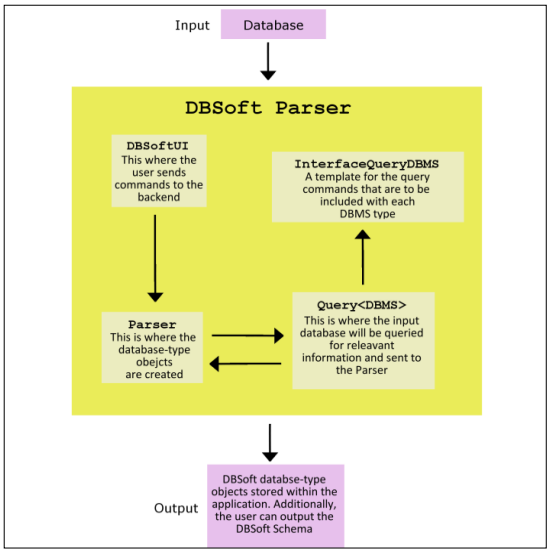
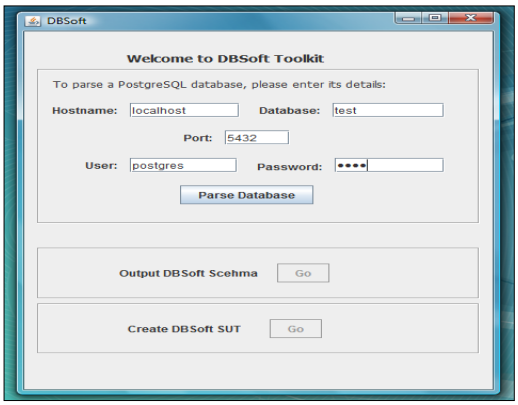


Figure 5. The Internal Build of the Parser

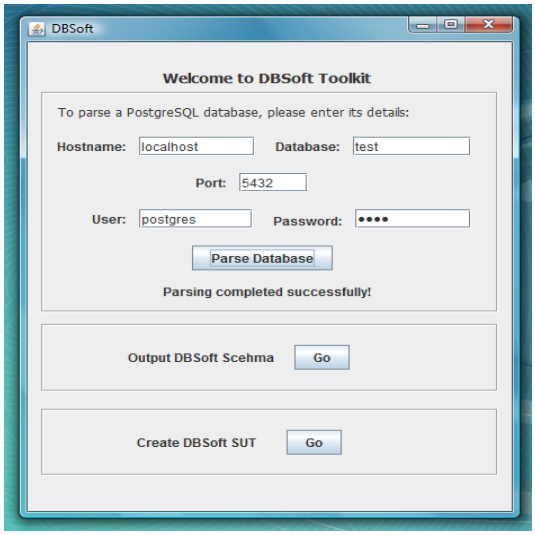


Snapshot 1. The User Interface of DBSoft

B. Schema Creation

At the end of parsing the input database, by means of both SUTBuilder and ParsePostgreSQL, and storing the relevant information in the DBSoft database-type objects, the DBSoft schema is finally created as shown in the Snapshot 2.

As mentioned before, this is based on tables in the [5] AGENDA DB. The main difference however is that the AGENDA DB is an actual DB that will be used in replicating the original input DB i.e., it will be queried for the information collected. This is contrary to the DBSoft schema that is an internal representation of a schema within the DBSoft Java application.



Snapshot 2. End of Parsing Message

The users have an option of outputting the DBSoft schema as an actual schema, but this will not make a difference to the replication of the DB, and the creation of test cases, as the DBSoft DB-type objects will be used in this point.

C. XML document Generation

To evaluate the proposed approach, we have used a simple PostgreSQL database application with two tables (Department and Employee) as shown in Figure 6.

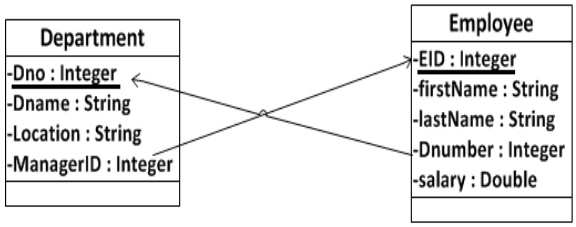


Figure 6. UML Diagram for Company Database Schema

```
<?xml version="1.0" ?>
<!-- An XML file generated for a test database named
     myTestDatabase  composed of two tables-->
<database >
<dbname>SQU Test Database</dbname>
<!--Department table definition -->
<table>
<name>Department</name>
<attributelist>
        <attribute>
                <name>Dno</name>
                <type>integer</type>
                <default> 1</default>
                 <minvalue>1</minvalue>
                <maxvalue>99</maxvalue>
        </attribute>
        <attribute>
                <name>Name</name>
                <type>String</type>
                <maxlength>30</maxlength>
        </attribute>
        <attribute>
                <name>location</name>
                 <type>String</type>
```

The DBSoft was used to extract the metadata for the database in Figure 6 from PostgreSQL database, and the corresponding XML file was generated as shown in Figure 7. The XML file is composed of a single database element <database> that contains one or more <table> elements (each corresponds to a relational database). Each <table> element is composed of <name> (relational table name), <attributelist> (list of table attributes) and <constraints> (list of constraints in the table such as the primary keys <primarykey> and foreign keys <foreignkey>).

Each <attributeList> element contains one or more <attribute> elements that corresponds to all attributes in a given relational database table. The <attribute> element is composed of serious XML element including attribute name <name>, attribute type <type>, default value <default>, maximum value <maxvalue>, minimum value <minvalue> and maximum field length <maxlength> elements.

The XML file can be used to generate test cases for the database using XML tags such as <types>, <maxlength>, <maxvalue>, <minvalue>, <default>,.., etc. Also, the XML file can be used to validate SQL queries such as checking the validity for the table names, attribute names, and constant ranges.

```
<!-- Employee table definition -->
<table>
<name>Employee</name>
<attributelist>
        <attribute>
                <name>EID</name>
                <type>integer</type>
                <minvalue>10000</minvalue>
                <maxvalue>99999</maxvalue>
        </attribute>
        <attribute>
                <name>firstName</name>
                 <type>String</type>
                <maxlength>25</maxlength>
        </attribute>
        <attribute>
                <name>lastName</name>
                 <type>String</type>
                <maxlength>50</maxlength>
        </attribute>
        <attribute>
                <name>Dnumber</name>
                 <type>integer</type>
        </attribute>
```

```
                <maxlength>20</maxlength>
                <default>Muscat, Oman </default>
        </attribute>
        <attribute>
                <name> ManagerID</name>
                <type>integer</type>
        </attribute>
</attributelist>
<constraints>
        <primarykey> Dno </primarykey>
        <unique>Name</unique>
        <foreignkey>
                <attribute> ManagerID <attribute>
                <reftable> Employee </reftable>
                <refattribute>EID<refattribute>
        <foreignkey>
</constraints>
</table>
        <attribute>
                <name>salary</name>
                <type>double</type>
                <minvalue>200</minvalue>
                <maxvalue>3000</maxvalue>
        </attribute>
</attributelist>
<constraints>
        <primarykey>Dnumber</primarykey>
        <foreignkey>
                <attribute> ManagerID <attribute>
                <reftable>Department</reftable>
                <refattribute>Dno<refattribute>
        <foreignkey>
</constraints>
</table>
</database>
```

Figure 7. XML File for a Company Database

VII. CONCLUSION AND FUTURE DIRECTION

Nowadays, the need for an automated tool in testing DB transactions is crucial and critical. DBs support large organization activities if not all, and hence would need to behave correctly to avoid errors and bugs.

DBSoft toolkit is proposed in this paper. It is an efficient, practical tool for DB testing. This is realized through the following milestones:

- Implementing the testing process has been realized with the DBSoft parser that collects the required information about the DB to be tested, specifically its metadata.
- The creation of test cases helps in enhancing the DBSoft tester to use a standard method of entering DB information, i.e., XML documents produced by the DBSoft parser. A range of different test cases will be generated and employed in checking the correctness of the DB.
- Another milestone is to enable the DBSoft tester to use test cases for regression testing on DBs, since it generates test cases that can be stored and run several times.

We expect DBSoft to be a startup for DB testing community towards the realization of a standard DBMSs testing process. Although, the complete process has not been presented in the paper, a stepping stone into it has been with the standardizing transaction of the DBSoft parser.

Further development first concerns with the development of all the components. Then, we foresee a DB testing method.

REFERENCES

[1] D., Chays, Y., Deng, P. G., Frankl, S., Dan, F. I., Vokolos and E. J., Weyuker, An AGENDA for Testing Relational Database Applications. Journal of Software Testing, Verification and Reliability, 14(1), PP.17–44, March 2004.

[2] D., Chays, Y., Deng, P. G., Frankl, S., Dan, F. I., Vokolos, and E. J., Weyuker, Demonstration of AGENDA Tool Set for Testing Relational Database, ICSE '03 Proceedings of the 25th International Conference on Software Engineering, PP. 802-803, Published by IEEE Computer Society, May 2003.

[3] Y., Deng, P., Frankl, and D., Chays, Testing database transactions with AGENDA, Proceedings of the 27th international conference on Software engineering (ICSE 05), ACM Press, PP. 88-96, May 2005.

[4] PostgreSQL. The PostgreSQL Global Development Group, [Online] http://www.postgresql.org/ [accessed: 12 July 2012].

[5] F., Haftmann, D., Kossmann, and E., Lo., A Framework for Efficient Regression Tests on Database Applications, The VLDB Journal — The International Journal on Very Large Data Bases, 16(1), PP.145-164, January 2007.

[6] F., Haftmann, D., Kossmann, and A., Kreutz, Efficient Regression Tests for Database Applications. Conference on Innovative Data Systems Research (CIDR), pp. 95-106, 2005.

[7] E., Lo, C., Binnig, D., Kossmann, M. T., Ozsu, and W.-K., Hon, A Framework for Testing DBMS Features. VLDB Journal, 19(2), pp.203–230, April 2010.

[8] N., Bruno and S., Chaudhuri, Flexible database generators, Proceedings of the 31st international conference on Very large data bases, 2005.

[9] C., Binnig, D., Kossmann, E., Lo, and M.T., Özsu., QAGen: generating query-aware test databases, Proceedings of the 2007 ACM SIGMOD international conference on Management of data, January 11-14, 2007.

[10] N. R., Lyons, An automatic data generating system for data base simulation and testing, ACM SIGMIS Database, 8(4), PP.10-13, 1977.

[11] T.Y., Chen, P.L., Poon, and T. H., Tse., A Choice Relation Framework for Supporting Category-Partition Test Case Generation, IEEE Transactions on Software Engineering, 29(7), PP.577-593, July 2003.

[12] H., Bati, L., Giakoumakis, S., Herbert, and A., Surna, A genetic approach for random testing of database systems, Proceedings of the 33rd international conference on Very large data bases, 2007.

[13] C., Mishra, N., Koudas, and C., Zuzarte, Generating targeted queries for database testing, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, PP. 499–510, 2008.

[14] J., Tuya, M.J.S., Cabal, and C.de la, Riva, Mutating database queries, Information and Software Technology, 49(4), PP.398-417, 2007.

[15] PostgreSQL JDBC Driver, The PostgreSQL Global Development Group, Last published: 6 February 2012 [Online] http://jdbc.postgresql.org/ [accessed: 20 August 2012].

Development of University Ontology for aSPOCMS

Sanjay K. Dwivedi
Babasaheb Bhimrao Ambedkar University, Lucknow, India
Email: skd200@yahoo.com

Anand Kumar
Babasaheb Bhimrao Ambedkar University, Lucknow, India
Email: anand_smsvns@yahoo.co.in

***Abstract*— In these days, ontology is the most popular and widespread formalism of knowledge representation. With the requirement of our system aSPOCMS (An Agent-based Semantic Web for Paperless Office Content Management System), we present the construction of university ontology to retrieve the information and accomplish the process of files of various sections by using the predefined workflow in university ontology. The aSPOCMS aims to provide the facility to manage and process the files/documents of various departments and sections of higher educational institutions (i.e. universities) in paperless environment. This paper reveals the conceptualization of university knowledge through construction of university ontology. Generalized structure of Indian universities and workflow processes have been taken for ontology development by describing the class hierarchy, and demonstrate the graphical view of ontology. We also demonstrate the ability of university ontology to execute intelligent query to retrieve the information.**

***Index Terms*— Semantic Web, Ontology, SemanticWorks tool, OWL, University Concepts**

I. Introduction

Ontology is most popular technology for knowledge representation in Semantic Web. The major reason is that the applications require more knowledge sharing and reuse. Here, we discuss the essential steps in optimal ontology development process of university domain. These steps may be favorable to construct the ontology of other domain.

A. Semantic Web

The inventor of today's web, Tim Berners-Lee introduced the concept of Semantic Web [1]. In his vision of Semantic Web, content located on web should be available, processible and understood by both people and machines. It is the extension of current web in which information is given well defined meaning [2], which makes it easily proccessible by machine. The well defined and linked data on web can be used for common understanding, effective discovery and reuse of particular knowledge across various applications. There are three major technologies: Resources Description Framework (RDF) [3][4][5], Resource Description Framework Schema (RDFS) [6][7] and Ontology Web Language (OWL) [8]. RDF is represented as triples statements which consist of a subject such as the resource, a predicate that is a property associated with resource and its object such as the value of the property. RDF Schema defines valid classes, properties for an unequivocal class, data type's properties, hierarchical relationships between classes or properties. OWL is a semantic markup language for sharing ontologies on the web and is designed for the use of software agents. OWL is used to describe the important concepts in a domain, essential properties of each concept and restrictions on properties such as property cardinality, property value type, domain and range of a property.

B. aSPOCMS System

The aSPOCMS (An Agent-based Semantic Web for Paperless Office Content Management System) [9] has been designed to provide the paperless environment to universities by processing electronic form of files or documents via predefined workflow of processes within university ontology. The varieties of information from different resources such as employees, departments, workflows and files of a typical university are involved in order to process the electronic form of files or documents.

The formalization of information represented in ontology can be easily interpreted by computer and the information resided in ontology can be processed on semantic level efficiently. Therefore, ontology is initiated in university domain to represent the facts and workflow.

The aSPOCMS is an agent-based Semantic Web system. It enables paperless office content management system that uses RDF, RDFS and OWL for metadata declaration and reasoning rules. The architecture of this system aSPOCMS is shown in figure 1. It has four major modules: communicator, access control, knowledge manager and reasoner. Communicator will provide the interface to users to communicate with the system. The access control has the capability to specify the authorizations over concepts defined in ontology. The user can annotate over concepts according to relationships, which are defined in ontology. The

doi:10.4304/jetwi.5.3.213-221

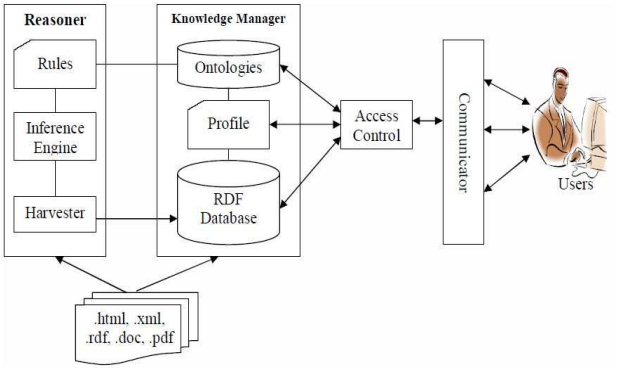


Figure 1. Architecture of aSPOCMS (adopted from [9]).

knowledge manager is the major component of the architecture as it will manage knowledge base of university and picking the knowledge by sorting, and structuring data according to the domain ontology. It has the RDF database, user's profile and ontologies. Ontology is the most prominent sub component of knowledge manager, which describes the conceptual terms and relations between these terms of university. The RDF database defines the metadata of conceptual terms of ontology.

The reasoner has the rules, an inference engine and harvester. The inference engine uses the rules to derive additional realistic knowledge from the university ontology. The harvester will harvest the additional knowledge in RDF triple format.

C. Ontology

The Semantic Web languages like OWL provide the facility to encode the ontology and an approach to integrate ontologies of multiple domains to support ontology sharing and mapping. The OWL language is divided into three syntax classes [10]. They are, OWL Lite , OWL DL and OWL Full in order of their increasing expressiveness. OWL Lite supports the users for a hierarchical classification of concepts and their simple constraint features. The advantage of this language is that it is easier to understand and implement than the other two; however, it restricts expressivity. OWL Lite has the lower formal complexity than OWL DL and OWL Full. OWL DL is the sublanguage of OWL Full which supports the maximum expressiveness without defeating the computational completeness. OWL DL is not fully compatible syntactically and semantically to RDF. The entire OWL languages are called as OWL Full in which all OWL language primitives use and allow combinations of the primitives in arbitrary ways with RDF and RDF Schema. This language comprises the possibility of changing the meaning of predefined primitives of RDF or OWL by applying the language primitives to each other. Therefore, OWL full formalism is used to construct the university ontology in this paper.

We used Altova SemanticWorks[11] to construct the ontology in OWL languages. The approach is depicted through creating university ontology based on student, employees, university structure, workflow of file processes and relationship between them.

II. SemanticWorks: Ontology Editor Tool

Altova SemanticWorks is a graphical RDF/OWL editor for building Semantic Web applications. It provides powerful, easy-to-use functionality for a visual creation and editing of RDF, RDF Schema (RDFS), OWL Lite, OWL DL, and OWL Full documents. There is no methodology associated to this tool. This editor is capable to manage the following files: N-triples, XML, RDF, RDFS and OWL. A screenshot of SemanticWorks with university ontology are viewed in figure 2. The figure shows some OWL class and their instances on university ontology.

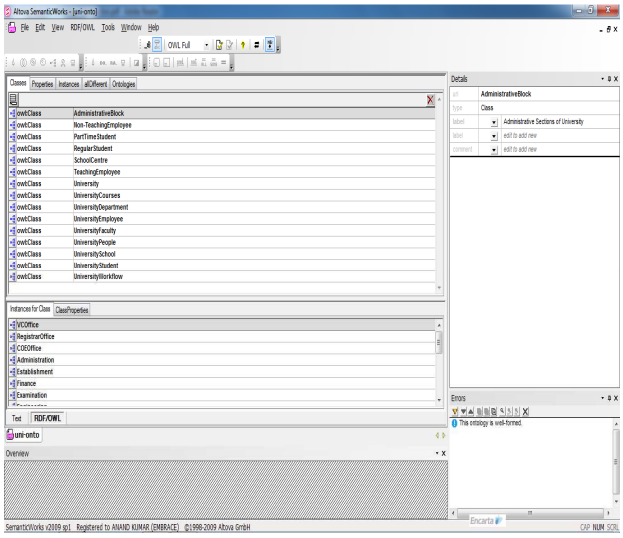


Figure 2. Note how the caption is centered in the column.

III. Related Work

A number of researcher's efforts are carried in development of ontology using various domains and purpose. We found some researches being carried on ontology development of universities. In the following subsections, we focus some of the prominent researches on university ontology:

S. Lovrenčić et al. [12] described university studies ontology in Croatia domain modeling and presented the way of university domain knowledge representation for study purpose and have shown that description logic ALC is enough to fulfill this task. They have developed the ontology of university studies for study content. University studies ontology has super classes, subclasses, their individuals and properties of study content of university. Formal documents related to university studies are identified by this ontology. The study content is represented as hierarchical structure, which is able to show the entire educational content, the sequence of learning and the structure of educational concepts such as super and sub classes. The advantage of description logic representation formalism is to provide direct possibility for further development of Web ontology.

Naveen Malviya et al. [13] aimed to focus on information of university not for human consumption only but also made available to machine consumption. Ontology is used by ontology developers to concentrate on conceptual terms and created university ontology

using Protégé. The authors have taken the example of Rajiv Gandhi Technical University Bhopal (India), for the ontology development with various aspects like super class and subclass hierarchy, creating a subclass instances for class illustration and query retrieval process.

Paul Kogut et al. [14] discussed the recent convergence of UML and ontologies. The modeling of ontology information is described in class diagrams and Object Constraint Language (OCL) constraints.

Boyce, S., & Pahl, C. [15] presented a method for domain experts rather than ontology engineers to develop ontologies for use in the delivery of courseware content and focused in particular on relationship types that allow to model rich domains adequately.

In contrast to the current status of university ontology research, in this paper university ontology and its logic reasoning is designed on the basis of generalized structure of Indian universities and the workflow processes of files/documents.

IV. Development of University Ontology

The developed university ontology here is in Indian perspective using OWL language and ontology editor tool. This ontology is the integration of four levels [16][17] i.e. top level ontology, domain ontology, task ontology and application ontology. We have taken the organizational structure of more than 10 Indian universities/institutes, 30 schools (or faculties) and 150 departments (or centers) for the purpose of ontology construction. The essential steps for development of university ontology with their implementation are described below:

A. STEP I: Classes and their hierarchy

Classes of the university are defined in this step. Hierarchy of class such as subclass and super class are also identified in this step. The major classes of university are represented in figure 3 and the hierarchical relationships between classes are represented in figure 4.

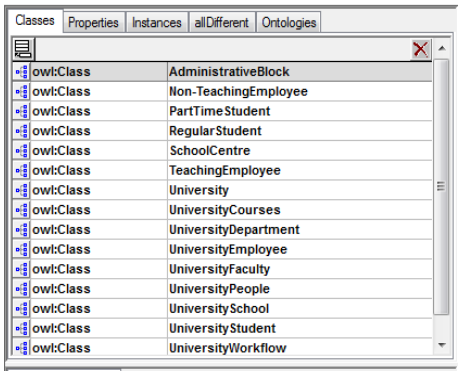


Figure 3. OWL Classes of University Ontology.

In figure 3, OWL class `AdministrativeBlock` has all sections related to administrative block of a university as resources. `Non-TeachingEmployee` has all the profile of employees of university. Similarly, remaining OWL classes have their relevant resources of the university. Figure 4 shows the graphical view of OWL classes and relationship among classes as properties. The OWL class `UniversityDepartment` (or its equivalent class `SchoolCentre`) is the subclass of `UniversityFaculty` (or its equivalent class `UniversitySchool`) using the syntax `rdfs:subClassOf` and `owl:equivalentClass` properties respectively. Similarly, OWL class `UniversityFaculty` is the subclass of `University` and equivalent class of `UniversitySchool` with `rdfs:subClassOf` and `owl:equivalentClass` properties.

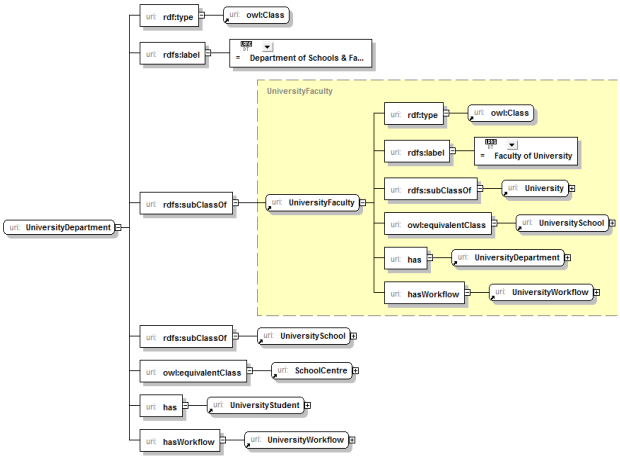


Figure 4. View of Hierarchical Relationships between Classes.

B. STEP II: Object properties of ontology

Object properties define the relationship between classes, through which we want to define among classes. These properties show the relationship between individual to individual. Some user defined object properties are shown in figure 5 which is used to

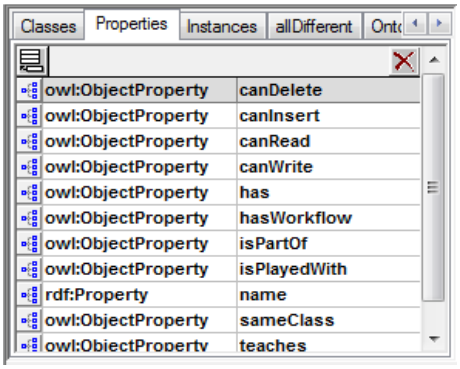


Figure 5. Defined Object Properties of University.

represent the relations between classes of university. The use of these properties is illustrated in figure 4. The OWL class `UniversityDepartment` is associated with `UniversityStudent` and `UniversityWorkflow` from the object properties `'has'` and `'hasWorkflow'` respectively.

C. STEP III: Metadata properties of ontology

Metadata properties show the relationship between individual and their data literal. In this step, we introduce the metadata properties of the resources of university

ontology. The metadata of a resource of the university is represented in figure 6.

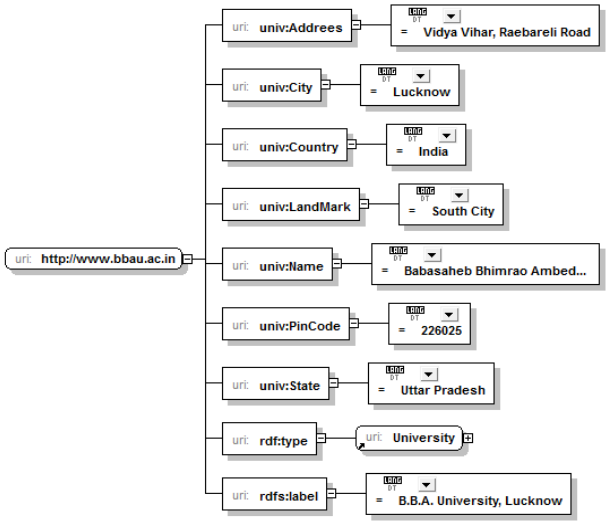


Figure 6. Metadata properties of the University.

The figure 6 represents the address of a university by using metadata properties, and the namespaces `'univ'` [18] and `'univwf'` [19] are used to define the workflow processes. The URI `http://bbau.ac.in` is represented a university and metadata of this resource and is described by metadata properties, which are shown in table I with literal values:

TABLE II.
METADATA PROPERTIES WITH LITERAL VALUES OF RESOURCE

Metadata Properties	Literal Values
univ:Address	Vidya Vihar, Raebareli Road
univ:City	Lucknow
univ:Country	India
univ:LandMark	South City
univ:Name	Babasaheb Bhimrao Ambedkar University
univ:PinCode	226025
univ:State	Uttar Pradesh

D. STEP IV: Property and Relationship

In order to define the link inside or between the classes, we use property to construct the relationship between individuals. The data properties are used to show the link between individuals to data type literal. The object properties also used to construct the relationship between individuals. The object properties domain and ranges can be defined as following spinet of XML code as examples.

```
<rdf:Description rdf:about="#sameClass">
        <rdf:type>
                <rdf:Description
        rdf:about="http://www.w3.org/2002/07/owl#Obj
        ectProperty"/>
        </rdf:type>
        <rdfs:domain>
                <rdf:Description
        rdf:about="#SchoolCentre"/>
        </rdfs:domain>
        <rdfs:range>
                <rdf:Description
        rdf:about="#UniversityDepartment"/>
        </rdfs:range>
</rdf:Description>
```

In the above list of code, object property `sameClass` has the domain and range as OWL classes `SchoolCentre` and `UniversityDepartment` respectively.

E. STEP V: Axioms of Ontology

The relationship between attributes and individuals of class can be described by axioms. Four axioms of classes are used. These are, the `existence of class`, `subclass`, `equivalent class` and `disjointwith`, which are constructed using the OWL syntax `rdf:id`, `rdfs:subClassOf`, `owl:equivalentClass` and `owl:disjointwih` respectively. Some axioms of university ontology are shown in figure 7.

ID	Axiom Type & Attribute Type	Arguments
32	SubClassAxiom	LOADED
32	SUB-CLASS	#:SchoolCentre
32	SUPER-CLASS	#:UniversitySchool
33	SubClassAxiom	LOADED
33	SUB-CLASS	#:SchoolCentre
33	SUPER-CLASS	#:UniversityFaculty
34	EquivalentClassesAxiom	LOADED
34	DESCRIPTIONS	(#:SchoolCentre #:UniversityDepartment)
37	SubClassAxiom	LOADED
37	SUB-CLASS	#:UniversityDepartment
37	SUPER-CLASS	#:UniversityFaculty
38	SubClassAxiom	LOADED
38	SUB-CLASS	#:UniversityDepartment
38	SUPER-CLASS	#:UniversitySchool
39	EquivalentClassesAxiom	LOADED
39	DESCRIPTIONS	(#:UniversityDepartment #:SchoolCentre)
40	ObjectPropertyAssertionAxiom	LOADED
40	SUBJECT	#:UniversityDepartment
40	REL-OBJECT-PROPERTY	#:has
40	OBJECT	#:UniversityStudent
41	ObjectPropertyAssertionAxiom	LOADED
41	SUBJECT	#:UniversityDepartment
41	REL-OBJECT-PROPERTY	#:hasWorkflow
41	OBJECT	#:UniversityWorkflow
44	EquivalentClassesAxiom	LOADED
44	DESCRIPTIONS	(#:UniversitySchool #:UniversityFaculty)

Figure 7. Some axioms of university ontology.

Axioms for attributes: The relations between attributes are described by the axioms of attribute. The axioms of attributes are listed in table II.

TABLE I.
AXIOMS WITH THEIR SYNTAX

S. No.	Relations	Syntax
1.	Relation of inclusion	rdfs:subPropertyof
2.	Equivalent	owl:equivalentProperty
3.	Inverse	owl:inverseOf
5.	Limitation of Function	owl:FunctionalProperty
6.	Inverse Function	owl:InverseFunctionalProperty
7.	Relation of Symmetry	owl:SymmetricProperty
8.	Transitive	owl:TransitiveProperty

Axioms of Instances: OWL provides two types of axioms between instances. One of these is the composition of members and value of attributes, in which firstly we classify the information and then describe the composition of each class and the value of its attribute. The other axiom defines whether the two instances are equivalent (`owl:sameAs`) or not (`owl:differentForm` and `owl:AllDifferent` etc). In our construction of university ontology, functional property is applied in object properties like `teachers` and `students` are associated with `department` with `'has'` property. In other way, `teachers` and `students` can be associated with `department` with `'is'` property such as:

"`Department` has `teachers` and `students`."

"`Teacher/student` is the member of `department`."

Here, we see `'is'` property as the inverse of `'has'` property.

F. STEP VI: The instance of ontology

After defining the efficient classes of ontology, we should select the relevant class to define the instances for the class. The `rdf:type` syntax is used to state the class of instances. we noted that more than one class can be associated with same instances and many instances can be associated with a particular class. The instances of various OWL classes of university ontology are defined in figure 8. Each instance can belong to many classes or same instance can belong to many classes.

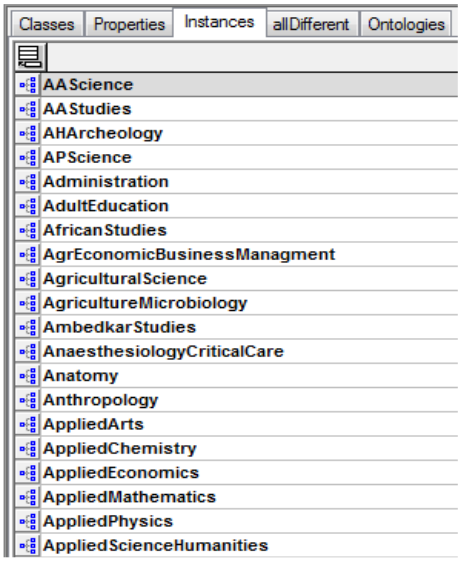


Figure 8. Instances of OWL classes of University.

G. STEP VII: Reasoning of Ontology

The reasoning is the most important part of ontology. Ontology reasoner i.e. RacerPro checks the consistency and correctness of concepts, roles, individuals, assertions, axioms, taxonomy, role hierarchy, query etc. and also find the logical contradictions implicitly described in university ontology, which are useful to cross check the various concepts and relationships of constructed ontology. We put forward the reasoning mechanism of university ontology in section 5.

V. Reasoning Mechanism of University Ontology

In this research, we grouped the reasoning mechanism into two categories. One, the ontology reasoning using description logic and the other is user-defined reasoning using first-order logic. Both categories are used to construct the university ontology by using OWL language.

A. Ontology Reasoning

Description logic (DL) allows us to specify a terminological hierarchy using a restricted set of first-order formulas [20]. To fulfill the important logical requirements, the equivalence of OWL description logic permits to exploit the considerable descriptive logic reasoning. These requirements associated with satisfyable of concept, including of OWL classes, class consistency and checking of instances. A partial set of reasoning rules that support OWL Full has been used to construct the university ontology, which are represented in table III.

TABLE III.
OWL Reasoning Rules

Property	Reasoning Rule
TransitiveProperty	(?P rdf:type owl:TransitiveProperty) ^ (?X ?P ?Y) ^ (?Y ?P ?Z) ⟹ (?X ?P ?Z)
subClassOf	(?X rdfs:subClassOf ?Y) ^ (?Y rdfs:subClassOf ?Z) ⟹ (?X rdfs:subClassOf ?Z)
subPropertyOf	(?X rdfs:subPropertyOf ?Y) ^ (?Y rdfs:subPropertyOf ?Z) ⟹ (?X rdfs:subPropetyOf ?Z)
disjointWith	(?X owl:disjointWith ?Y) ^ (?A rdf:type ?X) ^ (?B rdf:type ?Y) ⟹ (?A owl:differentForm ?B)
inverseOf	(?X owl:inverseOf ?Y) ^ (?A rdf:type ?X) ^ (?B rdf:type ?Y) ⟹ (?A owl:disjointWith ?B)
unionOf	(?X owl:unionOf ?Y) ⟹ (?A rdf:type ?X) ∨ (?B rdf:type ?Y)
intersectionOf	(?X owl:intersectionOf ?Y) ^ (?A rdf:type ?X) ⟹ (?B rdf:type ?Y)
equivalentClass	(?X owl:equivalentClass ?Y) ^ (?A rdf:type ?X) ^ (?B rdf:type ?Y) ⟹ (?A rdfs:subClassOf ?Y) ^ (?B rdfs:subClassOf ?X)

X, Y and Z are three OWL classes, A, and B are the instances of classes of any structure of university. We used these reasoning rules for our system. The examples of some properties are described below:

For `rdfs:subClassOf` property:

?X → `UniversityDepartment`
?Y → `UniversitySchool`
?Z → `University`

Reasoning Description: `UniversityDepartment` is the subclass of `UniversitySchool` and `UniversitySchool` is the subclass of `University` then `UniversityDepartment` is the subclass of `University`.

For `owl:disjointWith` property:
?X → `MathematicalScience` is the instance of `UniversitySchool` class.
?Y → `MedicalScience` is the instance of `UniversitySchool` class.
?A → `OperationalReasearch` is the instance of `MathematicalScience`.
?B → `ForensicMedicine` is the instance of `MedicalScience`.
Reasoning Description: If `MathematicalScience` is disjoint with `MedicalScience` and `OperationalResearch` is the RDF type of `MathematicalScience` and `ForensicMedicine` is the RDF type of `MedicalScience` then `OperationalResearch` is different form of `ForensicMedicine`.

For `owl:unionOf` property:
?X → `UniversityDepartment`
?Y → `UniversitySchool`
?A → `MedicalScience` is the instance of `UniversityDepartment`.
Reasoning Description: `UniversityDepartment` is the union of `UniversitySchool` then `MedicalScience` is RDF type of `UniversityDepartment` or `UniversitySchool`.

B. User-defined Reasoning

User-defined properties provide the more flexible reasoning mechanism, which is an enormous range of high-level reasoning within the entailment of first-order logic. The user defined properties are used to construct the relations between OWL classes of our university ontology, which also established the relational link with instances of OWL classes. Typical example of user defined properties and their reasoning rules are listed in table IV.

TABLE IV.
USER DEFINED REASONING RULES

User Defined Property	Reasoning Rule
has	(?X has ?Y) ^ (?Y has ?Z) ⟹ (?X has ?Z)
sameClass	(?X sameClass ?Y) ^ (?A rdf:type ?X) ^ (?B rdf:type ?Y) ⟹ (?X owl:equivalentClass ?Y)
teaches	(?D hasfaculty ?X) ^ (?D hasStudent ?Y) ⟹ (?X teaches ?Y)

The implementation of these properties with OWL classes is shown in figure 9.

X, Y and Z are three OWL classes, A, and D are the instances of classes of any structure of university. We used these reasoning rules for our system. Some examples of properties are shown below:

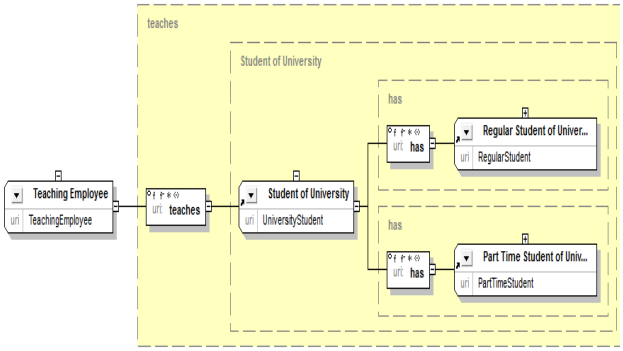


Figure 9. Implementation of User-defined Properties.

For `'has'` property:
?X → `University`
?Y → `UniversitySchool`
?Z → `UniversityDepartment`
Reasoning Description: `University` has `UniversitySchool` and `UniversitySchool` has `UniversityDepartment` then `University` has `UniversityDepartment`.

For `'sameClass'` property:
?X → `UniversityFaculty`
?Y → `UniversitySchool`
?A → `MedicalScience` is the instance of UniversityFaculty.
Reasoning Description: In this case, if `UniversityFaculty` is same class as `UniversitySchool` and `MedicalScience` is the RDF type of `UniversityFaculty` and `UniversitySchool` then both classes are equivalent class.

For `'teaches'` property:
?X → `TeachingEmployee`
?Y → `UniversityStudent`
?D → `UniversityDepartment`
Reasoning Description: If `UniversityDepartment` has faculty `TeachingEmployee` and has student `UniversityStudent` then `TeachingEmployee` teaches the `UniversityStudent`.

VI. RESULTS

We represent the results of ontology, which can show the correctness and consistency of constructed university ontology. For this, graphical views of various ontologies have been created (SemanticWorks) as shown in this section. Further, some DL rules have also been applied to extract the information from ontology. The hierarchy of OWL classes, subclasses and their relationship is shown in the graphical view of ontology. Metadata of these classes are also depicted in visualization of ontology.

A. Visulisation View

Here, we add some important classes of concepts and some important subclasses of generalized structure of universities. Visual view of some classes and their subclasses results using ontology editor tool is shown in figure 10. Asserted view depicts the classes, subclasses and their relationship, which we defined in the construction of university ontology.

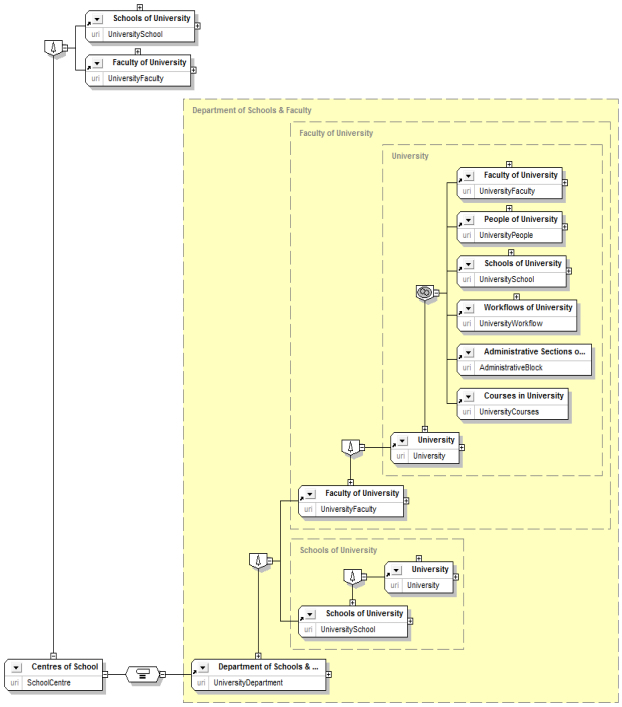


Figure 11. Graphical view of classes, subclasses and relationship.

In the figure, `SchoolCentre` is the equivalent class of `UniversityDepartment` and subclass of `UniversitySchool` and `UniversityFaculty`. The `UniversityDepartment` class is the subclass of `UniversityFaculty` and `UniversitySchool` classes, which are the subclass of `University` class. Furthermore, the `University` class is the union of (or has the members) following classes: `UniversityFaculty`, `UniversityPeople`, `UniversitySchool`, `UniversityWorkflow`, `AdministrativeBlock` and `UniversityCourses`.

Figure 11 presents the metadata of various OWL classes defined in university ontology. The relations of OWL class `UniversityPeople` is shown in table V.

TABLE V.
RELATIONS AND THEIR VALUES

Relations	Relational Value (Metadata)
rdf:type	Owl:class
rdfs:label	People of university
rdfs:subClassOf	University
has	UniversityEmployee, UniversityStudent class

Similarly, we can see the some other metadata of the

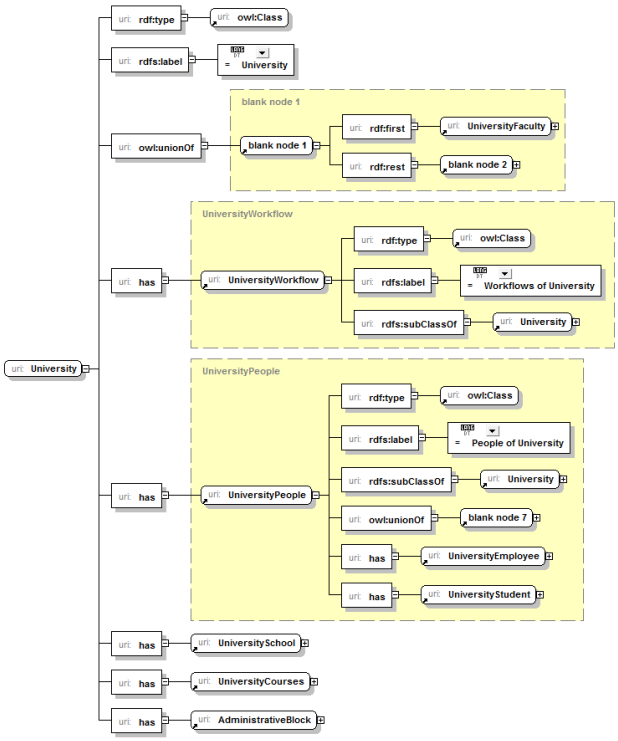


Figure 10. OWL classes and their metadata.

concepts in figure 11.

In these graphical views, we show that the classes, subclasses and their relationship, ontology reasoning, user-defined reasoning and metadata of classes etc. of university are similar to predefined organizational structure and information, which are represented in the construction of ontology. We can clearly see various concepts and their metadata of the universities, which can help to rectify the unusual concepts, relations and metadata etc.

B. Results of Query Retrieval

The DL query is used to retrieve any information about the concept of university from designed university ontology. The purpose of these queries is intended to check the correctness of concepts and information, which are designed at the time of construction of ontology.

We can provide class or any property name correctly to retrieve the information and reasoner will display related information about particular class or property. For example, if we want to retrieve various sections of administrative block of a university, then we must enter the class name correctly (upper or lower case) as created in the ontology construction. The results of some of the input queries we tested on university ontology using RacerPro [21] inference engine are listed in table VI with the description of expected results and actual results as verification. The results are as per our expectations (as shown in column 2 & 3 of table VI).

VII. CONCLUSION

The workflow processes of university domain are a challenging task for knowledge representation and

TABLE VI.
QUERY AND THEIR RESULTS BY USING RACERPRO REASONER

Input Query	Description of Expected Result	Actual Result
? (retrieve (?X) (?X #!:AdministrativeBlock) :abox file://E:/University%20Ontology/uni-onto.owl) **[retrieve all subclasses of OWL class 'AdministrativeBlock']**	InformationBureau, Legal, Estate, Planning, SC-STCell, PPPCell, ProctorialBoard, SportBoard, ComputerCentre, Hospital, Library, Engineering, Examination, Finance, Establishment, Administration, COEOffice, RegistrarOffice, VCOffice. **[all the section of Administrative Block]**	(((?X #!:InformationBureau))((?X #!:Legal))((?X #!:Estate)) ((?X #!:Planning))((?X #!:SC-STCell))((?X #!:PPPCell)) ((?X #!:ProctorialBoard))((?X #!:SportBoard)) ((?X #!:ComputerCentre))((?X #!:Hospital))((?X #!:Library)) ((?X #!:Engineering))((?X #!:Examination))((?X #!:Finance)) ((?X #!:Establishment))((?X #!:Administration)) ((?X #!:COEOffice))((?X #!:RegistrarOffice)) ((?X #!:VCOffice)))
? (individual-types #!:PhysicalEducation file://E:/University%20Ontology/uni-onto.owl) **[describe the types of individual 'PhysicalEducation']**	UniversityDepartment, SchoolCentre, UniversityFaculty, UniversitySchool, University. **[which types of the Physical Education Department]**	((#!:UniversityDepartment #!:SchoolCentre) (#!:UniversityFaculty #!:UniversitySchool) (#!:University) (*top* top))
? (describe-individual1 #!:Anthropology file://E:/University%20Ontology/uni-onto.owl) **[describe the individual 'Anthropology' of university ontology]**	Anthropology is the individual of university ontology as a department of university. This query describes about Anthropology Department.	(#!:Anthropology :assertions ((#!:Anthropology #!:UniversityDepartment)) :role-fillers nil :told-attribute-fillers nil :told-datatype-fillers ((#!rdfs:label ("Department of Anthropology"))) :annotation-datatype-property-fillers ((#!rdfs:label ("Department of Anthropology"))) :annotation-property-fillers nil :direct-types :to-be-computed
? (describe-concept #!:UniversityDepartment file://E:/University%20Ontology/uni-onto.owl) **[describe the concept 'UniversityDepartment' of university]**	This query provides the description of UniversityDepartment concept. The synonyms of this concept are SchoolCentre and parents are UniversityFaculty and UniversitySchool.	(#!:UniversityDepartment :told-primitive-definition (and #!:UniversitySchool (and #!:UniversityFaculty (and #!:UniversityFaculty (and #!:UniversitySchool (and #!:UniversityDepartment #!:SchoolCentre))))) :synonyms (#!:UniversityDepartment #!:SchoolCentre) :parents ((#!:UniversityFaculty #!:UniversitySchool)) :children ((*bottom* bottom)))

ontology development. This paper focused a way of university domain knowledge representation and ontology development, which makes such type of information like machine understandable format. The Semantic Web technologies fulfills the requirement of this work, where the represented information is understandable by machine and cooperate to human users for efficient result on intelligently described information. Some essential steps are described to construct and elaborate the reasoning mechanism of university ontology. In the reasoning mechanism, the reasoning is carried out according to the connotative relationships between concepts and shows the result according to DL rule. Altova SemanticWorks tool is used to create and edit the university ontology in visualized format. The ontology and user-defined reasoning mechanism of concepts, individuals, axioms and assertions have been elaborated and tested by inference engine.

REFERENCES

[1] Berners-Lee, T., "Semantic Web Road Map", W3C Design Issues. [Online]. Available: http://www.w3.org/DesignIssues/Semantic.html.
[2] Tim Berners-Lee, James Hendler and Ora Lassila. "The Semantic Web", Scientific American May 2001.
[3] Resource Description Framework. http://www.w3.org/RDF/.

[4] W3C, "RDF Primer", W3C Recommendation, Retrieved from, http://www.w3.org/TR/2004/REC-rdf-primer, February 2004.
[5] O. Lassila and R. R. Swick (editors). Resource description framework (rdf) model and syntax specification. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, Feb, 1999.
[6] D. Brickley and R.V. Guha (editors). Resource description framework (rdf) schema specification 1.0. http://www.w3.org/TR/2000/CR-rdf-schema-20000327/, March, 2000.
[7] W3C. RDF Schema Specification. http://www.w3.org/TR/PR-rdf-schema/, 1999.
[8] OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/.
[9] Sanjay K. Dwivedi and Anand Kumar, "An Agent-based Semantic Web for Paperless Office Content Management System", ICDM 2010: Proceedings of the 2010 Third International Conference on Data Management, pp. 352-360, Ghaziabad, India, March 2010.
[10] OWL: Web Ontology Language Overview. http://www.w3.org/TR/owl-feartures, 10 Feb 2004.
[11] SemanticWorks Semantic Web tool - Visual RDF and OWL editor, http://www.altova.com/semanticworks.html.
[12] S. Lovrenčić, Mirko Cubrilo. "University Studies Ontology – Domain Modeling", 11th International Conference on Intelligent Engineering Systems 2007. IEEE Xplore, pp. 55-58.
[13] Naveen Malviya, Nishchol Mishra and Santosh Sahu. "Developing University Ontology using Protégé OWL Tool: Process and Reasoning", International Journal of Scientific & Engineering Research, Volume 2, Issue 9, September 2011, pp. 1-8.
[14] P. Kogut, S. Cranefield, L. Hart, M. Dutra, K. Baclawski, M. Kokar, J. Smith, UML for Ontology development, Knowledge Engineering Review, 2002.
[15] Boyce, S., & Pahl, C. (2007). Developing Domain Ontologies for Course Content. Educational Technology & Society, 10 (3), 275-288.
[16] Kumar, A., Dwivedi S. K. "Ontology Exemplification for aSPOCMS in the Semantic Web", World Congress on Information and Communication Technologies (WICT) 2011. IEEE Xplore, pp. 473-478.
[17] Sanjay K. Dwivedi, Anand Kumar. "Ontology Exemplification and Modeling for aSPOCMS in the Semantic Web", International Journal of Computer Information Systems and Industrial Management Applications. MIR Labs USA, Volume 5 (2013) pp. 542-549.
[18] Sharmin Rashid Linta, Md. Maidul Islam & Md. Rakibul Islam (2012). "An Enhanced Model of E-Learning Management System Using Semantic Web Technology and Development of Universal Namespace for University Domain", International Journal of Computer Science Issues, Vol. 9, issue 2, No. 2, March 2012.
[19] Sanjay K. Dwivedi and Anand Kumar (2013). "Development of Universal Namespace for Workflow of University Domain for aSPOCMS", International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.1, January 2013, pp. 1-17.
[20] Zhang, D.Q.; Gu, T.; Pung, H.K.; "Ontology Based Context Modeling and Reasoning using OWL", Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, pp. 18-22, Singapore, March 2004.
[21] RacerPro Version 2.0, Retrieved from, http://www.racersystems.com/products/download/index.phtml.

Priority Recommendation System in an Affiliate Network

Zeeshan Khawar Malik, Colin Fyfe and Malcolm Crowe
University of The West of Scotland, Paisley, Scotland, UK
{zeeshan.malik, colin.fyfe, malcolm.crowe}@uws.ac.uk

***Abstract*—Affiliate Networks are the main source of communication between publishers and advertisers where publishers normally subscribe as a service provider and advertisers as an employer. These networks are helping both the publishers and advertisers in terms of providing them with a platform where they can build an automated affiliate connection with each other via these affiliate networks. The problem that is highlighted in this paper is the huge gap that exists between the publisher and advertiser in these affiliate networks and a solution is provided by proposing a priority recommendation system based on K-Means clustering algorithm. Every advertiser desires to have that type of publisher who is already practiced in his category of business or at least has the same skills and talent. This paper presents the concept of a recommendation system based on clustering the real-time data of all the existing transactions of publishers and advertisers of an affiliate network and based on the resulting POST-HOC classified data, a new publisher or advertiser will automatically be classified. Real-time data is provided by Affiliate Future a well-known company among all the affiliate networks. After carefully examining the data the most effective attribute is selected as the base attribute for clustering. The data is encoded into binary numbers for the purpose of clustering. More than one distance approaches are used and the most suitable one is selected for classifying the data.**

***Index Terms*—Affiliate Marketing, Clustering, Publisher, Advertiser, Sammon Mapping**

I. Introduction

Affiliate Networks like Linkshare, E-Junction and Affiliate Future are becoming the key platforms for all e-business people who want to search for a highly ranked and most effective publisher to market their product or service. Similarly on the other hand, publishers also use these affiliate networks to get connected with their choice of product so that they can better perform in terms of marketing and generating revenue. These affiliate networks have boosted the process of affiliate marketing.

Today affiliate marketing has become a key technique to market a product or service and to generate revenue in the shortest time possible [1]. Affiliate marketing has also shown a great impact on other ecommerce strategies as well in terms of generating revenue by making referrals in an n-tier commission-based mechanism [2]. More than one model has been introduced in affiliate marketing for the purpose of generating revenue that includes primarily percentage of sales model, pay per lead model, flat referral rate model, pay per email model, cost per view model and cost per click model [3]. The basic working in affiliate marketing is a process in which a publisher gets commission for selling an advertiser's product through its own platform and the advertiser confirms its sale by checking the backlink coming from publisher's own platform [4, 5, 6, 7] and [8]. The three oldest affiliate networks are

1) Linkshare

2) Be Free and

3) Cyberotica [18, 36]

Clustering is one of the most popular methods [9] for exploratory data analysis. The prime purpose is to group similar data into one cluster in such a way that data within the cluster are as similar as possible while data in different clusters are as dissimilar as possible. This technique is continuously being refined and considered as a highly studied area of AI, machine learning and statistics. Clustering is always a very important problem for marketing researchers as well in terms of grouping of persons, products, or occasions which may act as a basis for further analysis [10].

K-means clustering [12] is one of the most widely used clustering techniques in commercial environments and works very efficiently on high dimensional data as well [11]. It clusters the data based on initially defined prototypes for each cluster and based on the calculation of the sum of square distances of each point with all the defined prototypes assigns the point to that cluster where the distance is minimum. Then the prototype's positions are updated to be the average of all the data which has been assigned to that cluster.

II. Affiliate marketing

The concept of affiliate marketing was first introduced by the pioneering company Amazon, headed by Jeff Bezos in the late 20th century. The concept was so much appreciated that many online companies started adopting this technique of marketing to generate revenue in an n-tier mechanism [13, 14, 15]. Amazon has generated a lot

doi:10.4304/jetwi.5.3.222-229

of income through its Amazon Associates Affiliate Program [17]. There are different methods of pricing proposed by researchers under affiliate marketing but the most common methods that are being preferred by most of the merchants are

1) Cost per Click Model and

2) Cost per Sale Model [16].

The development of affiliate networks is one more reason for the popularity of affiliate marketing and in the absence of affiliate networks, there was no one before who could cross-check the validity of the service provided by the affiliates which resulted in a lot of scams [36]. This is one reason why few empirical researches have been done in this area because of the poor image of affiliate marketing due to inconsistent branding through many non-reliable affiliates and the lack of development of trust due to intrusive mass advertising [20, 4]. The prime objective of affiliate marketing remains the same, that is to generate revenue by selling products or services through additional outlets known as affiliates of the advertisers and in return the affiliates get commission for every sale produced. The pricing model is normally selected depending upon the affiliate model chosen. In the pay per sale model, the commission is given to the affiliate on each sale produced through its platform. In the pay per lead model, the advertisers reward affiliates for each new subscriber coming through their platform. In the pay per click model the advertiser rewards the affiliate for every click or every cost-per 1000 times impressions online users view advertisement [18, 2, 21]. The concept of making affiliates in the real world was first introduced by airlines, hotels and other tourism companies [22].

According to [18] the concept of affiliate marketing first originated in the year 1996 and that is the year when content analysis of 93 articles from three journals most related to marketing was done and the conclusion was that there was very little research of affiliate marketing and most of them failed to meet the current study requirement covering the development of affiliate marketing. The year 1999 is considered to be the year affiliate marketing opened its gates in the UK and resulted in the opening of companies like Commission Junction and Tradedoubler [23]. These affiliate networks are also termed top-tier affiliates which offer key services to merchants such as account tracking and management. Some researchers [24, 2] have also differentiated the affiliates into two types "first tier" and "second tier". The first tier affiliates are large-scale established affiliates having their own brand name and have a relatively large consumer base whereas the second tier affiliates are small-scale individuals who have their own individual platforms through which they offer services to merchants.

III. HARD CLUSTERING

K-means Clustering is

"A major clustering method producing a partition of the entity set into non-overlapping clusters along with within-cluster centroids. It proceeds in iterations consisting of two steps each; one step updates clusters according to the minimum distance rule, the other step updates centroids as the centres of gravity of clusters. The method implements the so-called alternating minimization algorithm for the square error criterion. To initialize the computations, either a partition or a set of all K tentative centroids must be specified" [37].

Let x_1, x_2, $x_3 \ldots x_n$, be the number of data points and c_1, c_2, $c_3 \ldots$, c_n be the cluster on the search space, K is assumed as the total number of clusters and n is assumed as the total number of data points. Let μ_1, μ_2, $\mu_3, \ldots$, μ_k, be the initially defined prototypes, then clustering of data points into K-means clusters J_k is determined by minimizing the sum of squared errors given in eq.(1).

$$J_k = \sum_{j=1}^{k} \sum_{i \varepsilon c_j} \left(x - \mu_j\right)^2 \qquad (1)$$

K-means is one of the most popular clustering algorithms among all the algorithms proposed in the literature for clustering: ISODATA [26, 27], CLARA [27], CLARANS [28], Focusing Techniques [29], P-CLUSTER [30], DBSCAN [31], Ejcluster [32], BIRCH [34] and GRIDCLUS [33] are all extensions of k-means. Algorithm is shown below which explains the basic working of k-means [12].

DIRECT K-MEANS

Initialize K prototypes $(m_1, m_2, m_3, \ldots, m_k)$ such that $m_j = i_l, j \varepsilon \{1,2,3,\ldots,K\}, l \varepsilon \{1,2,3,\ldots, n\}$
Each cluster C_j is associated with prototype m_j
Repeat
For each input vector i_l, where $l \varepsilon \{1, 2, 3, \ldots, n\}$
Assign i_l to the cluster C_j, with nearest prototype m_j.
i.e. $|i_l - m_{j*}| \leq |i_l - m_j|$, $j \varepsilon \{1, 2, 3, \ldots, k\}$
For each cluster C_j, where $j \varepsilon \{1, 2, 3, \ldots, k\}$
Update the prototype mj, to be the centroid of all samples currently in c_j so that $m_j = \sum_{i_l \varepsilon c_j} (i_l / |c_j|)$

Compute the error function

$$E = \sum_{j=1}^{k} \sum_{i \varepsilon c_j} |i_l - m_j|^2$$

Until E does not change significantly.

IV. DESCRIPTION

In this paper the k-means clustering algorithm [38] is selected for a real-time affiliate network data provided by a company named as "Affiliate Future" in the form of excel files. The data is related to the individual profiles and transactional information of advertisers and publishers.

Publisher's Profile Data

1) AffiliateID {Unique ID for each publisher}
2) AffiliateSiteID {More than one SiteID for single affiliate ID}
3) SiteName {Site Description}
4) SiteAddress {Site URL}

Publisher's Category Data

1) AffiliateSiteID {More than on SiteID for single affiliateID}
2) Category {Category Name}

Advertiser's Profile Data

1) MerchantID {Unique ID for each advertiser}
2) MerchantSiteID {More than on SiteID for single merchantID}
3) SiteName {Site Description}
4) SiteAddress {Site URL}

Advertiser's Category Data

1) MerchantSiteID {More than one SiteID for single merchant ID}
2) Category {Category Name}

Publisher and Advertiser's Transactional Information

1) MerchantID {Unique ID for each advertiser}
2) AffiliateID {Unique ID for each publisher}
3) MerchantSiteID {More than one SiteID for single merchant ID}
4) AffiliateSiteID {More than one SiteID for single affiliateID}
5) LogDate {Temporal information for each transaction}

There are a total of 46 categories for which all the publishers and advertisers are associated with :- 1) Adult 2) Arts and Craft 3) Auctions 4) Baby Gear 5) Books, Catalogues & Magazines 6) Business Services 7) Clothing & Accessories-Men's 8) Clothing & Accessories - Women's 9) Competitions, Freebies & Discounts 10) Computers 11) Dating 12) DVDs, Videos & Games 13) Eco-Friendly 14) Education 15) Experience 16) Financial & Legal 17) Food 18) Gadgets 19) Gaming 20) Gaming & Gambling 21) Gifts & Flowers 22) Health & Beauty 23) Home & Garden 24) Insurance 25) Internet Services 26) Jewellery 27) Latest Merchants 28) Loans 29) Mobile Phones & Accessories 30) Motoring 31) Music 32) Office Equipment 33) Outdoor Equipment 34) Pets 35) Posters & Memorabilia 36) Seasonal 37) Sports & Fitness 38) Telecommunications 39) Telecom 40) Toy Shops 41) Travel-Accommodation 42) Travel-Essentials 43) Travel-Flights 44) Travel-Holidays 45) Wedding & Celebrations 46) Wine & Drinks.

The data for both the publishers and advertisers are clustered on the basis of the above mentioned categories. A single publisher can be associated with more than one category and similarly with the advertiser's data. There is more than one publisher and advertiser associated with each category. The same approach for clustering has been used for both publisher's and advertiser's data. Further the clustering of only advertiser's data is explained. The total number of prototypes taken for clustering advertiser's data is 46 i.e. one for each category. The reason for taking 46 prototypes with each prototype belonging to each category is to classify the data in a way that each cluster resembles a category which are 46 in total. The data is first encoded into binary numbers. Total 46 bits are associated with each advertiser. Out of these 46 bits, '1' is for On and '0' is for off. A single advertiser can have more than one '1' in their total bits. This means that a single advertiser can be associated with more than one category. The initial values of half of the prototypes are shown in Table 1.

IV. Final Prototypes Result

More than one distance approach are used to cluster data using the k-means algorithm shown in Table 2.The most suitable proved to be the euclidean distance approach which is shown to be the most compatible distance approach for the k-means algorithm [12]. Some of the final values of the 46 dimensional prototypes is shown in the tables below.

TABLE 2.
DISTANCE APPROACH

Measure	Forms
Euclidean Distance Approach	$D_{ij} = \left(\sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^{1/2} \right)^{2}$
Manhattan Distance Approach	$D_{ij} = \sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert$
Cosine Similarity Approach	$S_{ij} = \cos\alpha = \frac{x_i^T x_j}{\lVert x_i \rVert \lVert x_j \rVert}$
Dot Product Approach	$D_{ij} = \lVert x_i . x_j \rVert$

TABLE 1.
INITIAL PROTOTYPES

Center	Advertisers	Category	Value
1	38	Adult	10000000000000000000000000000000 00000000000000000
3	814	Auction	00100000000000000000000000000000 00000000000000000
5	538	Books	00001000000000000000000000000000 00000000000000000
7	3506	Clothing	00000010000000000000000000000000 00000000000000000
9	583	Computers	00000000100000000000000000000000 00000000000000000
11	148	DVD's	00000000001000000000000000000000 00000000000000000
13	917	Eco-Friendly	00000000000010000000000000000000 00000000000000000
15	113	Experience	00000000000000100000000000000000 00000000000000000
17	161	Food	00000000000000000100000000000000 00000000000000000
19	127	Gaming	00000000000000000001000000000000 00000000000000000
21	161	Gifts	00000000000000000000100000000000 00000000000000000
23	343	Home	00000000000000000000000100000000 00000000000000000
25	2067	Internet	00000000000000000000000001000000 00000000000000000
27	932	Latest M.	00000000000000000000000000100000 00000000000000000
29	299	Mobile	00000000000000000000000000000000 10000000000000000
31	1409	Loans	00000000000000000000000000000000 00100000000000000
33	1985	Outdoor Eq.	00000000000000000000000000000000 00001000000000000
35	5460	Poster	00000000000000000000000000000000 00000010000000000
37	617	Sports	00000000000000000000000000000000 00000000100000000
39	4968	Telecom	00000000000000000000000000000000 00000000001000000
41	1383	Travel	00000000000000000000000000000000 00000000000010000
43	282	Flights	00000000000000000000000000000000 00000000000000100
45	1717	Wedding	00000000000000000000000000000000 00000000000000001

TABLE 3.
FINAL PROTOTYPE VALUES

Cluster	Prototype	Count	Category	Value
1	1	62	Adult	0.242
			Financial Legal	0.177
			Insurance	0.21
			Travel Holiday	0.371
2	2	4	Art and Craft	1
			Baby Gear	0.25
			Books and Mag.	0.25
5	5	7	Books and Mag	1
			Gifts and Flowers	0.413
			Home and Garden	0.413
			Jewellery	0.413
			Travel-Essentials	0.413
6	6	8	Business Services	1
			Travel Essential	0.125
			Travel Holiday	0.125
7	7	20	Adult	0.05
			Clothing and Acc.	0.25
			Latest Merchant	0.05
8	8	26	Adult	0.038
			Baby Gear	0.038
			Gifts and Flowers	0.077
			Latest Merchant	0.115
			Travel Essential	0.038
			Travel Holiday	0.038
9	9	14	Freebies and Discount	1
			Gaming and Gambling	0.071
13	13	5	Eco-Friendly	1
			Education	0.2
			Experience	0.2
14	14	7	Baby-Gear	0.143
			Education	1
			Travel-Holiday	0.143
15	15	9	Experience	1
			Travel-Essential	1
			Travel-Holiday	0.143
16	16	3	Financial & Legal	0.333
			Insurance	0.667
			Motoring	0.667
			Travel-Essentials	0.667
17	17	26	Food	0.346
			Gadgets	0.115
			Gifts and Flowers	0.731
			Health & Beauty	0.038
18	18	9	Clothing and Acc.	0.222
			Gadgets	1
			Gaming	0.111
			Jewellery	0.111
			Mobile Phone	0.111
43	43	6	Insurance	0.167
			Travel-Flights	1
			Travel-Holidays	0.167
45	45	3	Clothing and A.M.	0.333
			Clothing and A.W.	0.333
			Jewellery	0.667
			Wedding and Cel.	1

V. SAMMON MAPPING

It is always better to view a large-dimensional data as a 2-dimensional projection to facilitate visualization. One more reason of choosing this visualization technique is to better visualize the tightness and looseness of clusters in a 2-dimensional space. The Sammon mapping [39] is the most appropriate approach to transform a high-dimensional (n-dimensions) space to a space with lower dimensionality (q-dimensions) by finding N projected points in the q-dimensional space.

The basic idea of the Sammon mapping [40] and indeed all MDS methods is to arrange all the projected points in a 2-dimensional space in such a way that the distance among all the projected points remains almost the same as the distance among the data points in the original space.

In order to further elaborate the Sammon mapping mathematically. Let us suppose the distance between two points x_i and x_j such that (i≠j) in the n-dimensional space be denoted by d_{ij} and the distance between two projected points y_i and y_j in the q-dimensional space be denoted by d_{ij}' *then* the mapping from the higher dimension space to the lower dimension space is done by minimization of the Sammon stress function defined in equation (2).

$$E = \frac{1}{\lambda}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(d_{ij}-d'_{ij})^2}{d_{ij}}, \lambda = \sum_{i=1}^{N-1}\sum_{j=1+1}^{N} d_{ij} \quad (2)$$

For minimization of E, steepest descent procedure is most commonly used in which the new iteration y_{il} at iteration t+1 is given in equation (3).

$$y_{il}(t+1) = y_{il}(t) - \alpha\left[\frac{\frac{\partial E(t)}{\partial y_{il}}}{\frac{\partial^2 E(t)}{\partial^2 y_{il}}}\right] \quad (3)$$

where y_{il} is the l-co-ordinate of point y_i in the new space and α is a constant which sammon takes to be 0.3 or \$0.4\$. The partial derivatives in (3) are

$$\frac{\partial E(t)}{\partial y_{il}(t)} = -\frac{2}{\lambda}\sum_{k=1\,k\neq i}^{N}\left[\frac{d_{ki}-d'_{ki}}{d_{ki}d'_{ki}}\right](y_{il}-y_{ki}) \quad (4)$$

$$\frac{\partial^2 E(t)}{\partial^2 y_{il}(t)} = -\frac{2}{\lambda}\sum_{k=1\,k\neq i}^{N}\frac{1}{d_{ki}d'_{ki}}\left[(d_{ki}-d'_{ki})-\left[\frac{(y_{il}-y_{ki})^2}{d'_{ki}}\right]\right]\left[1+\frac{d_{ki}-d'_{ki}}{d_{ki}}\right](y_{il}-y_{ki}) \quad (5)$$

The graphs in Figure 1 show the visualization of 46 dimensional data before and after implementation of k-means clustering algorithm. Figure 1 top left shows the Sammon mapping of the data using category information supplied to us. The second diagram in that column shows a zoom in projection of the first taking only the central portion. The third diagram in that column shows a further zoom in. Ideally we would like to see groups of projected points which are categorised as the same type of merchant as lying close to each other (all the blue *s close to each other, and separate from all the red +s etc.) but this has not happened.

The right hand side of Figure 1 shows projections when we use categorical data from a prior k-means clustering. The projections are a bit different because a single merchant belonging to more than one categories have more than one projections. Here we do not say that a merchant belongs to a particular type but only a particular cluster. We see in the second and third diagrams (after zooming in) groups of points which are very close to one another and separate from other groups of points.

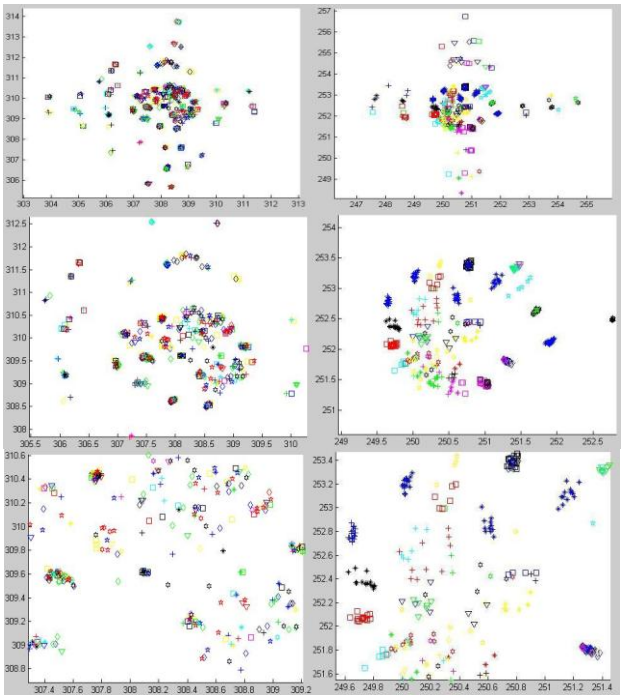


Figure 1. Top left: Visualization using the Sammon Mapping of the data. The class information is given by the pre-defined categories. Top right: the same projection but the class information is taken as the clusters from the k-means algorithm. Middle and bottom diagrams: zooming in on the top diagrams in each column.

Since the data is 46 dimensional and a single merchant is most of the time working in more than one category and also with more than one publisher, the resulting clusters also consist of more than one category of merchants classified in the same cluster. It can be further elaborated as for suppose a single merchant X is working at the same time in ``Art & Craft", ``Baby Gear" and ``Auction" categories and at the same time linked with more than one publishers working in these three categories so this data when classified will give the recommendation of all the publishers working in these three categories to the new merchants working in any one of the above mentioned categories. This is the main reason for the looseness of some of the clusters whereas in some clusters where the merchants are working in one category the data is tightly classified. The samples in cluster 1 are 62 related to ``Adult", ``Travel-holidays", ``Insurance" and ``Financial & Legal", cluster 2 are 4 related to ``Art & Crafts", ``Baby Gear", ``Books Catalogues and Magazines", cluster 3 are 3 related to ``Auctions", cluster 4 is 1 related to "Baby Gear", "Food", "Gifts & Flowers", "Latest Merchants" and "Wedding & Celebrations", cluster 5 are 7 related to ``Baby Gear'', "Food'',''Gifts & Flowers", "Latest Merchant" ,"Wedding & Celebrations", "Travel & Essentials", "Home & Garden", "Jewellery" and "Books", "Catalogues & Magazines", cluster 6 are 8 related to "Business Services", "Travel-Essentials" and "Travel-Holidays", cluster 7 are 20 related to "Clothing &

Accessories -Men's", "Adults", "Clothing & Accessories - Women's", cluster 8 are 26 related to "Adult", "Clothing & Accessories - Women's", "Latest Merchants", "Gifts & Flowers", "Wedding & Celebrations", "Travel-Essentials" and "Baby Gear", cluster 9 are 14 related to "Competitions", "Freebies & Discounts" and "Gaming & Gambling"', cluster 10 are 10 related to "Computers", cluster 11 are 8 related to "Dating", cluster 12 are 3 related to "DVDs, Videos & Games", cluster 13 are 5 related to "Eco-Friendly", "Education" and "Experience", cluster 14 are 7 related to "Education", "Baby Gear" and "Travel-Holidays", cluster 15 are 9 related to "Experience", "Travel-Essentials" and "Travel-Holidays", cluster 16 are 3 related to "Financial & Legal", "Insurance", "Latest Merchants" ,"Motoring" and "Travel-Essentials", cluster 17 are 26 related to "Food", "Gift & Flowers", and "Gadgets", cluster 18 are 9 related to "Clothing & Accessories-Women's", "Gadgets", "Gaming", "Mobile Phones & Accessories" and "Jewellery", cluster 19 are 9 related to "Gaming", "Competition Freebies & Discounts", "Gaming" and "Gaming & Gambling", cluster 20 are 30 related to "Games & Gambling" and "Latest Merchants", cluster 21 are 0, cluster 22 are 32 related to "Health & Beauty" and "Adult", cluster 23 are 42 related to "Home & Garden", "Baby Gear" and "Latest Merchants", cluster 24 are 0, cluster 25 are 2 related to ``Internet Services", cluster 26 are 2 related to "Jewellery", cluster 27 are 9 related to "Health & Beauty", "Latest Merchant", "Health & Beauty", "Travel & Essentials", "Latest Merchants" and "Motoring", cluster 28 are 10 related to "Loans", cluster 29 are 18 related to "Mobile Phones & Accessories" and "Travel-Essentials", cluster 30 are 0, cluster 31 are 4 related to "Latest Merchants" and "Music", cluster 32 are 2 related to "Latest Merchants" and "office Equipment", cluster 33 are 1 related to "Home & Garden", "Outdoor Equipment" and "Sports & Fitness", cluster 34 are 3 related to "Pets", cluster 35 are 1 related to "Posters & Memorabilia", cluster 36 are 1 related to "Gift & Flowers" and "Seasonal", cluster 37 are 15 related to "Sports & Fitness", "Gifts & Flowers" and "Latest Merchants", cluster 38 are 2 related to "Telecommunications", cluster 39 are 1 related to "Telecoms", cluster 40 are 16 related to "Motoring" and "Mobile Phones & Accessories", cluster 41 are 62 related to "Travel-Accommodation", "Travel-Holidays", "Travel-Essentials", "Travel-Flights", cluster 42 are 0, cluster 43 are 6 related to "Travel-Flights", "Insurance" and "Travel-Holiday", cluster 44 are 0, cluster 45 are 3 related to "Wedding & Celebrations", "Jewellery", "Clothing & Accessories-Men's", "Clothing & Accessories-Women's" and "Jewellery".

In this paper clustering is being done by using the robust method of k-means experimented on a real-time data provided by "Affiliate Future" and further recommendation process is functioned by an enhanced recommendation algorithm given below.

VI. PRIORITY RECOMMENDATION SYSTEM FOR AN AFFILIATE NETWORK

Function Priority-Recommendation ()

Take N Publishers (P_1, P2, P3,....., P_N) as the total number of publishers

Take K Advertisers (A_1, A_2, A_3,...., A_k) as the total number of Advertisers

Step 1

Retrieve the total Number of publishers linked with the input vector consisting of more than one advertiser

Step 2

For each input of vector Advertiser A_K where k ε {1, 2, 3 ,....,L} where L<=K

Calculate the Number of Advertisers linked with each publisher in an array linked[J] where J ε {1, 2,3,....N}
End For

/* Sorting the Linked Array using Bubble Sort Procedure from Higher to Lower */

Step 3

```
Repeat
    swapped = false
    For i = 1 to length (linked) - 1 inclusive
do:
      /* if this pair is out of order */
      if  linked[i-1] > linked[i]  then
        /* swap them and remember something changed */
        swap(linked[i-1],linked[i])
        swapped = true
      end if
    end for
until not swapped
```

In this way when the new publisher comes to select an advertiser, the advertiser currently working with higher number of publishers will be ranked the highest and so on. The same procedure will be applied for the advertisers as well looking to work with the most appropriate and trust worthy publishers. The Figure 2 shown below portrays the overall working of the proposed idea of clustering on an affiliate network data and priority recommendation process.

VII. CONCLUSION

In this paper, a recommendation system using k-means clustering is introduced in the new domain of affiliate networks. The experiments demonstrated that the recommendation system proposed using k-means clustering and priority recommendation algorithm will play a significant role in selecting the best suitable

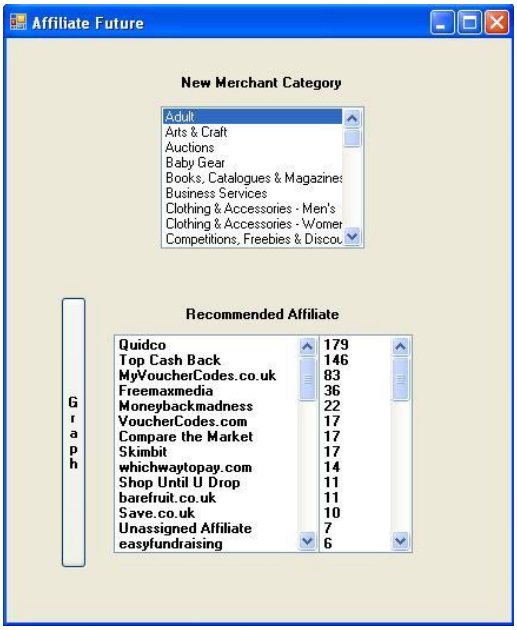


Figure 2. Snapshot of the Experiment

candidate in both the cases of publisher as well as advertisers.

ACKNOWLEDGMENT

The authors would like to gratefully acknowledge the careful reviewing of an earlier version of this paper which has greatly improved the paper

REFERENCES

[1] B., C., Brown, "The Complete Guide to Affiliate Marketing on the Web: How to Use and Profit from Affiliate Marketing Programs," *Atlantic Publishing Company*, 2009.

[2] D., L., Duffy, "Affiliate Marketing and its impact on e-commerce" *Journal of Consumer Marketing*, vol. 22(3), pp. 161–163, 2005.

[3] S., Bandyopadhay, J., Wolfe and R., Kini , "A Critical Review of Online Affiliate Models" *Journal of Academy of Business and Economics*, vol. 9(4), pp. 1, 2009.

[4] F., M., Del and M., Paul, "Reevaluating Affiliate Marketing" *Journal of Academy of Business and Economics*, vol. 9(4), pp. 1, 2009.

[5] Clare, G., "Affiliate Marketing [Electronic Version]" *New Media Age*, vol. 11, 2006.

[6] S., Goldschmidt, S., Junghagen and U., Harris "Strategic Affiliate Marketing" *Edward Elgar Publishing*, 2003.

[7] M., Haig, "The e-marketing handbook: An indispensable guide to marketing your product and services on the internet" *Kogan Page Limited*, 2001.

[8] D., Tweney, "Affiliate Marketing: the future of e-commerce or another hard sell" *Info World Electronic*, 1999.

[9] R., Xu and W., D., II "Survey of Clustering Algorithm" *Neural Networks, IEEE Transactions*, vol. 16(3), pp. 645-678, 2005.

[10] P., Girish and D., W., Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application", *Journal of Marketing Research*, pp. 134-148, 1983.

[11] C., Elkan, "Clustering with K-means: faster, smarter, cheaper", *SIAM International Conference on Data Mining*, 2004.

[12] J., Macqueen, "Some methods for classification and analysis of multivariate observations", *Proceedings of the fifth Symposium on Math, Statistics and Probability*, 1967.

[13] J., Dysart, "Click Through Customers", *Bank Marketing*, 2002

[14] L., Forex, "Affiliate Marketing Makes Headway", *Upside*, vol. 12(4), pp. 176, 2000.

[15] S., Oberndorf, "Get yourself affiliated", *Catalog Age*, vol. 16(9), pp. 63-64, 1999.

[16] B., Libai, E., Biyalogorsky and E., Gerstner, "Setting referral fees in affiliate marketing", *Journal of Service Research*, vol. 5(4), 303-315, 2003.

[17] A. Dhesikan, "Exploring Internet Marketing: A Whole New World of Opportunity", *Bachelor Project Report Worcester Polytechnic Institute*, 2012.

[18] S. Collins, "History of Affiliate Marketing", *Successful Affiliate Marketing for Merchants*, 2000.

[19] M., Anastasia, D., Roberto and B., David, "Unintended consequences in the evolution of affiliate marketing networks: a complexity approach", *The Service Industries Journal*, vol. 30(10), pp. 1707-1722, 2010.

[20] M., D., Paula, "The Web Generation Calls for Extra Image Vigilance", *Bank Technology News*, vol. 17(10), pp. 40-42, 2004

[21] G., Helmstetter and P., Metivier, "Affiliate Selling: Building Revenue on the Web", *John Wiley\ & Sons, Inc.*, 2000.

[22] C., Dale, "The Competitive Network of Tourism emediaries: New strategies new advantages", *Journal of Vacation Marketing*, vol. 9(2), pp. 109-118, 2003.

[23] J., Wallington and D., Redfearn, "IAB Affiliate Marketing Handbook", *Internet Advertising Bureau*, 2007.

[24] C., Dorobantescu, "6 Tips for Affiliate Managers", *Avantgate*, 2008.

[25] R., Leslie, "Clustering for data mining: A data recovery approach", *Psychometrika*, vol. 72(1), 2007.

[26] A., Jain and R., C., Dubes, "Algorithm for clustering data", *Prentice-Hall, Inc.*, 1988.

[27] L., Kaufman and P., J., Rousseew, "Finding groups in data: an introduction to cluster analysis", vol. 344, 2009.

[28] R., T., Ng and J., Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", *Proceedings of the 20th VLDB Conference*, pp. 144-155, 1994.

[29] M., Ester, H., Kriegel and X., Xu, "Knowledge Discovery in Large Spatial Database: Focusing Techniques for Efficient Class Identification", *Proc. 4th Int. Symp. On Large Spatial Databases*, pp. 144-155, 1995.

[30] Dan, J., P.,K., McKinley and A.,K., Jain, "Large-scale parallel data clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 871-876, 1998. [33]

[31] M., Ester, H., Kriegel, J., Sander and X., Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", *AAAI Press*, pp. 226-231, 1996.

[32] J., A., Garcia, J., A., Garcaposia, J., Fdez-valdivia, F., J., Cortijo and R., Molina, "A Dynamic Approach for Clustering Data", *Signal Processing*, 1994.

[33] E., Schikuta, "Grid-Clustering: An efficient hierarchical Clustering method for very large data sets", *Pattern Recognition, IEEE Proceedings of the 13th International Conference on,* vol. 2, pp. 101-105, 1996.
[34] Z., Tian, R., Raghu and L., Miron, BIRCH: An efficient data clustering method for very large databases, *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 1996.
[35] S., L., Brown and K., M., Eisenhardt, "The art of continuous change: linking complexity theory and timepaced evolution in relentlessly shifting organizations", *Administrative Science Quarterly*, pp. 1-34, 1997.
[36] D., Samosseiko, "The Partnerka – What is it, and Why should you care?", *Virus Bulletin Conference,* 2009.
[37] B., G., Mirkin, "Clustering for Data Mining, A Data Recovery Approach", *Chapman & Hall/CRC Taylor & Francis Group,* 2005.
[38] A., Enright, J., V., Dongen, S. and A. Christos, "An efficient algorithm for large-scale detection of protein families." *Nucleic acids research,* vol. 30(7), pp. 1575-1584, 2002.
[39] D., de Ridder *, R., P., W., Duin, "Sammon's Mapping using Neural Networks: A Comparison", *Pattern Recognition,* vol. 18, pp. 1307-1316, 1997.
[40] J., W., Sammon, JR., "A Non-Linear Mapping for Data Structures Analysis", *IEEE Transactions on Computers,* vol. 18(5), 1996.

Zeeshan Khawar Malik is currently a PhD Candidate at the University of The West of Scotland. He received his MS and BSCS (honors) degree from University of The Central Punjab, Lahore Pakistan, in 2003 and 2006, respectively. By profession he is an Assistant Professor in University of The Punjab, Lahore Pakistan currently on Leave for his PhD studies.

Colin Fyfe is a Personal Professor at The University of the West of Scotland. He has published more than 350 refereed papers and been Director of Studies for 23 PhDs. He is on the Editorial Boards of 6 international journals and has been Visiting Professor at universities in Hong Kong, China, Australia, Spain, South Korea and USA.

Malcolm Crowe obtained his D.Phil (Oxon) in Mathematics in 1978 and was appointed a Professor of Computing in 1985. He maintains a number of research interests including Enterprise Computing and Business Intelligence, and is Director of Studies or second supervisor to over a dozen PhD students at the University of the West of Scotland.

Adaptive Search and Selection of Domain Ontologies for Reuse on the Semantic Web

Jean Vincent Fonou-Dombeu
Department of Software Studies, Vaal University of Technology, South Africa
Email: fonoudombeu@gmail.com

Magda Huisman
School of Computer, Statistical and Mathematical Sciences, North-West University, South Africa
Email: Magda.Huisman@nwu.ac.za

***Abstract*—Ontology plays an important role in Semantic Web applications. However, building ontology remains challenging due to the time, cost an effort required. Several studies have proposed the reuse of existing ontologies when building new ones. However, some challenges remain: (1) locating relevant domain ontologies for reuse, (2) determining appropriate concepts for searching targeted ontologies and (3) understanding the discovered ontologies.This study presents an adaptive strategy for searching and selecting domain ontologies for reuse on the Semantic Web. The strategy relies on ontology-based and generic search engines, and predefined ontology features to locate existing domain ontologies and related data sources. The data sources provide ontologies' specific concepts that enable their easy location over the Semantic Web. Finally, a set of criteria including semantic coverage, codification language, modularity and open availability are used to select the best reusable set of ontologies for the domain. The application of the framework in the e-government domain demonstrated its feasibility and yielded promising results.**

Index Terms - Semantic Web, Ontology Search, Ontology Selection, Ontology Reuse, E-government.

I. Introduction

The Semantic Web is an evolution of the current web that provides meaning to web contents to enable their intelligent processing by computers. The meaning of web contents is represented with ontology and described formally in logic-based syntaxes to facilitate their integration and interoperability. As such, ontology is a key component of any semantic web application. Ontology is commonly defined as an explicit specification of a conceptualization [1] i.e., a model of the real world domain such as medicine, geographic information systems, physics, e-government and so forth; which is explicitly represented with existing objects, concepts, entities and relationships between them.

Building ontology in Semantic Web remains a challenging task due to the demand in time, cost an effort. The solution lies in the reuse of existing domain ontologies when building new ones [2][3][4][5][6][7]. In fact, ontology reuse may (1) reduce human efforts required to formalized new ontologies from scratch, (2) increase the quality of the resulting ontologies because the reused ontologies have already been tested, (3) simplify the mapping between ontologies built using shared components of existing ontologies, and (4) improve the efficiency of ontology maintenance [5].

However, existing domain ontologies are spread over the Internet and presented in different media including Semantic Web ontology files such as Resource Description Framework (RDF) and Web Ontology Language (OWL), text files (related research/project reports/published articles, program generated codes, etc.), Web pages, etc. Furthermore, ontology search engines enable the retrieval of ontology files based on keywords search; this presents some challenges: searching ontologies by keywords requires one to provide keywords that are likely to match those in the ontology files available in the indexes of the search engines [8]; but it is difficult to guess keywords of unknown domain ontologies; even if the domain ontology is known, it remains challenging to accurately guess keywords that are included in this ontology over the Internet. Moreover, semi-automatic and automatic ontology reuse solutions largely rely on ontology search engines for locating existing domain ontologies over the Semantic Web; consequently, they only focus on ontology files stored in the indexes of the search engines [4][9]; other data sources of existing ontologies such as related research/project reports, published articles, programme codes, Web page contents are left out. This results in many useful domain ontologies and information sources, including that of located ontologies, being ignored in these ontology reuse solutions; consequently, these solutions are directed towards experienced ontology engineers who are able to understand the located domain ontology files (RDF/OWL for example) to guide the process for building new ontologies.

The aforementioned challenges hinder the widespread reuse of existing domain ontologies and undermine the adoption of Semantic Web technologies in the respective domains. This study presents a framework for searching and selecting domain ontologies for reuse on the Semantic Web. The proposed framework may be applied in any application domains of Semantic Web such as e-commerce, e-business, e-learning, multimedia, e-government, etc., to identify and analyze existing domain ontologies for the purpose of knowledge sharing and reuse across domain specific Semantic Web applications. The framework uses an adaptive strategy that relies on ontology-based and generic search engines, and predetermined ontology features to locate existing domain ontologies and

doi:10.4304/jetwi.5.3.230-239

related data sources.The result is a list of candidate domain ontologies along with sets of data sources. The data sources of an ontology may include semi-structured and unstructured data such as research and project deliverable reports, related published articles, ontology codes, plain texts on project web sites , ontologies repositories, etc. These data sources disclose valuable information that may support the widespread reuse and evolution of corresponding domain ontologies. Examples of such information are: (1) the purpose(s) for which the ontology was built, (2) the methodology employed to build the ontology, (3) the full or partial ontology graph(s), (4) theoretical explanation of the meaning of concepts and axioms, (5) full or partial code of the ontology, and (6) detailed description of the use of the ontology in real world semantic-based projects, etc. [10][11][12][13]. This information is certainly valuable for any reuse tasks, including the automatic or semi-automatic ontology reuse which requires the ontology engineer to have prior knowledge of existing domain ontology to be able to comprehend and guide the process for building new ontologies through the reuse of existing ones [3][4][5][7]. More importantly, the collected ontologies' data sources provide Semantic Web developers with specific concepts of the targeted ontologies to enable their easy location over the Semantic Web; furthermore the data sources provide useful information for analyzing, understanding and reusing the existing domain ontologies.

Finally, various metrics including semantic coverage, open availability, codification language, and modularity are applied on the set of located candidate domain ontologies to evaluate and select the best reusable set of ontologies for the respective domain. The selected ontologies provide a good sharable and reusable conceptual representation and description of the domain. This may (1) promote their reuse across domain specific Semantic Web projects, (2) save the time and cost needed for building new ontologies from scratch in domain specific Semantic Web projects, (3) prevent inconsistency and confusion that may arise from multiple semantic representations of the same domain knowledge, and (4) strengthen the harmonization and adoption of Semantic Web technologies in the respective domain.

The proposed framework is simple and suitable to any Semantic Web developer who may like to search and locate existing domain ontologies on the Semantic Web, analyze, understand and reuse these ontologies in the process of building new ontologies either manually or with semi-automatic or automatic ontology reuse solutions [3][4][5][7]; this may promote the widespread reuse of existing domain ontology on the Semantic Web. The application of the framework in the e-government domain demonstrated its feasibility and yielded promising results.

The rest of the paper is structured as follows. Section 2 provides a formal specification of the framework of the search and selection strategy. The results of the application of the framework to the e-government domain is presented and discussed in Section 3. Section 4 discusses related studies and the last section concludes the paper.

II. Framework of the Search and Selection Strategy

Let's D be a domain of knowledge such as e-commerce, e-business, e-government, etc. The aim is to investigate available data sources on semantic web initiatives (real world projects, academic research works, etc.) in that particular domain. The sources of information may include technical research / deliverable reports, published articles, programming codes, data repositories, plaintext pasted on websites, etc.

Let's L_D be the set of identified semantic based data sources gathered in the domain D. L_D is defined as in Equation (1).

$$L_D = A \cup P \tag{1}$$

where, A is the set of data sources that are related to research carried out for academic purposes and P is the set of sources that are related to business projects for building semantic web applications. A and P are defined as in Equation (2) and (3).

$$A = \{a_i\}, 1 \leq i \leq N \tag{2}$$

$$P = \{p_j\}, 1 \leq j \leq M \tag{3}$$

where, N and M are the cardinalities of A and P respectively.

Let's l_{Dk} be the list of domain keywords to be used for the search of data sources, l_{Op} the list of ontology features required to guess the presence of any ontology activities in the data sources, and l_{Oc} the list of ontology specific concepts identified in data sources. Moreover, let's L_C and C be the list of data sources susceptible to content domain ontologies and the set of candidate domain ontologies respectively. C is defined as in Equation (4).

$$C = \{O_k\}, 1 \leq k \leq n \tag{4}$$

where, O_k is the candidate ontology number k; it is assumed that there are up to a number n candidate ontologies in the domain.

Finally, let's C_r be the set of predefined criteria for selecting an ontology for the domain D and d_o the final set of selected ontologies. To fulfill the goal of the framework which is to search and select domain ontologies in the domain D, the following tasks are manually or semi-automatically performed: online search, group data sources, analyze data sources, online specific search, find candidate domain ontology, and select domain ontology. A brief definition of each of these tasks is provided below.

- **Online Search** - This task uses ontology search engines such as Swoogle, Watson, OntoSearch, OntoSearch2, OntoKhoj [9], etc. and generic search engines such as Google, Google Schoolar, IEEE Explore, ISI Web of Knowledge, etc. to gather diverse data sources on existing semantic-based research and projects, based on the list of domain keywords in l_{Dk}. The result is the set L_D of all identified semantic-based data sources.
- **Group Data Sources** - The set L_D of all data sources is used in this task; evidences of relatedness are searched

in the data sources; this enable to group the data sources. Two data sources are related if they were produced under the same project or study. At this stage, the result is a collection of folders containing data sources related to the same semantic-based academic research (the set A) or real world semantic-based project (the set P).

- **Analyze Data Sources** - This task uses the set of targeted ontology features l_{Op} and either an element of A (a folder containing data sources related to an academic research project) or an element of P (a folder holding data sources pertaining to a real world semantic-based project). Ontology features include ontology graphs and concepts of the semantic web ontology languages such as RDF/RDFS and OWL. These concepts may have been used in a semi-formal definition of an ontology (simple definition of concepts and relationships in the form of texts), in the graphical representation of an ontology or within different axioms representing a formal ontology (machine generated codes). OWL constructs targeted could include Class, SubClassOf, EquivalentClass, DisjointWith, ObjectProperty, Property, Domain, Range, etc., whereas, RDF constructs could encompass Class, SubClass, SubProperty, Domain, Range, Object, Predicate, Type, Literal, etc. The result of this task is a list of ontology concepts l_{Oc}. l_{Oc} may be empty or not, depending on whether the targeted ontologies features in l_{Op} where found or not.
- **Online Specific Search** - The set of specific ontology concepts l_{Oc} obtained with the previous task is used in this task to perform further search with ontology search engines; aiming at finding the codes of the targeted ontologies. The results of the search are used to update the sets $a_i \subset A$ or $p_j \subset P$ of semantic-based data sources.
- **Find Candidate Domain Ontology** - This task consists of scrutinizing each data source $a_i \subset A$ or $p_j \subset P$ where ontology features were found to identified candidate domain ontology. The result is a candidate ontology O_k. O_k is added to the set of candidate ontologies C.
- **Select Domain Ontology** - A candidate ontology $O_k \subset C$ and the set of predefined criteria C_r for selecting domain ontologies in the domain D are used in this task. Further analysis of the ontology $O_k \subset C$ data sources is then performed to tell whether the candidate ontology $O_k \subset C$ meet the selection criteria. Based on the works in [14] and [15], it is suggested that the elements of the set C_r of predefined criteria for selecting a domain ontology $O_k \subset C$ be: codification language, semantic coverage, modularity and open availability. These criteria are defined below.

  - **Codification Language** This characteristic refers to the language employed for the formal representation of the ontology. In fact, it is expected that the codification language of a selected ontology be one of the standard ontology languages for the Semantic Web, such as Resource Description Framework (RDF) or Web Ontology Language (OWL).
  - **Semantic Coverage** The value of this characteristic is low, medium or high, thereby indicating the level of semantic richness of the ontology; the semantic richness is assessed based on the ontology features such as the number of concepts, supsumption (is.a), meronymy (part-of), etc.; in brief, a selected ontology should not be built as a simple taxonomy, it must further be formed of rich semantic features.
  - **Modularity** This characteristic tells whether the ontology is formed of a single or many components. An ontology with several modules enables: (1) easy reuse of smaller parts, (2) distributed and collaborative development, (3) smooth and efficient evolution, and (4) easy replacement of parts of the ontology [16].
  - **Open Availability** Here, it is shown whether the ontology is publicly available or not. The accessibility of the selected ontologies to the public is of prime importance as the major aim of the study is to foster the reuse of the selected domain ontologies in Semantic Web projects in the domain D.

In light of the above, the pseudo-code of the framework's algorithm is drawn in Table 1. In the next section, the framework described above and formalized in the algorithm in Table 1 is applied on the e-government domain.

III. Application in E-government

A. Online Search of Domain Ontologies

First of all, it became necessary to investigate and choose amongst existing ontology search engines those that are suitable for the task at hand. The researchers benefited from the work in [9]. In fact, in [9] a detailed comparative analysis of the commonly used [9] semantic web search engines including Swoogle, Watson, Sindice, Falcons and Semantic Web Search Engine; the study revealed that Swoogle and Watson are the state-of-the-art of all ontology search engines. Consequently, the Swoogle and Watson ontology search engines were adopted in this study. Thereafter, the following e-government domain keywords were chosen to perform the search in Swoogle and Watson search engines: government, citizen, service, business, tax, procurement, law, department, agency, civil servant, and life event.

These keywords were not exhaustive, but the aim was to perform the search and appreciate the nature of the results obtained. Furthermore, the abovementioned keywords were grouped into triplets as in Fig. 1 with the aim of improving the quality of the search results [9].

Although Swoogle and Watson search engines could return hits on OWL and RDF ontology files, some general problems surfaced. Firstly, searching ontologies by keywords requires one to provide keywords that are likely to match those in the ontology codes available in the indexes of the search engines [8]; but it is difficult to guess keywords of unknown domain

TABLE I
PSEUDO-CODE OF THE ONTOLOGY SEARCH AND SELECTION ALGORITHM

Inputs : $D; l_{Dk}; l_{Op}; C_r$
1. $L_D = Online\ search\ with\ domain\ keywords\ in\ l_{Dk}$
2. $A = Group\ academic - based\ data\ sources\ from\ L_D$
3. $P = Group\ real\ world\ projects\ data\ sources\ from\ L_D$
4. **For** All academic research a_i in A
5. $l_{Oc} = Analyse\ a_i\ data\ sources\ with\ the\ ontolgy\ features\ in\ l_{Op}$
6. **If** specific ontology concepts were found i.e. l_{Oc} isn't empty **Then**
7. $a_i = update\ a_i\ data\ sources\ with\ a\ specific\ online\ search\ with\ l_{Oc}$
8. **EndIf**
9. $O_k = analyse\ a_i\ data\ sources\ to\ identify\ corresponding\ domain\ ontology$
10. $C = update\ the\ set\ of\ domain\ ontologies\ C\ with\ the\ new\ ontology\ O_k$
11. **EndFor**
12. **For** All project p_j in P
13. $l_{Oc} = Analyse\ p_j\ data\ sources\ with\ the\ ontolgy\ features\ in\ l_{Op}$
14. **If** specific ontology concepts were found i.e. l_{Oc} isn't empty **Then**
15. $p_j = update\ p_j\ data\ sources\ with\ a\ specific\ online\ search\ with\ l_{Oc}$
16. **EndIf**
17. $O_k = analyse\ p_j\ data\ sources\ to\ identify\ corresponding\ domain\ ontology$
18. $C = update\ the\ set\ of\ domain\ ontologies\ C\ with\ the\ new\ ontology\ O_k$
19. **EndFor**
20. **For** All candidate domain ontologies O_k in C
21. $Use\ selection\ criteria\ in\ C_r\ to\ analyse\ O_k\ data\ source$
22. **If** O_k matches the selection criteria in C_r **Then**
23. $d_o = update\ the\ set\ d_o\ of\ selected\ domain\ ontologies\ with\ O_k$
24. **EndIf**
25. **EndFor**
Output : d_o

ontologies; even if the domain ontology is known, it remains challenging to accurately guess keywords that are included in this ontology over the Internet. Secondly, the number of hits returned for certain keywords entered in the search engines was high; then, it becomes impractical to click and visually assess each hit; furthermore, a large number of hits returned were not related to useful ontologies for the domain [9]. Finally, the ontology codes downloaded from the search did not provide enough information on the target ontologies; in general only concepts of the ontologies and their semantic structures (axioms) are provided in these codes. Although the Watson search engine could provide some Meta data such as the size of the ontology, its number of statements, classes, properties, individuals, etc. little information was provided in these ontology codes on the discovered ontologies such as the purposes and circumstances for which they were built, the available documentation such as the deliverable reports of projects in which they were built, the related published articles, etc. This information may provide important insights for analysing and reusing these ontologies. In fact, a good documentation on an existing ontology would certainly ease its reuse and evolution. In light of the above mentioned challenges, it becomes necessary to complement the results of the ontology search engines (Swoogle and Watson) with that of robust and generic search engines. To this end, a generic search was carried out in several search engines including ISI Web of Knowledge, IEEE Explore, Google Scholar and Google. The keywords employed were "e-government ontology" and "semantic e-government". These generic searches produced 202 e-government domain semantic-based documents presenting ontology codes, semantic-based published articles, research and projects' deliverable reports, and ontology repositories. These ontologies' data sources are grouped in the next subsection.

B. Group Data Sources

It was discovered that several documents downloaded with the generic searches were related to the same semantic-based projects or study. Then, a strategy based on the analysis of their contents was used to group related documents. To this end, each downloaded document was searched for the acknowledgement section. In fact, where found, the acknowledgement section provided information on the project or study in which the research was undertaken. Furthermore, the deliverable reports of various e-government projects, mainly European based projects, were scrutinized to discover more semantic-based e-government projects. As a result, all the documents downloaded were grouped into 21 folders, corresponding to 19 e-government projects and several academic studies. The analysis of the discovered ontology data sources is explained in the next subsection.

C. Analyse Data Sources

The semantic-based researches and projects documents downloaded in the previous task were further scrutinized to identify the projects and research studies which have employed ontology to address a particular aspect of e-government services delivery. This was done by checking ontology features in these documents. Let's recall that ontology features include ontology graphs and concepts of the semantic web ontology languages such as RDF/RDFS and OWL. These concepts may have been used in a semi-formal definition of an ontology (simple definition of concepts and relationships

Fig. 1. Triplets of Domain Keywords Employed for E-government Domain Ontologies Search

in the form of texts), in the graphical representation of an ontology or within different axioms representing a formal ontology (machine generated codes). The identified candidate ontologies were recorded along with their authors, date of publication and where applicable, the project in which they were developed. Out of the 19 semantic-based projects initially identified, 12 projects remained (See Table 2 and Table 3); the related published papers and reports provided enough evidence (conceptual part of domain ontology, informal description of domain ontology, and/or sample code of ontology) of ontology development in these projects. The next subsection performs a specific search using the specific ontology concepts discovered in the data sources.

D. Search Specific Ontology Codes

The ontology features discovered within the ontologies data sources in the previous task provided in some cases, specific concepts of the candidate ontologies. At this stage, some of these concepts were used in Swoogle and Watson ontology search engines to attempt to retrieve the full codes of these ontologies. Fig. 2 depicts the concept $lkif - core$ obtained from the data sources on the FEA-RMO ontology along with the OWL files of 4 FEA-RMO modules retrieved with the search in Swoogle; the URLs in Fig. 2 disclose that the ontology modules were developed under the Estrella e-government project. Furthermore, Table 5 shows selected e-government domain ontologies along with the Web links to their full OWL codes, retrieved from the Web with specific keywords search. Ontology selection is done in the next subsection.

E. Select E-government Domain Ontologies

With the list of candidate e-government domain ontologies in Table 2 and Table 3, their data sources including eventual full codes, predefined criteria such as codification language, semantic coverage, modularity and open availability [14][15] are applied to select the best set of ontologies for the e-government domain as in Table 4. The next subsection presents and discusses the complete results of the application of the framework in Section 2 in the e-government domain.

F. Results and Discussions

Table 2 and Table 3 list 62 discovered candidate e-government domain ontologies along with selected data sources on these ontologies as well as the e-government research and projects in which they were developed. This provide any e-government developer interested in reusing these existing domain ontologies with relevant information for analyzing, understanding and reusing these domain ontologies for building new ontologies, even with existing automatic and semi-automatic ontologies reuse solutions [3][4][5][7] that required ontology engineers to guide the process.

Further, Table 2 and Table 3 shows that most of e-government projects employ several domain ontologies for the Semantic Web development of e-government systems. Moreover, one can notice in Table 2 that some candidate ontologies are being repeated in different projects with the same name to serve the same purposes; for instance, the life-event ontology have been developed in 6 projects and the service ontology in 3 projects; this shows a lack of ontology reuse culture in the Semantic Web e-government development community.

Table 4 presents the candidate ontologies that were selected as the best set of ontologies for the e-government domain, based on their codification language, semantic coverage, modularity, and open availability as defined in the Section 2. A brief presentation of these selected e-government domain ontologies obtained from their data sources is provided below.

The selected e-government domain ontologies in Table 4 were developed within real world e-government projects in the United States [10], European countries [12][11][17][13], and Palestine [16]. This indicates that these ontologies have been well thought of, consistently designed and published. In particular:

- The LKIF-core ontology [12] describes the law and regulations that government the public administration domain through basic legal concepts; it is formed of 150 concepts and built with intensive semantic features (hyponymy, supsumption, etc.).
- The government ontology [16] is composed of 15 modules describing public administration entities such as address, bank, local government unit, natural and non-natural person, company, partnership company, shareholder company, driving licence, etc.; these set of ontologies model processes and enable systems interoperability

Fig. 2. Screenshot Showing how the LKIF-core Concept was used to Retrieve the Modules of the LKIF-core ontology from Swoogle

in e-government.

- The FEA-RMO [10] ontology is a set of 5 modules namely performance, business, services, technology and data reference models ontologies; these ontologies were developed to enable the interoperability of the US government's federal agencies; they basically provide common reference models for modelling federal agencies' business processes, thereby, supporting their interoperability.
- The SAKE ontology [11] is formed of 3 modules including: process and profile, information, and decision making quality ontologies; these ontologies were developed as support to an agile knowledge management system for e-government. In particular, the process and profile ontology models the business process and related activities that might involve a public administration user; it is formed of 47 concepts including input, output, date, creation-date, last-modification-date, process-model, and so forth and fully represented in an is-a hierarchy. The information ontology describes metadata such as subject, description, title, creator, publisher, format, location, and the like; overall, it contents 33 concepts describing storable information; these concepts were designed after a meticulous analysis of existing metadata standards and their harmonization. The decision making quality ontology models concepts that might be used as performance evaluation parameters of a process in a public administration organization; these concepts are in total 33 and include: metric, accountability, cost, quality, and many more.
- The GEA ontology [17][18] is a single abstract model that describes the public administration semantic as well as the overall e-government domain; it includes concepts such as governance-entity, political-entity, admin-level, service-provider, public-administration-service, law, outcome, and so forth. It is also used to enable the automatically mapping of citizens' needs to suitable public services.
- The life-event-ontology [13] is a single generic ontology model as well; it models the public administration services with 18 concepts related to life-events (e.g., get married, change address) of citizens with the public administration systems; these concepts include: public-service, input, output, profile, document, citizen, family-status, education-level, job-category, gender, and the like.

In light of the above, the selected e-government domain ontologies in Table 4 are largely formed of several modules that are publicly available; this may promote their reuse and evolution in the Semantic Web e-government development community [16].

Tables 5 provides the Web links to chosen data sources of the selected e-government domain ontologies in Table 4; these Web links are directed to either the ontology codes, deliverable reports or published research articles from projects in which these domain ontologies were developed. It is worth mentioning that in some cases, the ontology codes were not found with a keywords search in Swoogle and Watson search engines; instead, the full codes of some of the domain ontologies discovered were found in deliverable reports of corresponding projects with generic search engines; this shows the effectiveness of the adaptive search strategy presented in this study for locating domain ontologies and their data sources on the Semantic Web.

Furthermore, the deliverable and research reports of projects provided valuable information on the identified ontologies such as: (1) the purpose(s) for which the ontologies were built, (2) the methodologies employed to build the ontologies, (3) the full or partial ontology graphs, (4) theoretical explanations of the meaning of concepts and axioms, (5) full or partial codes of the ontologies, (6) detailed descriptions of the use of these ontologies in real world semantic-based projects, etc.

TABLE II
CANDIDATE E-GOVERNMENT DOMAIN ONTOLOGIES PART I

Code	Ontology	Selected Data Sources	Project
O_1	**DIP ontology**	Gugliotta et al. [19]	DIP
	Legacy ontology		
	Workflow ontology		
	Service ontology		
	Life-event ontology		
	E-government domain ontology		
O_2	**3 kinds of ontologies**	Sabucedo & Rifon [20]	Academic work
	Life-event ontology		
	Variable ontology		
	Legal document ontology		
O_3	E-government Business ontology	Xiao et al. [34]	Academic work
O_4	LKIF-core ontology	Breuker et al. [12]	Estrella
O_5	Social care ontology	Barthes & Moulin [21]	TerreGov
O_6	Life-event ontology	Sanati & Lu [22]	Academic work
O_7	FEA-RMO ontology	Allemang & Hodgson [10]	OSERA
	PRM ontology		
	BRM ontology		
	SRM ontology		
	TRM ontology		
	DRM ontology		
O_8	**Access-eGov ontology**	Hreno et al. [25]	Access-eGov
	Life-event ontology		
	Service profiles ontology		
	Domain ontology		
O_9	Life-event ontology	Todorovski et al. [13]	OneStopGov
O_{10}	Process document ontology	Puustjarvi [26]	Academic work
O_{11}	**SAKE ontology**	Butka et al. [11]	SAKE
	Public Administration ontology		
	Process and Profile ontology		
	Information ontology		
	Decision making quality ontology		
O_{12}	**OntoGov ontology**	Apostolou et al. [23], [24]	OntoGov
	Legal ontology		
	Organizational ontology		
	Life-cycle ontology		
	Domain ontology		
	Service ontology		
	Life-event ontology		
	Profile ontology		
	Web Service Orchestration ontology		
O_{13}	**3 kinds of ontologies**	Chen et al. [27]	Academic work
	E-government ontology		
	Regulatory ontology		
	Service ontology		
O_{14}	E-government services ontology	Fraser et al. [28]	SmartGov
O_{15}	GEA ontology	Goudos et al. [17]	SemanticGov

[10][12][11][13]; this may promote the reuse and evolution of the corresponding domain ontologies.

Finally, Table 6 provides the URLs of Web sites of e-government projects under which the selected ontologies in Table 4 were developed; these Web links may provide the interested reader access to more information on the selected e-government domain ontologies in Table 4. Related studies are discussed in the next section.

IV. RELATED WORK

In [9] the Swoogle ontology search engine is used to search multimedia ontologies on the Semantic Web; the search in Swoogle is based on domain keywords and their combinations; the data sources of the targeted multimedia ontologies are not considered for selecting ontologies specific keywords that are likely to improve the search results.

A strategy for searching biomedical ontologies is presented in [8]; the strategy relies on the keywords search in Swoogle; the keywords used are extracted from related Web pages retrieved with domain keywords search in Google; the data sources on the targeted domain ontologies that may help identifying ontologies specific concepts for the search are not considered.

In [4] an infrastructure for searching and reusing distributed ontologies is presented . The proposed infrastructure is composed of many ontology servers or nodes that store and maintain ontologies; a domain ontology to be searched is described in a meta-ontology with information such as the ontology author, ontology location and used ontology language; the meta-ontology is further improved with a list of ontology terms by matching each ontology concept to the WorldNet lexical semantic net; finally, the meta-ontology

TABLE III
CANDIDATE E-GOVERNMENT DOMAIN ONTOLOGIES PART II

Code	Ontology	Selected Data Sources	Project
O_{16}	**Real-estate transaction ontology**	Ortiz-Rodriguez & Villazon-Terrazas [29]	Reimdoc
	Real-estate ontology		
	Person ontology		
	Organizational ontology		
	Legislation ontology		
	Location ontology		
	Tax ontology		
	Contract model ontology		
	Jurisprudence ontology		
	Civil personality ontology		
	Real-estate transaction verification ontology		
O_{17}	**Government ontology**	Jarrar et al. [16]	Zinnar
	Address ontology		
	Association ontology		
	Bank ontology		
	Company ontology		
	Currency code ontology		
	Driving licence ontology		
	Legal person ontology		
	Local government unit ontology		
	Natural person ontology		
	Non Natural ontology		
	Partnership company ontology		
	Professional association ontology		
	Shareholding company ontology		
	Vehicle ontology		
	Vehicle engine ontology		

TABLE IV
SELECTED E-GOVERNMENT DOMAIN ONTOLOGIES

Code	Ontology	Codification Language	Semantic Coverage	Modularity	Open Availability
O_7	FEA-RMO ontology	OWL	High	5 domain ontologies	publicly available
O_4	LKIF-core ontology	OWL	High	15 domain ontologies	Publicly available
O_9	Life-event ontology	OWL	High	1 generic	Publicly available
O_{11}	SAKE ontology	OWL	High	3 modules	Publicly available
O_{15}	GEA ontology	OWL	High	1 generic model	Publicly available
O_{17}	Government ontology	Not publicly available	High	15 domain ontologies	Publicly available

is stored in an ontology registry, providing a compact representation for efficient search and reuse of related ontologies. However, to build a meta-ontology for searching targeted domain ontologies, the ontology engineer need to have prior knowledge of the targeted ontologies; but, it is unclear in the study how such prior knowledge could be acquired. The available data sources of ontologies in the domain could be of help to the ontology engineer in this case.

The underlying algorithms of ontology and semantic search engines including Swoogle, OntoSearch and OntoKhoj are presented in [30][31][32], respectively. However, the search in these search engines is based on keywords [8]; but, the scope of these studies do not address the issue of selecting relevant domain and specific ontology keywords for the search. This study performs a content analysis of ontology data sources based on predefined ontology features to guess specific concepts for searching domain ontologies on the Semantic Web.

Ontology editors such as Protégé allow the reuse of an existing ontology in another ontology being designed [6]; furthermore, the Web Ontology Language (OWL) offers the possibility to import an OWL ontology into a new ontology under development [33][6]; both ontology reuse solutions require the ontology engineer to have good knowledge and understanding of the existing domain ontologies to be integrated or imported; once more, locating existing domain ontologies and their data sources may be of assistance to the ontology engineer in these cases.

Other solutions for semi-automatic and automatic ontology reuse are presented in [4][5][7]. However, there remains some general challenges in these ontologies reuse solutions: (1) locating relevant domain ontologies for reuse [4], (2) determining appropriate concepts for searching targeted ontologies and (3) understanding the discovered ontologies. This study may be used as a pre-investigative task to existing semi-automatic and automatic ontology reuse solutions in the sense that it enables the ontology engineer to search and retrieve existing domain ontologies along with their data sources; this information may help the ontology engineer in analyzing, understanding and reusing the discovered ontologies. Furthermore, in [2] the authors described the process of reusing

TABLE V
SELECTED E-GOVERNMENT DOMAIN ONTOLOGIES AND WEB LINKS TO THEIR DATA SOURCES

Ontology	Links to Data Sources
Government ontology	http://zinnar.pna.ps/ontologyServer/
	http://www.jarrar.info/publications/JDF11.pdf
LKIF-core ontology	http://www.estrellaproject.org/lkif-core/lkif-core.owl
	http://www.estrellaproject.org/lkif-core/legal-role.owl
	http://www.estrellaproject.org/lkif-core/lkif-rules.owl
	http://www.estrellaproject.org/lkif-core/legal-action.owl
	http://www.estrellaproject.org/doc/D1.4-OWL-Ontology-of-Basic-Legal-Concepts.pdf
FEA-RMO ontology	http://protege.cim3.net/file/work/ontology/FEARMO/
	http://www.osera.gov/owl/2004/11/fea/brm.owl
	http://www.osera.gov/owl/2004/11/fea/prm.owl
	http://www.osera.gov/owl/2004/11/fea/srm.owl
	http://www.osera.gov/owl/2004/11/fea/trm.owl
Life-event ontology	http://islab.uom.gr/onestopgov/index.php?name=UpDownload&req=getit&lid=459
	http://islab.uom.gr/onestopgov/index.php?name=UpDownload&req=getit&lid=460
SAKE ontology	www.sake-project.org/fileadmin/filemounts/sake/DeliverableD6b.pdf
GEA ontology	http://islab.uom.gr/semanticgov/index.php?name=UpDownload&req=getit&lid=454
	http://islab.uom.gr/semanticgov/index.php?name=Web_Links&req=visit&lid=65

TABLE VI
URLS OF PROJECTS WEBSITES OF THE SELECTED E-GOVERNMENT DOMAIN ONTOLOGIES

Code	Ontology	Projects	Websites Links
O_7	FEA-RMO ontology	OSERA	http://osera.modeldriven.org/projects/fearmo.htm
O_4	LKIF-core ontology	ESTRELLA	http://www.estrellaproject.org/
O_9	Life-event ontology	OneStopGov	http://islab.uom.gr/onestopgov/
O_{11}	SAKE ontology	SAKE	http://www.sake-project.org/
O_{15}	GEA ontology	SemanticGov	http://islab.uom.gr/semanticgov/
O_{17}	Government ontology	Zinnar	http://zinnar.pna.ps/

and applying existing ontologies and concluded that reusing ontologies is far from an automatic process and requires significant effort from the knowledge engineer; this assertion is also supported in [3].

V. CONCLUSION

This study presents a framework that uses an adaptive technique based on ontology and generic search engines, and predefined ontology features to search and locate domain ontologies and their data sources over the Semantic Web. The predefined ontologies features are used to learn ontology specific concepts from the data sources; these concepts are further employed to improve the quality of the search results.

The application of the framework in the e-government domain permitted the discovery of 62 candidate e-government domain ontologies; furthermore the framework enabled the application of predefined criteria including semantic coverage, open availability, codification language, and modularity on the candidate ontologies to select the best reusable set of ontologies for the e-government domain. The selected ontologies provide a good sharable and reusable conceptual representation and description of the public administration domain as well as the electronic services delivery processes; this may promote their reuse across semantic-based e-government projects.

The study may be used as a pre-investigative task to existing automatic and semi-automatic ontologies reuse solutions which require the ontology engineers to have prior knowledge of the targeted ontologies to guide the process for building new domain ontologies from existing ones.

The framework of the study may be applied in any application domains of Semantic Web such as e-commerce, e-business, e-learning, multimedia, etc., to identify and analyze existing domain ontologies for the purpose of knowledge sharing and reuse across domain specific Semantic Web applications.

The future direction of the research will be to conceptualize and build a generic ontology model for the e-government domain through the reuse of the discovered domain ontologies.

REFERENCES

[1] T. R. Gruber, "Toward Principles for the Design of Ontologies used for Knowledge Sharing," *International Journal of Human-Computer Studies*, Vol. 43, pp. 907-928, 1993.

[2] M. Ushold, M. Healy, K. Williamson, P. Clark and S. Woods, "Ontology Reuse and Application," *In Proceedings of the 1st International Conference on Formal Ontology and Information Systems - FOIS'98*, Trento, Italy, pp. 179-194, 1998.

[3] Maedche, A. & Staad, S. (2001). Ontology Learning for the Semantic Web. IEEE Intelligent Systems, 16, 72–79.

[4] A. Maedche, B. Motik, L. Stojanovic, R. Studer and R. Volz, "An Infrastructure for Searching, Reusing and Evolving Distributed Ontologies," *In Proceedings of the World Wide Web Conference (WWW 2003)*, Budapest, Hungary, pp. 439-448, 2003.

[5] Y. Ding, D. Lonsdale, D. W. Embley, M. Hepp and L. Xu, "Generating Ontologies via Language Components and Ontology Reuse," *In Proceedings of the 12th International Conference on Applications on Natural Language to Information Systems (NLDB'07)*, Paris, France, pp. 131-142, 2007.

[6] P. Doran, V. Tamma and L. Lannone, "Ontology Module Extraction for Ontology Reuse: An Ontology Engineering Perspective," *In: International Conference on Information and Knowledge Management (CIKM'07)*, Lisboa, Portugal, pp. 61-69, 2007.

[7] M. d'Aquin, M. Sabou and E. Motta, "Reusing Knowledge from the Semantic Web with the Watson Plugin," *In Proceedings of the 7th International Semantic Web Conference (ISWC)*, Karlsruhe, Germany, 2008.

[8] H. Alani, N. Noy, N. Shah, N. Shadbolt and M. Musen, "Searching Ontologies Based on Content: Experiments in the Biomedical Domain," *In Proceeding of the Fourth International Conference on Knowledge Capture (K-Cap)*, Whistler, BC, Canada, pp.55-62, 2007.

[9] G. A. Atemezing, "Analyzing and Ranking Multimedia Ontologies for their reuse," *MSc Dissertation*, Unversidad Politechnica de Madrid, Madrid, Spain, 2010.

[10] D. Allemang, R. Hodgson and I. Polikoff, "Federal Reference Model Ontologies (FEA-RMO)," *White Paper*, 2005.

[11] P. Butka, A. Gabor, A. Ko, M. Mach, S. Ntioudis, A. Papadakis, N. Stojanovic, R. Vas and T. Zelinsky, "Semantic-enable, Agile, Knowledge-based e-Government (SAKE)," *Deliverable No. 3*, 2006.

[12] J. Breuker, R. Hoekstra, A. Boer, K. Van der berg, G. Sartor, R. Rubino, A. Wyner, T. Bench-Capon and M. Palmirani, "OWL Ontology of Basis Legal Concepts (LKIF-Core)," *Deliverable 1.4*, 2006.

[13] L. Torodovski, M. Kunstelj, D. Cukjati, M. Vintar, I. Trochidis, E. Tambouris, OneStopGove: D13 Life-event Reference Models, Deliverable No. 13, 2007.

[14] R. Fitterer, U. Greiner and F. Stroh, "Towards Facilitated Reuse of Ontology Results from European Research Projects - A Case Study," *In: 16th European Conference on Information Systems (ECIS)*, Galway, Ireland, pp. 1929-1940, 2008.

[15] A. Esposito, M. Zappatore and L. Terricone, "Evaluating Scientific Domain Ontologies for the Electronic Knowledge Domain: A General Methodology," *International Journal of Web & Semantic Technology*, Vol. 2, 1-18, 2001.

[16] M. Jarrar, A. Deik and B. Farraj, "Ontology-Based Data Process Governance Framework The Case of e-Government Interoperability in Palestine," *In IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA'11)*, Campione, Italy, pp. 83-98, 2011.

[17] S. K. Goudos, V. Peristeras and K. Tarabanis, "Mapping Citizen Profiles to Public Administration Services Using Ontology Implementations of the Governance Enterprise Architecture (GEA) models,"*In Proceedings of the 3rd Annual European Semantic Web Conference*, Budva, Montenegro, pp. 25-37, 2006.

[18] S. K. Goudos, V. Peristeras, N. Lutas and K. Tarabanis, "A Public Administration Domain Ontology for Semantic Discovery of e-Government Services," *In Proceedings of the 2nd IEEE Conference on Digital Information Management 2007 (ICDIM 2007)*, Lyon, France, pp. 260-265, 2007.

[19] A. Gugliotta, L. Cabral, J. Domingue and V. Roberto, "A Conceptual Model for Semantically-Based E-government Portal," *In Proceedings of the International Conference on e-Government 2005 (ICEG 2005)*, Ottawa, Canada, 2005.

[20] L. A. Sabucedo and L. A. Rifon, "Semantic Service Oriented Architectures for E-government Platforms," *American Association for Artificial Intelligence*, 2006.

[21] J. P. Barthes and C. Moulin, "Impact of e-Government on Territorial Government Services," *Deliverable No. 1.4*, 2005.

[22] F. Sanati, J. Lu, "Multilevel Life-event Abstraction Framework for E-government Service Integration," *In Proceedings of the 9th European Conference on E-government 2009 (ECEG 2009)*, London, UK, pp. 550-558, 2006.

[23] D. Apostolou, L. Stojanovic, T. P. Lobo, J. C. Miro,and A. Papadakis, "Configuring E-government Services Using Ontologies," *IFIP International Federation for Information Processing*, Springer Boston, Vol. 2005, pp. 141-155, 2005.

[24] D. Apostolou, L. Stojanovic, T. P. Lobo and B. Thoensen, "Towards a Semantically-Driven Software ngineering Envirionment for E-government," *IFIP International Federation for Information Processing*, M. Bohlen (Eds), Vol. 3416, pp. 157-168, 2005.

[25] J. Hreno, P. Bednar, K. Furdk and T. Sabol, "Integration of Government Services using Semantic Technologies," *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 6, pp. 143-154, 2011.

[26] J. Puustjarvi, "Using Knowledge Management and Business Process in E-government," *In Proceedings of the Information Integration and Web-based Applications and Services 2006 (iiWas2006) Conference*, Yogyakarta, Indonesia, pp. 331-339.

[27] D. Chen, G. Nie and P. Liu, "Research Knowledge Sharing of E-government Based on Automatic Ontology Mapping," *In Proceedings of the 6th Wuhan International Conference on E-Business*, Business, China, pp.105-111, 2008.

[28] J. Fraser, N. Adams, A. Mckay-Hubbard, A. Macintosh and R. Canadas, "A Framework for e-Government Services," *Deliverable No. 71*, 2003.

[29] Ortiz-Rodriguez, F. & Villazon-Terrazas, B. (2006). EGO Ontology Model: Law and Regulation Approach for E-government. In Proceedings of the Workshop on Semantic Web for E-government 2006, Workshop at the 3rd European Semantic Web Conference (pp. 13-23). Budva, Serbia and Montenegro.

[30] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi and J. Sachs, "Swoogle: A Search and Metadata Engine for the Semantic Web," *In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, Washington, DC, USA, pp. 1-8, 2004.

[31] Y. Zhang, W. Vasconcelos, D. Sleeman, "Ontosearch: An ontology search engine," *In Procceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Cambridge, UK, pp. 1-12, 2004.

[32] C. Petel, K. Supekar, Y. Lee and E. K. Park, "OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking and Classification," *In Proceedings of the Workshop On Web Information And Data Management*, New Orleans, Lousiana, USA, pp. 58-61, 2003.

[33] J. Z. Pan, L. Serafini, Y. Zhao,"Semantic Import: An Approach for Partial Ontology Reuse," *In Proceedings of the 1st Workshop on Modular Ontologies (WoMO06)*, Athens, GA, USA, pp. 1-12.

[34] Y. Xiao, M. Xioa and H. Zhao, "An Ontology for E-government Knowledge Modelling and Interoperability," *In Proceedings of the IEEE International Conference on Wireless Communications, Networking and Mobile Computing, (WiCOM 2007)*, Shanghai, pp. 3600-3603, 2007.

Web Usage Mining: An Analysis

Mehak
ME (Computer Science & Engineering), University Institute of Engineering & Technology, Panjab University, Chandigarh, India
Mehakjain_1988@yahoo.co.in

Mukesh Kumar
Assistant Professor, Computer Science & Engineering Department, University Institute of Engineering &Technology, Panjab University, Chandigarh, India
Mukesh_rai9@yahoo.com

Naveen Aggarwal
Assistant Professor, Computer Science & Engineering Department, University Institute of Engineering &Technology, Panjab University, Chandigarh, India
navagg@gmail.com

***Abstract*—Web usage mining is research area in web mining. Web mining is an activity that focuses to discover new, relevant and reliable information and knowledge by examining the structure, content and usage of web. The major focus is on learning about web users and their interaction with websites. Web log files generated on web servers are used in order to extract web usage of different users. There are three types of web repositories: web server log, proxy server log, browser log. Analysing web logs for usage can not only provide important information to websites developers but also help in creating adaptive web sites.**

In this paper we discuss various sources of information for WUM, Methodology of web usage mining techniques which involves Data collection , Data pre-processing , knowledge discovery and knowledge analysis. Various applications of WUM are personalization, prefetching and caching, support to design and E-commerce. Major application of web usage mining is to predict future accesses. Thus, the result obtained after web usage mining can be used to improve the performance of prefetching and caching.

***Index terms*—Web usage mining, Methodology, pre-processing, clustering, classification, applications.**

I. INTRODUCTION:

World Wide Web is a huge repository of data. It has become one of the most important repository for storing, sharing and to distribute information. The expansion of web is very rapid which has provided a great opportunity to study user and system behaviour by exploring web access [5].

Data mining is the process that attempts to discover

Patterns in large data. Applying data mining techniques on web data to discover knowledge has been defined as WEB MINING [3]. It can be viewed as an extraction of structure from an unlabeled, semi structured dataset containing the characteristics of users or information respectively.

Data that is actually mined is varied and different approaches have been followed. Some researchers have applied mining techniques on the web logs maintained by the servers so as to discover user access and traversal path [3].

Web mining is categorized in three types:

A. *Web content mining:* It is the scanning and mining of text, pictures of a Web page to determine the relevance of the content to the search query. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

B. *Web structure mining:* Web structure mining is a tool used to identify the relationship between Web pages linked by information or direct link connection. The motive of web structure mining is generating structured summaries about information on web pages/webs.

C. *Web usage mining*: This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web page. Web server gathers this information automatically into the Access Log File.

Typical Sources of Data [1]:

1. Data generated automatically is stored in different types of log files such as server access logs, referrer logs, and client-side cookies.
2. E-commerce and product-oriented user events.
3. User profiles and user ratings
4. Meta-data, page attribute, page content, site structure.

doi:10.4304/jetwi.5.3.240-246

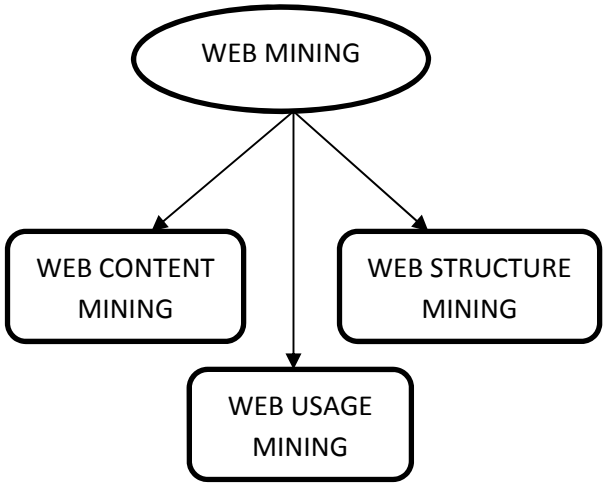


Figure 1: Types of Web Mining.

New tools promising to apply data warehousing and mining techniques on web logs have entered in the market. These include surfAid, speedTracer from IBM, bazaar analyser etc [3].

II. Web Usage Mining

Web usage mining is used to analyse web log files to discover user accessing patterns of web pages [13]. Web usage mining is a main research area in Web mining focused on learning about Web users and their interactions with Web sites. Web usage mining is the discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers. A *web log* is a listing of page reference data. A web server log file contains requests made to the web server recorded in chronological order. It is at times referred to as *clickstream* data as each entry corresponds to a mouse click [7].

Information Obtained through web usage mining [15]:

A. Number of Visitors: It is the count of users who navigates to your website and browses one or more pages on your site.

B. Visitor Referring Website: The referring website gives the information or URL of the website which referred the particular website in consideration.

*C. Visitor Referral Website***:** The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

D. *Number of Hits***:** This number usually signifies the number of times any resource is accessed in a Website.

*E. Time and Duration***:** This information in the server logs give the time and duration for how long the Website was accessed by a particular user.

*F. Path Analysis***:** Path analysis gives the analysis of the path a particular user has followed in accessing contents of a Website.

*G. Visitor IP address***:** This information gives the Internet Protocol (I.P.) address. It is the address of the visitors who visited the website.

*H. Browser Type***:** This information provides the data of the kind of browser that was used for accessing the web site.

*I. Cookies***:** A message given to an online browser by an online server. The browser stores the message during a document known as cookie. The message is then sent back to the server whenever the browser requests a page from the server. The purpose of cookies is to spot users and probably prepare tailor-made sites for them.

*J. Platform***:** This info provides the kind of OS etc. that was accustomed access the web site.

III. Methodology of Web Usage Mining.

A web server log file contains requests made to the web server. These requests are recorded in chronological order. The popular log file formats are the Common Log Format (CLF) and extended CLF.

As shown in Figure 3[1].

Web Usage Mining includes following steps: Data Collection, Data Pre-processing, Knowledge Discovery and Pattern Analysis. As shown in Figure 4[8] and Figure 2[1].

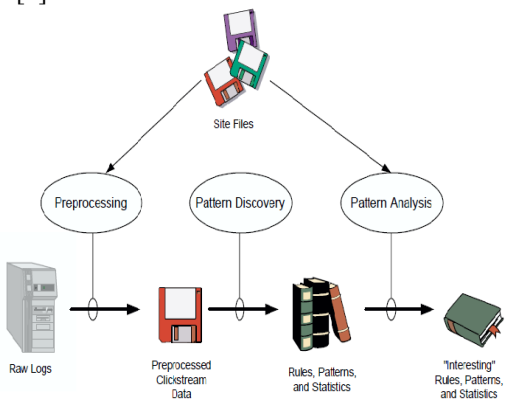


Figure 2: Basic Steps of Web Usage Mining [1]**.**

A. *Data Collection***:**

Web Usage Mining applications are based on data collected from three mainsources [13]: (i) web servers, (ii) proxy servers, and (iii) web clients [2].

i. ***Server Side:*** Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of

the request, the request line exactly as it came from the client, etc.

ii. ***Proxy Side:*** A Web proxy acts as an intermediate level of caching between client browsers and Web servers. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of group of users accessing huge groups of web servers.

iii. ***Client Side***: Most of the users have tendency to open several pages simultaneously and in between, use some non-browsing applications such as MS-word, Excel etc. for their own personal work, in such cases data recorded in server log only shows the requested time of the web pages and cannot help us to find out which web page and for how long has been really browsed on client machine. Usage data can be tracked on the client side by using JavaScript, java applets, or even modified browsers. These techniques avoid the problems of users sessions identification and the problems caused by caching (like the use of the back button). However, these approaches rely heavily on the users' cooperation and rise many issues concerning the privacy laws, which are quite strict.

*B. Data Pre-processing***:**

Some databases are insufficient, inconsistent and include noise. The pre-treatment of data is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine.

Steps involved in data pre-processing are shown with the help of block diagram below Fig 5.

i. Data Cleaning [5]:

Data cleaning is the process where irrelevant records are removed. The main aim of web usage mining is to fetch the traversal pattern; following two kinds of record are unnecessary and should be removed [5].

a. The records having filenames suffixes of GIF, JPEG, CSS and so on, which can be found incs_uri_stem field of record.

b. By examining the status field of every record in the web log, the record with status code over 299 and below 200 are removed.

ii. User and Session Identification [5]:

The main task in this step is to identify different user session from access log. A referrer-based method is used for identifying sessions. The different IP addresses distinguish different users.

<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>

Figure 3: Common Log Format [1]**.**

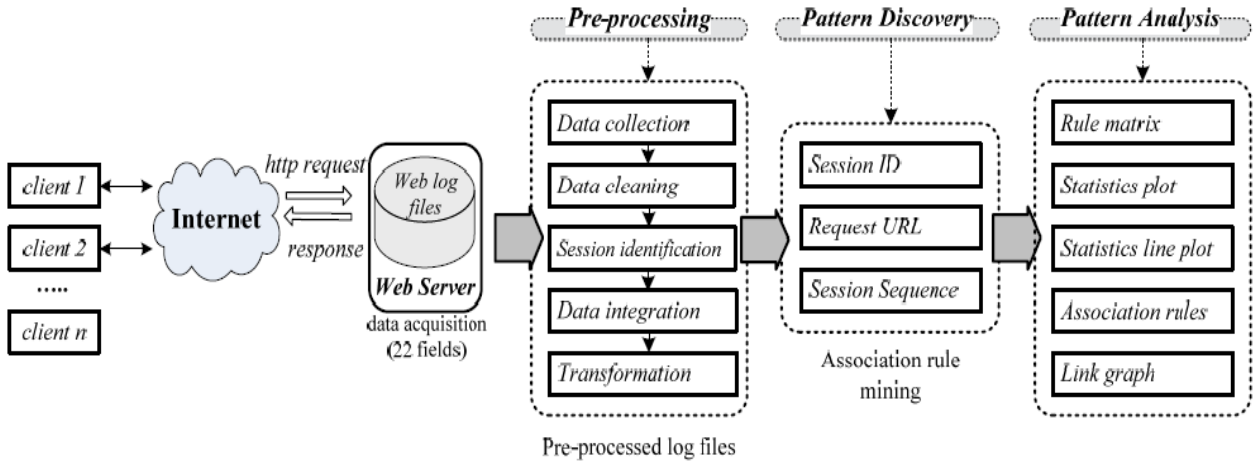


Figure 4: Algorithm Scheme for Web Usage Mining [8].

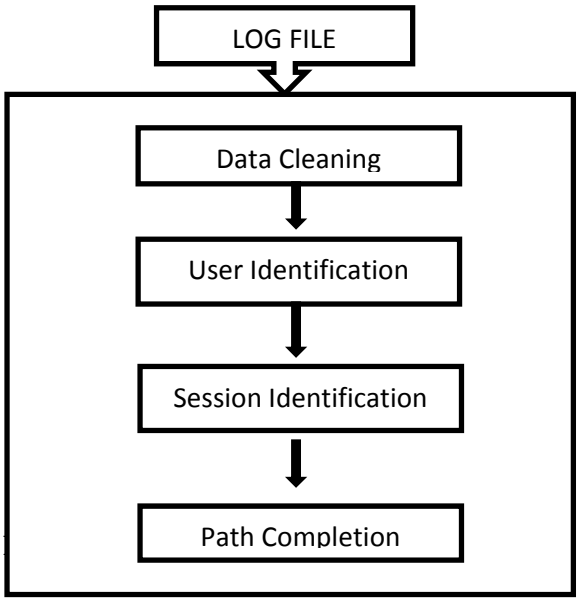


Figure 5: Steps for Data Pre-processing

a. If the IP addresses are same, then information regarding different browsers and operating systems given by client IP address and user agent indicate different users.

b. If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The ReferURL (cs_referer) is checked, a new user session is identified if the URL in the ReferURL– field hasn't been accessed previously, or there is a large interval between the accessing time of this record.

iii. Path Completion:

Client- or proxy-side caching can often result in missing access references to those pages or objects that have been cached. For Example, if Page A is returned by the user during the same session, the second time when page A will be accessed again, no request is made to the server and it will result in viewing the previously downloaded version of A that was cached on the client-side. This results in the second reference to A not being recorded on the server logs. Missing references due to caching can be heuristically inferred through path completion which relies on the knowledge of site structure and referrer information from server logs [13].

Path Completion should be used acquiring the complete user access path. The incomplete access path of every user session is recognized based on user session identification. If in a start of user session, Referrer as well URL has data value, delete value of Referrer by adding '-'. Web log pre-processing helps in removal of unwanted click-streams from the log file and also reduces the size of original file by 40-50% [5].

Tools used for Pre- processing [6]:

Active Server Pages (ASP) is one of the popular scripting languages used for developing web-based application. This study focuses on this language in order to develop the application that can manipulate the server logs. To access the server logs from windows 2000, the *.dll file named *logscrpt.dll* is used to load the class object MSWC.IISLog. The MSWC.IISLog class contains several *methods* and *properties* that can be used either to retrieve log entries or write log entries [6].

In order to perform pattern mining and generalized association rules, a tool was written using Active Server Pages (ASP) to perform pre-processing techniques.

The algorithm for pre-processing is shown below [6] Figure 6:

Several attributes are ignored and the interesting fields are included in the database [6]. The algorithm that implements this function is written as [6] Figure 7:

```
Const ForReading = 1
Const ForWriting = 2
Sub ReadLog( Physical-Path, ModeFile-1,
TypeOfLogFile, ModeFile 2,StrTypeOfLogFormat)
RecordCounter = 0
Set LogReader = Server.CreateObject("IISLog")
LogReader.OpenLogFile
        LogFilePath, ModeFile-1, TypeOfLogFile,
        ModeFile-2, StrTypeOfLogFormat
LogReader.ReadLogRecord
While NOT LogReader.EndOfLogRecord
Retrieve Log Attributes
RecordCounter = RecordCounter + 1
LogReader.ReadLogRecord
Loop
LogReader.CloseLogFile
End Sub
```

Figure 6: Pre-processing Algorithm.

Approaches used for data pre-processing:

i. Pre-Processing Using Xml [5]

XML (Extended Mark-up Language) provides a structure to the records which are present in web logs. Data Pre-processing can be done using XML. Hence, understanding of web logs becomes easier. Steps involved in pre-processing using above approach are:

a. Using XML parsers DOM tree structure is created from Logs recorded in the web log.
b. Next step is user identification and session identification is same as given basic algorithm of data pre-processing.
c. Finally, the path completion helps to complete and format the paths in user session, so that these paths can be further used for analysis.
d. After the above steps, transfer the records which are present in XML file into Knowledge base.

Transfer server logs to database:

```
Declare Variables
Set DB
=Server.CreateObject("ADODB.Connection
")
Set RS
=Server.CreateObject("ADODB.Recordset"
)
ConnStr = {MsAccess Driver}
DB.OpenConnStr
RS.OpenTableName, ActiveConnection,
Add Data
RS.Update
Set Rs = Nothing
DB.Close
```

Figure7:Algorithm for transfer of log file.

ii. Pre-Processing Using Text File [5]

Data pre-processing is applied on records which are present in the web log file. Steps for pre-processing are:

a) Web log file contains log records in unprocessed form.
b) Before applying cleansing process, attributes in the text file needs to be separated using delimiter as space. These spaces help in identifying exact position of attributes/fields.
c) Steps 3 & 4 are same as in above approach.
d) After the above steps, transfer the records which are present in text file into Knowledge base.

C. Knowledge Discovery :

This is the key component of the Web usage mining. Various techniques are used to discover rules or patterns such as Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns etc.

i. Statistical Analysis :

Knowledge about visitors to a Web site is extracted with the use of Statistical techniques. Different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path can be performed by analyzing the session file. The web system report can be potentially useful for improving the system performance, enhancing the security of the System, facilitation the site modification task, and providing support for marketing decisions simply by analysing the statistical information in the report [13].

ii. Association Rules:

Association rule generation can be used to relate pages that are most often referenced together in a single server session [13].

In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. Figure 8 shows the item set generation for a set of transactions.

Most common approaches to association discovery are based on the Apriori algorithm.

This algorithm finds groups of items (page-views appearing in the pre-processed log) occurring frequently together in many transactions (i.e., satisfying a user specified minimum support threshold). Such groups of items are referred to as **frequent item sets**. Association rules which satisfy a minimum confidence threshold are then generated from the frequent item sets.

The support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. When loading a page from a remote site, association rules can be used as a trigger for prefetching documents to reduce user perceived latency.

iii. Clustering:

Clustering is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. [13].

Clustering of pages (or items) can be performed based on the usage data (i.e., starting from the user sessions or transaction data), or based on the content features associated with pages or items (keywords or product attributes).

In the case of ***content-based clustering***, the result may be collections of pages or products related to the same topic or category. In ***usage-based clustering***, items that are commonly accessed or purchased together can be automatically organized into groups.

Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users.

iv. Classification:

Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of given the class or category. Classification can be done by using supervised learning algorithms such as decision trees, naive Bayesian classifiers, *k*-nearest neighbour classifiers, and Support Vector Machines.

Classification techniques play an important role in Web analytics applications for modelling the users according to various predefined metrics.

For example, given a set of user transactions, the sum of purchases made by each user within a specified period of time can be computed. A classification model can then be built based on this enriched data in order to classify users into those

Transactions
A, B, D, E
A, B, E, C, D
A, B, E, C
B, E, B, A, C
D, A, B, E, C

Size 1		Size 2		Size 3		Size 4	
Itemset	Supp.	Itemset	Supp.	Itemset	Supp.	Itemset	Supp.
A	5	A,B	5	A,B,C	4	A,B,C,E	4
B	5	A,C	4	A,B,E	5		
C	4	A,E	5	A,C,E	4		
E	5	B,C	4	B,C,E	4		
		B,E	5				
		C,E	4				

Figure 8: Web Transaction and resulting Itemsets (minsup = 4) [16]

Who have a high propensity to buy and those who do not, taking into account features such as users' demographic attributes, as well their navigational activities.

v. *Sequential Patterns [2]***:**

Sequential Patterns are used to discover frequent sub sequences among large amount of sequential data. In web usage mining, sequential patterns are exploited to find sequential navigation patterns that appear in users' sessions frequently.

The typical sequential pattern has the form [14]: the 70% of users who first visited A.html and then visited B.html afterwards, in the same session, have also accessed page C.html. Sequential patterns might appear syntactically similar to association rules; in fact algorithms to extract association rules can also be used for sequential pattern mining.

[7] Presents a comparison of different sequential pattern algorithms applied to WUM.

D. Pattern Analysis:

The need behind pattern analysis is to filter out uninteresting rules or patterns from the set. Common form of pattern analysis consists of a knowledge query mechanism such as SQL [17]. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match certain hyperlink structure.

The common techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying, and Usability Analysis. Visualization techniques are useful to help application domains expert analyse the discovered patterns.

IV. APPLICATIONS OF WEB USAGE MINING[2] :

The general goal of Web Usage Mining is to gather interesting information about user's navigation patterns. This information can be used later to improve the web site from the users' viewpoint. The results produced by the mining of web logs can used for various purposes [13]:

A. To personalize the delivery of web content;
B. To improve user navigation through prefetching and caching;
C. To improve web design; or in e-commerce sites
D. To improve the customer satisfaction.

A. Personalization of Web Content **:**

Web Usage Mining techniques can be used to provide personalized web user experience. For instance, in real time, it is possible to predict the user behaviour by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users.

B. Prefetching and Caching:

The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. With the use of weblogs that store user's access history can be used predict future accesses.

Typically, Web Usage Mining can be used to develop proper prefetching and caching strategies so as to reduce the server response time.

C. *Support to the Design* :

Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications. [12]. Adaptive Web sites represent a further step. In this case, the content and the structure of the web site can be dynamically reorganized according to the data mined from the users' behaviour.

D. *E-commerce* :

Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies. Web usage mining techniques can also be useful in Customer Relationship Management (CRM). The issues specific to business such as customer attraction, customer retention, cross sales, and customer departure are mainly in focus.

CONCLUSION:

Web Usage mining is a technique used to mine the logs available on the server. The various advantages of Web Usage Mining is to improve the structure of web page, improving the system performance and also prefetching and caching. Prefetching and pre-caching is the techniques to reduce the user's perceived latency while accessing a web page through web server. The user's access history can be used to predict its future accesses.

REFERENCES

[1] Rajni Pamnani and PramilaChawan, "Web Usage Mining: A Research Area in Web Mining", Department of computer technology, VJTI University, Mumbai, 2013.

[2] Federico Michele Facca and Pier Luca Lanzi, "Recent Developments in Web Usage Mining Research", In the proceedings of 5th international conference on Data Warehousing and Knowledge Discover, Prague, 2003.

[3] Karuna P. Joshi, Anupam Joshi, Yelena Yesha, and Raghu Krishnapuram, " Ware housing and Mining Web logs" , ACM, Pp. 63-68, 1999.

[4] Paulo Batista and Mario J. Silva, "Mining Web Access Logs of an On-line Newspaper", In the proceedings of 12th International Meeting of the Euro Working Group on Decision Support Systems, 2002.

[5] Ms.Dipa Dixit and Ms. M Kiruthika, "Preprocessing of web logs" International Journal on Computer Science and Engineering (IJCSE) Volume 02, Issue no. 07, Pp. 2447-2452, ISSN 0975-3397, 2010.

[6] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi and Mohamad Farhan Mohamad Mohsin, " Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", In proceedings of World Academy of Science, Engineering and Technology, Volume 48, December 2008.

[7] Behzad Mortazavi-Asl, "Discovering and mining user web-page traversal patterns" Master's thesis, Simon Fraser University, 2001.

[8] ResulDaş, İbrahim Türkoğlu, "Extraction of Interesting Patterns through Association Rule Mining For Improvement of Website Usability", Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 2006.

[9] John R. Punin , Mukkai S. Krishnamoorthy , Mohammed Javeed Zaki, LOGML: Log Markup Language for Web Usage Mining, Revised Papers from the Third International Workshop on Mining Web Log Data Across All Customers Touch Points, p.88-112, August 26, 2001.

[10] Configuration File ofW3C httpd, 1995. http://www.w3.org/Daemon/User/Config/.

[11] W3C Extended Log File Format, 1996. http://www.w3.org/TR/WD-logfile.html.

[12] Dr. G. K. Gupta, "Introduction to Data Mining with Case Studies", PHI Publication, Third edition, ISBN – 9788120330535, 2009.

[13] Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD, Volume 1, Issue 2, January 2000.

[14] Eleni Stroulia Nan Niu and Mohammad El-Ramly. Understanding web usage for dynamic web-site adaptation: A case study. In Proceedings of the Fourth International Workshop on Web Site Evolution (WSE'02) IEEE, Pp. 53–64, 2002.

[15] Aniket Dash and Liju Robin George, "Web Usage Mining: An Implementation", National Institute of Technology, Rourkela, Master's report, 2010.

[16] Haizheng Zhang, Myra Spilipoulou, Bamshad Mobasher, C. lee Giles, Andrew McCallum, Olfa Nasraoui, Jaideep Srivastava, John Yen: "Advances in Web Mining and Web Usage Analysis" 9th International Workshop on Knowledge Discovery on the Web, WebKDD 2007, and 1st International Workshop on Social Networks Analysis, SNA-KDD 2007, San Jose, CA, USA, August 12-15, 2007. Revised Papers.Lecture Notes in Computer Science 5439, Springer 2009, ISBN 978-3-642-00527-5

[17] Sathya Babu Korra, Saroj Kumar Panigrahy, and Sanjay Kumar Jena , "Web Usage Mining: An Implementation view", In the proceedings of Advances in Computing, Communication and Control International Conference, ICAC3 2011, Mumbai, India, volume 125, Pp. 131-136, 2011

Prevention of Losing User Account by Enhancing Security Module: A Facebook Case

M. Milton Joe
Assistant Professor, Department of Computer Application,
St. Jerome's College, Nagercoil, Tamilnadu, India.
m.miltonjoe@gmail.com

Dr. B. Ramakrishnan,
Associate Professor, Department of Computer Science and Research Centre,
S.T. Hindu College, Nagercoil, Tamilnadu, India.
ramsthc@gmail.com

Dr. R.S. Shaji
Professor, Department of IT, Noorul Islam University, Nagercoil, Tamilnadu, India.
shajiswaram@yahoo.com

***Abstract*— The blooming development of internet technologies, lead to the growth of Online Social Networks (OSNs) day by day. There are many Online Social Networks (OSNs) came on internet but still very few could get the attraction of the users forever. The most attracted and world level leading Online Social Networks (OSNs) is Facebook, Twitter, and Google Plus and so on. All these Online Social Networks (OSNs) allow its users to create profiles and share any information on their profile, which could be viewed by other users of the same network user. These Social Networks allow users to form friendship with other people and make them to get connected in an easiest way. The major advantages of these Online Social Networks (OSNs) are getting connected with friends (wherever they are in the world), thoughts and ideas can be shared on online and feedback could be obtained as soon as possible. Every user of Social Networking sites will have many more confidential and private data on their account. However the Security and Quality of Service (QoS) are the major constraints must be always maintained in all the social networks. Most of the Online Social Networks (OSNs) provides security to the data available on the user account and provides good Quality of Service (Qos). Obviously the security constraint must be maintained not only the data available on user account but also must be maintained to the user account completely, which ultimately improves the efficiency of Quality of Service (Qos). In this paper, we evaluate and propose a new model to enhance the security for improving quality of service in online social networks.**

***Index Terms*— Online Social Networks (ONSs), Security, Quality of Service (QoS), Authentication, Random Number.**

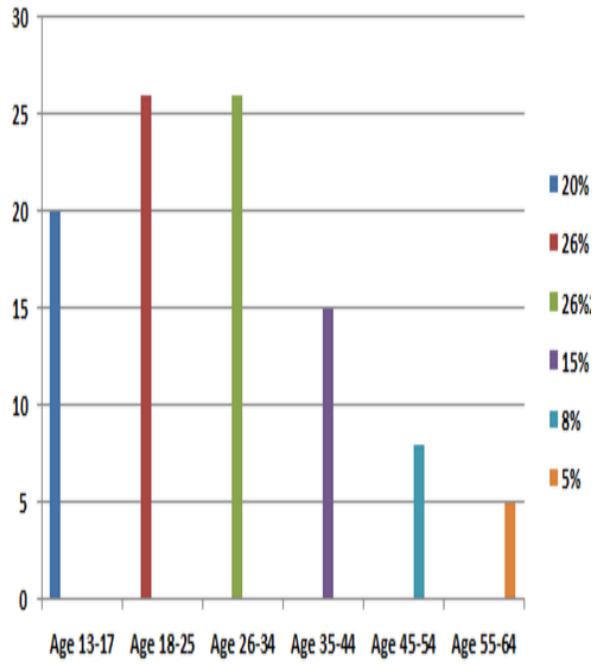
I. INTRODUCTION

The most common and easiest way to connect with friends, relatives and with other people is internet. The development of internet gradually made people to get connected and share ideas with one another in the form of establishing a network communication. The web 2.0 technology developments made this form of communication in the name of Online Social Networks (OSNs) [1], which made people to create profiles and share information available on their pro file to other users. There are many Online Social Networks (OSNs) do exist such as Facebook, Twitter, MySpace, and Google Plus but very few could stand among internet users and became popular. The active users of these Online Social Networks (OSNs) especially Facebook, and twitter almost crossed more than one billion [2] [3]. Among all the social networking websites, Facebook attracted many users in and around the world. Facebook needs a registration to open an account; the registration could be done with an E-mail id. Facebook was founded February 2004 and operated by Facebook community [4]. Facebook crossed over one billion active users during the month of September 2012, and more than half of whom use on mobile devices [4]. Facebook was founded by Mark Zuckerberg with his college roommates and fellow Harvard University students Eduardo Saverin, Andrew McCollum, Dustin Moskovitz and Chris Hughes [4]. Users of Facebook can create personal profile and add others as their friends by giving friend request and the other user must accept the friend request to become friend with them. The user of Facebook can share information and send message to the other user's message box or simple post the information on the wall of the user. The active user of Facebook can chat with the other active user but both of them must be friends on

doi:10.4304/jetwi.5.3.247-256

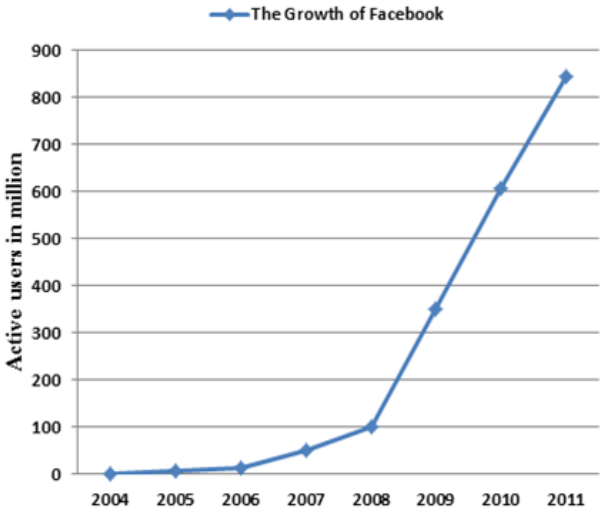
Facebook. Even though the user goes offline simply offline message can be sent to the user.

A January 2009 Compete.com study ranked Facebook as the most used social networking service by worldwide monthly active users[4]. Every day many new users are opening account on Facebook to get connected with their friends. The growth of Facebook also eventually increased day by day. The major reason for the growth of Facebook is, because it is very easy to use and it allows its user to post pictures and videos on their profile in a simplified way, which attracts most of the users. Users can comment on a picture as well as on any post made by the user in simplified manner, which could be viewed and notified to the others users immediately and reply back to the comment at once. According to a May 2011 *Consumer Reports* survey, there are 7.5 million children under 13 with accounts and 5 million under 10, violating the site's terms of service [4]. The Facebook site was mostly covered with the children and youngster of the all-around world.



Graph 1 Facebook users by age in percentage

The above graph 1 indicates the users of Facebook by age, the graph shows most of the users of Facebook come under the age 18 to 34. Facebook allows its users to set their own privacy policy, which enables the users to set whom should view the specific information of their profile and the information shared by them. Facebook has various useful privacy settings to prevent the private information of a particular user account. For instance, in Facebook we can set who could view our friends, whereas in all the other social networking websites the friend list is public. Also Facebook allows its users to go offline from chatting for a particular user, on other hand the user will be shown as available to other users except the particular user for whom it was set as offline from chatting. Such a fabulous facility made Facebook popular in a short period of time.



Graph 2 The growth of Facebook

Graph 2 shows the growth and popularity of Facebook by year wise from the year 2004 to 2011 [4]. As the graph represents Facebook achieved a big victory within a short period of time. As stated above even though Facebook became very popular among internet users, still it has certain limitations. Facebook provides full privacy policy to the information available on the user account alone. All the data available on a user account on Facebook is highly securable but it fails to provide the same security to the entire user account. However this limitation should be avoided to increase the usability and provide good Quality of Service (QoS) to all its users. In this paper, we provide an alternative method instead of the present method in order to provide security on user account and improve the Quality of Service (QoS).

II. Related Work

The users, who are aware of security, are able to set privacy policy to their profile object [5]. That is, users can divide their total number of friends into various groups such as schoolmates, college mates, family members and user can set privacy options to each group depending upon the priority. This type of different privacy setting to each group restricts one group members from viewing the personal photo of the user, while other group member can view the photo [5]. For grouping of friends effectively previous research was carried out based on clustering technology, which aid users in grouping their friends more efficiently [5]. When the growth of online social networking came into effect the third party applications started to access the information from the user profile of the online social networking websites. The third party applications also posted some sort of information about their applications on user profile. So ultimately the access control on user profile by the third party application must be securable [1].The earlier research had focused on mainly user-

to-user interactions in social networks, and seems to ignore the third party applications [1]. To control the third party application an access control framework is presented [1]. The framework is based on enabling the user to specify the data attributes to be shared with the application and at the same time be able to specify the degree of specificity of the shared attributes [1]. This mechanism controlled the third party application from preventing the private information of the user profile. Other related work has analyzed both privacy risks associated with information disclosure in social networks, and developed initial mechanisms to protect against some involuntary information disclosure and proposed a framework for deriving a "privacy score" to inform the user of the potential risks to their privacy created by their activities and activities with other users within the social network [7].However the previous research areas concentrated on friends grouping policy to enable different privacy setting to provide, which data should be accessed by which group of users and similarly work carried out to prevent the third party application from accessing the user data from the profile. A new framework was developed to give more security to the data available on user profile. All these privacy and security mechanisms provide protection only to the profile data of the user not to the entire user account. However every user account must be secured in all the ways from fraudulent access. We propose a new method to prevent the user account from being lost access by the owner of the corresponding user.

III. Problem Statement

The most popular Online Social Networks (ONSs) Facebook allow internet users to create personal profile and post photos, videos, share information and send message to friends. However this famous Online Social Network website has a security mechanism, which most of the times makes the corresponding author or the owner of the account to lose the access to the account. Let us consider the following scenario:

(1) Usually a user uses his/her mobile phone to login to Facebook. We know already more than half of Facebook users use mobile device to access Facebook [4]. However the user access Facebook on his/her mobile device for long days regularly and one day wishes to change the mobile device, which has more facility to get the access easier. When the device is changed and user tries to login in with the new device the Facebook application will prompt as follows:

(2) On the other hand if a user does not access his/her account for few months and tries to access his/her account with a new device; or unauthorized access is found; Facebook will prompt as follows:

- Your account is temporarily locked.
- Someone recently tried to log into your account from unrecognized device or location. Please verify if it was you who tried to log in.

The same case is also applicable when the user changes the laptop or desktop too. This type of security mechanism has major drawback, which is evaluated below.

A. Limitations of Existing Model

Once the account of a user is locked, it prompts the user to authenticate the user originality in the following ways.

- Provide your birthday.
- Identify the photos of friends.

Or try logging into Facebook from a device you have logged in before.

Case 1 - Provide your birthday: In every online social networking websites, the users may not wish to provide their real birthday especially VIP members like Politicians, actors, actress and others too. In this scenario they may give some dummy birthday details when they create the account on Facebook and they may not remember that birthday forever. When the access to the account is lost due to change of device and asking them to prove user originality by providing birthday is really a risk to the users.

Case 2 - Identify the photos of friends: As we all know, everyday many users will post more photos on Facebook and the user may not remember which photo was posted by which user. Let us assume a user does not log into his/her account for one month. So the user will not know whoever has posted photos and which photo posted during that one month. Identifying the photos of friend's means, a photo will be displayed and the user has to choose the correct friend (user) who had posted that photo among the list of friends shown. However the displayed photo will also be from the days while the user did not log into his/her account. So identifying the friends by random photos will be tougher and really risk to the users.

Case 3 - Device used before: The final way to logging into Facebook from a device you have logged in before. In this scenario, if the user has sold his mobile device to someone else or had lost his/her mobile device then how the user could could log in with the device used before. It also will be highly a risk to the users.

As presented above, all the present mechanism provided by Facebook to prove the originality of the user has a major drawback, risk and makes the original user to lose his/her account permanently. If the user has lost his account permanently, he/she will lose entire friend circle and all the information associated with account. However a new mechanism is needed, which enhance the security for improving the quality of service.

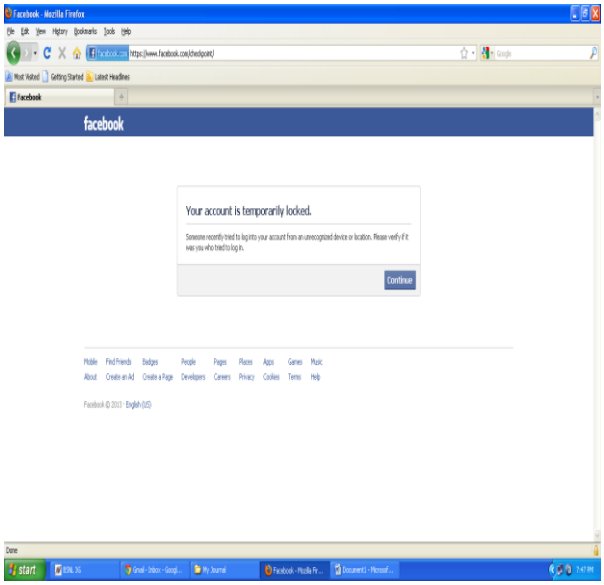


Fig 1 Account Temporarily Locked.

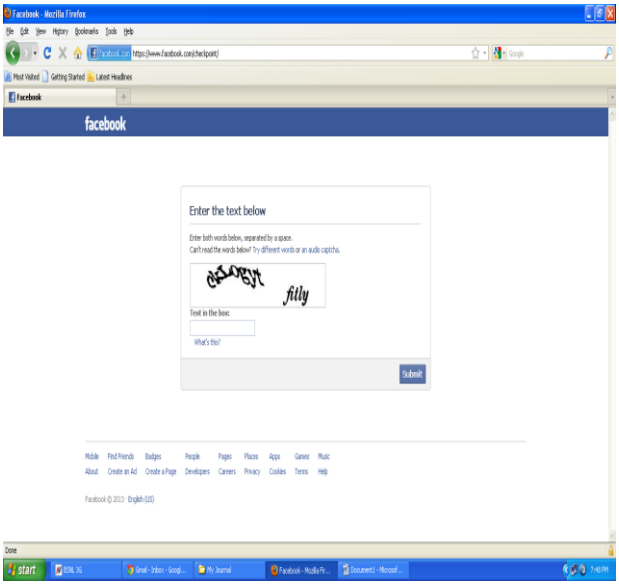


Fig 2 Enter the Text Below

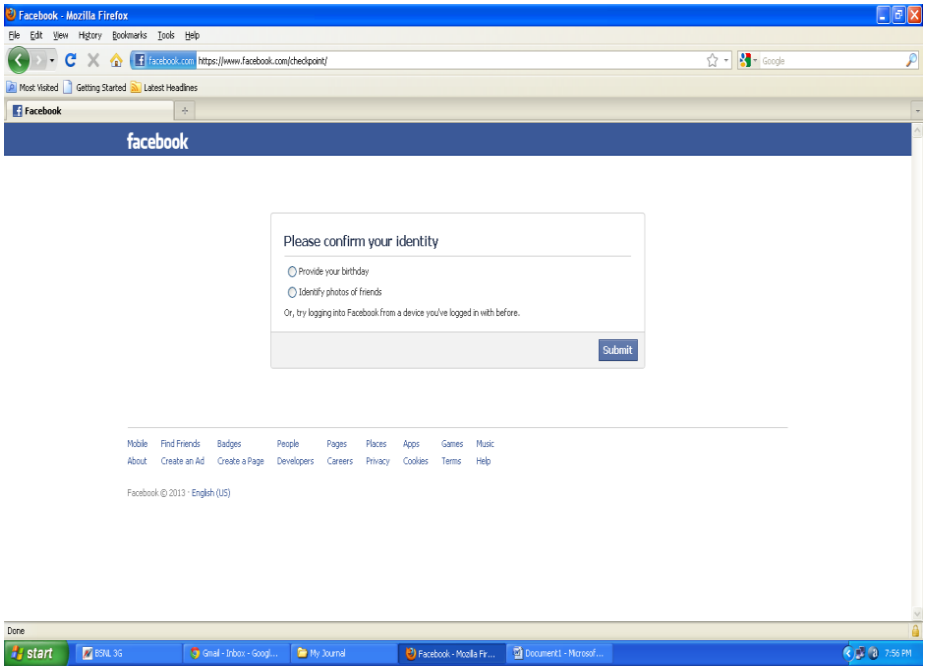


Fig 3 Confirm your identity

The above showm diagrams Fig 1, Fig 2, and Fig 3 represents the concept of Facebook to prove the originality of the user when the account is temporarily locked. However all these mechanism are not efficient and convenient for the internet users at all times. A new mechanism is ultimately needed to enhance the security for improving the quality of services in Online Social Networks (OSNs). The new system that we propose in this paper will always improve the efficiency and will be very much convenient to the internet users and also the proposed system will maintain the security constraints along with improved quality of service always.

B. Data Flow in Existing Model:

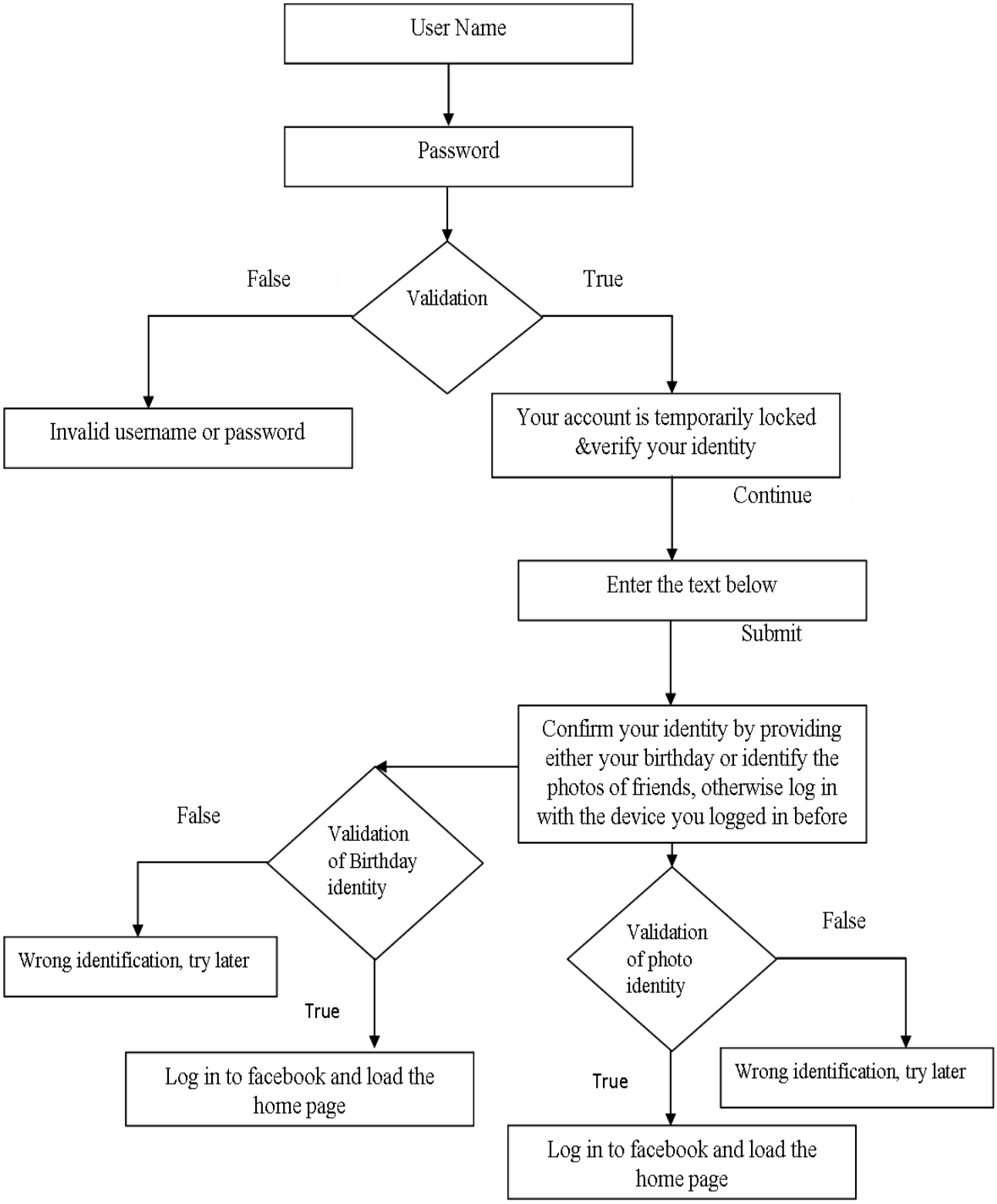


Fig 4 Data Flow in Existing Model

IV. Proposed Model

A. *Data Flow in Proposed Model:*

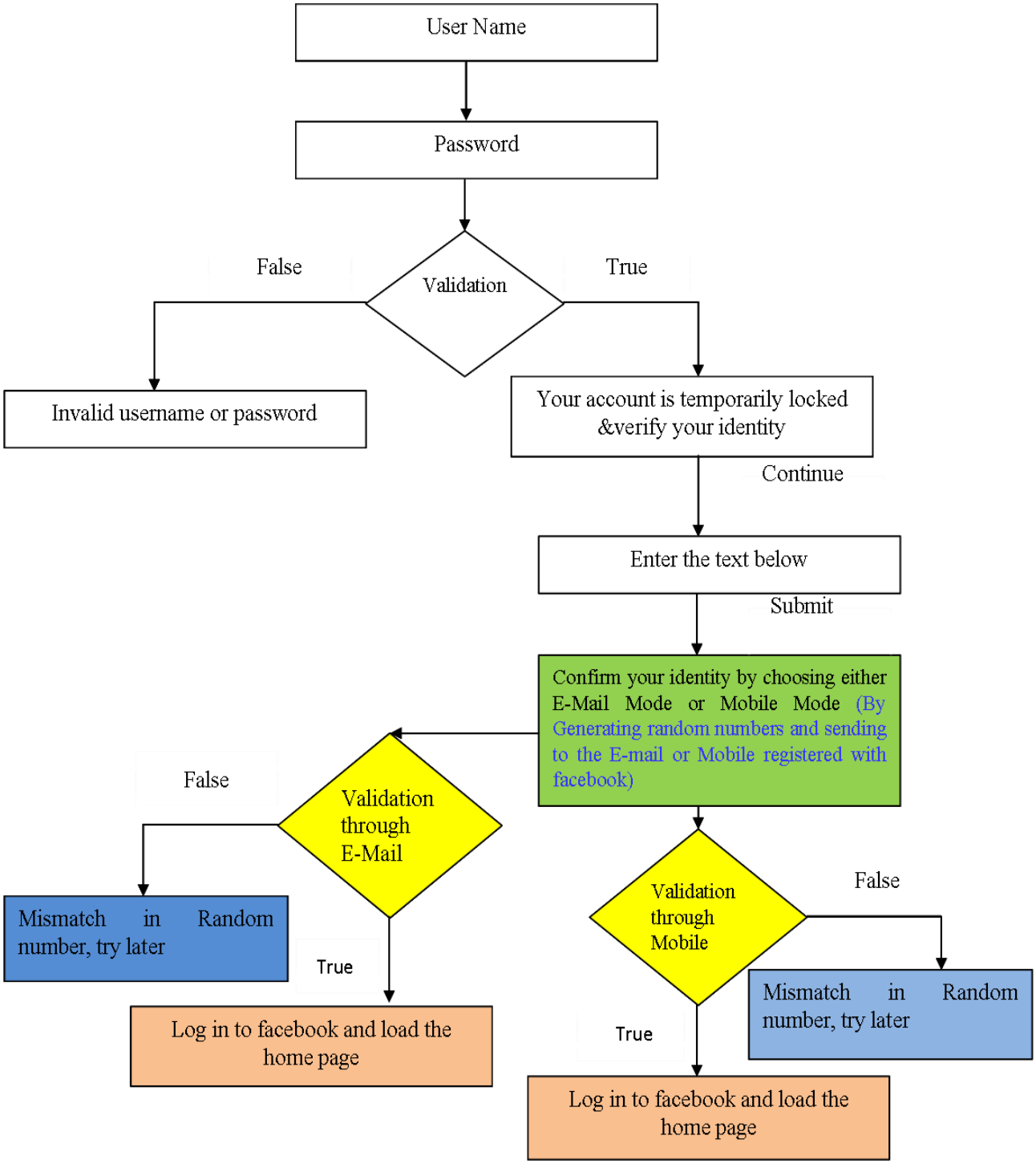


Fig 5 Data Flow in Proposed Model

The improvement of Quality of Service (QoS) by enhancing security mechanism in the proposed scheme can be evaluated as follows: when the user encounters the problem, while log in to his/her account such as the account is temporarily locked as shown below:

- Your account is temporarily locked.
- Someone recently tried to log into your account from unrecognized device or location. Please verify if it was you who tried to log in.

The current verification mechanism leads to poor performance and so in the proposed scheme we evaluate a new verification methodology, which maintains security constraint and improves quality of service always. The proposed verification scheme is shown and illustrated below.

- Validation through Mobile.
- Validation through E- Mail.

Case 1 – Validation through Mobile: In the proposed model, we propose the Facebook community organization should make its users give their mobile number compulsorily, while registering for new account creation. However the already existing users also must be asked to give their mobile number before logging into their account. If a user wishes to change his/her mobile number later that also must be allowed at any time. Later, whenever the account of any user is locked temporarily and the user chooses the validation through mobile, randomly generated number will be forwarded to the user's mobile number, which is associated with the corresponding user account. Once the user receives the randomly generated number on mobile must give the same number as input in the form, where the number is asked for validation. If both the numbers are same, the account is verified successfully and the user can have access to his/her account.

Case 2 – Validation through E- Mail: Account can be created in Facebook with an existing E- Mail id. However when the user account is locked temporarily and user chooses the validation through E- Mail, randomly generated number will be forwarded to the user's E- Mail id, which is associated with the corresponding user account. Once the user receives the randomly generated number on E- Mail must give the same number as input in the form, where the number is asked for validation. If both the numbers are same, the account is verified successfully and the user can have access to his/her account.

The above described to validation process methods replaces the existing methods available presently. Our proposed scheme uses random generated numbers to validate the user account and prove the user originality. This new scheme will enhance the security of the user account and also it improves the quality of service to Facebook users always. The implementation of our proposed scheme will make Facebook users not to lose access to their accounts and they will get connected with their friends, families and so on.

V. Implementation

The implementation of the proposed scheme can be carried out by generating random numbers. A random number is a number generated by a process, whose outcome is unpredictable, and which cannot be sub sequentially reliably reproduced [8]. The generation of random numbers cannot be guessed by previous value. This random number generation is used in most of the applications especially web applications. Here, in the proposed model we generate random numbers using the programming language JAVA. Java language is fully object oriented and highly securable language. Java language supports for generating random numbers by its packages. The Random class must be imported to generate random number in Java programming language as shown below.

- import java.util.Random;

Next the random object must be created to generate the random numbers. This random object allows the programmer to generate the simple random number generator. The predefined methods associated with the random object will generate the random numbers within the range of values.

(1) Random rand = new Random (); } Random object creation

(2) nextInt() & nextLong() } Methods of Random object

Let's consider the following Java programming that generates the random number, which cannot be predicted by the user.

```
import java.util.Random;
{
class number
{
   public static void main(String args[])
   {
  Random rand= new Random();
       for(j=0;j<1;j++)
       {
           System.out.println(rand.nextInt());
           System.out.println();
       }
   }
}
```

The above program will produce the random numbers as shown below, for each execution different number will be generated as shown below.

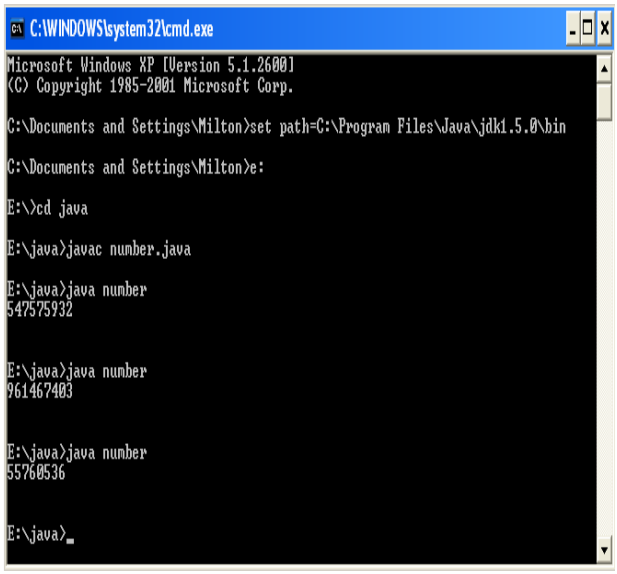


Fig 6 Random Numbers

The fig 6 shows the three different values generated by the same programming code for its three executions. However the same program can be executed as many times as need for the random numbers. The execution of for loop can be maximized if it is needed for the application. This random number can be used in our proposed model to enhance the security and also to improve the quality of service in online social networking.

In the proposed model, when the user account is temporarily locked the user must gone through the validation process to prove the use originality. The validation process can be done in the following two ways.

- Validation through Mobile.
- Validation through E- Mail.

If the user chooses the validation through mobile, the generated random number can be sent to the user's mobile device that is registered with the Facebook account and the received verification code is asked to be given in the form of Facebook application to validate the user. The user will be validated if the entered data on the form matches with the generated random number.

If the user chooses the validation through E- Mail, the generated random number can be sent to the user's E- Mail that is registered with the Facebook account and the received verification code is asked to be given in the form of Facebook application to validate the user. The user will be validated if the entered data on the form matches with the generated random number. Thus the proposed model uses random number generation methods to prove the user originality forever. This validation process is simple and will be very much convenient to the Facebook users. The implementation of the proposed scheme will make the Facebook users get connected with their friends and prevent them from losing access to their account.

VI. Results

The following screenshots represent the execution of our proposed data flow model, which enhance the security for improving the quality of service in online social networks.

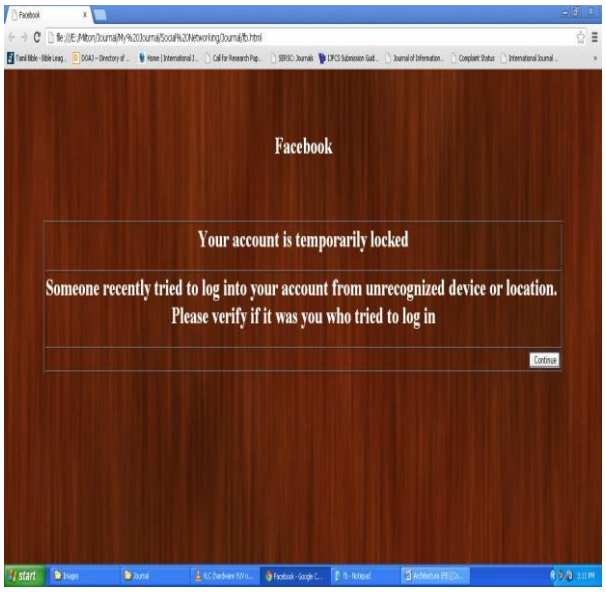


Fig 7 Account Temporarily Locked.

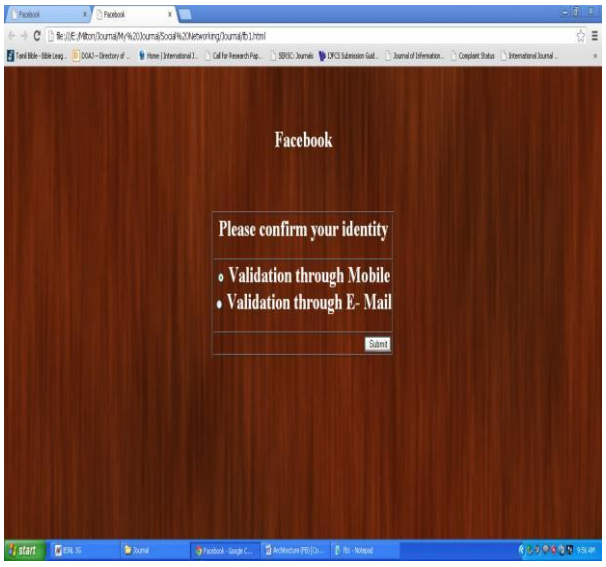


Fig 8 Validation Process

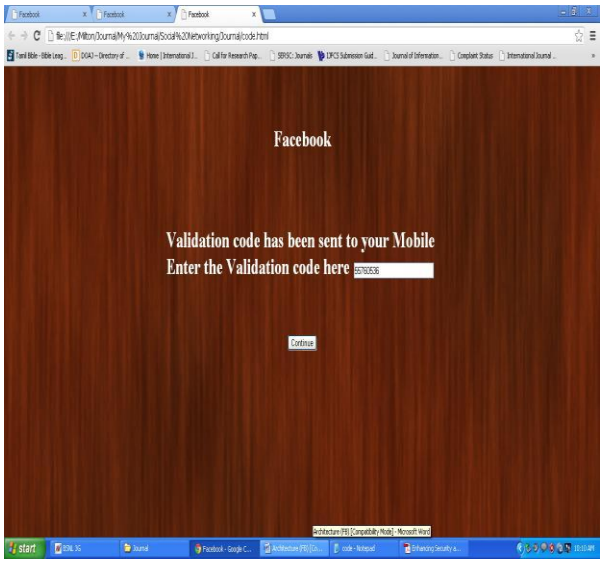


Fig 9 Validation through Mobile

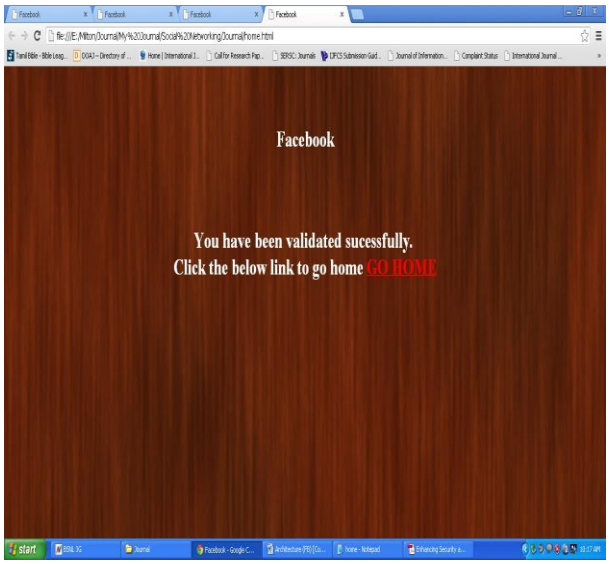


Fig 10 User validated successfully

The fig 7 shows the user account temporarily locked, when the user changes his/her device after long period of time being used the same device. The user account will be also temporarily locked, if some unauthorized user tries to access another user's account. Fig 8 represents the proposed data model concept of validating a user, after his/her account is temporarily locked. The proposed validation process can be carried out by either validation through Mobile or validation through E- Mail. The fig 9 shows the validation process of a user to prove the user originality through mobile. When the user chooses the validation process through mobile, automatically generated random number will be sent to the user's mobile device that is registered with the Facebook account. Once the random number is sent to the user's mobile device the user will be asked to give the validation code as input in the form as shown in the fig 9. After entering the validation code in the form, the entered code is matched with the generated code for the particular user. If both codes are matched, the user is validated successfully and the temporary lock is released and the user is allowed to go to his homepage as shown in the fig 10. If mismatch is found between codes, the user is not validated successfully and the user will not be allowed to go his/her homepage. The same process is applied, when the user chooses the validation process through E- Mail. Once the user have chosen the validation process through E- Mail, the automatically generated random number is sent to the user's E- Mail id that is registered with Facebook account and the user will be asked to enter the validation code in the Facebook form to check the user originality. If both codes are matched, the user is validated successfully and the temporary lock is released and the user is allowed to go to his homepage as shown in the fig 10. If mismatch is found between codes, the user is not validated successfully and the user will not be allowed to go his/her homepage.

CONCLUSION

In this paper, we have studied the comprehensive characteristics of Facebook application and we found Facebook application is lacking to produce security constraint to entire user account by its existing data flow model, which degrades the quality of service in Facebook application. We have proposed a new data flow model for Facebook application, which ultimately enhance the security constraints for improving the quality of service in facbook application. Our proposed model is evaluated and expected outcome is obtained at security level and improving quality of service.

REFERENCES

[1] Mohamed Shehab a, Anna Squicciarini b, Gail-Joon Ahn c, Irini Kokkinou "Access control for online social networks third party applications" Elsevier- Computers & Security 31 (2012) 897-911.
[2] Facebook, Facebook Statistics, March 2011. <http://www.Facebook.com/press>.
[3] Twitter, Twitter Numbers, March 2011. <http://blog.twitter.com/2011/03/numbers.html>.
[4] Facebook, http://en.wikipedia.org/wiki/Facebook
[5] Gorrell P. Cheek, Mohamed Shehab "Policy-by-Example for Online Social Networks" SACMATO 12 JUNE 20-22, 2012 NEWARK, NEW JERSY, USA, ACM YEAR 2012.
[6] Liu Kun, Terzi Evimaria. A framework for computing the privacy scores of users in online social networks. In: ICDM 2009, the ninth IEEE international conference on data mining, pp.288e297; December 2009.
[7] H. Kim, J. Tang, R. Anderson, Social authentication: harder than it looks, in: Proceedings of the 2012 Cryptography and Data Security Conference, 2012.
[8] M. Gjoka, M. Kurant, C.T. Butts, A. Markopoulou, Walking in Facebook: a case study of unbiased sampling of osns, in: Proceedings of IEEE INFOCOM '10, San Diego, CA, 2010.
[9] A.S. Yuksel, M. E. Yuksel, and A. H. Zaim. An approach for protecting privacy on social networks. In Proceedings of 5th International Conference on

Systems and Networks Communications, Washington, DC, USA, 2010. IEEE Computer Society.

[10] A.Acquisti and J. Grossklags. Privacy and rationality in individual decision making. IEEE Security and Privacy, 3(1):26{33}, 2005.

AUTHORS

Mr. M. Milton Joe was born in Nagercoil, Tamilnadu, India in 1988. He received his B.Sc Computer Science degree from Bharathidasan University, India in 2009 and he received his MCA degree from Anna University, India in 2012. Thereafter he worked as Assistant Professor in Meenakshi Academy of Higher Education and Research (Meenakshi University) for a year. Presently he is working as Assistant Professor at St. Jerome's College in Nagercoil, Tamilnadu, India. He has authored five research papers in reputed International Journals. His research topic includes Network Security, Network Communication, Vehicular Network and Social Networks.

Dr. B. Ramakrishnan is currently working as Associate Professor in the Department of Computer Science and research Centre in S.T. Hindu College, Nagercoil. He received his M.Sc degree from Madurai Kamaraj University, Madurai and received Mphil (Comp. Sc.) from Alagappa University Karikudi. He earned his Doctorate degree in the field of Computer Science from Manonmaniam Sundaranar University, Tirunelveli. He has a teaching experience of 26 years. His research interests lie in the field of Vehicular networks, mobile network and communication, Cloud computing, Green computing, Ad-hoc networks and Network security.

Dr. R.S. Shaji received his M.Tech in Computer Science and Engineering from Pondicherry University and PhD from Manonmaniam Sundaranar University. Presently he is working as Professor in Noorul Islam University. He has seven years of research experience and published more than twenty international journals. His research interests include Mobile and pervasive Networks.

Automatic Text Summarization System for Punjabi Language

Vishal Gupta
UIET, Panjab University, Chandigarh, India
Email: vishal@pu.ac.in

Gurpreet Singh Lehal
Department of Computer Science, Punjabi University, Patiala, India
Email: gslehal@gmail.com

Abstract— **This paper concentrates on single document multi news Punjabi extractive summarizer. Although lot of research is going on in field of multi document news summarization systems but not even a single paper was found in literature for single document multi news summarization for any language. It is first time that this system has been developed for Punjabi language and is available online at: http://pts.learnpunjabi.org/. Punjab is one of Indian states and Punjabi is its official language. Punjabi is under resourced language. Various linguistic resources for Punjabi were also developed first time as part of this project like Punjabi noun morph, Punjabi stemmer and Punjabi named entity recognition, Punjabi keywords identification, normalization of Punjabi nouns etc. A Punjabi document (like single page of Punjabi E-news paper) can have hundreds of multi news of varying length. Based on compression ratio selected by user, this system starts by extracting headlines of each news, lines just next to headlines and other important lines depending upon their importance. Selection of sentences is on the basis of statistical and linguistic features of sentences. This system comprises of two main steps: Pre Processing and Processing phase. Pre Processing phase represents the Punjabi text in structured way. In processing phase, different features deciding the importance of sentences are determined and calculated. Some of the statistical features are Punjabi keywords identification, relative sentence length feature and numbered data feature. Various linguistic features for selecting important sentences in summary are: Punjabi-headlines identification, identification of lines just next to headlines, identification of Punjabi-nouns, identification of Punjabi-proper-nouns, identification of common-English-Punjabi-nouns, identification of Punjabi-cue-phrases and identification of title-keywords in sentences. Scores of sentences are determined from sentence-feature-weight equation. Weights of features are determined using mathematical regression. Using regression, feature values of some Punjabi documents which are manually summarized are treated as independent input values and their corresponding dependent output values are provided. In the training phase, manually summaries of fifty news-documents are made by giving fuzzy scores to the sentences of those documents and then regression is applied for finding values of feature-weights and then average values of feature-weights are calculated. High scored sentences in proper order are selected for final summary. In final summary, sentences coherence is maintained by properly ordering the sentences in the same order as they appear in the input text at the selective compression ratios. This extractive Punjabi summarizer is available online.**

Index Terms—**Punjabi text summarizer, extractive summarization, named entity recognition, keywords identification, headlines identification**

I. Introduction

Automatic text summarization [1] [2] deals with reducing the source-text into a shorter version preserving its contents and overall meaning. Generally there are two phases of text summarization [3] systems: 1) Pre-Processing-phase [4] represents the source text in structured way. 2) In Processing phase [5] [6] [7] different features deciding the importance of sentences are determined and calculated. Scores of sentences are determined using equation of feature weights and high scored sentences in proper order as of input text are extracted for final summary. This paper describes single document multi news Punjabi extractive summarizer. It is text extraction based summarization system which is used to summarize the single Punjabi document with multi news by retaining the relevant sentences based on statistical and linguistic text features. Punjab is one of Indian states and Punjabi is its official language. For Punjabi language, it is the only summarizer available as no other Punjabi summarizer exists. This summarizer is available online at: http://pts.learnpunjabi.org/ and has two phases. 1) Pre processing phase [4][13] includes finding boundary of Punjabi sentences, Elimination of Punjabi-stop-words, Stemmer for Punjabi nouns and proper names, Allowing input restrictions to input text, Elimination of duplicate sentences and normalization of Punjabi noun words in noun morph. 2) In processing phase, different features deciding the importance of sentences are determined and calculated. Some of the statistical features are Punjabi keywords identification, relative sentence length feature and numbered data feature. Various linguistic features for selecting important sentences in summary are: Punjabi-headlines identification, identification of lines just next to headlines, identification of Punjabi-nouns, identification

doi:10.4304/jetwi.5.3.257-271

of Punjabi-proper-nouns, identification of common-English-Punjabi-nouns, identification of Punjabi-cue-phrases and identification of title-keywords in sentences etc. Sentence-feature-weight equation is applied for finding the final-scores of sentences. Weights of each feature are calculated using weight learning methods. Top ranked sentences in proper order are selected for final summary at selective compression ratios.

There is very complex derivational morphology for English language but not in case of Punjabi. As compared to English, Punjabi has rich system of inflectional morphology. Usually an English verb has five distinct inflectional forms. Different forms of a verb 'go' in English are go, gone, going, goes and went. But a Punjabi verb can have an average forty eight forms based on gender, tense, aspect value, number and person in any sentence. Moreover there are up to two causative forms for some of Punjabi verbs and further there will be on an average forty eight forms for each such causative form. Punjabi language is entirely different from other languages of world based on its syntax and grammar. For Punjabi summarizer, there is need to develop lexical resources for Punjabi because these resources are not available. The architectural diagram of Punjabi summarizer is given in Figure 1.

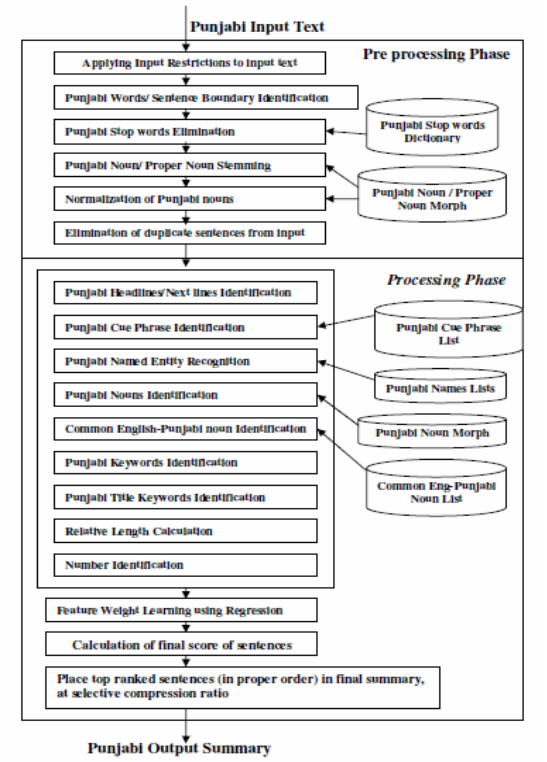


Figure 1. Overall architecture of Punjabi summarizer

II. PRE PROCESSING PHASE

Pre-Processing-phase represents the source text in structured way. Pre processing phase [4][13] includes finding boundary of Punjabi sentences, elimination of Punjabi-stop-words, stemmer for Punjabi nouns and proper names, allowing input restrictions to input text, elimination of duplicate sentences and normalization of Punjabi noun words in noun morph. The architectural diagram for Pre Processing Phase is given in Figure 2.

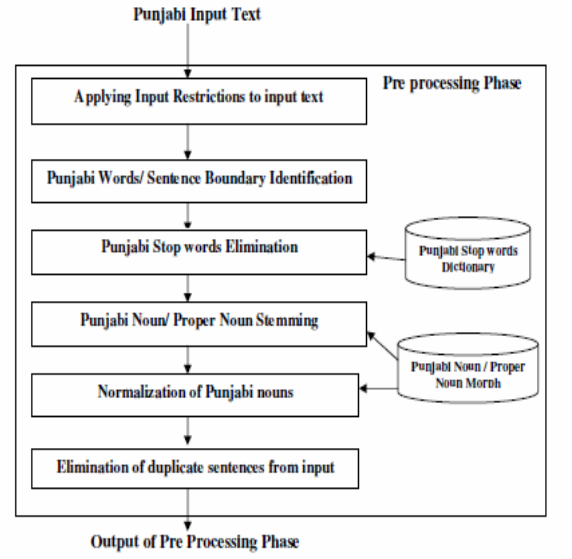


Figure 2. Pre processing phase of Punjabi summarizer

Different sub phases of pre-processing phase of Punjabi text summarization system are given below:

A. Boundary Identification Punjabi Words and sentence

From the Punjabi text, remove the punctuation mark characters like; . " " : - -- space character, tab space and so on for finding individual Punjabi words and sentence boundary is identified by presence of vertical bar ।, question mark ?, exclamation sign !, enter key, new line character etc at the end of sentence.

B. Applying Input Restrictions

Punjabi Text Summarization system allows Unicode based Gurmukhi text as input. Gurmukhi is the most common script used for writing the Punjabi language. Majority of input characters should be of Gurmukhi, otherwise error will be printed. From the input text, calculate length of Gurmukhi characters, punctuation mark characters, numeric characters, English characters and other characters. If number of Gurmukhi characters are less than equal to number of punctuation characters or number of numeric characters or number of English characters or number of other characters then error message is produced, otherwise if number of English characters or number of other characters are greater than equal to 10% of total input characters length, then error is produced "Can not accept the input!!!".

C. Punjabi Stop Words Elimination

Punjabi stop words are high frequency words appearing in Punjabi text like: ਹੈ hai "is", ਨੂੰ nūṃ "to",

ਨਾਲ nāl "with", **ਤੋਂ** tōṃ "from" and **ਦੇ** dē "of" etc. We need to delete these stop words from the source text, otherwise, those sentences which contain them may get importance unnecessarily. We have prepared Punjabi-stop-words-list by developing frequency-list from Punjabi-corpus. Punjabi corpus is taken from popular Punjabi newspaper Ajit and its thorough analysis is done. There are around 11.29 million words and 2.03 lakh unique words in this corpus. We have found 615 Punjabi stop words after analyzing the unique words of Punjabi corpus. The frequency of Punjabi stop words in corpus is 5.267 million words, which is equal to 46.64% of the corpus. Sample input and output for stop words elimination phase:

ਘਰੇਲੂ ਗੈਸ ਦੀ ਸਮੱਸਿਆ ਪਹਿਲ ਦੇ ਆਧਾਰ ਤੇ ਹੱਲ ਹੋਵੇਗੀ-ਥਿੰਦ

"Problem of domestic gas will be solved on priority basis-Thind"

In the input text **ਦੀ**, **ਦੇ**, **ਤੇ** and **ਹੋਵੇਗੀ** are Punjabi stop words. Sample output text after removing the stop words is:

ਘਰੇਲੂ ਗੈਸ ਸਮੱਸਿਆ ਪਹਿਲ ਆਧਾਰ ਹੱਲ -ਥਿੰਦ

"Problem domestic gas solved priority basis-Thind"

D. Punjabi Stemmer for Nouns/Names

The objective of any stemmer [19] [20] is to get the root of those words which are not in their basic forms and are not present in morph/dictionary. After stemming, if word is found in morph/dictionary [21], then it is correct word, otherwise it can be name or some incorrect word. In case of Punjabi stemmer [4][12][13] for nouns/ names, objective is to find root words and then root words are checked in Punjabi morph for nouns and in Punjabi names dictionary. After analyzing the Punjabi corpus, 18 suffixes were found for Punjabi nouns/names like ◌ਾਂ āṃ , ◌ਿਆਂ iāṃ, ◌ੂਆਂ ūāṃ and ◌ੀਆਂ īāṃ etc. and different rules for Punjabi noun/name stemming have been developed. Some outputs of stemmer for Punjabi nouns/names for different suffixes are:

ਲੜਕੀਆਂ laṛkīāṃ "girls" → **ਲੜਕੀ** laṛkī "girl" with suffix **◌ੀਆਂ** īāṃ, **ਮੁੰਡੇ** muṇḍē "boys" → **ਮੁੰਡਾ** muṇḍā "boy" with suffix **◌ੇ** ē, **ਫਿਰੋਜ਼ਪੁਰੋਂ** phirōzpurōṃ → **ਫਿਰੋਜ਼ਪੁਰ** phirōzpur with suffix **◌ੋਂ** ōṃ and **ਫੁੱਲਾਂ** phullāṃ "flowers" → **ਫੁੱਲ** phull "flower"with suffix **◌ਾਂ** āṃ etc.

The algorithm of Punjabi language stemmer [12] for nouns and proper names proceeds by segmenting the source Punjabi text into sentences and words. For each word of every sentence follow following steps:

Step 1: If suffix of current-Punjabi-word is **◌ਆਂ** āṃ (in case of **◌ੂਆਂ** ūāṃ, **◌ਿਆਂ** iāṃ and **◌ੀਆਂ** īāṃ), **◌ੇ** ē (in case of **◌ੀਏ** īē), **◌ੋ** ō (in case of **◌ੀਓ** Īō), **◌ਾ** ā (in case of **◌ੀਆ** ā, **◌ੀਆ** īā), **◌ਾਂ** vāṃ, **◌ੀ** ī , **◌ਾਂ** āṃ, **◌ੀਂ** īṃ, **ਜ/ਜ਼/ਸ** ja/z/s and **◌ੋਂ** ōṃ then delete the respective suffix from end and then go to Step 4.

Step 2: Else If current-word ends with **◌ੇ** ō, **◌ਿਓ** iō, **◌ੇ** ē, **◌ਿਆ** iā and **◌ਿਉਂ** iuṃ then delete the respective suffix and add kunna at the end and then go to Step 4.

Step3: Else current word is some unknown name or incorrect word.

Step 4: Stemmed Punjabi word is searched in Punjabi-noun morph/ names-dictionary. If it is found, It is Punjabi noun or Punjabi-name.

Algorithm Input: **ਮੁੰਡੇ** muṇḍē "boys" and **ਫੁੱਲਾਂ** phullāṃ "flowers"

Algorithm Output: **ਮੁੰਡਾ** muṇḍā "boy and **ਫੁੱਲ** phull "flower"

Punjabi stemming algorithm for nouns/names has been tested over fifty single-document-multi-news documents of Punjabi news corpus and its accuracy is 87.37%. This overall accuracy of Punjabi stemmer is ratio of correctly stemmed words to the total stemmed words by stemmer. Similarly the accuracy of each individual rule of stemmer is ratio of correct results under that rule to total results produced under that rule. Three types of errors can occur in case of Punjabi stemmer: 1) Dictionary errors 2) Violation of stemming-rules 3) Syntax mistakes. In case of dictionary errors, after stemming, root word is not found in Punjabi noun-morph/names dictionary, but in reality it is Punjabi noun/ proper name. In syntax errors, there is some syntax mistake while typing the Punjabi word, but actually it lies under any of stemming-rules. Overall stemming-errors, due to spelling mistakes is 0.45%, due to dictionary mistakes is 2.4% and due to rules violation is 9.78%.

Examples of errors due to rules violation are as follows:
Punjabi word **ਹਲਕੇ** halkē "light weight" is adjective and **ਬਦਲੇ** badlē "in lieu of" is adverb. These words are not found in Punjabi noun-morph/ names dictionary, but they lie under **◌ੇ** ē stemming-rule which treats them noun after stemming, but it is not true.
Examples of dictionary errors are as follows:
Some Punjabi words like **ਪ੍ਰਦੇਸਾਂ** pradēsāṃ "foreign"and **ਮੁਨਾਫ਼ਿਆਂ** munāphaiāṃ "profits" are actually nouns but are not present in noun morph or Punjabi dictionary. These words lie under **◌ਾਂ** āṃ rule and **◌ਿਆਂ** iāṃ rule of Punjabi stemmer and after performing noun stemming their root

words ਪ੍ਰਦੇਸ pradēs “foreign”and ਮੁਨਾਫ਼ਾ munāphā “profit” are also missing from Punjabi noun morph or Punjabi dictionary due to which these words are not considered as nouns by Punjabi stemmer.

Examples of syntax errors are as follows:
Some times we wrongly type the spellings of certain Punjabi noun words like ਚਿੜੀਆ chiṛīā “sparrow”and ਆਕ੍ਰਤੀ ākrtī “shape”but their correct spellings are ਚਿੜੀਆ chiṛīā “sparrow”and ਆਕ੍ਰਿਤੀ ākritī “shape” respectively, due to this these words are not found in Punjabi noun morph or Punjabi dictionary.

E. Normalization of Punjabi Nouns

This sub phase works on spelling normalization issues for Punjabi nouns, thereby resulting in multiple spelling variants for the same noun word. It is first time that Punjabi Normalizer [13] has been developed for Punjabi nouns. Problem with Punjabi is the non-standardization of Punjabi spellings. Many of the popular Punjabi noun words are written in multiple ways. For example, the Punjabi words ਚੰਡੀਗੜ੍ਹ chaṇḍīgaṛh “chandigarh”, ਪ੍ਰਕਾਸ਼ prakāsh “light”, ਜ਼ਿਲ੍ਹਾ jailhā “district” and ਖ਼ਿਆਲ khaiāl “idea” can also be written as ਚੰਡੀਗੜ chaṇḍīgaṛ “chandigarh”, ਪਰਕਾਸ਼ parkāsh “light”, ਜ਼ਿਲਾ zilā “district” ਜਿਲਾ jilā “district” ਜਿਲ੍ਹਾ jilhā “district” and ਖਿਆਲ khiāl “idea” respectively. To overcome this problem, input Punjabi text and Punjabi noun morph has been normalized for different spelling variations of Punjabi noun words. Punjabi noun morph is having 37297 noun words. The text has been normalized for the various for the various characters like ◌ੱ aadak, ◌ੰ bindi at top, Punjabi foot character ◌੍ for ਰ ra, ਵ v and ਹ ha and ◌਼ bindi at foot for ਸ਼ sha, ਖ਼ k̲ha, ਗ਼ ġa, ਜ਼ za, ਫ਼ fa, and ਲ਼ ḷa.

The algorithm for normalization of Punjabi nouns proceeds by copying noun_morph into another table noun_morph_normalized. For each noun word in table noun_morph_normalized follow the following steps:

Step 1 : Replace all the occurrences of ◌ੱaadak with null character.

Step 2 : Replace all the occurrences of ◌ੰ Bindi at top with null character.

Step 3 :Replace all the occurrences of ◌੍ Punjabi foot characters with any of suitable ਰ (ra) or ਵ (v) or ਹ (ha) characters.

Step 4 :Replace all the occurrences of ◌਼ bindi at foot with null character.

Step5:Noun_morph_normalized is now normalized

Step 6: End of algorithm

Algorithm Input: ਟੱਬ ṭabb , ਰਕਮੀਂ rakmīṃ, ਆਕ੍ਰਿਤੀ ākritī and ਖ਼ਿਆਲ khaiāl

Algorithm Output: ਟਬ ṭab, ਰਕਮੀ rkamī, ਆਕਰਿਤੀ ākritī and ਖਿਆਲ khiāl

After exhaustive analysis of Punjabi news corpus it is found that there is very less spelling variation in Punjabi nouns. Only 1.562% nouns show variation in their spellings. Out of these 1.562% words, percentage of words having one, two or three variations in the Punjabi news corpus are 99.95%, 0.046 % or 0.004% respectively. Figure 3 shows that rules for Bindi at foot and Aadak have maximum applicability. The least used rule is for Bindi at top and rule for foot characters is having usage of 22% in standardization of Punjabi nouns.

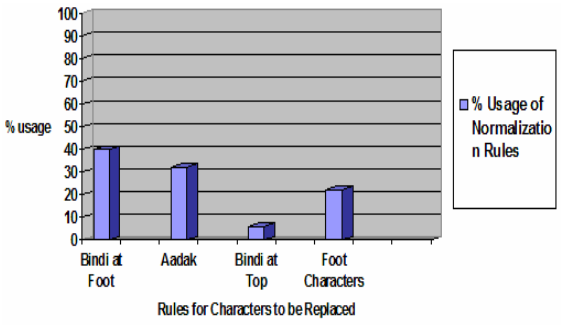


Figure 3. %Usage of normalization rules

F. Elimination of Duplicate Sentences

Duplicate sentences are the redundant sentences which need to be deleted otherwise these can get the influence unnecessarily and due to which certain other important sentences will not be displayed in the summary. In our system, duplicate sentences are deleted from input by searching the current sentence in to the sentence list which is initially empty. If current sentence is found in sentence list then that sentence is set to null otherwise it is added to the sentence list being the unique sentence. This elimination prevents duplicate sentences from appearing in final summary. An exhaustive analysis has been done on fifty Punjabi multi news documents for determining the frequency of duplicate sentences and it is discovered 9.60% sentences are duplicate. Average frequency of a duplicate sentence in a Punjabi document is three and maximum frequency is four. Out of 9.6% duplicate sentences from fifty Punjabi news documents, there are 5.4% sentences with minimum frequency two, 2.29% sentences with average frequency three and 1.91% sentences with maximum frequency four.

III. PROCESSING PHASE

In processing phase [22], different features deciding the importance of sentences are determined and calculated. Feature-weight equation is applied for finding the final-scores of sentences. Weights of each feature are

calculated using regression based weight learning method. Figure 4 describes the processing-phase of Punjabi text summarizer.

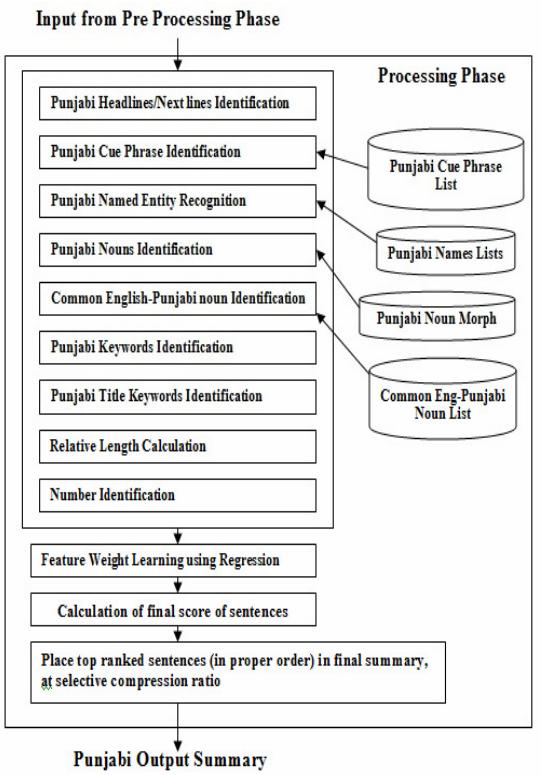


Figure 4. Processing phase

Features are of two types statistical and linguistic features. Some of the statistical features are Punjabi keywords identification, relative sentence length feature and numbered data feature. Various linguistic features for selecting important sentences in summary are: Punjabi-headlines identification, identification of lines just next to headlines, identification of Punjabi-nouns, identification of Punjabi-proper-nouns, identification of common-English-Punjabi-nouns, identification of Punjabi-cue-phrases and identification of title-keywords in sentences etc. Scores of sentences are determined from sentence-feature-weight equation:

$w_1f_1+w_2f_2+ w_3f_3+\ldots\ldots\ldots\ldots....w_nf_n$

Where f_1, f_2, f_3…… f_n are different features of sentences calculated in the different sub phases of Punjabi text summarization system and w_1, w_2, w_3…… w_n are the corresponding feature weights of sentences. Weights of features are determined using mathematical regression. Using regression, feature values of some Punjabi documents which are manually summarized are treated as independent input values and their corresponding dependent output values are provided. In the training phase, manually summaries of fifty news-documents are made by giving fuzzy scores to the sentences of those documents and then regression is applied for finding values of feature-weights and then average values of feature-weights are calculated. High scored sentences in proper order are selected for final summary. In final summary, sentences coherence is maintained by properly ordering the sentences in the same order as they appear in the input text at the selective compression ratios. The sub phases for Processing-phase [22] are as follows:-

A. Identification of Headlines and Next lines

It is first time that an automatic system for identification of multi news headlines and lines just next to headlines of a single document has been developed for Punjabi language. Headlines are highly important in news documents and are always part of final summary. There can be very important information in the next-line to headline, so next-line usually becomes part of final summary. In Punjabi-news-corpus with 957553 sentences, the frequency-count of these headlines/next lines is 65722 lines, which is 6.863% of the news-corpus. In Punjabi a sentence usually ends with '**।**' vertical bar, ?, or ! etc. and in Punjabi headlines identification system, if current sentence does not ends with punctuation marks like '**।**' vertical bar, ?, or ! etc. but ends with enter key or new line character then set the headline flag for that line to true. If the next subsequent line of this headline ends with punctuation marks like '**।**' vertical bar, ? or ! etc. but does not ends with enter key or new line character then set the next line flag to true for that line. An in depth analysis of results of headlines detection system and next lines identification system has been done over fifty Punjabi news documents taken randomly from Punjabi news corpus. Headline sentences are assigned very high score equal to 10 and their headline flags are set to true. The accuracy of Punjabi headline identification system is 97.43%. This accuracy is tested over 50 Punjabi single/multi-news documents. There are 2.57% errors due to the reason that some times name of author and location name are written in second line after the headline with enter key as last character, so it may be wrongly picked as second headline which is wrong. For example consider the following single news document as input:

ਅੱਸੀ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿਖਾਇਆ

ਉਜਾਗਰ ਸਿੰਘ (ਬਰਨਾਲਾ)

"Eighty girls were taught the sewing work

Ujagar Singh (Barnala)"

In the above news, second line containing author name and location name **ਉਜਾਗਰ ਸਿੰਘ (ਬਰਨਾਲਾ)** "Ujagar Singh (Barnala)" will also be treated as headline along with first line because it also ends with enter key, but this line is not important and should not come in summary. The accuracy of next lines identification system is 98.57% which is tested over fifty Punjabi single/multi news documents. Errors of 1.43% are due to the same reason of multiple headlines may come in a single news, due to which, some times next line may be missing from summary. Suppose the case in which there are multiple headlines in single news and we need only two lines in summary at 10% compression ratio and further suppose

second headline is not important as it is only containing name of author and location. But in this case next line will be missing from summary and second headline will be wrongly placed in summary. Next-lines are always assigned higher weight-age equal to 9 and their next line flags are set to true.

B. Punjabi Cue Phrase Identification

Cue Phrases are some important terms in text documents like: finally, conclusion, conclude, summary and summarize etc. Sentences containing cue-phrases are given more weight-age for summary than other sentences. We have prepared a list of Punjabi cue-phrases for assigning more weight-age to sentences containing these cue-phrases. Problem with Punjabi is non-standardization of Punjabi spellings. Many of the popular Punjabi words are written in multiple ways. As for example, the word **ਵਿਚ** "in" can be written both with and without addak, so both of these forms have been included in cue phrases. For example **ਅੰਤ ਵਿੱਚ/ ਅੰਤ ਵਿਚ** "in the end" and **ਸੰਖੇਪ ਵਿੱਚ/ ਸੰਖੇਪ ਵਿਚ** "in brief" etc. For those sentences containing cue phrase/cue phrases, their cue phrase flag is set to true. The frequency count of cue phrases is 58708 words in Punjabi news corpus which covers 0.52% of this corpus.

C. Punjabi Named Entity Recognition

Punjabi rule based named entity recognition system is first of its kind developed and implemented for identifying proper names in Punjabi text [9]. There was no other Punjabi rule based NER system was available prior to our NER. Different gazetteer lists are used in it like prefix-list, suffix-list, middle-name-list, last-name-list and names-list for checking whether the given Punjabi word is name or not. Gazetteer lists are developed by doing analyses of Punjabi news-corpus.

For checking if next-word in Punjabi-name or not, Prefix-list includes different prefixes of Punjabi names. like **ਸ਼੍ਰੀ.** "Mr.", **ਸ਼੍ਰੀਮਤੀ** "Mrs.", **ਸ.** "sardar", **ਪ੍ਰਿ.** "Prin." and **ਡਾ:** "Dr." etc. There are fourteen prefixes identified from the Punjabi-news-corpus. We have developed prefix-list by making freq-list from corpus. The freq-count of prefix-words is 17,127 which includes 0.15% of the corpus. Suffix-list includes various suffixes of names like **ਪੁਰੀ** "puri", **ਪੁਰਾ** "pura", **ਜੀਤ** "jit" and **ਪੁਰ** "pur" etc for checking if current-Punjabi-word is name or not. There are fifty suffixes identified from the Punjabi-news-corpus. We have developed suffix-list by making freq-list from corpus. The freq-count of suffix-words is 225306 which includes 1.99% of the corpus.

Punjabi-middle-names-list includes various middle-names of persons like **ਕੁਮਾਰੀ** "kumari", **ਕੌਰ**"kaur" and **ਕੁਮਾਰ** "kumar" etc for checking if that word is name or not. There are 08 middle-names identified from Punjabi-news-corpus. We have developed middle-names-list by making freq-list from corpus. The freq-count of middle-name words is 97907 which includes 0.8672% of the corpus. Punjabi-last-names-list includes various last-names of persons like **ਗੋਇਲ** "goel", **ਗੁਲਾਟੀ** "gulati" and **ਖੁਰਾਨਾ** khurānā "khurana" etc for checking if that word is name or not. There are 310 last-names identified from Punjabi-news-corpus. We have developed last-names-list by making freq-list from Punjabi-corpus. The freq-count of last-name words is 69268 which includes 0.6135% of the corpus. For finding importance of sentences, proper names are very much useful. There are 17598 proper-names identified from the Punjabi-news-corpus. Punjabi-proper- names-list covers 13.84% of words from Punjabi-news-corpus. The value of Punjabi-names-feature is calculated by taking ratio of number of Punjabi-names in a sentence to the length of that sentence and value of this feature for a sentence lies between 0 to 1.

The Algorithm for rule based Punjabi NER has been published in [9] and it increments the NER score of current sentence by 01 if current Punjabi-word matches with any word from any of prefix-list or suffix-list or names-list or middle-names-list or last-names-list. Punjabi NER has been tested over fifty Punjabi-news-documents with Precision=89.32%, Recall=83.4%, F-score=86.25% and 13.75% errors. There are no errors in prefix rule. There are 1% errors in suffix rule for example **ਕਰਿਆਣਾ** "grocery" and **ਅਫਸਰ** aphsar "officer" are not found in Punjabi nouns-morph/dictionary but both of them fall under suffix-rule which treats them as Punjabi-names which is false. There are 0.25% errors in middle-name-rule for example in a proper-name **ਕੌਰ ਸਿੰਘ** "Kaur Singh" both middle-names are together as a single name, but they lie under middle-name-rule which makes NER score equal to 2 in this case. There are 10% errors in last-name-rule for example in case of a Punjabi-names **ਕਰਤਾਰ ਸਿੰਘ ਜੰਗੀਆਣਾ** "Kartar Singh Jangiana" and **ਬੰਤਾ ਸਿੰਘ ਬੰਟੀ** "Banta Singh Banti" their last names **ਜੰਗੀਆਣਾ** "Jangiana" and **ਬੰਟੀ** "Banti" are not in last-names-list but are part of proper-names-list so their NER score will be wrongly incremented to 2 in each case. There are 0.25% errors in proper-names-rule for example a Punjabi word **ਨਿਹਾਲ** "Nihal" some times is treated as Punjabi-name and some times is treated in different sense, but because it is part of proper-names-list so it will be always treated as a proper-name by Punjabi NER. Remaining 2.25% errors are because of those Punjabi-names which do not fall under any of NER rules for example **ਗਰੀਣ ਐਵਿਨਿਊ** "Green Avenue" or because of those Punjabi-words which are some times treated as Punjabi-names and some times treated as Punjabi-nouns for example a Punjabi word

ਬਹਾਦਰ "brave" is some times treated as Punjabi-name and some times treated as noun.

D. Punjabi Nouns/ Common English-Punjabi Nouns Identification

Sentences Possessing Punjabi-Nouns [1] are given more weight-age. Punjabi words are searched in noun morph or stemming is done for possibility of nouns. There are 37297 nouns in Punjabi-noun-morph [10]. The score of this feature is ratio of number of Punjabi-nouns in a sentence to length of that sentence. The range of value of this feature is between 0 to 1. The frequency of Punjabi nouns is 16.56% of words in Punjabi-news-corpus. The accuracy of Punjabi noun identification phase is 98.43% which is tested over 50 Punjabi-news-documents of Punjabi-corpus. Errors of 1.57% are due non existence of certain Punjabi nouns in noun morph and due to stemming errors of Punjabi noun stemmer [4].

In these days, there is common usage of English-words in Punjabi text. Consider a Punjabi sentence "**ਟੈਕਨਾਲੋਜੀ ਦੇ ਯੁੱਗ ਵਿਚ ਮੋਬਾਈਲ**" "In the era of mobile technology" This sentence includes **ਮੋਬਾਈਲ** "mobile" and **ਟੈਕਨਾਲੋਜੀ** "technology" as common English-Punjabi nouns. Majority of such terms are not found in Punjabi dictionary or Punjabi noun morph. In text summarization, Sentences containing common English-Punjabi nouns are assigned more weight-age. A small offline module has been developed to generate the database for common English-Punjabi nouns by analyzing the Punjabi news corpus along with frequency of these common English-Punjabi nouns into Punjabi news corpus. Punjabi-words are searched in Common-English-Punjabi-nouns-dictionary for possibility of common English-Punjabi nouns. This value of this feature is calculated by ratio of number of common-English-Punjabi-nouns in a sentence to the length of that sentence. The range of value for this feature lies from 0 to 1 for a sentence. From Punjabi news corpus, frequency count of common-English-Punjabi-nouns is 18245, which covers 6.44% of Punjabi-corpus. The accuracy of common-English-Punjabi-noun identification phase is 95.12% which is tested over 50 Punjabi news-documents of Punjabi-corpus. Errors of 4.88% are due non existence of certain common English-Punjabi nouns in database of common English-Punjabi nouns.

E. Punjabi Keywords/Title Keywords Identification

Keywords are thematic words containing important information. Keywords are helpful in deciding the sentence importance. Punjabi keywords identification system is first of its kind system developed and implemented as prior to it no other Punjabi keywords identification system was available. Algorithm for Punjabi Keywords Identification [7] [11]:

Step 1:- Set noun flag to true for those words of input text which are found in Punjabi noun morph.

Step 2:- For each Punjabi word w, find its TF-ISF-Score which is calculated by multiplying TF(w,s) with ISF(w). Where TF(w,s) is the frequency of word w in sentence s, and the inverse sentence frequency ISF(w) = log(|Sl/ SF(w)). Sentence-frequency SF(w) is the frequency of sentences containing word w. Store top ranked words (with high TF-ISF-Scores) with Punjabi_noun_flag= true in a priority queue.

Step 3:- Delete top 20% of Punjabi-noun-words from the priority queue, which are candidates for keywords in this phase.

The Precision, Recall and F-Score of Punjabi keywords identification system are 80.4%, 90.6% and 85.2% respectively which are calculated by analyzing the results of keywords identification system over fifty Punjabi news documents. Errors of 14.8% are because of absence of some Punjabi-nouns in noun-morph or dictionary errors or syntax mistakes in input text or due to violation of stemming-rules. In case of dictionary errors, after stemming, root word is not found in Punjabi noun-morph/names dictionary, but in reality it is Punjabi noun/ proper name. In syntax errors, there is some syntax mistake while typing the Punjabi word, but actually it lies under any of stemming-rules. Examples of errors due to rules violation are: Punjabi word **ਹਲਕੇ** halkē "light weight" is adjective and **ਬਦਲੇ** badlē "in lieu of" is adverb. These words are not found in Punjabi noun-morph/ names dictionary, but they lie under ੇ ē stemming-rule which treats them noun after stemming, but it is not true.

Title lines are the headlines of single/multi news documents. Sentences containing Title-keywords [5] are given more weight-age. For obtaining Title-keywords, stop words are removed from title-lines with headline_flag= true. This feature-score is calculated as ratio of unique title-keywords in a sentence to the total number of title-keywords. The efficiency of Punjabi title keywords identification is 97.48% which is calculated over fifty Punjabi single/multi news documents of Punjabi corpus. Errors of 2.52% are due to the reason that some of stop words may be left in the title line as Punjabi stop words list is not exhaustive and contains 615 Punjabi stop words.

F. Punjabi Sentence Relative Length Feature

Short Punjabi sentences are avoided for including in final summary as often they contain less information [5]. But lengthy sentences can have lot of important information. This value of this feature is calculated as ratio of frequency of words in current sentence to the words frequency of largest sentence. The value of this feature is always less than or equal to one.

Punjabi-Sentence-Length-feature-Score= frequency of words in current sentence / words frequency of largest sentence.

G. Numeric Data Identification Feature

For text summarization, those sentences containing contain numeric data [5] are assigned more weight-age Numeric digits, Gurmukhi and Roman numerals are considered as numeric data. The value for this feature is determined by dividing the frequency of numeric data in current sentence by the length of that sentence.
Number-feature-score= Frequency of numeric data in current sentence/ Sentence Length

H. Calculation of Scores of Sentences and Producing Final Summary

Final scores of sentences are determined from sentence-feature-weight equation.

w1f1+w2f2+ w3f3+……wnfn Where f1, f2, f3……fn are different features of sentences calculated in the different sub phases of Punjabi text summarization system and w1, w2, w3……wn are the corresponding feature weights of sentences. We have applied regression [5] [8] model for estimating the weights of text features for Punjabi text summarization system. A relation between inputs and outputs is established. Regression can be represented in the matrix notation as below:

$$\begin{bmatrix} Y_0 \\ Y_1 \\ . \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & .... & X_{010} \\ . & . & .... & . \\ . & . & .... & . \\ X_{m1} & X_{m2} & .... & X_{m10} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ . \\ w_m \end{bmatrix}$$

Where
$[Y]$ is fuzzy output vector having values between 0 to 1 based on importance of sentences given manually for fifty documents.
$[X]$ is the input matrix (feature parameters) for different features having values between 0 to 1.
$[w]$ is weight matrix of system (with weights w1, w2…….w10 in the given equation)
In the training corpus, m denotes total number of sentences.
Weight w of a particular feature k (k=1 to 10) with input matrix x and fuzzy output matrix y can be calculated as follows:-

$$w = \frac{\sum_{i=01 \text{ to } m} (x_i - \text{mean}(x))(y_i - \text{mean}(y))}{\sum_{i=01 \text{ to } m} (x_i - \text{mean}(x))^2}$$

From the above equation, weights of each of ten features of Punjabi text summarization have been calculated. Table I shows the results of weight learning using regression.

TABLE I.
WEIGHT LEARNING RESULTS USING REGRESSION

Features	Learned weights
Sentence relative length feature	0.31
Punjabi Keywords identification feature	0.29
Number feature	2.54
Headline feature	10
Lines just next to headline feature	9
Punjabi noun feature	0.42
Punjabi proper noun feature	0.75
Common English-Punjabi noun feature	1.29
Punjabi Cue phrase feature	1
Punjabi Title keywords feature	1.8

From results of weight learning in Table I, we concludes that three most important features of Punjabi Text Summarizer are identification of Punjabi-headlines, identification of next-lines and identification of numeric data. Top ranked sentences in proper order are selected for final summary. In final summary, sentence coherence is maintained by properly ordering the sentences in the same order as they appear in the input text at the selective compression ratios.

Algorithm for single document multi news Punjabi Text Summarization System:-
Algorithm starts by splitting the input Punjabi text into sentences and words. Initially scores of every sentence is set to 0.

Step I: Delete the duplicate sentences from input text by searching the current sentence in the sentence list which is initially empty. For each sentence check the following condition: If current sentence is found in sentence list then Current sentence is set to null being the duplicate sentence. Else Current sentence is added to the sentence list being the unique sentence. Follow steps II to step XII for every word in sentences.

Step II: Delete stop words from every sentence in input.

Step III: Calculate the noun-score of sentence, if current Punjabi-word is noun.

Step IV: Calculate common-English-Punjabi-noun score of sentence, if current Punjabi-word is common-English-Punjabi noun.
Step V: Calculate proper-name-score of sentence, If current Punjabi-word is proper-name.

Step VI: Apply stemmer [12] for Punjabi Noun/Proper Names for those words which are not found in nouns-morph/ common-English-Punjabi-noun-list/ proper-names-dictionary and go to step III.

Step VII: Calculate numeric-feature-score of sentence, if current Punjabi-word is any number like 45.

Step VIII: Set the headline-flag= true for current Punjabi-word, if it is part of headline.
Step IX: Set the next-line-flag= true for current Punjabi-word, if it is part of line just next to headline.

Step X: Calculate the score of keyword feature for current Punjabi-word using TF-ISF technique.

Step XI: Set cue-phrase-flag for current Punjabi-word to true, if it is cue-phrase.

Step XII: Calculate title-keyword-feature-score of current word, if it is title keyword.

Step XIII: Calculate the relative-length-feature-score of all the sentences.

Step XIV: Calculate the weight-age of each feature by applying regression using sentence-feature-weight-equation.

Step XV: Calculate final-scores of all the sentences by applying sentence-feature-weight-equation.

Step XVI: Select the top scored sentences at given compression ratios i.e. at 10%, 30%, 50% C.R. etc.

Step XVII: Final summary is formed by arranging top scored sentences in ascending order of their position in input text at selective compression ratios. In this step coherence of sentences is maintained in final summary.

Algorithm Input-

First News: -

ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ 'ਚ ਵਿਕਾਸ ਕਾਰਜਾਂ 'ਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ-ਭਾਨਾ

ਸ਼ਹਿਣਾ, 8 ਜਨਵਰੀ (ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਹਲਕਾ ਵਿਧਾਇਕ ਸੰਤ ਬਲਵੀਰ ਸਿੰਘ ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ ਹੇਠ ਹਲਕੇ ਦੇ ਵਿਕਾਸ ਕਾਰਜਾਂ ਵਿਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ ਹੈ। ਇਹ ਸ਼ਬਦ ਭਗਵਾਨ ਸਿੰਘ ਭਾਨਾ ਯੂਥ ਆਗੂ ਤੇ ਸੰਮਤੀ ਮੈਂਬਰ ਨੇ ਪਿੰਡ ਨਾਨਕਪੁਰਾ ਵਿਖੇ ਸ਼ਗਨ ਸਕੀਮ ਦੇ ਚੈਕ ਦੇਣ ਸਮੇਂ ਸੰਬੋਧਨ ਕਰਦਿਆਂ ਆਖੇ। ਭਗਵਾਨ ਸਿੰਘ ਨੇ ਕਿਹਾ ਕਿ ਸ਼ਗਨ ਸਕੀਮ ਲਈ ਰਹਿੰਦੇ ਪਰਿਵਾਰਾਂ ਲਈ ਛੇਤੀ ਹੀ ਬਾਕੀ ਦੀ ਰਾਸ਼ੀ ਜਾਰੀ ਕੀਤੇ ਜਾਣ ਦੀ ਹਲਕਾ ਵਿਧਾਇਕ ਨੇ ਹਾਮੀ ਭਰੀ ਹੈ ਅਤੇ ਸੰਮਤੀ ਰਾਹੀਂ ਵੀ ਪਿੰਡਾਂ ਲਈ ਸਬਮਰਸੀਬਲ ਪੰਪ ਤੇ ਗਰਾਂਟਾਂ ਦਿੱਤੀਆਂ ਜਾ ਰਹੀਆਂ ਹਨ। ਇਸ ਸਮੇਂ ਸੁਖਦੇਵ ਸਿੰਘ ਸਰਪੰਚ ਨਾਨਕਪੁਰਾ, ਪਵਨ ਕੁਮਾਰ, ਗੁਰਚਰਨ ਸਿੰਘ ਜ਼ੈਲਦਾਰ, ਕੇਰ ਸਿੰਘ ਪੱਖੋਕੇ, ਗੁਰਤੇਜ ਸਿੰਘ ਘੋਨਾ, ਜੰਗ ਸਿੰਘ ਪ੍ਰਧਾਨ ਟੈਕਸੀ ਯੂਨੀਅਨ, ਜਗਸੀਰ ਸਿੰਘ, ਕਰਤਾਰ ਸਿੰਘ ਪੱਖੋਕੇ ਆਦਿ ਆਗੂ ਵੀ ਹਾਜ਼ਰ ਸਨ।

Second News:-

ਅਧਿਆਪਕ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਨੂੰ ਸ਼ਰਧਾਂਜਲੀਆਂ ਭੇਟ

ਧਨੌਲਾ, 8 ਜਨਵਰੀ (ਨਿੱਜੀ ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਸੁਖਵਿੰਦਰ ਸਿੰਘ ਵੜੈਚ ਦੇ ਹੋਣਹਾਰ ਅਧਿਆਪਕ ਪੁੱਤਰ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਦੀ ਅੰਤਿਮ ਅਰਦਾਸ ਗੁਰਦੁਆਰਾ ਪਾਤਸ਼ਾਹੀ ਨੌਵੀਂ ਵਿਖੇ ਹੋਈ। ਇਸ ਮੌਕੇ ਵੱਖ-ਵੱਖ ਸਖ਼ਸ਼ੀਅਤਾਂ ਨੇ ਸ਼ਰਧਾ ਦੇ ਫੁੱਲ ਭੇਟ ਕੀਤੇ। ਜਥੇਦਾਰ ਸਾਧੂ ਸਿੰਘ ਰਾਗੀ ਸਾਬਕਾ ਚੇਅਰਮੈਨ ਮਾਰਕੀਟ ਕਮੇਟੀ ਭਦੌੜ ਨੇ ਕਿਹਾ ਕਿ ਸਾਡੇ ਕੋਲ ਅਜਿਹੀ ਦੁਖਦਾਈ ਮੌਤ 'ਤੇ ਮਾਪਿਆਂ ਕੋਲ ਭਾਣਾ ਮੰਨਣ ਨੂੰ ਕਹਿਣ ਲਈ ਸ਼ਬਦ ਵੀ ਨਹੀਂ। ਅਜਿਹੀ ਹੋਣਹਾਰ ਔਲਾਦ ਦਾ ਬੇ-ਵਕਤ ਚਲੇ ਜਾਣਾ ਬਹੁਤ ਦੁਖਦਾਈ ਹੈ। ਜਗਤਾਰ ਸਿੰਘ ਜੰਗੀਆਣਾ ਪ੍ਰਚਾਰਕ ਸ਼੍ਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ ਅੰਮ੍ਰਿਤਸਰ ਨੇ ਕਿਹਾ ਕਿ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਤੇ ਧਨੌਲਾ ਵਾਸੀਆਂ ਨੂੰ ਹੀ ਨਹੀਂ, ਬਲਕਿ ਨਾਨਕੇ ਪਿੰਡ ਜੰਗੀਆਣਾ ਨੂੰ ਬਹੁਤ ਮਾਣ ਸੀ। ਉਸ ਵਿਚ ਨਿਆਣਿਆਂ ਵਾਲੀ ਚੰਚਲਤਾ ਘੱਟ ਸੀ ਅਤੇ ਸਿਆਣਿਆਂ ਵਾਲੀ ਲਿਆਕਤ ਵਧੇਰੇ ਸੀ। ਸੰਤ ਬਲਵੀਰ ਸਿੰਘ ਘੁੰਨਸ ਹਲਕਾ ਵਿਧਾਇਕ ਭਦੌੜ, ਸ: ਭੋਲਾ ਸਿੰਘ ਵਿਰਕ ਮੈਂਬਰੀ ਕੌਮੀ ਜਨਰਲ ਕੌਂਸਲ ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ ਦਲ, ਜਥੇਦਾਰ ਬਲਦੇਵ ਸਿੰਘ ਚੂੰਘਾਂ, ਜਥੇਦਾਰ ਅਮਰ ਸਿੰਘ ਬੀ.ਏ. ਮੈਂਬਰ ਸ਼੍ਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ ਅੰਮ੍ਰਿਤਸਰ, ਜਥੇਦਾਰ ਭਰਪੂਰ ਸਿੰਘ ਧਨੌਲਾ ਸਾਬਕਾ ਚੇਅਰਮੈਨ, ਗੁਰਵੀਰ ਸਿੰਘ ਗੁਰੀ ਯੂਥ ਕਾਂਗਰਸੀ ਆਗੂ, ਇਕਬਾਲ ਸਿੰਘ ਜੰਗੀਆਣਾ ਸੰਮਤੀ ਮੈਂਬਰ, ਗਮਦੂਰ ਸਿੰਘ ਮਾਨ ਸਰਪ੍ਰਸਤ ਆੜ੍ਹਤੀਆ ਐਸੋਸੀਏਸ਼ਨ ਧਨੌਲਾ, ਗੁਰਨਾਮ ਸਿੰਘ ਸਿੱਧੂ ਸਾਬਕਾ ਪ੍ਰਧਾਨ ਨਗਰ ਕੌਂਸਲ ਧਨੌਲਾ, ਗੁਰਚਰਨ ਸਿੰਘ ਕਲੇਰ ਪ੍ਰਧਾਨ ਆੜ੍ਹਤੀਆ ਐਸੋਸੀਏਸ਼ਨ, ਭਰਪੂਰ ਸਿੰਘ ਸਾਬਕਾ ਐਮ.ਸੀ. ਸੁਰਿੰਦਰ ਸਿੰਘ ਸੱਦੋਵਾਲੀਆ ਜ਼ਿਲ੍ਹਾ ਪ੍ਰਧਾਨ ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ ਦਲ ਅੰਮ੍ਰਿਤਸਰ, ਚੇਤਨ ਸਿੰਘ ਮੂੰਮ, ਲਾਭ ਸਿੰਘ ਮੱਝੂਕੇ, ਜੰਗ ਸਿੰਘ ਜੰਗੀਆਣਾ, ਕਰਮਜੀਤ ਸਿੰਘ ਨੀਟਾ ਮੈਂਬਰ ਜ਼ਿਲ੍ਹਾ ਪ੍ਰੀਸ਼ਦ, ਹਰਨੇਕ ਸਿੰਘ ਸਾਬਕਾ ਪ੍ਰਧਾਨ ਸਹਿਕਾਰੀ ਸਭਾ ਜੰਗੀਆਣਾ, ਮਨਮੋਹਨ ਸਿੰਘ, ਗੁਰਤੇਜ ਸਿੰਘ ਸਰਪੰਚ, ਰਾਜ ਸਿੰਘ ਨੈਣੇਵਾਲੀਆ, ਗੁਰਪ੍ਰੀਤ ਸਿੰਘ ਕਲੱਬ ਆਗੂ, ਰਾਮ ਸਿੰਘ ਢੀਂਡਸਾ, ਗੁਰਮੀਤ ਸਿੰਘ ਸ਼ਹਿਣਾ, ਬੂਟਾ ਸਿੰਘ ਬੁਰਜ, ਭਗਵਾਨ ਸਿੰਘ ਭਾਨਾ, ਹਰਵਿੰਦਰ ਸਿੰਘ, ਯਾਦਵਿੰਦਰ ਸਿੰਘ ਵਾਲੀਆ ਐਮ.ਸੀ., ਜਗਤਾਰ ਸਿੰਘ ਕਲੇਰ ਪ੍ਰਧਾਨ ਸਹਿਕਾਰੀ ਸਭਾ ਧਨੌਲਾ, ਗੁਰਪ੍ਰੀਤ ਸਿੰਘ ਚੀਮਾ ਆਦਿ ਨੇ ਹਾਜ਼ਰੀ ਲਵਾਈ।

Third News:-

ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਹੋਈ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਗਰੀਨ ਐਵੀਨਿਊ ਦੇ ਪਾਰਕ ਨੇੜੇ ਨਾਨਕਸਰ ਗੁਰਦੁਆਰਾ ਵਿਖੇ ਹੋਈ, ਜਿਸ ਵਿਚ ਸਰਬ ਸੰਮਤੀ ਨਾਲ ਬੂਟਾ ਸਿੰਘ ਚੌਹਾਨ ਅਤੇ ਸੁਰਜੀਤ ਸਿੰਘ ਦਿਹੜ ਸਰਪ੍ਰਸਤ, ਡਾ: ਉਜਾਗਰ ਸਿੰਘ ਮਾਨ ਪ੍ਰਧਾਨ, ਡਾ: ਅਮਨਦੀਪ ਸਿੰਘ ਟੱਲੇਵਾਲੀਆ ਅਤੇ ਕੁਲਵੰਤ ਸਿੰਘ ਧਿੰਗੜ ਮੀਤ ਪ੍ਰਧਾਨ, ਜਨਰਲ ਸਕੱਤਰ ਪਾਲ ਸਿੰਘ ਲਹਿਰੀ, ਸਹਾਇਕ ਜਨਰਲ ਸਕੱਤਰ ਲੈਕਚਰਾਰ ਸੁਖਮਿੰਦਰ ਸਿੰਘ ਸ਼ਹਿਣਾ, ਜਥੇਬੰਦਕ ਸਕੱਤਰ ਬਿੰਦਰ ਖੁੱਡੀ ਕਲਾਂ, ਸੁਦਰਸ਼ਨ ਗੁੱਡੂ ਤੇ ਅਵਤਾਰ ਸਿੰਘ ਸੰਧੂ, ਪ੍ਰਚਾਰ ਸਕੱਤਰ ਅਸ਼ੋਕ ਭਾਰਤੀ ਅਤੇ ਬੰਤ ਸਿੰਘ ਬਰਨਾਲਾ ਵਿੱਤ ਸਕੱਤਰ ਲਵਪ੍ਰੀਤ ਚਾਸ ਮਸ਼ਾਰਿਕ ਤੇ ਸਹਾਇਕ

ਸਕੱਤਰ ਲਛਮਣ ਦਾਸ ਮੁਸਾਫ਼ਿਰ ਤੇ ਸਹਾਇਕ ਵਿੱਤ ਸਕੱਤਰ ਬਲਵਿੰਦਰ ਸਿੰਘ ਠੀਕਰੀਵਾਲਾ ਚੁਣੇ ਗਏ। ਚੋਣ ਉਪਰੰਤ ਡਾ: ਉਜਾਗਰ ਸਿੰਘ ਮਾਨ ਨੇ ਦੱਸਿਆ ਕਿ ਇੱਕੀ ਮੈਂਬਰੀ ਕਾਰਜਕਾਰਨੀ ਦਾ ਐਲਾਨ ਅਗਲੀ ਸੂਚੀ ਵਿਚ ਕੀਤਾ ਜਾਵੇਗਾ।

Fourth News:-

ਅੱਸੀ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿਖਾਇਆ-ਸਿੱਧੂ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਮਾਲਵਾ ਸੱਭਿਆਚਾਰਕ ਅਤੇ ਵੈਲਫੇਅਰ ਕਲੱਬ ਬਰਨਾਲਾ ਵੱਲੋਂ ਚਲਾਏ ਜਾ ਰਹੇ ਸਿਲਾਈ ਸੈਂਟਰ ਦੀਆਂ ਦਸ ਵਿਦਿਆਰਥਣਾਂ ਨੂੰ ਸਿਖਲਾਈ ਸਰਟੀਫਿਕੇਟ ਵੰਡੇ ਗਏ। ਇਸ ਮੌਕੇ ਬੋਲਦਿਆਂ ਟਰੱਸਟ ਦੇ ਚੇਅਰਮੈਨ ਗੁਰਜਿੰਦਰ ਸਿੰਘ ਸਿੱਧੂ ਪ੍ਰਧਾਨ ਸਾਬਕਾ ਸੈਨਿਕ ਵਿੰਗ ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ ਦਲ ਅਤੇ ਸੈਂਟਰ ਸੰਚਾਲਕ ਜਗਸੀਰ ਸਿੰਘ ਚੌਹਾਨ ਨੇ ਦੱਸਿਆ ਕਿ ਨਿਸ਼ਕਾਮ ਤੌਰ 'ਤੇ ਇਹ ਸੈਂਟਰ ਪਿਛਲੇ ਦਸ ਸਾਲ ਤੋਂ ਚੱਲ ਰਿਹਾ ਹੈ ਅਤੇ 70 ਵਿਦਿਆਰਥਣਾਂ ਉਕਤ ਵਿਦਿਆਰਥਣਾਂ ਤੋਂ ਇਲਾਵਾ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿੱਖ ਕੇ ਆਪਣੀ ਰੋਜ਼ੀ-ਰੋਟੀ ਕਮਾਉਣ ਦੇ ਯੋਗ ਹੋ ਸਕੀਆਂ ਹਨ। ਸਮਾਗਮ ਵਿਚ ਉਚੇਚੇ ਤੌਰ 'ਤੇ ਪੁੱਜੇ ਟਰੱਕ ਯੂਨੀਅਨ ਬਰਨਾਲਾ ਦੇ ਪ੍ਰਧਾਨ ਕੁਲਵੰਤ ਸਿੰਘ ਕੰਤਾ ਨੇ ਸੰਸਥਾ ਦੇ ਕੰਮਾਂ ਦੀ ਸ਼ਲਾਘਾ ਕੀਤੀ ਅਤੇ 21 ਸੌ ਰੁਪਏ ਸਹਾਇਤਾ ਲਈ ਵੀ ਦਿੱਤਾ। ਇਸ ਮੌਕੇ ਨਗਰ ਕੌਂਸਲ ਬਰਨਾਲਾ ਦੇ ਪ੍ਰਧਾਨ ਸ: ਪਰਮਜੀਤ ਸਿੰਘ ਢਿੱਲੋਂ, ਕੌਮੀ ਤਰਕਸ਼ੀਲ ਆਗੂ ਬਲਵਿੰਦਰ ਬਰਨਾਲਾ, ਬੌਬੀ ਬਾਂਸਲ ਸਮਾਜ ਸੇਵੀ, ਸਾਬਕਾ ਚੇਅਰਮੈਨ ਸੁਖਮਹਿੰਦਰ ਸਿੰਘ ਸੁੱਖੀ, ਜਥੇਦਾਰ ਜਰਨੈਲ ਸਿੰਘ ਭੋਤਨਾ ਅਤੇ ਹਰਪਾਲਇੰਦਰ ਸਿੰਘ ਰਾਹੀ, ਸੁਖਜੀਤ ਕੌਰ ਸੁੱਖੀ, ਮਾਰਕੀਟ ਕਮੇਟੀ ਬਰਨਾਲਾ ਦੇ ਚੇਅਰਮੈਨ ਕਰਨੈਲ ਸਿੰਘ ਠੁੱਲੀਵਾਲ, ਸੈਨਿਕ ਵਿੰਗ ਦੇ ਸਰਕਲ ਪ੍ਰਧਾਨ ਕੈਪਟਨ ਬੂਟਾ ਸਿੰਘ ਸਹੋਤਾ, ਕੈਪਟਨ ਮਹਿੰਦਰ ਸਿੰਘ ਮਾਨ, ਪੰਜਾਬੀ ਗਾਇਕ ਜੈਸੀ ਬਾਜਵਾ ਤੋਂ ਇਲਾਵਾ ਹੋਰ ਬਹੁਤ ਸਾਰੀਆਂ ਸੰਸਥਾਵਾਂ ਦੇ ਆਗੂਆਂ ਨੇ ਆਪਣੀ ਹਾਜ਼ਰੀ ਲਵਾਈ। ਇਸ ਵੇਲੇ ਤਿੰਨ ਗਰੀਬ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਮਸ਼ੀਨਾਂ ਵੀ ਭੇਟ ਕੀਤੀਆਂ ਗਈਆਂ।

The English Translation of above four multi news of single document is given below:-

First News:-

Under leadership of Ghunnas development activities are highly accelerated-Bhana

Shhina, 8 January (motivational letters) – Under the leadership of local MLA Sant Balbir Singh Ghunnas development activities in the constituency are highly accelerated. These words are spoken by Bhagwan Singh Bhana youth leader and Committee member in the village Nankpura while distributing the cheques of shagun scheme. Bhagwan Singh said that local MLA has agreed to release the remaining amount soon to rest of families for shagun scheme and submersible pumps & grants are also given to the villages through Committee. On this occasion the Nankpura Sarpanch Sukhdev Singh, Pawan Kumar, Gurcharan Singh jaildar, Kaur Singh Pakhoke, Gurtej Singh Ghona, Jang Singh head taxi union, Jagsir Singh, Kartar Singh pakhoke etc. leaders were also present.

Second News:-

Tributes paid to teacher Gurdeep Singh Vdaich

Dhnaula, January 8 (private motivational letters) - The final prayer for outstanding teacher Gurdeep Singh Vdaich son of Sukhwinder Singh Vdaich was held at Gurudvara ninth Kingdom. On this occasion, different dignities presented flowers of worship. Jathedar Sadhu Singh Ragi former chairman market committee Bhadaur said that they had no words to console his parents for this painful death other than obeying the God's will. Untimely demise of such a promising child is very painful. Jagtar Singh Jangiana preacher Shiromani Gurdwara Parbandhak Committee Amritsar said not only Dhnaula residents but also the maternal village Jangiana were proud of Gurdeep Singh Vdaich. He was having less restlessness of children and more wisdom like elders. Sant Balvir Singh Ghunnas local MLA Bhadaur, sardar Bhola Singh Virk member general council Shiromani Akali Dal, Jathedar Baldev Singh Chungan, Jathedar Amar Singh B.A. member committee Shiromani Gurdwara Parbandhak Committee Amritsar, Jathedar Bharpoor Singh Dhnaula ex chairman, Gurvir Singh Guri youth congress leader, Iqbal Singh Jangiana committee member, Gamdoor Singh Mann leader broker association Dhnaula, Gurnam Singh ex head city council Dhnaula, Gurcharan Singh Kaler broker association, Bharpoor Singh former M.C. Surinder Singh Saddowalia district head Shiromani Akali Dal Amritsar, Chetan Singh mumm, Labh Singh Majjhuke, Jang Singh Jangiana, Karamjit Singh Neeta member district prishad, Harnek Singh former head co-operative assembly Jangiana, Manmohan Singh, Gurtej Singh chairman of panchayat, Raj Singh Nainewalia, Gurpreet Singh club leader, Ram Singh Dhindsa, Gurmeet Singh Shhina, Boota Sungh Burj, Bhagwan Singh Bhana, Harwinder Singh, Yadwinder Singh Walia M.C. Jagtar Singh Kaler head co-operative assembly Dhnaula,Gurpreet Singh Cheema etc. were present.

Third News:-

Eelection held for literature discussion forum Barnala

Barnala, January 8 (Staff Reporter) – Election of literature discussion forum held at Gurdwara Nanksar near the park of Green Avenue, in which unanimously Buta Singh Chauhan and Surjit Singh Dehd patron, Dr Ujagar Singh Mann head, Dr: Amandeep Singh Tallewalia, Kulwant Singh Dhingad circle head, General Secretary Pal Singh Lahiri, assistant general secretary lecturer Sukminder Singh Shhina, organisational secretary Binder Khuddi kalan, Sudarshan Guddu, Avtar Singh Sandhu, publicity secretary Ashok Bharti, Bant Singh Barnala, finance secretary Lakshman Das Musafir and assistant financial Secretary Balwinder Singh Thikriwala were elected. After election Dr Ujagar Singh Mann said that executive list of twenty one members will be announced in the next list.

Fourth News:-

Eighty girls were taught the sewing work-Sidhu

Barnala, January 8 (Staff Reporter) – the sewing centre run by malwa cultural and welfare club Barnala distributed the training certificates to ten students. While speaking on this occasion, trust charman Gurjinder Singh Sidhu head former military wing Shiromani Akali Dal and centre administrator Jagsir Singh Chauhan said that this centre has been running for the past ten years without desire for reward and besides the above mentioned students 70 more students have been able to earn their living after learning the sewing work. Specially reached on this occasion the head of truck union Barnala Kulwant Singh Kanta praised the tasks of organization and gave rupees twenty one hundred for help. On this occasion, other than Barnala city council's head Sardar Paramjit Singh Dhillon, the national logical leader Balwinder Barnala, Boby Bansal social servant, former chairman Sukmhinder Singh Sukhi, Jathedar Jarnail Singh Bhotna, Harpalinder Singh Rahi, Sukhjit Kaur Sukhi, chairman of market committee Barnala Karnail Singh Thuthiwal, military wing circle leader Captain Boota Singh Shota, Captain Mahender Singh Maan, Punjabi singer Jassy Bajwa many more leaders of organizations were present. During this sewing machines presented to three poor girls."

Algorithm Output at 30% Compression Ratio:-

ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ ਚ ਵਿਕਾਸ ਕਾਰਜਾਂ ਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ-ਭਾਨਾ

ਸ਼ਹਿਣਾ, 8 ਜਨਵਰੀ (ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਹਲਕਾ ਵਿਧਾਇਕ ਸੰਤ ਬਲਵੀਰ ਸਿੰਘ ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ ਹੇਠ ਹਲਕੇ ਦੇ ਵਿਕਾਸ ਕਾਰਜਾਂ ਵਿਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ ਹੈ।

ਅਧਿਆਪਕ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਨੂੰ ਸ਼ਰਧਾਂਜਲੀਆਂ ਭੇਟ

ਧਨੇਲਾ, 8 ਜਨਵਰੀ (ਨਿੱਜੀ ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਸੁਖਵਿੰਦਰ ਸਿੰਘ ਵੜੈਚ ਦੇ ਹੋਣਹਾਰ ਅਧਿਆਪਕ ਪੁੱਤਰ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਦੀ ਅੰਤਿਮ ਅਰਦਾਸ ਗੁਰਦੁਆਰਾ ਪਾਤਸ਼ਾਹੀ ਨੌਵੀਂ ਵਿਖੇ ਹੋਈ।

ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਹੋਈ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਗਰੀਨ ਐਵੀਨਿਊ ਦੇ ਪਾਰਕ ਨੇੜੇ ਨਾਨਕਸਰ ਗੁਰਦੁਆਰਾ ਵਿਖੇ ਹੋਈ, ਜਿਸ ਵਿਚ ਸਰਬ ਸੰਮਤੀ ਨਾਲ ਬੂਟਾ ਸਿੰਘ ਚੌਹਾਨ ਅਤੇ ਸੁਰਜੀਤ ਸਿੰਘ ਦਿਹੜ ਸਰਪ੍ਰਸਤ, ਡਾ: ਉਜਾਗਰ ਸਿੰਘ ਮਾਨ ਪ੍ਰਧਾਨ, ਡਾ: ਅਮਨਦੀਪ ਸਿੰਘ ਟੱਲੇਵਾਲੀਆ ਅਤੇ ਕੁਲਵੰਤ ਸਿੰਘ ਧਿੰਗੜ ਮੀਤ ਪ੍ਰਧਾਨ, ਜਨਰਲ ਸਕੱਤਰ ਪਾਲ ਸਿੰਘ ਲਹਿਰੀ, ਸਹਾਇਕ ਜਨਰਲ ਸਕੱਤਰ ਲੈਕਚਰਾਰ ਸੁਖਮਿੰਦਰ ਸਿੰਘ ਸ਼ਹਿਣਾ, ਜਥੇਬੰਦਕ ਸਕੱਤਰ ਬਿੰਦਰ ਖੁੱਡੀ ਕਲਾਂ, ਸੁਦਰਸ਼ਨ ਗੁੱਡੂ ਤੇ ਅਵਤਾਰ ਸਿੰਘ ਸੰਧੂ, ਪ੍ਰਚਾਰ ਸਕੱਤਰ ਅਸ਼ੋਕ ਭਾਰਤੀ ਅਤੇ ਬੰਤ ਸਿੰਘ ਬਰਨਾਲਾ, ਵਿੱਤ ਸਕੱਤਰ ਲਛਮਣ ਦਾਸ ਮੁਸਾਫ਼ਿਰ ਤੇ ਸਹਾਇਕ ਵਿੱਤ ਸਕੱਤਰ ਬਲਵਿੰਦਰ ਸਿੰਘ ਠੀਕਰੀਵਾਲਾ ਚੁਣੇ ਗਏ।

ਅੱਸੀ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿਖਾਇਆ-ਸਿੱਧੂ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਮਾਲਵਾ ਸੱਭਿਆਚਾਰਕ ਅਤੇ ਵੈਲਫੇਅਰ ਕਲੱਬ ਬਰਨਾਲਾ ਵੱਲੋਂ ਚਲਾਏ ਜਾ ਰਹੇ ਸਿਲਾਈ ਸੈਂਟਰ ਦੀਆਂ ਦਸ ਵਿਦਿਆਰਥਣਾਂ ਨੂੰ ਸਿਖਲਾਈ ਸਰਟੀਫਿਕੇਟ ਵੰਡੇ ਗਏ।

The English Translation of above output is as follows:

Under leadership of Ghunnas development activities are highly accelerated-Bhana

Shhina, 8 January (motivational letters) – Under the leadership of local MLA Sant Balbir Singh Ghunnas development activities in the constituency are highly accelerated.

Tributes paid to teacher Gurdeep Singh Vdaich

Dhnaula, January 8 (private motivational letters) - The final prayer for outstanding teacher Gurdeep Singh Vdaich son of Sukhwinder Singh Vdaich was held at Gurudvara ninth Kingdom.

Eelection held for literature discussion forum Barnala

Barnala, January 8 (Staff Reporter) – Election of literature discussion forum held at Gurdwara Nanksar near the park of Green Avenue, in which unanimously Buta Singh Chauhan and Surjit Singh Dehd patron, Dr Ujagar Singh Mann head, Dr: Amandeep Singh Tallewalia, Kulwant Singh Dhingad circle head, General Secretary Pal Singh Lahiri, assistant general secretary lecturer Sukminder Singh Shhina, organisational secretary Binder Khuddi kalan, Sudarshan Guddu, Avtar Singh Sandhu, publicity secretary Ashok Bharti, Bant Singh Barnala, finance secretary Lakshman Das Musafir and assistant financial Secretary Balwinder Singh Thikriwala were elected.

Eighty girls were taught the sewing work-Sidhu

Barnala, January 8 (Staff Reporter) – the sewing centre run by malwa cultural and welfare club Barnala distributed the training certificates to ten students."

As can be seen from output of algorithm of Punjabi summarizer, at 30% compression ratio, mainly the headlines and next lines have been retrieved and at 50% compression ratio, more detailed summary is produced including headlines, lines just next to head lines and other important lines.

IV. RESULTS AND DISCUSSIONS

Punjabi summarization system has been tested over fifty Punjabi multi news documents (Data set containing 6185 sentences and 72689 words) from Punjabi news-corpus. We have applied four Intrinsic measures of summary evaluation 1) F-Score 2) Cosine Similarity 3) Jaccard Coefficient and 4) Euclidean distance and two extrinsic measures of summary evaluation 1) Question Answering Task and 2) Keywords Association Task for

Punjabi multi news documents. Firstly we have produced gold summaries (reference summaries) of these 50 Punjabi multi news articles. For making the gold summaries, three human experts have been assigned the task of producing the manual summaries separately of these 50 documents at 10%, 30% and 50% compression ratios. Finally gold summaries (reference summaries) are produced by including mostly common sentences of three manual summaries produced by three human experts at their respective compression ratios.

As First measure of intrinsic summary evaluation, we have calculated F-Score [23] at respective compression ratios 10%, 30% and 50% for Punjabi news documents and Punjabi stories as follows:

$$\text{F-Score} = \frac{(2 * \text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

$$\text{Recall} = \frac{\text{Number of correct sentences retrieved by system}}{\text{Total number of sentences retrieved by human expert}}$$

$$\text{Precision} = \frac{\text{Number of correct sentences retrieved by system}}{\text{Total number of sentences retrieved by system.}}$$

As second measure of intrinsic summary evaluation [24], we have calculated Cosine Similarity between our system produced summary and gold summary at respective compression ratios for Punjabi news documents and Punjabi stories. Using Cosine-similarity-measure, documents are treated as term-vectors and the similarity of two documents corresponds to correlation between the vectors. Given two documents and vectors A and B are the term frequency vectors of these documents for term set T= $\{t_1,\ldots\ldots t_m\}$ Cosine similarity between two vectors is calculated as follows:

$$\text{COSINE\_SIMILARITY}(A, B) = \text{Cos}(\Theta) = (A \cdot B) / (|A|\,|B|) = \sum A_i \times B_i / \sqrt{\sum (A_i)^2} \times \sqrt{\sum (B_i)^2}$$

where i= 1 to n

Each dimension denotes the term with its frequency in the document and is non negative. The value of cosine similarity is non-negative and lies from 0 to 1. If cosine-similarity for two documents is closer to one, it means these two documents are very much similar to each other. For dissimilar type of documents cosine similarity is approaching towards zero. We have computed Cosine-similarity between our gold summary (reference summary) and summary produced by our Punjabi summarization system.

As third measure of intrinsic summary evaluation, we have calculated Jaccard-coefficient between our system produced summary and gold summary at respective compression ratios for Punjabi news documents and Punjabi stories. The Jaccard-coefficient measures similarity as the intersection divided by the union of the objects. Given two documents and vectors A and B are the term frequency vectors of these documents over the term set T= $\{t_1,\ldots\ldots, t_m\}$ then Jaccard-coefficient is calculated as follows:

$$\text{Jaccard Coefficient} = \text{SIM}(A, B) = (A \cdot B) / (|A|^2 + |B|^2 - A \cdot B) = (A \cdot B) / (\sqrt{\sum (A_i)^2} \times \sqrt{\sum (A_i)^2} + \sqrt{\sum (B_i)^2} \times \sqrt{\sum (B_i)^2} - A \cdot B)$$ Where i= 1 to n

Where each dimension represents a term with its frequency in the document. The value of Jaccard-coefficient-measure ranges from 0 to 1. If value of Jaccard-coefficient is approaching towards one then two documents are almost similar. If value of Jaccard-coefficient is approaching towards zero then two documents are dissimilar.

As fourth measure of intrinsic summary evaluation, we have calculated Euclidean distance between our system produced summary and gold summary at respective compression ratios for Punjabi news documents and Punjabi stories. Measuring distance between text documents, given two documents with their key term frequency vectors X_{ik} and X_{jk} respectively, where k= 1 to n key terms. The Euclidean distance of the two documents is defined as follows:

$$\text{Euclidean distance}(X_{ik}, X_{jk}) = \left(\sum (X_{ik} - X_{jk})^2 \right)^{1/2}$$ for k=1 to n key terms.

The results of intrinsic summary evaluation are shown in Table II. and Figure 5 at respective compression ratios.

TABLE II.
RESULTS OF INTRINSIC SUMMARY EVALUATION

Compression Ratio (In %)	Intrinsic Summary Evaluation for Punjabi News Documents			
	Avg. F-score (In %)	Avg. Cosine Similarity	Avg. Jaccard Coeff.	Avg. Euclidean Distance
10%	97.87	0.98	0.97	0.12
30%	95.32	0.96	0.95	0.32
50%	94.63	0.95	0.94	0.56

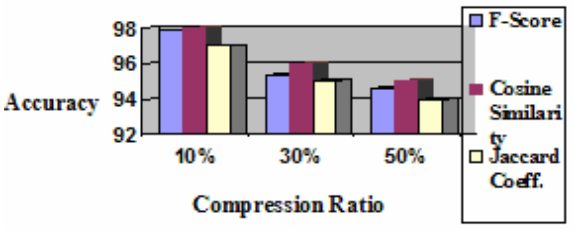


Figure 5. Intrinsic summary evaluation task

As can be seen from Table II and Figure 5, at 10% compression ratio, Average F-Score, Average Cosine Similarity and Average Jaccard Coefficient values are very high and Average Euclidean distance is very low because at 10%, usually few important sentences are extracted including headlines and next lines. Headlines and next lines are sufficient to describe the complete news document. The values of average F-Score, average cosine similarity and average Jaccard Coefficient are in descending order of compression ratios for Punjabi news

documents. Average value of Euclidean distance is in ascending order of compression ratios for Punjabi news documents. Few errors are due to presence of those sentences which contain many names of persons, but actually these sentences are not important.

Extrinsic measures [25] of summary evaluation are task oriented. We have performed question answering task and keywords association task as extrinsic measures of summary evaluation at compression ratios 10%, 30% and 50% respectively for Punjabi multi news documents. For performing the task of question answering, firstly three human experts have been given fifty multi news documents and then they jointly prepared five questions for each of fifty documents. Then answers of these questions are looked into system produced summary. For each correct answer, counter for number of correct answers is incremented by one for that document. Accuracy for performing task of question answering is calculated as follows:

Accuracy= No. of correct answers/ Total No. of questions asked

In keywords association task, keywords are the key terms which can represent the theme of whole document. For performing this task, firstly five keywords (gold keywords) have been extracted from source text by human experts and then these gold keywords have been associated with the summary produced by summarization system. Accuracy for performing task of keyword association is calculated as follows:

Accuracy= No. of gold keywords present in summary/ Total No. of gold keywords

The results of extrinsic summary evaluation are shown in Table III and Figure 6 at respective compression ratios.

TABLE III.
RESULTS OF EXTRINSIC SUMMARY EVALUATION

Compression Ratio (In %)	Extrinsic summary evaluation for Punjabi multi news documents	
	Accuracy of Question Answering Task (In %)	Accuracy of Keywords Association Task (In %)
10%	78.95	80.13
30%	81.38	92.37
50%	88.75	96.32

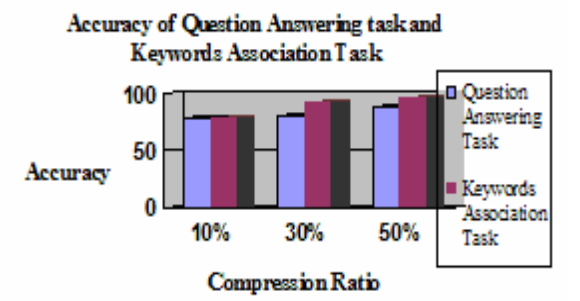


Figure 6. Extrinsic summary evaluation task

As can be seen from Table III and Figure 6 that at 10% compression ratio, Performance of multi news Punjabi Text Summarization System is low because news documents are usually short and at 10% CR, mainly headlines and lines just next to headlines are extracted which are not sufficient to give all answers of question-answering task. At 30% compression ratio, Punjabi Text Summarization System is able to give answers of 81.38% questions for Punjabi news documents. At 50% compression ratio, Punjabi Text Summarization System is able to give answers of 88.75% questions. Task of question answering is performed very well at 50% compression ratio with summary produced by Punjabi Text Summarization system because summary produced is enough to give answers of majority of questions.

At 10% compression ratio, average of 80.13% gold keywords are found in summary produced by Punjabi Text Summarization System for fifty Punjabi news documents. At 30% compression ratio, average of 92.37% gold keywords are found in summary produced by Punjabi Text Summarization System for fifty Punjabi news documents. At 50% compression ratio, average of 96.32% gold keywords are found in summary produced by Punjabi Text Summarization System for fifty Punjabi news documents. The accuracy percentage for the task of keywords association is low at 10% compression ratio because at 10% compression ratio, summary usually contains headlines and next lines and only few gold keywords are found in headlines and next lines. But at 50% compression ratio, the task of keywords association is performed very well because summary produced is enough to cover majority of gold keywords. The snap shot of Single document multi news Punjabi summarization system is given in Figure 7.

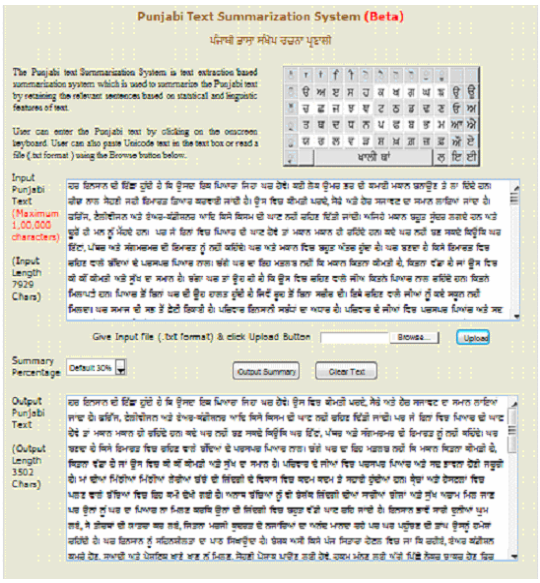


Figure 7. Web based online Punjabi text summarization system

V. COMPARISON OF PUNJABI TEXT SUMMARIZER WITH EXISTING INDIAN SUMMARIZERS

TABLE IV.
PERFORMANCE COMPARISON

Summarization Systems	Performance Comparison	
	Accuracy (In %)	Test Used
Single document multi news Punjabi Summarization System	For Multi News Single documents: F-Score = 95.32% Cosine Similarity= 96% Question Answering task with accuracy=81.38% (At 30% Compression Ratio)	Intrinsic and Extrinsic Summary Evaluation
Bengali Summarizer using Textual Images [14]	56%	Efficiency
Bengali Summarizer using Text Extraction [15]	84% (At 40% Compression Ratio)	Efficiency
Topic based Bengali Opinion Summarizer[16]	69.65%	F-Score
Multi Lingual Summarizer for English, Hindi, Gujarati & Urdu [17]	82%	Efficiency
Document Summarizer for Kannada [18]	For Literature: 70% For Entertainment: 80% For Sports: 76%	Efficiency

We can seen from Table IV, that performance of Punjabi Text Summarizer is reasonably good as compared with performance of other existing summarizers for Indian languages.

VI. CONCLUSONS

Single-document multi-news Punjabi Summarization system is first of its kind Punjabi summarizer and is available online at http://pts.learnpunjabi.org/. We have developed a number of lexical resources from scratch used in Punjabi text summarization such as Punjabi stemmer, Punjabi nouns normalizer, Punjabi named entity recognition, Punjabi Keywords Identification, Punjabi proper names list, common English-Punjabi nouns list, Punjabi stop words list, Punjabi suffix and prefix list, Punjabi cue phrase list etc. We have done thorough analysis of Punjabi corpus, Punjabi dictionary and Punjabi noun-morph for developing these resources. These Punjabi resources have been developed for the first time and these might be helpful for developing other NLP applications for Punjabi language.

REFERENCES

[1] F. Kyoomarsi, H.Khosravi, E. Eslami, P.K. Dehkordy, "Optimizing text summarization based on fuzzy logic", *In Seventh IEEE/ACIS International Conference on Computer and Information Science*, University of Shahid Bahonar Kerman, UK,pp. 347-352, 2008.

[2] V. Gupta and G.S. Lehal, "A Survey of Text Summarization Extractive Techniques", *International Journal of Emerging Technologies in Web Intelligence* vol.2, no.3, pp. 258-268, 2010.

[3] J. Lin, "Summarization", *In Encyclopedia of Database Systems*, Springer-Verlag Heidelberg, Germany, 2009.

[4] V. Gupta and G.S. Lehal, "Pre processing Phase of Punjabi Language Text Summarization", *In International Conference on Information Systems for Indian Languages Communications in Computer and Information Science*, Springer-Verlag Berlin Heidelberg, pp. 250–253, 2011.

[5] M.A. Fattah and F. Ren, "Automatic Text Summarization" *In World Academy of Science Engineering and Technology* vol. 27, pp.192-195, 2008.

[6] K. Kaikhah, "Automatic Text Summarization with Neural Networks", *In IEEE international Conference on intelligent systems*, Texas, USA, pp.40-44, 2004.

[7] J.L. Neto, A.D. Santos, C.A.A. Kaestner and A.A. Freitas, "Document Clustering and Text Summarization", *In 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, London, pp.41-55, 2000.

[8] V. Gupta and G.S. Lehal, "Features Selection and Weight learning for Punjabi Text Summarization", *In International Journal of Engineering Trends and Technology*, vol. 2, issue.2, pp. 45-48, 2011.

[9] V. Gupta and G.S. Lehal, "Named Entity Recognition for Punjabi Language Text Summarization", *In International Journal of Computer Applications*, vol.33 no.3, pp.28-32, 2011.

[10] M.S. Gill and G.S. Lehal, "Part of Speech Tagging for Grammar Checking of Punjabi", I*n The Linguistic Journal* vol.4, no.1, pp.6-21, 2009.

[11] V. Gupta and G.S. Lehal, "Automatic Keywords Extraction for Punjabi Language", *In International Journal of Computer Science,* vol. 8, issue 5, pp.327-331, 2011.

[12] V. Gupta and G.S. Lehal, "Punjabi language stemmer for nouns and proper nouns", *In the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP) IJCNLP*, Chiang Mai, Thailand, pp.35-39, 2011.

[13] V. Gupta and G.S. Lehal, "Complete Pre processing Phase of Punjabi Language Text Summarization", *In International Conference on Computational Linguistics COLING-2012*, IIT Bombay, India, pp.199-205, 2012.

[14] U. Garain, A.K. Datta, U. Bhattacharya and S.K. Parui, "Summarization of JBIG2 Compressed Indian Textual Images", *In Proceeding of 18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, Kolkata, India, 2006.

[15] K. Sarkar, "Bengali text summarization by sentence extraction", *In Proceedings of International Conference on Business and Information Management ICBIM'12*, NIT Durgapur, pp.233-245, 2012.

[16] A. Das and S. Bandyopadhyay, "Topic-Based Bengali Opinion Summarization", *COLING'10*, Beijing, China, pp.232–240, 2010.

[17] A. Patel, T. Siddiqui and U.S. Tiwari, "A language independent approach to multilingual text summarization", *In Proceedings of IEEE international conference RIAO2007*, Pittsburgh PA, U.S.A, 2007.

[18] R. Jayashree, M.K. Srikanta, K. Sunny, "Document Summarization in Kannada using Keyword Extraction", *In Proceedings of AIAA'11,CS & IT 03*, pp.121–127, 2011.

[19] M.Z. Islam, M.N. Uddin and M. Khan, "A light weight stemmer for Bengali and its Use in spelling Checker", *In Proceedings of 1st International Conference on Digital Comm. and Computer Applications (DCCA 2007)*, Irbid, Jordan, PP.19-23, 2007.

[20] A. Ramanathan and D. Rao, "A Lightweight Stemmer for Hindi", *In Workshop on Computational Linguistics for South-Asian Languages, EACL'03*, 2003.
[21] G. Singh, M.S. Gill and S.S. Joshi, "Punjabi to English Bilingual Dictionary", Punjabi University Patiala, India, 1999.
[22] V. Gupta and G.S. Lehal, "Automatic Punjabi Text Extractive Summarization System", *In International Conference on Computational Linguistics COLING-2012*, IIT Bombay, India, pp.191-198, 2012.
[23] H. Nanba and M. Okumura, "Some Examinations of Intrinsic Methods for Summary Evaluation Based on the Text Summarization Challenge", *In Proceedings of international conference on language resources and evaluation LREC'02*, pp.739-746, 2002.
[24] A. Huang, "Similarity Measures for Text Document Clustering", *In the Proceedings of New Zealand Computer Science Research Conference*, Christchurch New Zealand, pp.49-56, 2008.
[25] M. Hassel, "Evaluation of Automatic Text Summarization", *Licentiate Thesis*, Stockholm, Sweden, pp.1-75, 2004.

AUTHORS' INFORMATION

Dr. Vishal Gupta is Senior Assistant Professor in Computer Science & Engineering, University Institute of Engineering & Technology, Panjab University Chandigarh, India. He did his BTech. in Computer Science & Engineering from Shaheed Bhagat Singh College of Engineering & Technology, Ferozepur, Punjab in 2003. He did his M.Tech. and completed his Ph. D. in Computer Science & Engineering from Punjabi University Patiala in 2005 and 2013 respectively. He is among University toppers. He is winner of Young Scientist Award-2013 in Engineering & Technology at Punjab Science Congress. He has written around 41 research papers in reputed international and national journals and conferences. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research paper on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10^{th} and 12^{th} classes of Punjab School education board.

Professor Gurpreet Singh Lehal received undergraduate degree in Mathematics in 1988 from Panjab University, Chandigarh, India, and Post Graduate degree in Computer Science in 1995 from Thapar Institute of Engineering & Technology, Patiala, India and Ph. D. degree in Computer Science from Punjabi University, Patiala, in 2002. He joined Thapar Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he joined Department of Computer Science at Punjabi University, Patiala. He is actively involved both in teaching and research. His current areas of research are- Natural Language Processing and Optical Character recognition. He has published more than 25 research papers in various international and national journals and refereed conferences. He has been actively involved in technical development of Punjabi and has to his credit the first Gurmukhi OCR, Punjabi word processor with spell checker and various transliteration software. He was the chief coordinator of the project "Resource Centre for Indian Language Technology Solutions- Punjabi", funded by the Ministry of Information Technology as well as the coordinator of the Special Assistance Programme (SAP-DRS) of the University Grants Commission (UGC), India. He was also awarded a research project by the International Development Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi Transliteration Solution for Networking.

JAPL: the JADE Agent Programming Language

Mohamed BAHAJ, Abdellatif SOKLABI

FSTS/ Department of Mathematics and Computer Science, University Hassan I

Settat, Morocco

mohamedbahaj@gmail.com

abd.soklabi@gmail.com

***Abstract*— This article describes JADE Agent Programming Language (JAPL) that allows fast and efficient implementation of intelligent behaviors into mobile agents, based on three logical, FIPA speech acts, and a part of complex procedural script for actions. It integrates the ontologies and defines communication services. Rather than rely on a library of plans, JAPL allows agents to plan from first principles. It also describes how to program the multiple JADE behaviors using JAPL instructions and how to compile JAPL to JAVA classes.**

***Keywords*— Mobile agents, JADE, Agent programming language, agents communication, MAS**

I. INTRODUCTION:

JADE (Java Agent Development Environment) is the system for mobile agents, most used in the world. It is adapted to the rules of FIPA [10] and is programmed as a basic object-oriented programming language Java, but the development of a complex application is not yet clear and it requires a lot of concentration, effort and time [4].

In the literature, many programming languages for mobile agents ware created as: Cougaar [6], MetateM [7], AgentSpeak (L) [8], [15], MadKit [12], 3APL [9], [13], Golog [14], [17] and SWAM [3]. These languages are used in the prototyping phase, in order to provide a high level application design-based to mobile agents, who try to follow the architecture Belief- Desire-Intension [16, 15].

In this paper, we present the programming language of JADE agents: "JADE Agent Programming Language". It was designed to implement many complex applications by assigning high mental level concepts to mobile agents, so they reach a new abstraction level that allows their programming as reactive behaviours instead of programming as meaning. In addition to these features, the agents will have the ability to schedule other tasks in order to accomplish the tasks that have been assigned.

First of all, section 2 presents broad overview of the different items JADE Agent Programming Language. Secondly, section 3 talks about the state of the art. Thirdly, section 4 describes its different features in some detail. In particular, we highlight knowledge representation. The following section contains the programming treatment agent's behaviours. The sixth section finalizes this part with a focus on the service concept, and the last section we wrap up with some conclusions.

II. JADE AGENT PROGRAMMING LANGUAGE OVERVIEW:

JAPL is the extension of ADL (Action Description Language) [18] on the mobile agents of JADE. ADL allows the use of quantifiers and conditional expressions in the description of operators to describe classical planning. In ADL, the focus of the quantifiers is over finite domains and disjunction that are not allowed in the expression of effects, because it would bring non-deterministic actions. ADL is a good representation of formal references in classical planning. JAPL is a computer language that standardizes description of planning problems. Also it is a PDDL (Planning Domain Description Language [15]), which allows it to specify the possible operators, the relationships and environmental constraints, the initial state and the goal states.

JAPL provides open world semantics. It includes four essential elements; plans element, rules, ontologies and services. One of the main features for agents is that they can communicate via services. Taking into consideration the view of the agent about the execution, a service call is handled in the same way as the execution of internal actions. This is possible because the services have the same structure as actions. They obtain; pre-conditions and post-conditions, and the body (which contains the actual code to be executed) is reduced to service calls, which allows us to integrate advanced features such as safety, in JADE. In addition, the programming of complex communication scenarios becomes easier, because all messages are processed in a clearly defined framework. In the following sections we will detail some of JAPL's different regions, namely knowledge representation, the behavior of agents, and communications.

doi:10.4304/jetwi.5.3.272-277

III. STATE OF THE ART:

In the literature there are many programming languages of agents, represent a family of programming languages which allows developers to create high-level abstractions and structures which are necessary for the implementation of mobile agents. It is possible to classify them in two categories, those based on logic as AgentSpeak (L) [15], [8], 3APL [9], [13], Golog [17], MetateM [7] and SWAM [3], and those based on object-oriented programming language Java such as Cougaar, [6] and MadKit [12]. These languages are mostly in the prototype stage, and provide high-level concepts that implement some notion of BDI [16], [1].

AgentSpeak (L) programming language is the most developed programming language, and is an article comparing this language to other languages [15]. ALAS and IndiGolog are the newest programming languages of mobile agents. ALAS is a fast, effective, simple and powerful programming language of reactive agents. It has been designed to support the execution of agents in heterogeneous environments, and allow easy use of features of agents, such as mobility [2]. IndiGolog is a programming language for autonomous agents that detect their environment and plan their tasks. IndiGolog supports the execution of high-level programs, gives programmers the ability to create high-level and non-deterministic programs, tests agent tasks and provide a declarative specification of the domain in calculated situations [5]. However, all these languages are widespread and are not suitable for the distinction of all mobile agents systems. We used the strong point in these programming languages for mobile agents and created a simple and efficient language, while respecting the particularity of JADE.

IV. THE KNOWLEDGE REPRESENTATION:

JAPL has been designed to respond to the need of JADE, for a flexible and dynamic language, that allows for the migration of agents and services at any time and which makes the local information's validity very short. So every programming system that supports dynamic behavior needs to address the issue of synchronization and information sharing. So research in the domain of transaction management tries to find solutions to the issue of synchronization [11].

Our approach, however, is to integrate the idea of uncertainty about the bits of information in the presentation of knowledge to allow programmers to manage incorrect or expired information. We integrated the concept of uncertainty using calculated situations which have three assessed logics. The third truth value is added to the predicate and cannot be evaluated with the information available to a particular agent. Thus, a predicate can be explicitly evaluated as unknown. It is an integral part of language, and the programmer is forced to deal with uncertainty in the development of a new agent. Therefore, JAPL can handle incomplete or incorrect information explicitly.

JAPL allows knowledge bases that are most used in the other languages to be defined. Each object that refers to the language needs to be defined in ontology. JAPL implements strong typing, because the variables flow in classes rather than sets of speech. It should be noted here that the alphabet JAPL consists of variables, functions, symbols, actions, quantifiers, connectors, and punctuation symbols. We also use "?" (for testing), "!" (for realization), "and" (for sequencing) and "1" (for implication). Classes are represented in a tree structure. Each node represents a class with a set of attributes. Classes are defined as follows Fig. 1:

```
(Class object classname)
```

Fig.1: Example of JAPL class representation.

JAPL allows multiple inheritances. The classes inherit all attributes of all ancestors. Therefore, the problem of names conflicts are not posed, since the attribute names have been expanded to include the class structure. In addition to classes, JADE Agent Programming Language can define methods and comparisons. The interpretation of the methods is given by the operational semantics. In practice, methods are coded in JAVA.

Complex actions describe functional capabilities of agents. They in turn can call Java methods, or use JAPL. There are different types of complex actions or invocations of services. They all have the same structure; they consist of three main elements, in addition to the name of the action Fig. 2:

```
(action ActionName pre PreCondition eff Effect Body)
```

Fig. 2: Example of JAPL actions representation.

An action is called using the following code Fig. 3:

```
(call role-interface action-id [variable])
```

Fig. 3: JAPL action call example

V. AGENT BEHAVIOUR:

A. *SimpleBehaviour and Action selection:*

As JADE Agent Programming Language is intended to be interpreted in a BDI-like architecture, it incorporates the notion of achievement agent tasks SimpleBehaviour. The SimpleBehaviours of the agent are implemented as

simple commands that the agent tries to respond to once they are activated. Each agent that carries one SimpleBehaviour starts to run by trying to appropriate recovery actions that replay at this SimpleBehaviour. These actions can be either simple scripts or services that are provided by other agents. For this selection, there is no difference between actions that can be performed locally and services. The final selection is made by comparing the orders listed in SimpleBehaviour with the effects of all the known actions of the agent. Comparisons and formulas are made in order to check their compatibility. If they are compatible, the values of variables of SimpleBehaviour agents are linked to the corresponding variables in the action. After the end of the action, the results are written to the original variables of the agent task's orders are evaluated to ensure that the objective of SimpleBehaviour is actually reached. If not, the agent reformulates the composition of its actions, and tries to reach the goal of SimpleBehaviour with other actions.

B. *Reactive Behaviour:*

JADE Agent Programming Language sets rules that control the execution of reactive behaviors of the agent. More specifically, a rule may give the agent a task, whenever a certain event occurs. Rules are applied simply, consisting of a condition and two actions, one of which is executed when the condition is true and the other when it is false Fig. 4.

```
(rule <name>
(var variables-declaration)
conjunction
(true 1)
(false 0)
)
```

Fig. 4: Example of a JAPL rule.

Precisely, whenever an object is added, deleted or modified in the facts, conditions that correspond to the type of the object in the fact are tested and executed. If the result of the test is unknown, no action is taken. For efficiency reasons, the restriction rules that apply to the types of objects have been designed to make it as simple as possible. Actions can themselves be tasks of a new agent or a call to an agent class.

C. *CompositeBehaviour composition and activities:*

In general, the concept of having beliefs, desires and intentions are translated into knowledge belief bases, tasks of agent and a composition library. This allows us to create some principle principles. All these languages require a library of fully developed composition. Overall execution cycle thus consolidates the internal and external states via conjunction function with one or more compositions, which are partially or fully implemented.

In order to achieve a complete composition, the partial compositions must be ordered consistently. As scheduling can be computationally expensive, the algorithm ensures that the main sequence of actions that depend on each other is satisfied. Actions that are executed in parallel are not checked for consistency. Elements of the composition may take a number of forms, which include actions or services as we will detail in the next section.

The agent task execution is discussed as follows. The locally known elements of the actions composition are compared to the agent task. If it finds one with the pre-conditions satisfied, it runs. If not, the composer tries to find others who can satisfy the prerequisites. If the task agent or certain pre-conditions cannot be met with local composers, a request is sent to the Directory Facilitator (DF). If the action is found in the composer, the Directory Facilitator will execute the action. If not a new action is created and executed Fig. 5.

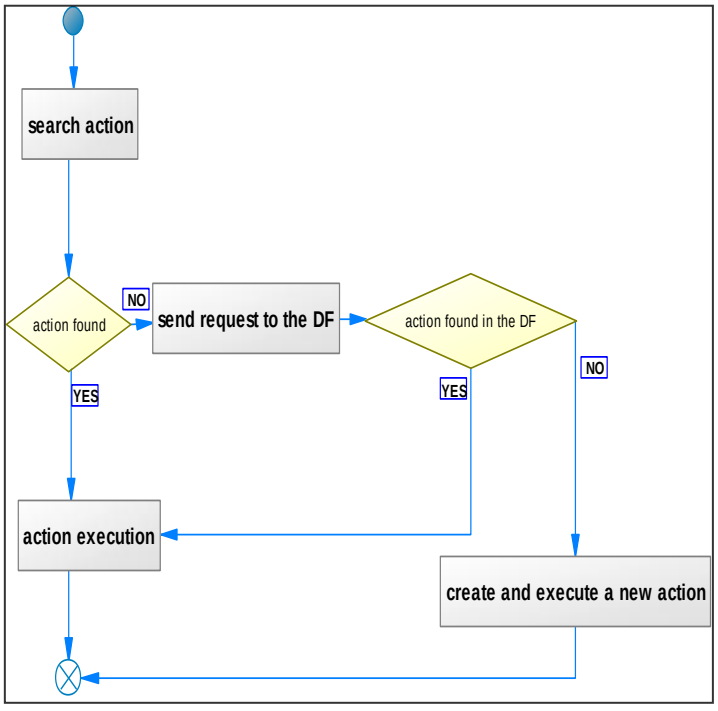


Fig. 5 agent task execution algorithm

Actions may be atomic elements, or they may be scripts. JAPL provides keywords for sequential execution, parallel or conditional, and even the creation of new agent tasks, which then lead to the composition of the new shares.

D. *SequentialBehaviour and ParallelBehaviour:*

JADE Agent Programming Language can define behaviour using sequential blocks that run sequentially, i.e. that all instructions are executed one sequentially. If any

of the statements fail for any reason, the entire block fails. Note that the entire composition does not necessarily fail (Fig. 6).

```
(sequential
(bind ?sender (obj Agent (name
"sender")))
(bind ?receiver (obj Agent (name
"reiceiver")))
(eval (and
(att name ? sender?str)
(not (att name ? receiver?str))
)
)
)
```

Fig. 6 an example of a SequentialBehaviour JAPL

JADE Agent Programming Language offers the possibility of defining Behaviour using parallel blocks running simultaneously, allowing each to be treated separately. If one of the statements must wait for any reason, the other can still be executed without delay. However, the entire execution fails if one of the statements fails, regardless of whether others have been completed or not (Fig. 7).

```
(parallel
(bind ?sender (obj Agent (name
"sender")))
(bind ?receiver (obj Agent (name
"reiceiver")))
(eval (and
(attribut name ? sender?string)
(not (attribut name ? receiver?string))
)
)
)
```

Fig. 7: Example of a ParallelBehaviour JAPL

VI. THE SERVICE CONCEPT:

To allow JADE to be more interoperable with other SMAs, the agents created by JAPL should only communicate using service calls, instead of having an implicit representation of features that can be used by other agents.

All interactions between agents are guided by a generic service. Thus, a service describes an act that the agent performs on behalf of another agent. Services are specified and defined using those conditions and effects. The interaction service always occurs between two agents. An agent must be either the user or the supplier during this operation. The provider is the agent that has some expertise to offer. The other agent in the interaction may act as the service user.

To initiate a service call, the agent must have failed to fulfill a task using only the actions that are available; this includes services that agents provide. If such a situation occurs, the agent sends a request to the DF, which responds with a list of services that could fulfill the task. Then the agent choose a service, and inform the Directory Facilitator, which sends back a return list, but this time a list of agents which provide the requested service.

VII. APPLICATION:

To compile the instructions to create the help in JAPL we had to create a Plugin compilation based on eclipse Plugins. In our case we used the 4.2.1 version of Eclipse. And to build ontology, we call it compilation Plugin. Which we integrate with JAR files JADE (Fig. 8 and Fig. 9).

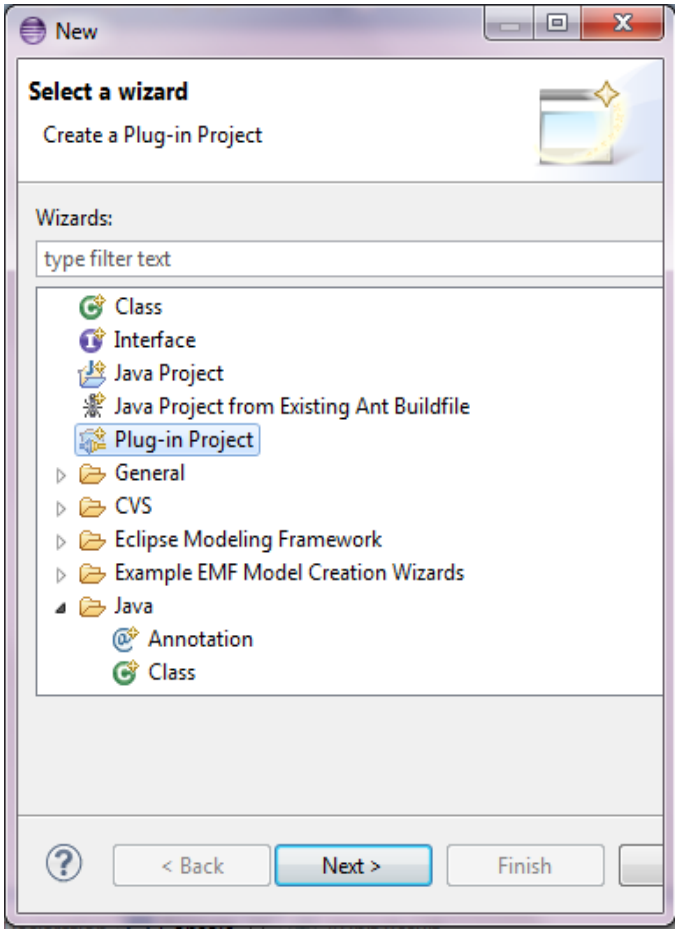


Figure 8: creating a Plugin using Eclipse

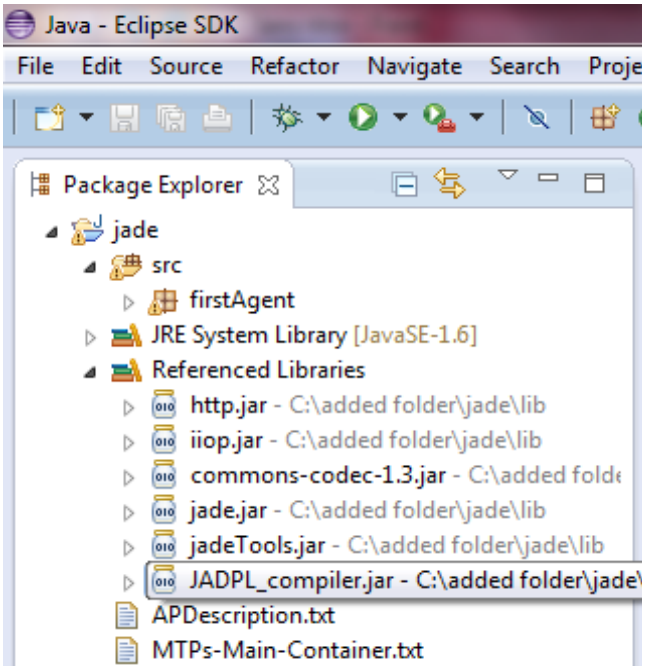


Fig. 9: integration JAPL compilation plugin in the JADE project library.

The compiler takes JAPL input and creates Java classes. These classes must be compiled subsequently using JAVAC. The generated classes implement some JADE interfaces architecture to insure this compilation.

Created classes are:

• OntologyName.java. This class contains the basic structure for classes and their attributes.

• OntologyNameIinterface.java. Java interface contains constants that define types of ontology for each class and attributes.

• OntologyNameMethod.java This class contains the JAVA code for such methods that were written by the programmer without modification.

More details on the use of language JAPL will be available in a users' guide, which soon will be posted on the web soon.

VIII. CONCLUSION:

In this article, we presented JAPL. On the basis of tertiary logic, it provides constructs for describing ontologies, protocols and services, and complex actions for the system of mobile agents JADE. Programmers can use JAPL composition of actions composer. Thus, internal actions and invocations of services are handled transparently to the planning component.

We are confident that JADE agent programming language is likely to be more fruitful than all old languages, in bridging the gap between theory and practice in the development of the JADE mobile agents. Further we are confident that it will push the research in both the pragmatic and theoretical aspects of BDI agents.

REFERENCES

[1] Dennis, M. Fisher, P. Webster, R. Bordini "*Model checking agent programming languages*" in Automated Software Engineering, March 2012, Volume 19, Issue 1, Pages 5-63

[2] D. Mitrovic, M. Ivanovic and M. Vidakovic "*Introducing ALAS: A Novel Agent-Oriented Programming Language* " in NUMERICAL ANALYSIS AND APPLIED MATHEMATICS ICNAAM 2011: International Conference on Numerical Analysis and Applied Mathematics , AIP Conf. Proc. 1389, Pages 861-864, 19–25 September 2011

[3] M. Crasso, C. Mateos, A. Zunino, M. Campo "*SWAM: A logic-based mobile agent programming language for the Semantic Web*" in Expert Systems with Applications Volume 38, Issue 3, March 2011, Pages 1723–1737.

[4] F. Bellifemine, G. Caire, T. Trucco, G. Rimassa "*JADE PROGRAMMER'S GUIDE*" last update: 08-April-2010. JADE 4.0

[5] Giuseppe De Giacomo, Yves Lespérance, Hector J. Levesque, Sebastian Sardina "IndiGolog: a High-Level Programming Language for Embedded Reasoning Agents" Multi-Agent Programming: 2009, Pages 31-72

[6] Helsinger, M. Thome and T. Wright "*Cougaar: A scalabe, distributed multi-agent architecture*" in IEEE SMC04, 2004.

[7] M. Fisher, C. Ghidini and B. Hirsch "*Programming groups of rational agents*" in Fourth International Workshop. Volume 2359 of LNAI. 2004, Pages 16–33.

[8] R. Bordini, J. Hubner et A. Jason: "*a Java Based AgentSpeak Interpreter Used with SACI for Multi-Agent Distribution over the Net*" in 5th edn, 2004.

[9] M. Dastani "3APL Platform" Utrecht University, 2004.

[10] FIPA: Fipa acl message structure specification, 2002.

[11] R. Kotagiri, J. Bailey, P. Busetta "*Transaction oriented computational models for multi-agent systems*" in 13th IEEE International Conference on Tools with Artificial Intelligence, IEEE Press, 2001, Pages 11–17.

[12] O. Gutknecht, J.Ferber "The *madkit agent platform architecture*" in Technical Report, Laboratoire d'Informatique, de Robotique et de Micro électronique de Montpellier, 2000.

[13] Hindriks, K.V., Boer, F.S.D., der Hoek, W.V.and J.J.: Meyer "*Agent programming*" in 3apl. Autonomous Agents and Multi-Agent Systems 2 (1999), Pages 357–401.

[14] G. Giacomo, Y. Lesperance, H. Levesque " *Congolog, a concurrent programming language based on the situation calculus*" Technical report, University of Toronto,1999.

[15] A. Rao: *AgentSpeak(L): "BDI agents speak out in a logical computable language*" in 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'96,. Volume 38 of Lecture Notes in Computer Science, 1996, Pages 42–55

[16] S. Rao *"BDI Agents: From Theory to Practice*" Technical Note 56. Australian Artificial Intelligence Institute, April, 1995.

[17] M. Huntbach, N. Jennings and G. Ringwood. "*How agents do it in stream logic programming*" in Proceedingsof the International Conferenceon Multi-Agent Systems, San Francisco, USA, June, 1995.

[18] E. Pednault "*ADL and the state transition model of action*" in logic and computation, 1994, Pages 467-512.

Bahaj Mohamed was born in 1964, in ouezzane, Morocco. He got his PhD in Applied Mathematics, from University of Pau, France, in 1993. He is now working as a Professor at the Department of Mathematics & Computer Sciences, University of Hassan 1er, Faculty of Sciences & Technology of Settat, Morocco. His research interests include pattern recognition, Load Balancing & Controls of mobiles agents, Semantic web & Ontology in MAS.

Soklabi Abdellatif was born in 1985, in El JADIDA, Morocco. He had a license degree in computer engineering in 2009 and a master's degree in computer systems and networks in 2011. Now he is a PhD researcher in mobiles agents and web services in Department of Mathematics & Computer Sciences, University of Hassan 1er, Faculty of Sciences & Technology of Settat, Morocco. His research interests include, Load Balancing & Controls of mobiles agents, Interoperability between different MAS.

A Direction-Based Vertical Handoff Scheme

Abdelnasser Banihani,[a] Mahmoud Al-Ayyoub[a] and Ismail Ababneh[b]
[a] Jordan University of Science and Technology, Irbid, Jordan
Email: abed_banihani@yahoo.co.uk, maalshbool@just.edu.jo
[b] Al al-Bayt University, Mafraq, Jordan
Email: ismael@aabu.edu.jo

***Abstract*— In heterogenous wireless networks, as the users move across the coverage regions of possibly-different wireless networks, they will have to switch between them. The procedure followed to determine when and how a mobile user should switch between networks of different types is known as the vertical handoff scheme. Several vertical handoff schemes have been proposed in the literature, but few of them employ the geographical nature of this problem like we do in this paper. The scheme we propose here takes the user's direction of movement into account when choosing the most suitable candidate for the handoff. When compared with existing schemes, our proposed scheme shows significant reductions in the number of lost connections and the number of unnecessary handoffs.**

***Index Terms*— Heterogenous Wireless Networks, Vertical Handoff Scheme, Mobility Model**

I. Introduction

OVER the past couple of decades, the demands of mobile users have increased significantly and the nature of these demands has shifted from making simple voice calls to running applications with high bandwidth requirements. Satisfying these demands for mobile users is a very challenging problem that requires taking advantage of the many available networks (of different types). Such heterogenous wireless networks have different access technologies, architectures, protocols, operators and users [1]. Examples include the Wideband Code Division Multiple Access (WCDMA) Networks and the Wireless Local Access Networks (WLANs). Such variations have made it challenging to deal with heterogeneous wireless networks, organize them, and enable effective interaction and information sharing to provide mobile users with high quality connections. The different service demands of mobile users have made the Always Best Connected (ABC) concept important so as to allow a mobile user to get connections and services using the devices and access technologies that best suit the mobile user's communication needs as the user crosses different geographical regions covered by the different networks [2].

The *Handoff* is the process of moving a user's communication session from an access device (such as an Access Point (AP) or a Base Station (BS)) to another (in most cases, to an adjacent one) to guarantee uninterrupted communication [1]. In other words, a handoff is defined as changing the frequency, time slot and spreading code of the channel used without effecting the active session [3]. The handoff process aims at guaranteeing

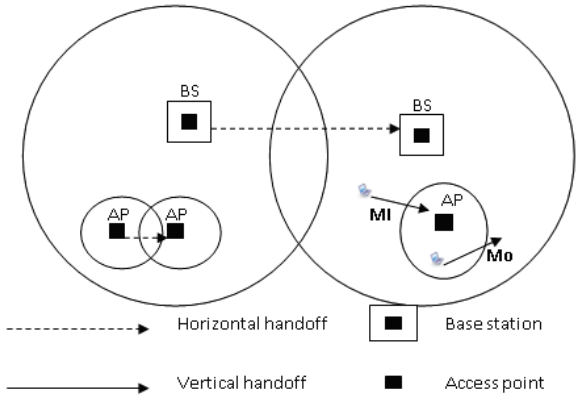


Figure 1. The different types of handoff [5]. The figure shows horizontal handoffs (between two base stations (BSs) and between two access points (APs)) as well as vertical handoffs (where mobile users move into (MI) or move out (MO) of the AP's coverage region causing handoffs between the BS and the AP).

that a mobile user's application work properly while the user is moving from one location to another. Two types of handoff have been in use: the intra-technology handoff (*horizontal* handoff) and the inter-technology handoff (*vertical* handoff). Examples of both types of handoff can be seen in Figure 1. The figure shows the two main scenarios considered in vertical handoff. The first one is moving out (MO) of the preferred network and the second one is moving into (MI) a preferred network. Note that when switching to a different network, there may be a preferred network to switch to among the list of candidate networks (e.g., WLAN is normally preferred over a Universal Mobile Telecommunications System (UMTS) network [4]).

A handoff process can be divided into three phases: the initiation phase (radio link transfer), the decision phase and the execution phase [4]. During the initiation phase, information about access technologies, mobile users, environment and neighbors is collected. Examples of such information include Received Signal Strength (RSS) from other neighbors, Signal to Interference and Noise Ratio (SINR), distance from access devices, direction and velocity of mobile users, etc. This information will be used in the decision phase to select the best new network for handoff. This will be the main topic of this paper.

Several parameters have been proposed in the liter-

doi:10.4304/jetwi.5.3.278-286

ature for use in vertical handoff algorithms [4]. Examples include RSS, SINR, connection time, handover latency, available bandwidth, power consumption, user preferences, monetary cost, and security. In this work, we focus on reliability. When a handoff request by a mobile user fails (request is denied due to unavailability of free channels at the chosen handoff candidate) the user is disconnected. Such disconnections are intolerable in cellular networks. In fact, users prefer networks with lower bandwidth if they are more reliable (i.e., have lower disconnection probability) [6]. Additionally, we focus on providing better QoS guarantees by reducing the number of unnecessary handoffs [6].

Except for a few works, existing schemes ignore the geographical nature of this problem unlike our scheme. In our scheme, we incorporate the mobile user's movement direction in the handoff decision. By doing so, we can decrease the probability of an unnecessary handoff. Another benefit of this approach is reducing the number of handoffs to candidates with "central" locations that are close to the movement trajectories of many mobile users. This will decrease the load on these candidates, and thus, decrease the number of disconnections. These intuitive arguments of why our scheme will outperform other schemes are supported by the experiments discussed in Section V.

The paper is organized as follows. In the following section, we discuss the related works before describing the system model we use and our assumptions we make in Section III. We present our scheme in Section IV and show its performance advantage over existing schemes in Section V. Finally, we conclude our paper and discuss futures directions of this work in Section VI.

II. Related Works

Due to its importance, several vertical handoff schemes have been proposed in the literature. Below, we review these schemes.

A. RSS-Based Schemes

Zahran et al. [5, 7] proposed an adaptive lifetime-based vertical handoff (ALIVE-HO) scheme. This scheme uses the RSS to estimate the expected period of time during which the mobile user's need can be served from WLAN taking into account delay, authentication, and service initiation. Application Signal Strength Threshold (ASST) is defined as the RSS needs of applications to perform their services. In [5], a framework is proposed to evaluate the performance of the ALIVE-HO scheme. The simulation results show the tradeoff between resource utilization and the user received QoS. The authors show that by introducing the lifetime metric, the algorithm adapts to application requirements and user mobility, reducing the number of unnecessary handoffs, and improving the average throughput provided to the user because the algorithm increases the connected-duration and decreases the number of dropped users.

Yan et al. [8, 9] proposed a scheme to minimize the unnecessary handoffs and to improve the overall network utilization based on a traveling distance prediction method within a WLAN cell. The scheme uses RSS measurements to predict the time that user will spend within a WLAN cell. Their performance analysis showed that the main advantage of this scheme is that it minimizes the probability of handoff failures and unnecessary handoffs whenever the predicted traveling distance inside the WLAN cell is smaller than the distance threshold value.

Mohanty et al. [10] proposed a vertical handoff management scheme to support smooth vertical handoff management in next generation wireless systems. A cross-layer (layer 2 + layer 3) vertical handoff management protocol (CHMP) uses two RSS values from measurements of the current RSS and a dynamic RSS threshold, which is calculated by estimating user speed and predicting the handoff signaling delay of possible handoffs between a WLAN and 3G cellular networks.

Yang et al. [11] proposed a Multi-dimensional Adaptive SINR based Vertical Handoff scheme (MASVH) scheme. This scheme tries to balance the effect of SINR, required user bandwidth, user traffic cost and network utilization to improve handoff decisions by taking into account the effect of multi-attributes QoS support. The simulation results show that MASVH improves system performance by enhancing the throughput and decreasing the failed handoff probability as well as the user's traffic cost.

B. Bandwidth-Based Schemes

Ayyappan and Kumar [12] proposed a QoS-based vertical handoff scheme that depends on the available bandwidth and the user's service requirements to make vertical handoff decision between WLANs and Wireless Wide Area Networks (WWANs).

Yang et al. [6] proposed a bandwidth-based vertical handoff scheme for WLAN and WCDMA networks. This scheme uses the effect of combined SINR as a main criterion for making handoff decisions. It converts the SINR value at the access network to an equivalent value at a target network so that the handoff algorithm can determine achievable bandwidths from both access networks so as to make handoff decisions considering QoS requirements.

Ayyappan et al. [13] proposed an SINR-based vertical handoff scheme for QoS in heterogeneous wireless networks. This scheme uses SINR to improve the QoS in heterogeneous wireless networks as compared with the RSS-based vertical handoff scheme. This scheme uses SINR in calculating the throughput using Shannon's capacity theorem. The handoff is initiated when the mobile user receives a higher equivalent SINR from another network. The user connects to the network that provides better QoS. Simulation results show that the proposed SINR-based vertical handoff scheme provides higher overall system throughput as well as fewer dropped connections.

C. Other Schemes

Xia et al. [14] proposed a novel fuzzy logic vertical handoff scheme with the assistance of differential distance and a pre-decision method. This scheme makes handoff decisions between WLAN and Universal Mobile Telecommunications Systems (UMTS). The scheme consists of the following parts:

- The predictor of a Forward Differential Distance Algorithm (FDPA) that is used to get the expected next RSS.
- A Pre-Decision (PD) method applied before the handoff decision to filter unnecessary data (i.e., mobile users with high mobility or less RSS from using the WLAN) to improve the vertical handoff decision.
- The Fuzzy logic based Normalized Quantitive Decision (FNQD) method implemented to quantitatively evaluate the performance of candidate networks.

This scheme takes into account some network parameters, including velocity, current RSS, predicted RSS and available bandwidth. At the end, the optimized vertical handoff decision is made by comparing the performance evaluation values of candidate networks.

Other schemes of [15, 16] use geographical information to make vertical handoff decisions, regardless of whether this information was gathered by a GPS device or a physical layer support. In [15], the mobile user compares the distance to its current AP with the distances to the APs of neighbor cells. When the user is moving away from the current AP, it calculates the time it exits the cell. If it determines that it will be out of the cell several scans later, it decides to perform a handoff and searches for the nearest AP. If it can find an AP closer than the current AP, it switches to this AP.

The authors of [16] suggest using a location-based scheme where the mobility model of the user is used to predict its next location L after a certain period. The scheme then finds a serving AP of the location L and if it is different from the current AP, it initiates a handoff to that AP.

Finally, Chi et al. [17] proposed an analytical model for vertical handoff that uses the distance to the AP as well as Wrong Decision Probability (WDP) and the Handover Probability (HP). This vertical handoff scheme assumes that there are two networks with overlapping coverage areas. A handoff is initiated if the probability of unnecessary handoff is less than a certain threshold or when the difference in the bandwidth between the two networks is less than another threshold.

III. System Model

We now discuss the system model used in this work. We start with the signal propagation model and then go into the mobility model.

A. Signal Propagation Model

In this work we consider WCDMA networks and WLANs. Below, we discuss how to compute the Received Signal Strength (RSS) for each network type [5, 6, 13].

In WCDMA Networks. Before going into RSS computation, let us discuss the Signal to Interference plus Noise Ratio (SINR) and the Path Loss (PL). The SINR received at mobile user i when associated with WCDMA Base Station (BS) j can be represented as follows.

$$\gamma_{ij} = G_j P_j / N + \sum (G_j P_j) - G_j P_j, \tag{1}$$

where G_j, P_j and N denote respectively the channel gain between user i and BS j, the transmission power of BS j and the background noise at i. For mobile user i and BS j, the PL in dB is computed as follows.

$$\begin{aligned} PL_{ij} &= 135.41 + 12.49 \log(f_j) - 4.99 \log(h_j) \\ &+ (46.84 - 2.34 \log(h_j)) \log(d_{ij}), \end{aligned} \tag{2}$$

where d_{ij}, f_j and h_j respectively are the distance between i and j in kilometers, the frequency in MHz and the effective antenna height in meters. Now, the RSS for a BS j at mobile user i is expressed in dBm as follows.

$$RSS_C = P_j + G_j - PL_{ij} - A_j \tag{3}$$

Where P_j, G_j, PL_{ij} and A_j respectively are j's transmission power in dBm, the transmitted antenna gain in dB, the total path loss in dB, and the connector and cable loss in dB.

In WLAN Networks. Similar to the above, we start with SINR and PL before going into the RSS. The SINR received at mobile user i when associated with WLAN Access Point (AP) k can be computed as follows.

$$\gamma_{k,i} = G_k P_k / N + \sum (G_k P_k), \tag{4}$$

where G_k, P_k and N denote respectively the channel gain between mobile user i and AP k, the transmitting power of AP k and the background noise at i. For mobile user i and BS j, the PL in dB is computed as follows.

$$PL_{ik} = L + 10n \log(d_{ik}) + S \tag{5}$$

where L, n, d_{ik} and S respectively are the constant power loss, the path loss exponent with values between 2 and 4, the distance between i and k, the shadow fading which is modeled as Gaussian with mean $\mu = 0$ and standard deviation σ with values between 6 and 12 dB depending on the environment. Now, the RSS for a AP k at mobile user i is expressed in dBm as follows.

$$RSS_W = P_k - PL_{ik} \tag{6}$$

Where P_k and PL_{ij} respectively are the transmission power in dBm and the total path loss in dB.

B. Mobility Model

In addition to the popular Random Waypoint model (RWP), we propose a variation of RWP to help us gain a better understanding of the characteristics of our scheme. Below, we discuss both models.

The Random Waypoint (RWP) Model is widely used due to its simplicity [18]. In this model, the users are randomly distributed in the network. Each user randomly

selects a destination and moves towards it in a straight line with constant velocity chosen uniformly from a predefined range, $[v_{\min}, \ldots, v_{\max}]$. When the user reaches the destination, it stops for a duration known as the pause time before choosing another destination and repeating the above steps.

The Random Waypoint with Changing Probability (RWPCP) Mobility Model is similar to the RWP model except that the former allows the user to change its direction of movement and velocity as it moves towards the destination.

IV. Direction-based Scheme for Vertical Handoff (DSVH)

As mentioned above, the proposed scheme makes use of many factors while handling the handoff process. First, a handoff decision is triggered whenever the RSS drops below a predefined threshold. Next, the access device which the user will be handed off to is selected as follows.

1) The scheme generates a candidate list of access devices that achieves the RSS threshold.
2) The scheme checks the movement direction of the mobile terminal by considering a cone with an angle of 2θ around the current movement direction (see Figure 2). Only access devices that cover this cone will be considered as future candidates. In other words, all access devices that do not cover the cone are excluded from the handoff candidate list. In case none of the candidates reside in the cone, then the scheme moves to step 4. See the appendix for more details.
3) Each time slot (see Figure 2)the scheme measures the RSS value for the candidate access devices (RSS_{NEW}) and compare it to the previous time slot RSS value (RSS_{OLD}). If the RSS_{NEW} is lower than RSS_{OLD}, that means that the signal is getting weaker with the passage of time, ergo, the mobile terminal is moving away from the access device. The scheme eliminates from the candidate list all access devices that the mobile terminal is moving away from. In case the mobile terminal is moving away from all of the candidates, then the scheme moves to step 4.
4) Finally, the scheme selects the closest access device to the movement direction line (see Figure 2) of the mobile user.

From our experiments, we found that choosing $\theta = 30^\circ$ gives the best results; the time slot is set to one second.

V. Simulation Results

In this section, we present and analyze the experiments conducted to evaluate the performance of our proposed scheme, DSVH. We compared DSVH with the SINR-based scheme since it is one of the newest schemes and it is known to have higher throughput and lower dropping ratio compared with other handoff schemes (see Section II for more details).

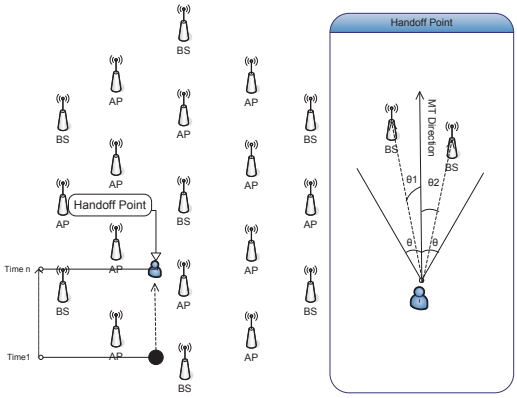


Figure 2. The handoff process in DSVH.

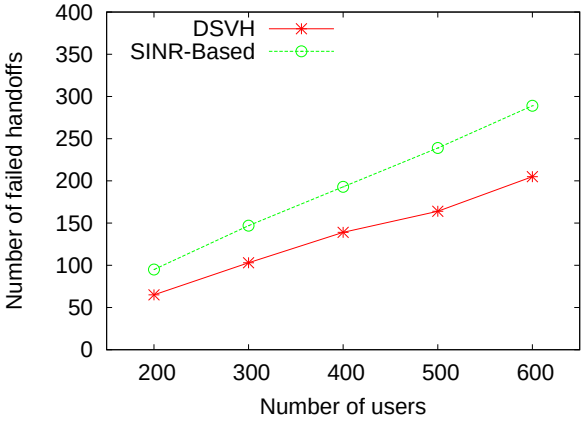
TABLE I.
Simulation Parameters.

Parameter	Values
Simulation area	5000×5000 m
Number of APs	12
Number of BSs	7
RSS threshold (WCDMA to WLAN)	-80 dBm
RSS threshold (WLAN to WCDMA)	-85 dBm
Antenna height of BS	30 m
AP transmitter power	20 dBm
BS transmitter power	33 dBm
Cable loss	5 dB
Channel gain	33 dBm
Operating frequency	894 MHz
Background noise power for WLAN	-96 dBm
Background noise power for WCDMA	-104 dBm
Bandwidth for WCDMA	5 MHz
Total noise or interference power over	16 dB

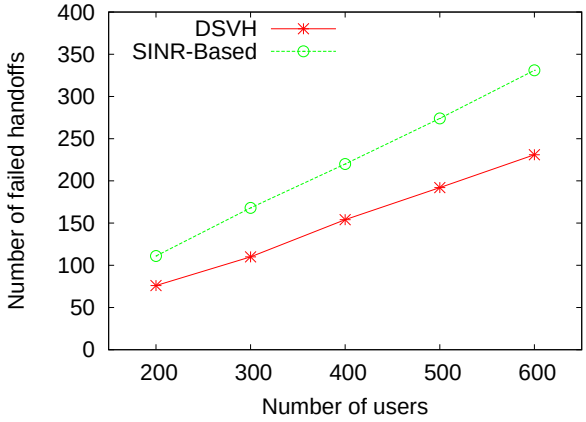
We consider a network of 7 BSs and 12 APs distributed in an area of 5000×5000 m. Previous works [6, 11, 13] have carefully placed the BSs/APs to maximize the performance of their scheme (see the left side of Figure 2). We compare the DSVH scheme with SINR-based under this fixed topology as well as a more generic topology where the BSs/APs are uniformly distributed. In the experiments below, we vary the number of mobile users between 200 and 600. The users are randomly distributed across the network area. At the beginning, each user is connected to the BS/AP with the highest SINR value. Table I summarizes the different configuration values we used in the simulations. These values were previously used with the SINR-based scheme of [6, 13].

Two metrics were used to compare the performance of DSVH and SINR-based schemes as follows.

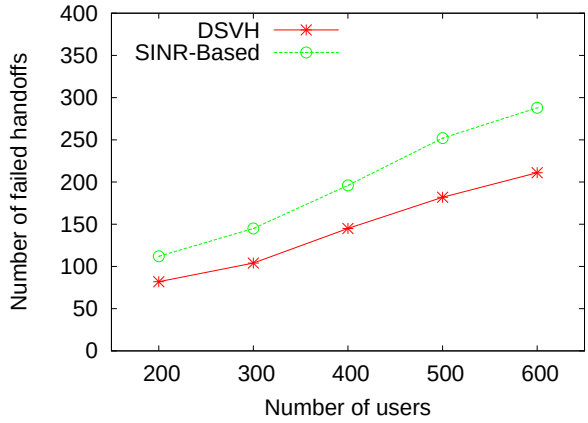
- Number of failed handoffs: when a handoff request by a mobile user fails (request is denied due to unavailability of free channels at the chosen handoff candidate) the user is disconnected. Such disconnections are intolerable in cellular networks. In fact, users prefer networks with lower bandwidth if they are more reliable (i.e., have lower disconnection probability) [6].
- Number of handoffs: Reducing the number of handoffs is generally preferred as frequent handoffs affect the network's throughput and reduce QoS [6].



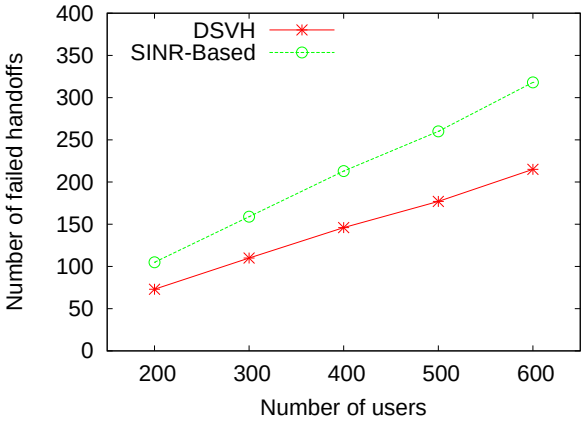
(a) Fixed topology; 5000×5000 m area; RWP model



(b) Fixed topology; 4000×4000 m area; RWP model



(c) Fixed topology; 5000×5000 m area; RWPCP model



(d) Fixed topology; 4000×4000 m area; RWPCP model

Figure 3. Comparison of the number of failed handoffs by DSVH and SINR-based schemes under various settings.

The simulation results presented in Figures 3 show the number of failed handoffs for both the DSVH and the SINR-based schemes under different settings. From these figures, we can clearly see that the DSVH outperforms the SINR-based algorithm in every setting.

In Figures 3(a) and 3(c), we test both schemes under the two mobility models discussed in Section III-B. The average improvement of DSVH over SINR-based under the RWPCP model is 31%, whereas the average improvement under the RWP model is 27%. This is due to the fact that the multiple direction changes allowed by the RWPCP model give more advantage to the DSVH since it uses a more involved algorithm for picking the best handoff candidate (see Figure 4 and the discussion associated with it). From these results, we predict that if we were to take a more realistic mobility model, the improvement ratio is likely to be higher. We are currently investigating this conjecture and the results will be part of our future work.

There are many insights related to why our proposed scheme, DSVH, outperforms the SINR-based scheme. Figure 4 depicts one such scenario. In the figure, when the user (or the Mobile Terminal (MT)) reaches the first handoff point (the red point). The SINR-based scheme will handoff to the BS that has the best SINR value, which is BS_3. Moreover, as the MT moves towards its destination, it reaches the second handoff point (the green point), and a second handoff takes place. The SINR-based scheme will handoff to the BS with the best SINR value which is BS_2. On the other hand, the DSVH scheme will behave differently. When the MT reaches the first handoff point (the red point), the DSVH scheme will nominate the access devices that reside in the cone of the MT's movement direction. So, only BS_2 will be an option for handoff and the MT will handoff to it. When the MT reaches the green point, the RSS value of BS_2 will not drop under the threshold and a second handoff will not take place. Informally speaking, since the coverage region in which the MT spends the longest period of time is the one closest to its movement direction, DSVH's selection will reduce the number of unnecessary handoffs.

As for why DSVH causes a smaller number of disconnections compared to the SINR-based scheme, it can

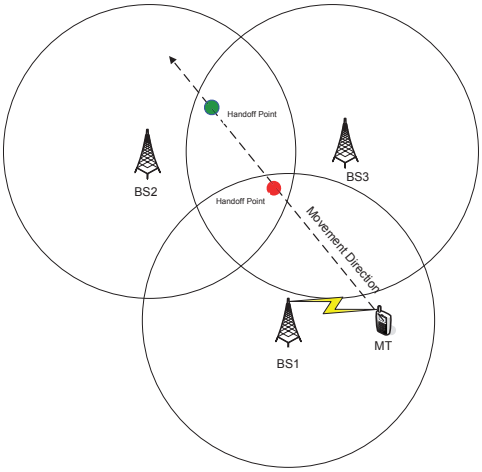


Figure 4. Example of when DSVH is better than SINR-based scheme.

be justified as follows. Consider the central region of the network and the set of BSs/APs within it. While passing through this region, the SINR-based scheme will prefer these BSs/APs due to their physical proximity to the MT. Thus, most of the MTs passing through this region will try to connect to the same small set of BSs/APs causing a high probability of disconnection. On the other hand, these BSs/APs may not necessarily be the closest to the movement trajectories for many MTs, and hence, DSVH will have no reason to give them any preference over the other BSs/APs. This will lead to a more balanced load distribution and lower probability of disconnection.

Figures 3(b) and 3(d) show that DSVH is better when decreasing the area to 4000×4000 m. The average improvements in Figures 3(b) and 3(d) are 31% and 34%, respectively. This is mainly due to the fact that reducing the area affects both the density of the network and the mobility of the users (in the sense that the users will have more frequent movement changes). This also means that the set of handoff candidates will be larger and the SINR-based scheme will choose the candidate with the best SINR value which is more likely to be out of the MT's movement direction. On the other hand, the DSVH scheme will have an advantage since it chooses the handoff candidate that is closest to the line of movement and thus requires a smaller number of handoffs (see Figure 4).

Until now, we have been using a network topology with the fixed BS/AP locations depicted in the left side of Figure 2. Note that the BSs are placed on a triangular grid and the APs are placed in the middle of the overlap regions of the coverage areas of the BSs. Such placement is in favor of the SINR-based scheme. In Figure 5(a), we use a uniform distribution of the BSs/APs. The results show that under such distribution, the average improvement gain of DSVH over the SINR-based is about 39%. Now, if we increase the number of BSs/APs (see Figure 5(b)), the average improvement gain jumps to 46%.

The plots in Figure 6 show how the number of handoffs is affected by the increase in the number of users under the various scenarios discussed above. Similar trends appear in these plots as in the ones of Figure 3; however the improvement ratios are smaller. Note that throughout Figures 6, where we consider a fixed topology, the improvement ratio is around 13%. However, when we consider uniform distributions of the BSs/APs (Figure 7(a)), the improvement ratio rises to 15%. Moreover, when the number of BSs/APs is increased to 10 and 15, respectively, the improvement ratio jumps to 18% (see Figure 7(b)).

VI. Conclusion and Future Work

In this work, we propose a new vertical handoff scheme based on the direction of the user's movement. Through extensive simulations, we show that the proposed scheme, DSVH, outperforms the SINR-based scheme, which is known to be better than other schemes [6], in terms of the number of failed handoffs. We also show that DSVH reduce the number of unnecessary handoffs.

In the future, we plan to use the user's movement history to predict its trajectory. This should enable the handoff algorithm to make better decisions especially when dealing with cases where the user keeps changing its movement direction drastically in a zig-zag fashion. Moreover, we are planning to use more realistic mobility models as well as network topologies taken from real locations of BSs/APs.

References

[1] X. Yan, Y. Ahmet Şekercioğlu, and S. Narayanan, "A survey of vertical handover decision algorithms in fourth generation heterogeneous wireless networks," *Computer Networks*, vol. 54, no. 11, pp. 1848–1863, 2010.

[2] E. Gustafsson and A. Jonsson, "Always best connected," *Wireless Communications, IEEE*, vol. 10, no. 1, pp. 49–55, 2003.

[3] X. Yan, Y. Sekercioglu, and S. Narayanan, "Optimization of vertical handover decision processes for fourth generation heterogeneous wireless networks," Ph.D. dissertation, PhD thesis, Australia, Monash University, 2010.

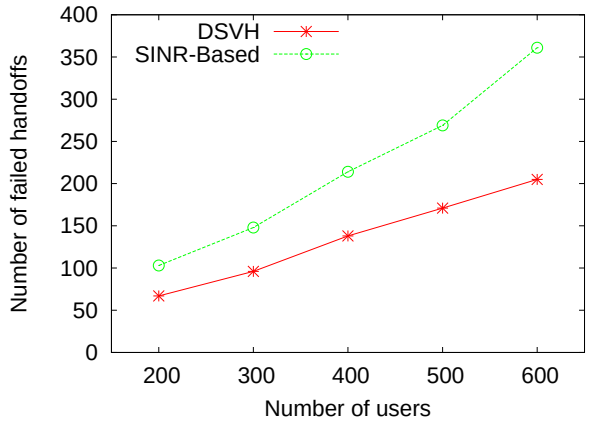
[4] K. Ayyappan and P. Dananjayan, "Rss measurement for vertical handoff in heterogeneous network," *Journal of Theoretical and Applied Information Technology*, vol. 4, no. 10, pp. 989–994, 2008.

[5] A. Zahran and B. Liang, "Performance evaluation framework for vertical handoff algorithms in heterogeneous networks," in *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 1, 2005, pp. 173–178.

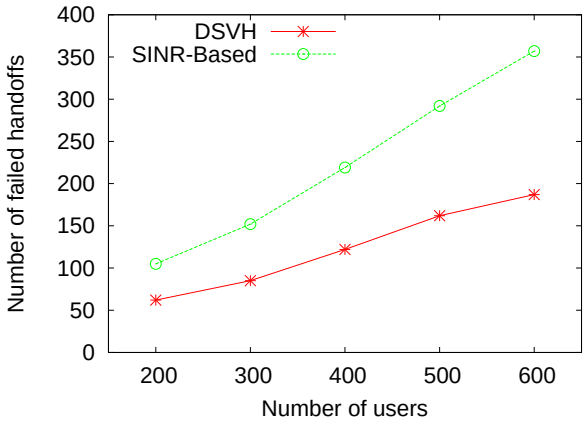
[6] K. Yang, I. Gondal, B. Qiu, and L. Dooley, "Combined SINR based vertical handoff algorithm for next generation heterogeneous wireless networks," in *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*, 2007, pp. 4483–4487.

[7] A. Zahran, B. Liang, and A. Saleh, "Signal threshold adaptation for vertical handoff in heterogeneous wireless networks," *Mobile Networks and Applications*, vol. 11, no. 4, pp. 625–640, 2006.

[8] X. Yan, N. Mani, and Y. Cekercioglu, "A traveling distance prediction based method to minimize unnecessary handovers from cellular networks to wlans," *Communications Letters, IEEE*, vol. 12, no. 1, pp. 14–16, 2008.

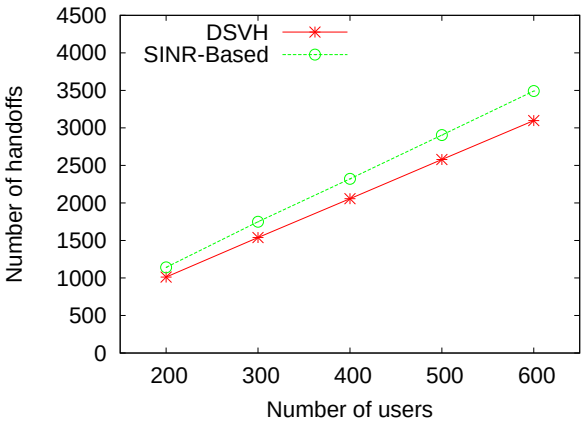


(a) Random topology with 7 BSs and 12 APs ; 5000×5000 m area; RWP model

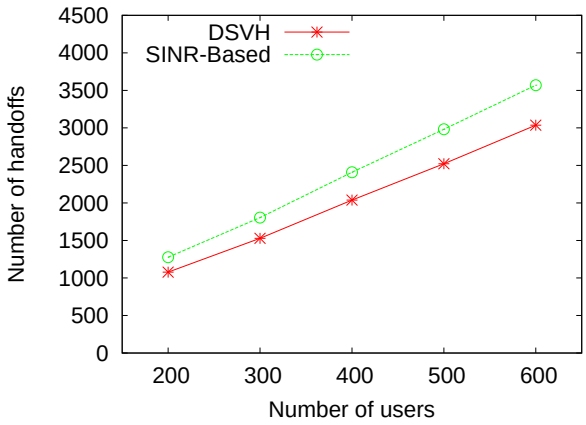


(b) Random topology with 10 BSs and 15 APs ; 5000×5000 m area; RWP model

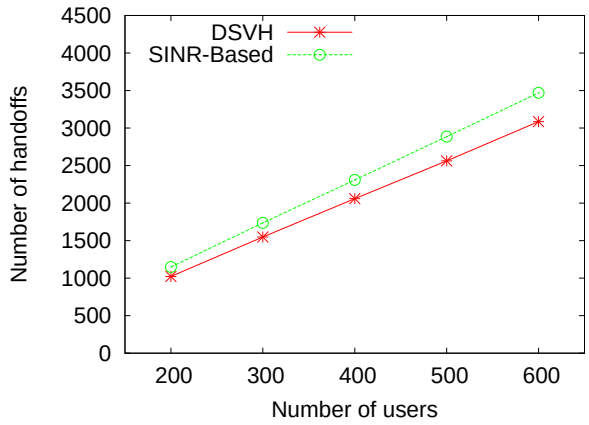
Figure 5. Comparison of the number of failed handoffs by DSVH and SINR-based schemes in random topologies with different densities.



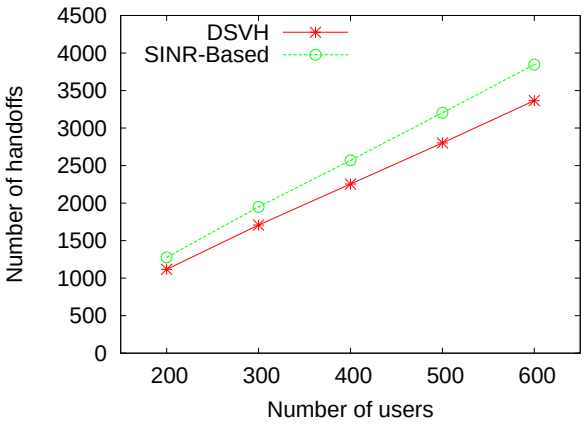
(a) Fixed topology; 5000×5000 m area; RWP model



(b) Fixed topology; 4000×4000 m area; RWP model

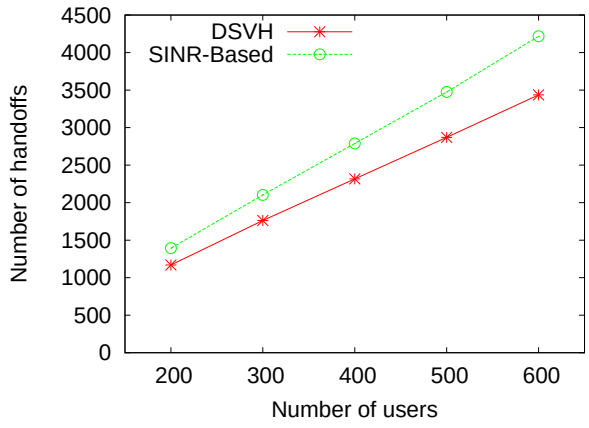


(c) Fixed topology; 5000×5000 m area; RWPCP model

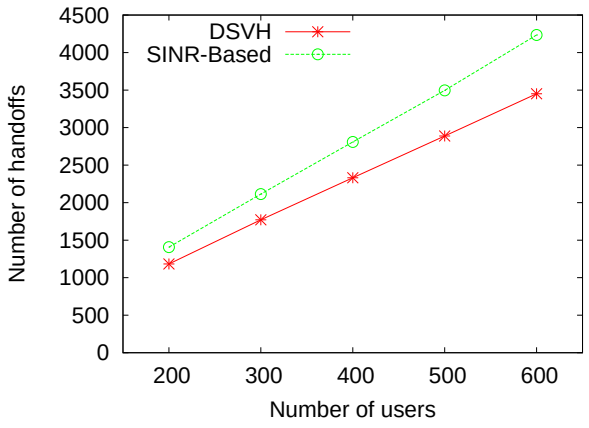


(d) Fixed topology; 4000×4000 m area; RWPCP model

Figure 6. Comparison of the number of handoffs by DSVH and SINR-based schemes under various settings.



(a) Random topology with 7 BSs and 12 APs ; 5000×5000 m area; RWP model



(b) Random topology with 10 BSs and 15 APs ; 5000×5000 m area; RWP model

Figure 7. Comparison of the number of handoffs by DSVH and SINR-based schemes in random topologies with different densities.

[9] X. Yan, Y. Sekercioglu, and N. Mani, "A method for minimizing unnecessary handovers in heterogeneous wireless networks," in *World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a*, 2008, pp. 1–5.

[10] S. Mohanty and I. Akyildiz, "A cross-layer (layer 2+3) handoff management protocol for next-generation wireless systems," *Mobile Computing, IEEE Transactions on*, vol. 5, no. 10, pp. 1347–1360, 2006.

[11] K. Yang, I. Gondal, and B. Qiu, "multi-dimensional adaptive sinr based vertical handoff for heterogeneous wireless networks," *Communications Letters, IEEE*, vol. 12, no. 6, pp. 438–440, 2008.

[12] K. Ayyappan and R. Kumar, "QoS based vertical handoff scheme for heterogeneous wireless networks," *Proceedings of the International Journal of Research and Reviews in Computer Science (IJRRCS'10)*, vol. 1, no. 1, pp. 1–6, 2010.

[13] K. Ayyappan, K. Narasimman, and P. Dananjayan, "Sinr based vertical handoff scheme for qos in heterogeneous wireless networks," in *Future Computer and Communication, 2009. ICFCC 2009. International Conference on*, 2009, pp. 117–121.

[14] L. Xia, L. Jiang, and C. He, "A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method," in *Communications, 2007. ICC'07. IEEE International Conference on*, 2007, pp. 5665–5670.

[15] M. Lott, M. Siebert, S. Bonjour, D. von Hugo, and M. Weckerle, "Interworking of WLAN and 3G systems," *Communications, IEE Proceedings-*, vol. 151, no. 5, pp. 507–513, 2004.

[16] J. Zhang, H. Chan, and V. Leung, "Wlc14-6: A location-based vertical handoff decision algorithm for heterogeneous mobile networks," in *Global Telecommunications Conference, 2006. GLOBECOM'06. IEEE*, 2006, pp. 1–5.

[17] C. Chi, X. Cai, R. Hao, and F. Liu, "Modeling and analysis of handover algorithms," in *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*, 2007, pp. 4473–4477.

[18] J. Broch, D. Maltz, D. Johnson, Y. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, 1998, pp. 85–97.

Abdelnasser Banihani received his M.S. degree in computer science from the Jordan University of Science and Technology, Irbid, Jordan, in 2012.

Mahmoud Al-Ayyoub received his B.S. degree in computer science from the Jordan University of Science and Technology Irbid, Jordan, in 2004. He received his M.S. and Ph.D. degrees in computer science from the State University of New York at Stony Brook, Stony Brook, NY, USA, in 2006 and 2010, respectively. He is currently an assistant professor at the Computer Science Dept at the Jordan University of Science and Technology, Irbid, Jordan.

Ismail Ababneh received an Engineer degree from Ecole Nationale Superieure d'Electronique et d'Electromechanique de Caen, Caen, France, in 1979. He received his M.S. degree in Software Engineering from Boston University, Boston, MA, USA, in 1984. He received his Ph.D. degree in Computer Engineering from Iowa State University, Ames, Iowa, USA, in 1995. He is currently the dean of the Information Technology College in Al al-Bayt University, Mafraq, Jordan. His research interests include processor allocation and job scheduling in highly parallel computers, mobile ad hoc networks, interconnection networks for parallel computers, and distributed computing.

APPENDIX

We now discuss the details of Step 2. Specifically, we are discussing how can we decide whether an access technology device resides in the cone (as shown in Figure 2) or not. We will show this for only one case (the one depicted in Figure 8) since it is easy to generalize this to all other cases. In the figures, the current position of the MT is the point A with coordinates (x_A, y_A). m is the length of line AB which is equal to the base station coverage distance. The point B coordinates can be computed as $(x_B, y_B) = (x_A, y_A + m)$. Since $\theta = 30^\circ$

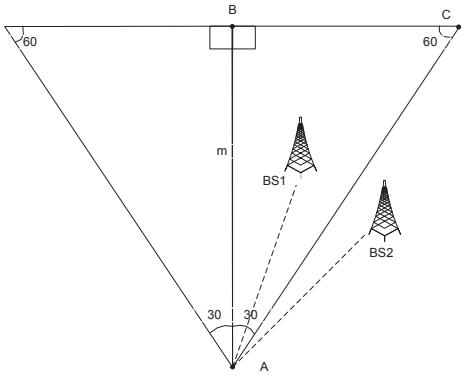


Figure 8. Movement direction cone determination.

(as mentioned above) we can use $\tan\theta = \frac{BC}{m} = 0.577$ to get $BC = 0.577m$. Thus, the coordinates for point C are $(x_C, y_C) = (x_A + BC, y_A + m)$. Now, the slope of the line AC is $Slope_{AC} = \frac{y_C - y_A}{x_C - x_A}$. Since the coordinates of each access technology device i are known, (x_i, y_i), we can compute the slope of the line Ai and if $|Slope_{AC}| < |Slope_{Ai}|$, then i resides outside the cone (see BS_2 in the figure). Otherwise, i resides inside the cone (see BS_1 in the figure).

Proposing a New Structure for Web Mining and Personalizing Web Pages

Hamid Alinejad-Rokny * [a, b]
[a] CVR, Faculty of Medicine, The University of New South Wales, Sydney, NSW, Australia
[b] School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW, Australia
Emails: H.Alinejad@ieee.org and Hamid.AlinejadRokny@UoN.edu.au

Mostafa Keikhay Farzaneh [c], Amir Goran Orimi [d], Mir Mohsen Pedram [e], Hojjat Ahangari Kiasari [f]
[c] Department of Industrial Engineering, Zahedan Branch, Islamic Azad University, Zahedan, Iran
[d] University of Mazandaran, Iran
[e] Department of Computer Engineering, Faculty of Engineering, Kharazmi University, Karaj, Tehran, Iran
[f] University of Hertfordshire, UK

***Abstract*—During users visiting a Web Server Access, data is stored. This information can be used broadly. Using this information we can obtain users' preferences and use them to personalize Web pages. Web mining is a new discussion that has been proposed to manage the web pages. In fact Web mining is the application of data mining techniques to discover patterns of user interests is the Data Web. In this article we provide a structure for web mining.**

***Index Terms*—Web Mining, Personalization of Web pages, Recommendation systems, Comprehensive site.**

I. INTRODUCTION

Today, the Web's global environment is the largest source of human information [4]. Currently, more than 124 million domains registered in cyberspace [5]. And every day the amount of domains increases. Despite the exponential growth speed, reading and understanding the information content remains constant [6]. Thus, we need automated tools and methods that can help to increase the speed, and allow users to access information on their favorite [1, 2, and 7]. One of these methods is called Web Mining.

In many resources Web exploration is defined as "the usage of data mining techniques for extracting information from the Web" [8 and 9]. One of Web exploration sub fields is web mining. Web mining is the application of data mining techniques to discover patterns of user interests from the Web Data [8 and 10]

One of the Web mining applications is the context of "Personalization of Web pages". For example, by comparing user's navigation patterns extracted from the log files the behavior of the user is predicted in real time [11]. One of the applications of this technique to real systems is recommendation systems. These systems are a specific type of information filtering systems, which recommend various items. (Such as movies, music, books, web pages, etc.) [12].

Nowadays the recommendation system is an important part of the user associated with web applications. E-commerce recommendation system is now proven that it can be effective in increasing profits and attract customers. [1, 2, 3 and 13]. In this article we propose a Web site structure (an online store), which is able to comply with the user and is customized for every user.

II. RELATED WORKS

In the field of web pages personalization, there are many algorithms and structures. Below are some examples of works.

Fabian Abel et al in Article [14] have provided a recommendation system based on a special Forums rule. SHEN Hui-Zhang presented a model of personalized web pages for web mining with hidden Markov model and dynamic clustering in Article [4]. In Article [15] Daniel Mican implemented a recommendation system called the WRS. Although the system has strengths as finding rules about less used pages and minor dependency between recommendation time and the number of saved records, but the important thing that should be noted is the lack of proper recommendations to users, especially when the information is not enough.

In Article [16] Minghao Lu, introduced structure that can evaluate the activities of the user (dynamic information), and also can scan profiles of the web site or Web site content. Thus its accuracy is higher than other structures, but this structures disadvantage is that it is slow. In this structure too much time spent that it depends on processing algorithm is used for Web Mining.

Qingtian Han and others in [17] has been set general view of a web mining algorithm for e-commerce website,

doi:10.4304/jetwi.5.3.287-295

this algorithm is simple and can be implemented in a conventional website. It is also a strong point and a weak point for it.

Our proposed structure is highly flexible. Simplified version of the structure can be used in small Store, and on the other hand, this structure can be implemented in big shops that have many users and products

The only difference is the algorithms used for grouping elements and some details.

III. Basic Concepts

A. Types of Personalization Web Pages

There are three general ways to personalize web pages of architecture and algorithms view [19]:

1. Rule based personalization systems on the
2. Content based personalization systems
3. Complex personalization systems

In the rule based personalization systems, site administrators define rules, according to these rules users classify and their web will personalize [20]. The main drawback of this system is the constant information. This means that users must specify their interests and be entered into a bunch of rules. This may cause false information [21].

In content based personalization systems, for each user one user identifier ID exist that contains a description of the goods or items that the user already has expressed interest [19]. In fact, the system uses the ID and the similarity of the goods with user's interest, and other commodities of interest to the user can be predicted. Some limitation of this system is that it has bad performance when enough data don't exist [22]. In complex Personalization systems, there is an attempt to solve some problems in the previous two systems.

This system focuses the same users and their choice will be examined and According to the choices and the user's interest in a certain score is assigned to each product [23]. KNN classifier is used for finding same users. For more information about this algorithm see [24].

B. Sources of Used Information

Information on the Server side

The first source for personalizing user's web pages is Access Logs on the server. When users visit a Web Server access data are stored in a file named Access Logs [25]. An example of this file is shown in Figure 1. As you can see this file stores any request to the server with the IP address of the requesting user, restored data and its date [26]. However there are three flaws in this information. First, per opening each page of a website ten lines of information in this file may be stored.

Second, the IP number might be used by several people, also some ISP, allocate different IP, for each user request in a session.

And third, it is possible some requests respond from the browser cache or proxy server these requests are not stored in the access logs file [18].

```
127.0.0.1 – [29/oct/2010 : 10:29:37 +0330] "Get /xampp/ HTTP/1.1" 302 - "-" "Opera/9.50 (Windows NT 5.1; U; en)"
127.0.0.1 – [29/oct/2010 : 10:29:37 +0330] "Get /xampp/ HTTP/1.1" 200 6545 "http:// locahost/xampp/splash.php"
127.0.0.1 – [29/oct/2010 : 10:29:37 +0330] "Get /xampp/ HTTP/1.1" 200 6338 "http:// locahost/xampp/splash.php"
127.0.0.1 – [29/oct/2010 : 10:29:41 +0330] "Get /xampp/ HTTP/1.1" 200 4357 "http:// locahost/xampp/splash.php"
127.0.0.1 – [29/oct/2010 : 10:29:41 +0330] "Get /xampp/ HTTP/1.1" 302 - "-" "http:// locahost/xampp/Opera/9.50 (Windows NT 5.1;)"
127.0.0.1 – [29/oct/2010 : 10:29:41 +0330] "Get /xampp/ HTTP/1.1" 302 - "-" ""http:// locahost/xampp/splash.php"
127.0.0.1 – [29/oct/2010 : 10:29:41 +0330] "Get /xampp/ HTTP/1.1" 200 4383 "http:// locahost/xampp/splash.php"
127.0.0.1 – [29/oct/2010 : 10:29:41 +0330] "Get /xampp/ HTTP/1.1" 200 1165 "http:// locahost/xampp/splash.php"
127.0.0.1 – [29/oct/2010 : 10:29:42 +0330] "Get /xampp/ HTTP/1.1" 302 - "-" "http:// locahost/xampp/splash.php"
127.0.0.1 – [29/oct/2010 : 10:29:42 +0330] "Get /xampp/ HTTP/1.1" 200 1254 "http:// locahost/xampp/splash.php"
```

Figure 1. A sample of access log file

Other types of server-side information are information that can be obtained during user registration. This information will be stored on a server, then it can be used to personalize the content and structure of the website [21].

Information on proxy side

A proxy server is a server that plays role of intermediary between the user (or organization) and the Internet. This will increase the security of organization. A proxy has ability to add security controls and cache services [27]. The proxy servers like common file servers Sent requests is stored in Access Logs [28], thus it can be used as the source of this information to personalize web pages.

Client side information

One of the information that is stored on the client computer and its browser is cookie. This cookie is a text that can be used for verifying the server settings storage, Contents of card, current session ID in server or other data that can be saved [29]. On the client side (with the user's knowledge) applications can be installed that evaluate user performance and send data to the server to be used to personalize web pages.

Or capabilities of Java Applet can be used in to run a server application on the client computer (using Java Virtual Machine) [30].

Steps of web pages personalization

In the most of references of web personalization three steps listed, which is described below [21, 31 and 32] since much time is needed for the first two steps they are done offline [15].

Preparation data (data gathering)

At this step, data is collected and some refining is done on it. For example, broken Data in Access Logs files are deleted [32].

Exploring data

At this step data mining techniques (such as clustering, classification, and return [33] are used to explore the relationships between pages, users, and also pattern on the using Web [21].

Decision making

This step uses the results obtained in the previous step, and personalizes the pages according to user's requirements and interests. This phase is performed online, i.e. when the user visits the website this step will be implemented and the results are shown to the user [31].

IV. Website Structure

The proposed structure, is intended for an online store. This store is provided with various products. There are also banners on the pages for advertising and also it can send comments and recommendations to the system user by email. In this website, we have used a mixed personalization system and there is no need to Access Logs file, rather after user identification visited page's information is stored in the database. (Excessive requests to the database can be avoided with storing the data in a temporary interface.) This information is stored in the user's last five sessions. The overall structure of the website is shown in figure 2. This structure will be described in the next sections.

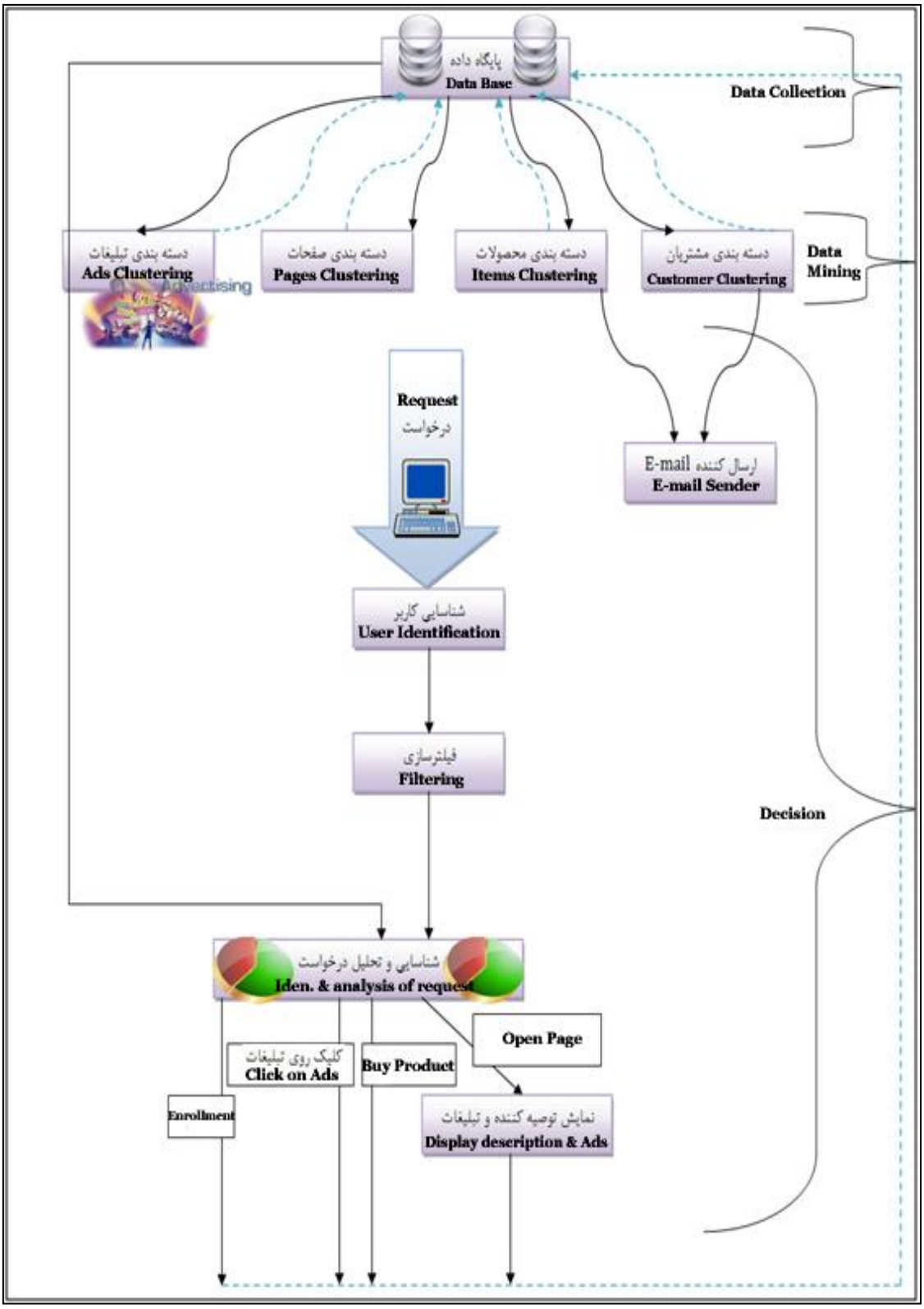


Figure 2. Structure of store

A. Client apply and Identification

For each request that website receive, first the customer must be identified. This website's customers can register on the website in order to make it easier to personalize web pages. However, customers who do not register are identified with the IP address. Of course, for some not registered customers the conventional and non-Personalized screen

is displayed, this includes the people that connected via proxy and don't have IP or they are machine. (Such as search engines or other reptiles). In the structure, the "filtering" is considered. to identify these people. For these people, a message is displayed on all pages and they will be invited to register. During registration the user may be asked, 'If you want to choose the best product which one you choose? "After selecting one of the products by the user it is stored in the database to be used later for personalization. Description of this process will be explained further. Also, during registration the user will be asked whether he/she want to receive E-Mail via the Web site or not. If he/she is interested, this is recorded in the database to introduce products purchased by similar users to him/her by E-Mail. On entrance the client a session ID assigned to him so that his activity is recognized in a session. After registering customers are detected via cookies in browser and consequently the Customer's previous visits can be accessed via the database.

B. Database

In the database, all customers, products, advertisements, and pages information can be saved. This information is used to personalize pages. Tables stored in the database are shown in figure 3.

Costumer table

Costumer name or IP	Costumer settings	5 last sessions	Favorite product	Purchased products ID	Introduced products ID	last email date for Introduced products ID	Clicked Advertises ID	Costumer Category

Products table

Product name	Product ID	Product Description	Product Category

Advertisements table

Advertisement name	Advertisement ID	Advertisement Description	Advertisement Category

Pages table

Page name	Page ID	Page Category

Figure 3. Tables of E-Shop's data base

Each customer is related to a particular category that by the "customer category" is determined. Purchased products ID, favorite product ID and clicked advertisements ID for each user is stored. For the last five sessions of a client the list of viewed pages are stored. (In the 5 last sessions column) Suppose web site have m page that named url1 to url_m respectively, then the data field (named U) is an m-bit number that:

$$\mathbf{U_i} = \begin{cases} \mathbf{0} \leftarrow \boldsymbol{if\ url_i\ has\ not\ been\ visited\ by\ customer} \\ \mathbf{1} \leftarrow \mathbf{if\ url_i\ has\ been\ visited\ by\ customer} \end{cases}$$

Formula (1)

The client settings (whether to receive emails or not) is stored in the field "Customer setting". Last email date is saved in the field "last email date for Introduced products ID" and the list of introduced product is stored in the field "Purchased products ID". Customer table's fields that have multi-valued attributes are stored in another table.

C. Classification of Customers, Products, Pages and Advancements

The "Costumer Category", "product category", "Pages Category" and "advertisements Category" are Classifiers of the items. This work is done offline, within a predetermined time. Overall three matrices are required for the Classifications, which is presented in the following:

Products matrix:

$$\mathbf{P_{i,j}} = \begin{cases} \mathbf{0 \leftarrow if\ the\ product_j\ not\ purchased\ by\ customer_i} \\ \mathbf{1 \leftarrow if\ the\ product_j\ purchased\ by\ customer_i} \\ \mathbf{2 \leftarrow if\ product_j\ is\ customer_i's\ favorite\ product} \end{cases} \quad \textit{Formula (2)}$$

Pages matrix:

$$\mathbf{U_{i,j}} = \begin{cases} \mathbf{0 \leftarrow if\ url_j\ \textit{has}\ not\ been\ visited\ by\ customer_i} \\ \mathbf{1 \leftarrow if\ url_j\ has\ been\ visited\ by\ customer_i} \end{cases} \quad \textit{Formula (3)}$$

Advertisements matrix:

$$\mathbf{A_{i,j}} = \begin{cases} \mathbf{0 \leftarrow if\ customer_i\ has\ not\ been\ clicked\ on\ advertisement_j} \\ \mathbf{1 \leftarrow if\ customer_i\ has\ been\ clicked\ on\ advertisement_j} \end{cases} \quad \textit{Formula (4)}$$

These matrices can be easily made from the data in the database. After the construction the matrices, for advertisements, pages, and products classification it is sufficient that classify column A, U and P of matrix them according to their similarity. For classification, KNN algorithm can be used or simply first select an item and then verify other elements similarity to it.

$$\mathbf{sim}(\mathbf{product_i}, \mathbf{product_j}) = \frac{\sum_{x=1}^{\text{number of users}} (P_{x,j} \text{ and } P_{x,i}) \text{ or not}(P_{x,j} - P_{x,i})}{\text{number of users}} \quad \textit{Formula (5)}$$

$$\mathbf{sim}(\mathbf{url_i}, \mathbf{url_j}) = \frac{\sum_{x=1}^{\text{number of users}} \mathbf{not}(P_{x,j} - P_{x,i})}{\text{number of users}} \quad \textit{Formula (6)}$$

$$\mathbf{sim}(\mathbf{advertisement_i}, \mathbf{advertisement_j}) = \frac{\sum_{x=1}^{\text{number of users}} \mathbf{not}(A_{x,j} - A_{x,i})}{\text{number of users}} \quad \textit{Formula (7)}$$

In the Above statement the "and" operator is a logical operator that returns one only if the two inputs are non-zero. "not" operator, converts the zero input to one and one input to zero, "or" operator returns a 1 when one of its inputs is non-zero.

Then order items at the similarity degree (descending) and put $\sqrt{\boldsymbol{number\ of\ items}} - \mathbf{1}$ of the elements in the set of selected elements and so we continue to $\sqrt{\boldsymbol{number\ of\ items}}$ classes are achieved that each has $\sqrt{\boldsymbol{number\ of\ items}}$ members.

For customers classification all of the three matrices are required, it is better to amend the three matrices and build a unique matrix so that the significance of columns of the matrix be considered. Unique matrix is obtained from equation 1 formula. (Multiplied by 2 have been added for significance of the products purchased by the customer in relation to other items.)

$$Cij = \begin{cases} P_{ij} & j \leq \text{number of products} \\ P_{ij\text{- number of products}} & (\text{number of products}) < j \leq (\text{number of products}) * 2 \\ U_{ij\text{- (number of products*2)}} & (\text{number of products}) * 2 < j \leq (\text{number of products}) * 2 + \text{number of urls} \\ A_{ij\text{- (number of products*2 + number of ruls)}} & (\text{number of products}) * 2 + \text{number of ruls} < j \leq (\text{number of products}) * 2 + \text{number of urls} + \text{number of advertisements} \end{cases}$$

Figure 4. Equivalent matrices mixed up

After that matrix acquired customers can be classified using KNN algorithm or use this formula to determine the degree of similarity between users:

$$\textbf{Total} = \textbf{umber of products} * \mathbf{2} + \textbf{number of urls} + \textbf{number of advirtisements}$$

$$\mathbf{sim}\left(\mathbf{customer_i}, \mathbf{customer_j}\right) = \frac{\sum_{x=1}^{\mathbf{Total}}\left(\left(\mathbf{C_{x,i}} - \left(\mathbf{C_{x,i}}\ \mathbf{and}\ \mathbf{C_{x,j}}\right) + \left(\mathbf{C_{x,j}} - \left(\mathbf{C_{x,i}}\ \mathbf{and}\ \mathbf{C_{x,j}}\right)\right)\right) * \mathbf{2}\right)}{\mathbf{Total}} + \frac{\sum_{x=1}^{Total} not(C_{x,i} - C_{x,j})}{\mathbf{Total}} \qquad \textit{Formula (8)}$$

At this time, the same method can be used to classify customers.

In this way we divided customers, products, advertising, and web pages into similar categories. Now, the results are stored in the database.

D. Application Analysis, and Display of the Customer Personalized Profile

Each request is given to the website the customer can be detected by "user identification" and identified by the filtering is done on the client. For example, reptiles, and users who are connected through a proxy, by the "filter" are identified.

Then the request will be determined If customer clicked on an advertisement, advertisement ID is stored in the database by customer click. If a product is bought, the product ID is stored database. But if the application is opening a page the products that have not been brought or pages that have not seen it should be advised to him. To do this we can use different strategies.

1. Find a bunch of pages or more products that customers have expressed interest in it and report to him, or products found on those.(If there is enough information about this customer).
2. Considered the customers similar to current customer and show the products or pages of their interest to him.

The advertisements should be displayed at opening page. Suitable advertisements display can be took from the above ways. After the page has seen its bit on the database becomes one.

Also, the "Sender E-mail" is used to send E-mail to users who have a request during registration. After adding a product to stores for introducing it via email it have to be bought by 30% of users of a category, then the product is introduced to other people in category that didn't buy it and have an E-mail request.

However, user group may be changed, in the last E-mail date and introduced product ID for that user is stored in database to prevent a burst of E-mail and post duplicate products.

V. EVALUATION

Web pages personalization is mandatory is attended nowadays that so far many various structures have been provided for it. The proposed structure fixes some of the existing problems in other structures and provides accuracy and speedup together. In this structure various solutions provided for elements classification so that according to the environment of structure implementation balance between the accuracy and speed had been established.

The provided structure in addition to the told advantage is developable in all of the Web application language (such as PHP, ASP and ...). Also, due to the fact that the users are detected through the register and also through number IP, possibility to personalize pages for most users is provided. for the storage of users ' sessions the structure suggests a method that makes so many information is not stored in the database (potentially redundant data or information redundancy can't arise) as well as the information stored is sufficient to personalize Web pages. This Web site contains a recommendation system and also the ability to personalize home page and index page of products based on user interests.

The introduced structure due to the environment and database that is developed can be different in efficiency. But generally this structure act smarter with increasing the information contained in the database and has better results to show to the customer, but rather more work is done for the classification of elements.

Using the formulas provided for classification of elements due to the fact that they were mostly using Boolean operators could improve the speed of the process. With regard to the evaluation saving the last 5 completed sessions for each user is affordable for web sites with less than 1000 users and close to 1000 pages. With the more number of users or pages, it is recommended that only the ID of the visited pages be saved.

An important point is that the structure is notable, its a simple version of this flexibility is the structure of a Web shop can be used in other small girdo can be used in several products, users and bezrgkah stores darndniz this structured implementation. the only difference, is used in the algorithm for the category of elements.

An important point that is notable in the structure is its flexibility. A simple version of this structure can be used in a small web shop as it can be developed in big stores that have several products and users. The only difference is in the algorithms used for the classification of elements.

VI. CONCLUSION

As mentioned above, web mining, is one of the branches of web browsing to uncover the user's interests. Uncovering these interests could have many applications, such as that it can be used to personalize Web pages. One of the reasons for the importance of personalizing Web pages is due to high growth in the information contained on the Web page that makes it hard to access the useful information. A successful electronic commerce Web site, should have a particular behavior with the oldest users who purchase products. In this paper, we develop a

structure that provides Web store a step more close to success. If customers have everything that they want available, the store sales and profit rate rise considerably.

REFERENCES

[1] A. Ghiasi, "Marketing", Journal of Management, vol. 81, no. 59, pp. 74-88, 2007.

[2] Behkamal Sanaye Plastic Khozestan, "Extract qualitative properties of software E-commerce", Journal of Technology Management, vol. 88, no. 2, pp. 19-34, 2005.

[3] M. Taleb, "Factor in the maturity of an organization's approach to e-business models using FCM", Journal of Information Technology Management, vol. 88, no. 2, pp. 85-102, 2008.

[4] H. Zhang, J. Shen, Z. Zhong-zhi and A. Yang, "Web Mining Model for real-time webpage personalization", Management Science and Engineering, ICMSE '06, 2006.

[5] Domain Counts & Internet Statistics, "Domain Tools", www.domaintools.com/internet-statistics, 2010..

[6] C. Shahabi and F. Banaei-Kashani, "Efficient and Anonymous Web-Usage Mining for Web Personalization", INFORMS Journal on Computing, vol. 15, no. 2, pp. 123-147, 2008.

[7] P. Galeas, "Web Mining", http://www.galeas.de/webmining.html.

[8] J. Srivastava, P. Desikan and V. Kumar, "Web Mining-Accomplishments & Future Directions", University of Minnesota, 2012.

[9] C. Zhang and L. Zhuang, "New Path Filling Method on Data Preprocessing in Web Mining", Computer and Information Science, vol. 1, no. 3, pp. 112-115, 2008.

[10] R. Kosala and H. Blockeel, "Web Mining Research: A Survey", Katholieke Universiteit Leuven, SIGKDD Explorations, 2003.

[11] M. Motiei, "Fuzzy Intrusion Detection System via Data Mining Technique With Sequences of System Calls", Journal of Information Assurance and Security, vol. 5, np. 12, pp. 224-231, 2010.

[12] J.B. Schafer, J. Konstan and R. Riedl, "Recommender Systems in E-Commerce", In E-COMMERCE 99, Denver, Colorado, 1999.

[13] W. Hill, L. Stead, M. Rosenstein and G. Furnas, "Recommending and evaluating choices in a virtual community of use", In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, pp. 194-201, 1995.

[14] F. Abel, "A Rule-Based Recommender System for Online Discussion Forums", Proceeding AH '08 Proceedings of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 12-21, 2008.

[15] D. Mican and N. Tomaei, "Association-Rules-Based Recommender System for Personalization in Adaptive Web-Based Applications", Current Trends in Web Engineering; Lecture Notes in Computer Science, vol. 6385, pp. 85-90, 2010.

[16] Minghao, L., Web Personalization Based on Association Rules Finding on Both Static and Dynamic Web Data, University of Toronto, 2005.

[17] H. Qingtian, G. Xiaoyan and W. Wenguo, "Study on Web Mining Algorithm Based on Usage Mining", Computer-Aided Industrial Design and Conceptual Design, CAID/CD 2008. 9th International conference, 2008.

[18] C. Shahabi and F. Banaei-Kashani, "A Framework for Ecient and Anonymous Web Usage Mining Based on Client-Side Tracking", Proceeding WEBKDD, International Workshop on Mining Web Log Data Across All Customers Touch Points, pp. 113-144, 2002.

[19] B. Mobasher, "Data Mining for Web Personalization", Chicago, De Paul University, 2007.

[20] B. Mobasher, "Web Usage Mining and Personalization", Chicago, De Paul University, 2004.

[21] O. Nasraoui, "World Wide Web Personalization. Olfa Nasraoui", www.webmining.spd.louisville.edu, 2003.

[22] M. Pazzani and D. Billsus., "Content Based Recommendation Systems", Lecture Notes in Computer Science Pazzani, vol. 4321, pp. 2031-2045, 2007.

[23] F. Heylighen, "Collaborative Filtering". Principia Cybernetica Web, www. pespmc1.vub.ac.be/collfilt, 2005.

[24] T., Saravanan, "A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. God", Your Book Is Great, www. saravananthirumuruganathan.wordpress.com, 2010.

[25] M. Yannis, M. Morzy, T. Morzy. A. Nanopoulos, M. Wojciechowski and M. Zakrzewicz, "Indexing Techniques for Web Access Logs", Web Information Systems , 2004.

[26] Log Files, Apache HTTP Server, www.attpd.apache.org/docs/2.0/logs, 2009.

[27] What is proxy server, www.whatis.techtarget.com, 2011.

[28] Web Proxy Server 2.0 Log File Format, Microsoft, www.support.microsoft.com, 2003.

[29] HTTP cookie, the free encyclopedia, Wikipedia, www.en.wikipedia.org/wiki/HTTP_cookie, 2010

[30] Applets, Sun Microsystems, www.java.sun.com/applets, 2005.

[31] B. Mobasher, R. Cooley and J. Srivastava, "Automatic Personalization Based on Web Usage Mining", Chicago, DePaul University, 2000.

[32] M. Hasan Nejad and S. Soltani, "Web Usage Mining: A way for Web Structure improvement", Iranian Data Mining Congress, 2008.

[33] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, 1996.

[34] O. Nasraoui, H. Frigui, A. Joshi and R. Krishnapuram, "Mining Web Access Logs Using Relational Competitive Fuzzy Clustering", In Proceedings of the Eight International Fuzzy Systems Association World Congress, 1999.

[35] Ch. Willy, "Web site personalization", IBM, 2001.

[36] M. Spiliopoulou, B. Mobasher, B. Berendt and M. Nakagawa, "A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis", INFORMS Journal of Computing, Vol. 15, no. 2, pp.171-190, 2003.

Hamid Alinejad-Rokny is a member of *School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW, Australia*. He is the author/co-author of more than 65 publications in technical journals and conferences. He served on the program committees of several national and international conferences. He was *Guest Editor-Chief* for special issue at *IJFIPM*. Also He is *Deputy Editor-Chief* at *International Journal of Software Engineering And Computing* and he is editorial board member at *IJSEI, IJFIPM, JETWI, IJSCIP, IJCSCS, IJCNT and IJEIS.* His research interests are in the areas of Data Mining, Bioinformatics, Artificial Intelligence and Biological Computing.

Mir Mohsen Pedram is a member of Kharazmi University, Karaj, Tehran, Iran. He has many papers in international journals and conferences. His research interests are in the areas of Data Mining, Bioinformatics, Artificial Intelligence.

Applying Clustering Approach in Blog Recommendation

Zeinab Borhani-Fard [a]
[a] School of Computer Engineering, University of Qom, Qom, Iran

Behrouz Minaei [b]
[b] School of Computer Engineering, Iran University of Science and Technology Tehran, Iran

Hamid Alinejad-Rokny* [c]
[c] School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW, Australia
Emails: H.Alinejad@ieee.org and Hamid.AlinejadRokny@UoN.edu.au

***Abstract*—The web has met a significant growth in using weblogs during the recent years. According to the large amount of information in the weblogs, bloggers are facing difficulties to find blogs with similar thoughts and orientations and their popular information. While there is a vast overload of information for blogs, it necessitates having a blog recommender system. Collaborative filtering is a well-known technique in recommender systems. This technique extracts the relations between users and items in according to its neighbor's ratings, and since users have rated just a small part of data, sparsity makes problems for collaborative filtering. This problem leads to an inaccurate comparison among users, and consequently it decreases the accuracy of collaborative filtering algorithms. The use of clustering technique decreases data sparsity and it improves system scalability. We have used clustering to recommend the blog while the blog have reciprocal role, and each blog is both as a user and as an item in the network. In this paper, we use graph clustering based on users' information about social network and we propose blog recommendation framework to get recommendations. Experiments on ParsiBlog [1] data indicated that application of clustering technique with collaborative filtering is better performed that traditional collaborative filtering algorithms, PageRank and etc. A comparison between PageRank algorithm and clustering application showed that graph clustering in recommender system could makes better results in terms of accuracy, quickness and scalability.**

***Index Terms*—Blog networks, Collaborative filtering, Hybrid recommendation system, Graph clustering.**

I. INTRODUCTION

During recent years, blog have changed into a remarkable social media on the internet that enable users to broadcast content on the web consisting thoughts completely personal or Private. Facility of blog contents broadcast likewise willingness for thoughts development is becoming to promote blogs fast and continuously growth. Nowadays there are hundreds of million blogs all over the world that still being increased quickly. A blog is a website consisting data entries (so-called post) having reverse date sequence, and is written and maintained by a blogger who uses a specific tool. Since each blog or blog entry may have links to other blogs and web pages, blog link structure can be considered as a social network.

Recommender systems apply some ideas of users groups to help this individual efficiently to identify their favorite topics amongst vast options. Techniques are divided into three types, content-based recommender system, collaborative filtering recommender system and hybrid recommender systems [1]. Collaborative filtering systems provide the recommendations based on ratings by users set to the active user. Content-based recommender system uses items features (like movie director, actors, etc.) to get recommendations. Hybrid techniques generate recommendations with combining CF methods and content-based recommending methods.

Methods in Collaborative filtering can be divided to memory-based, model-based and hybrid [6]. One of memory-based CF problems is that it must compute similarity between each user (item) with all other users (item) to define their neighbors. This problem is not working in social network or blog recommendation that have equal items and users and numbers of users are very large. To cope with traditional CF technique or memory-based CF weak points, we applied clustering approach to gain more precision, speed and efficiency. Using clustering techniques reduce data sparsity and improve systems scalability because similarity computation is performed only for the users of the same cluster.

doi:10.4304/jetwi.5.3.296-301

Computation of costly and complex clustering is performed off-line. Using clustering methods in the model needs to once more clustering graph and update the model now and then.

Blog recommender system differs from other recommender systems, in several ways. First, the goal of recommending of product, movie, music, news, web page, travel and tourism for all kinds of services, electronic sale and even virtual community is different. It is important to find features of recommendation goals, whereas inappropriate use of recommendation may reflect negative effects. Second, blog recommender system is a provider, and in contrast with meanings, bloggers are dynamic and recommendation changes quickly and blog recommendation mechanism must be more adjustable and flexible than the rest. Blog are human-oriented in other words, blog content are highly subjective and mind-oriented to recommend [4].

This paper is organized as following. Related works for blog recommender system and clustering application in recommender systems are provided in section 2. Section 3 deals with blog recommendation framework in detail that we proposed based on clustering approach. Experiments evaluation and results of applied framework and comparison with other methods are show in section 4. The paper ends up in our conclusion and objectives for future actions.

II. RELATED WORKS

Because of massive content provided by the blogs, and since most bloggers are non-professional users with difficulties for finding their suitable and favorite blogs, blogs recommending systems have recently attracted researchers' attentions.

In some aspects, meaning of blog ranking is similar to blog recommendation. Abbasi et.al[8] used a personalized PageRank method for blog recommendation . Fujimura and et.al attributed some scores to each blog entry via weighing in based on authority and hub scores on the basis of eigenvector computations [13]. Our study is related blog recommender system and network clustering.

In blog recommending systems domain, different studies were performed both on the basis of blogs content and blogs social network. In Hayes and et.al research, the analysis is performed on the type of suitable recommender strategy for blog, which in their study is applied tags, post subject for blog recommender system [9]. A blog recommendation mechanism is offered in [4] that combines trust model, social relation and semantic analysis. Garc á etal. [3] provide a framework to connect data semantic to web pages links on the basis of special ontology. A blog recommender system that called iTrustU is being offered based on collaborative filtering and multi-facet society [15]. A personalized recommender system is offered in Hart et.al [14] based on tags.

Clustering methods are used in several CF recommender systems to reduce dimensions, data sparsity and to increase scalability. A CF system based on k-means clustering is applied to cope with data sparsity [17]. A CF proposes on the basis of iterative clustering method that extracts internal links of users and items [10]. In this model, users and items are clustered by k-means to solve scalability problem, one part of items are selected by experiencing different clustering algorithms then recommendation are observed separately.

III. BLOG RECOMMENDATION FRAMEWORK

Our proposed blog recommendation framework is explained in this section. We are done the steps of this framework on the ParsiBlog data which is one of the blog hosts in Persian. We produced blog directed graph based on the blogs that each blogger has indicated in his blog roll as favourite blogs. We select favourite blogs in Parsiblog domain. Parsiblog graph has 21305 nods and 257316 Edges. Figure1 shows our blog recommendation framework consisting of two main phases; data preparation and model implementation.

A. Data Preparation

Data preparation contains three stages: data pre-processing, network clustering, data post-processing.

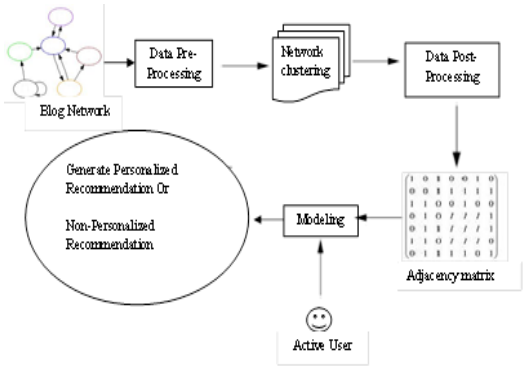


Fig.1. This is the caption for the figure. If the caption is less than one line then it needs to be manually centered.

Pre-processing

We have omitted nodes with no outgoing links in network or in other words with zero out degree, because it's not possible to have any recommendation for these nodes.

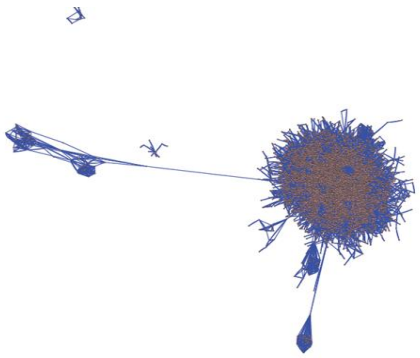


Fig.2. Social Network for Parsiblog

Most networks consisted of strong connected components that have a component with too many nodes. To reduce data sparsity, we can select only strong connected components with bigger nodes numbers. In this work, we selected strong connected components at least with 10 vertices. So, in this stage the derived graph consists of 9065 nodes and 222216 Edges. The biggest strong component has 8933 nodes. Figure 2 shows the blogs social network.

Network Clustering

Clustering is very important stage in our study, because it determines neighbors of active user. All of the next computations depend on clusters. It is important to select suitable clustering algorithm, since using different clustering techniques will have different results, and using a specific clustering algorithm may even decrease recommendations precision.

Cluster is a collection of data object that the members of the same cluster are similar and differ from other cluster members. Clustering methods are divided into three groups: partitioning method, density-based methods and hierarchical methods [7]. Term of cluster in graph is also called community in some papers. Today one of the main network subjects being mostly noticed and studied is communities' structure in network, and its goal is collecting vertices in to groups; so these groups will have larger density of edges inside the groups among the others. There are different algorithms to find such communities. During recent years, new algorithms are proposed. Newman and Girvan proposed an algorithm using Edge Betweenness as a metric to identify communities' boundaries [5]. The algorithm complexity is O(m3) on sparse graphs, while regarding available hardware. It limits the algorithm application to the networks having at least thousands of nodes.

We have used FastGreedy algorithm proposed by Claust et.al [13]. Algorithm complexity in the worst case is O(mdlog n) that d indicates depth of dendrogram, and m shows number of edges and n show number of vertices in the network. But algorithm complexity is O(nlog2n) for sparse graphs. Since most blog networks are sparse graphs so algorithm will be performed in linear time.

$$\mathbf{Q} = \sum_{\mathbf{i=1}}^{\mathbf{q}} (\mathbf{e_{ii}} - \mathbf{a_i^2}) \qquad (1)$$

$$\mathbf{a_i} = \sum_{\mathbf{j=1}}^{\mathbf{q}} \mathbf{e_{ij}} \qquad (2)$$

With $\mathbf{e_{ij}}$ consists of edges that connect vertices of community i to vertices of community j and q show number of clusters. $\mathbf{e_{ii}}$ consists of edges that connect nodes of cluster i to each other. This algorithm is a greedy implementation of hierarchical clustering algorithm. Algorithm consist of finding changes in q magnitude which is obtained merging each couple of communities and selecting the biggest one and at last doing the related mergence. Empirically, modularity more than 0.3 is a good index for suitable community structure in a network [2].

We have used igraph package [11] in R open source software for clustering implementation. We identified 192 clusters in Parsiblog graph and modularity amount was 0.372.

Post processing

This stage consists of clusters refinement to increase model accuracy. In this stage we have identified clusters with very few members as an outlier and omitted them.

We selected clusters at least with 50 members. Having done such operations on the clusters, clusters number declined in to 6 and there were 8435 nodes in the network. Table 1 shows the general information about features of primary blog graph, blog graph after pre-processing and after post- processing to be compared. With comparing network features, we can see that number of graph edges in each stage have not any remarkable reduction, but density, clustering coefficient, graph degree are increased. Figure 3 shows the distribution of clusters size (magnitude) and distribution of strong components size.

TABLE1.

THIS IS THE CAPTION FOR THE TABLE. IF THE CAPTION IS LESS THAN ONE LINE THEN IT IS CENTERED. LONG CAPTIONS ARE JUSTIFIED TO THE TABLE WIDTH MANUALLY.

	Vertices#	***Edge #***	***Degree Avg.***	***Density***	**Clustering Coefficient**
Initial Network	21305	257316	24.1554	0.0005669	0.31747
Pre-Processing Network	9065	222216	49.0272	0.0027042	0.37995
Post-Processing Network	8435	218207	51.7384	0.0030669	0.37663

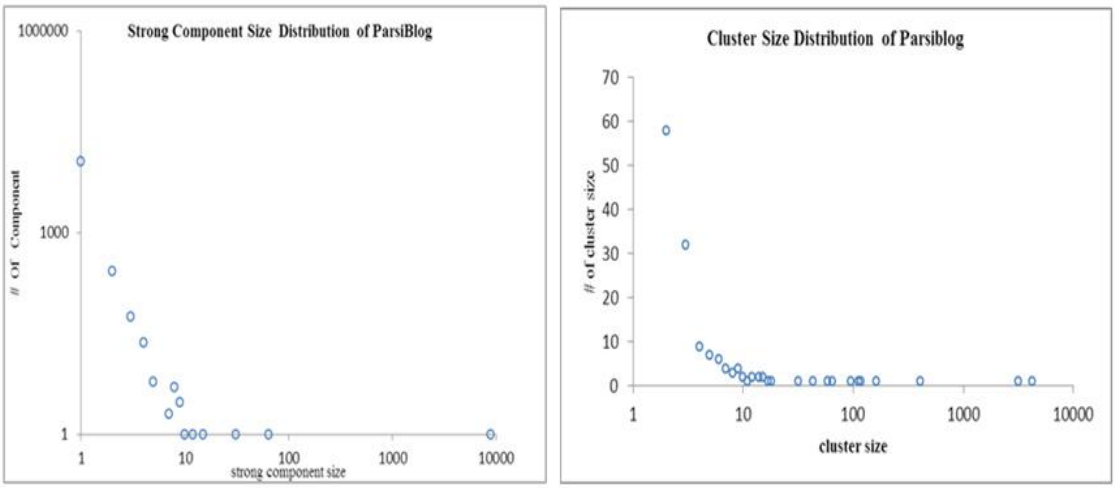


Fig.3. Strong component and cluster size distribution in Parsiblog

B. Model Implementation

Model implementation consists of two steps; model construction and generate recommendations. To implementation the model, one must convert data set to an adjacency matrix based on directed graph of blogs relations.

$$\boldsymbol{A[u,v]} = \begin{cases} \mathbf{1}, & \boldsymbol{if\ u\ links\ to\ v} \\ \mathbf{0}, & \boldsymbol{else} \end{cases} \qquad (3)$$

We used blogs link (blog roll) as the bloggers' favorite items rating, in this paper. So item-blog matrix is a asymmetric, square and binary one in which number of users and items are equal, and each blogger is an item and also a user. Each blog external links list shows the items preferred by blogger. Adjacency matrix in Parsiblog network has 8435 rows and 8435 columns.

Model construction

Development and design of models such as machine learning and data mining algorithms provides the system with learning opportunity to identify complex patterns based on train data, and then create intelligence recommendation for test data or real world data which is based on learner's models. In model construction, we predict a class to which each blogger belongs.

To construct the model, we divided adjacency matrix or data set into 70% train data obtained in the graph to classify 6 main clusters. We used C5 algorithm in Clementine software for classification. Accuracy of train data set was 81.03%. Test data set will be used for efficiency evaluation and accuracy of recommender system. Mean accuracy of test data was 79.60% after repeated practices.

Generate recommends

Generating recommend actions can be done in personalized and non-personalized format.

In non-personalized recommendations, some cases are recommended to the blog regardless of his characteristics that the most famous method is on the basis of ranking. Generally, there are three suitable approaches (input degree, HITS, PageRank) to rank nodes on the network. For each blogger regarding the cluster to which he/she belongs, we recommend the blog the k-Top highest rank node (blog) in that cluster.

In personalized, it's better to use personalized information to recommend the user. One of most well-known personalized recommending methods is collaborative filtering. We recommend each blog regarding the cluster to which he/she belongs an N-Top recommendation by collaborative filtering. The advantage of this method to traditional CF is that there is no need to compute active user similarity with the whole network users, and computing the users' similarity of the same cluster is enough.

We used PageRank algorithm to generate non-personalized recommendation and we also used collaborative filtering algorithm on the basis of a memory-based collaborative filtering method. In this method, we use cosine similarity standard to compute similarity. At the end of this section, Page Rank algorithm and collaborative filtering method based on neighborhood as our experiments basics is introduced. We named get of non-personalized recommendation as clustPR and get of personalized recommendation as clustCF.

PageRank algorithm: This algorithm is the most well-known link analysis algorithm offered in 1998 and it was

applied in Google search engine [12]. Assigning weight to each page, the algorithm sorts out search results based on the weight. Suppose that a random walker is searching through the created graph by Internet pages. Entering each site, the walker selects each of outgoing links with equal probability.

So, different pages with different weight would be seen. The main and valid page in PageRank is the one to which other valid and important pages offered link. This criterion indicates the popularity of each page through the whole graph, and it can be defined recursively as follows:

$$\boldsymbol{pagerank(u) = \frac{1-p}{N} + p \sum_{v \in S_u} \frac{pagerank(v)}{outdegree(v)}} \quad (4)$$

P is damping factor that in most cases it equals 0.85. $\boldsymbol{S_u}$is a set of all pages linked to page u and outdegree(v) shows the whole output pages of v.

Memory-based collaborative filtering: Memory-based CF algorithms use all or a sample of user-item data to create a prediction. Each user is a part of a group of individual with similar interests. Priorities predications in new items are produced for the blogger by determining what a new user's neighbor is nominated [6]. Neighborhood-based CF algorithm [16] is a memory-based CF algorithm, containing below stages:

Computing similarity or weight $\mathbf{w}_{i,j}$ between active user/item i and active user/item j. Neighborhood formation: selecting K item/user having most similarities with active item/user. Offering N-Top recommendation by weighed-in average neighbors' item/user is obtained. There are different methods for similarity or weight computation between users and items such as Pearson Correlation, cosine similarity, etc.

IV. EXPERIMENTAL EVALUATION

To evaluate clustCF and clustPR, we compare them with PageRank algorithm and traditional collaborative filtering algorithm.

A. Data Set

Evaluation is applied on data set of Parsiblog graph. Construct of Parsiblog graph described in pre-processing in section 3. Pre-processing data are used for PageRank algorithm and Traditional CF algorithm, Post-processing data is used for clustCF method and clustPR method that we propose in generate recommends in section 3. ClustCF is for personalized recommends and ClustPR is for non-personalized recommends.

B. Evaluation Metrics

To evaluate our offered framework; we applied recall and precision metric which were defined in information retrieval. These metrics are defined in blogs as follows:

$$\boldsymbol{Precision = \frac{|Favorite\ blogs \cap Recommended\ blogs|}{|Recommended\ blogs|}} \quad (5)$$

$$\boldsymbol{Recall = \frac{|Favorite\ blogs \cap Recommended\ blogs|}{|Favorites\ blogs|}} \quad (6)$$

C. Experimental Results

To perform such an evaluation, we selected 1000 nodes randomly in PR methods and traditional collaborative filtering (CF), and we computed recall and precision average. We obtained 20-Top recommendations for each user of this candidate set in CF algorithm, and we computed average for recall and precision.

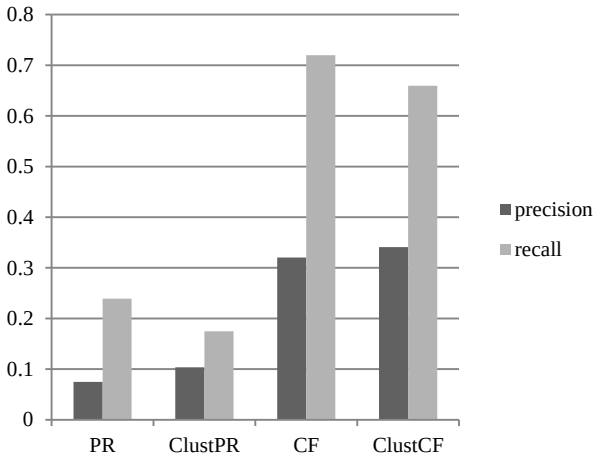


Fig.4. Comparison of algorithms (precision and recall)

We computed recall and precision average for 10 test data set in clustPR and clustCF methods, and then we computed total average for recall and precision. Figure 4 shows average of recall and precision for four algorithms.

Results indicate that clustPR as a non-personalized recommendation increases precision but recall decrease because of network clustering. ClustPRcan increase precision but this amount is less than personalized recommendation methods.

In clustCF method, amounts of precision are larger than traditional CF method but its amount is not big and recall is smaller because of network clustering.

V. CONCLUSION AND FUTURE WORKS

In this paper, we offered a blog recommendation framework that makes use of clustering approach to generate recommendation. A complex clustering network was used on blogs social network to find similar users group. Then we used neighborhood-based CF algorithm to generate recommendation in each cluster. We tried our experiments on the real world data set. We also did it for non-personalized recommendations to demonstrate that our framework with clustering approach increases accuracy for recommendations.

In future works, we intend to recommend a framework that can assign bloggers into several clusters (overlapping cluster). Overlapping clusters can depict real world conditions that bloggers participate in different communities. To do that, we intend to assign bloggers into several clusters with combining blog social network data and content-based recommender systems. In this study, we also used blogs links (blog roll) as an item-user rating matrix that is a binary matrix. In future studies, we

are going to consider the other links such as post-to-post, comment-to-post, and then combining them and obtaining the strength of blog relationship, we will compare the results by using non-binary matrix.

REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6):734–749.

[2] Clauset, A., Newman, M. E. J., and Moore, C. 2004. Finding community structure in very large networks. Phys. Rev. E, 70(6):66111

[3] Garcı́a-Crespo Á., Colomo-Palacios R., Gómez-Berbı́s J.M.l, Garcı́a-Sánchez F.2010.SOLAR: Social Link Advanced Recommendation System , Future Generation Computer Systems 26 (3): 374_380.

[4] Li Y.M., Ching-Wen C. 2009.A synthetical approach for blog recommendation: Combining trust, social relation and semantic analysis, Expert Systems with Applications 36 (3): 6536–6547.

[5] Newman M. E. J., Girvan M.2004. Finding and evaluating community structure in networks, Phys. Rev. E, 69(2): 26113

[6] Su X., Khoshgoftaar T.M. 2009. A Survey of Collaborative Filtering Techniques, Advances in Artificial Intelligence, 2009 (January 2009) Hindawi Publishing Corp. New York, NY, United States.

[7] Han J., KamberM.2001. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, Calif, USA.

[8] Abbasi, Z. and Mirrokni, V.S. 2009. A Recommender System Based on Local Random Walks and Spectral Methods. LNCS 5439, pp. 139–153.

[9] Hayes C., Avesani P., Bojars U.2007. An Analysis of Bloggers, Topics and Tags for a Blog Recommender System, Lecture Notes in Computer Science4737, 1-20.

[10] Jiang, X., Song, W., and Feng, W. 2006. Optimizing collaborative filtering by interpolating the individual and group behaviors. In APWeb. Lecture Notes in Computer Science3841,2006, 568-578

[11] Nepusz T., Csardi G. 2007. igraph Reference Manual, Technical Report, CRAN repository.

[12] Page L., Brin S., Motwani R., Wingord T. 1999. The PageRank Citation Ranking: Bringing Order to the Web, Technical Report. Stanford University.

[13] Fujimura, K., Inoue, T., Sugisaki, M. 2005. The Eigen Rumor Algorithm for Ranking Blogs, In proceeding of WWW 2005, May 10--14, 2005, Chiba, Japan.

[14] Hart M., Johnson R., Stent A.2009., iTag: a personalized blog tagger, In Proceedings of the third ACM conference on Recommender systems.

[15] Peng T .C, T. Chou S.c. 2009. iTrustU: a blog recommender system based on multi-faceted trust and collaborative filtering, In Proceedings of the 2009 ACM symposium on Applied Computing (New York, USA).

[16] Sarwar B.M., Karypis G., Konstan J.A, Riedl J. 2001. Item based collaborative filtering recommendation algorithms, in Proceedings of the 10th International Conference on World Wide Web (WWW ’01), Pp. 285–295.

[17] Xue, G., Lin, C., and Yang, Q. 2005. Scalable collaborative filtering using cluster-based smoothing, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.

Hamid Alinejad-Rokny is a member of *School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW, Australia*. He is the author/co-author of more than 65 publications in technical journals and conferences. He served on the program committees of several national and international conferences. He was *Guest Editor-Chief* for special issue at *IJFIPM*. Also He is *Deputy Editor-Chief* at *International Journal of Software Engineering And Computing* and he is editorial board member at *IJSEI, IJFIPM, JETWI, IJSCIP, IJCSCS, IJCNT and IJEIS.* His research interests are in the areas of Data Mining, Bioinformatics, Artificial Intelligence and Biological Computing.

Behrouz Minaei-Bidgoli obtained his Ph.D. degree from Michigan State University, East Lansing, Michigan, USA, in the field of Data Mining and Web-Based Educational Systems in Computer Science and Engineering Department. He is working as an assistant professor in Computer Engineering Department of Iran University of Science & Technology, Tehran, Iran. He is also leading at a Data and Text Mining research group in Computer Research Center of Islamic Sciences, NOOR co. developing large scale NLP and Text Mining projects for Persian and Arabic languages. He is the author/co-author of more than 60 publications in technical journals and conferences**.**

Automatic Extraction of Place Entities and Sentences Containing the Date and Number of Victims of Tropical Disease Incidence from the Web

Taufik Fuadi Abidin
Department of Informatics, College of Science, Syiah Kuala University
Banda Aceh, Aceh, 23111, Indonesia
Email: tfa@informatika.unsyiah.ac.id

Ridha Ferdhiana[1)] and Hajjul Kamil[2)]
[1)] Department of Statistics, College of Science, Syiah Kuala University
[2)] Department of Nursing, College of Medical, Syiah Kuala University
Banda Aceh, Aceh, 23111, Indonesia
Email: {ridha.ferdhiana, hajjul.kamil}@unsyiah.ac.id

***Abstract*—Many tropical disease incidences, such as leprosy, elephantiasis, malaria, dengue fever, are reported in online news portals. Online news portals are valuable data sources for creating a tropical disease repository if the information such as the location of the incidence, date of occurrence, and the number of victims can be automatically extracted from news articles. This paper describes approaches to extract that information from the Web. We introduce a rule-based algorithm to identify and extract the locations of the incidence and use Support Vector Machine (SVM) to determine the sentences containing the date of occurrence and the number of victims. Our experiments show that, the accuracy of the rule-based algorithm to identify the location entities is 99.8%, while the accuracy of the classifier to determine the sentences that contain one or more places of the incidence is 82%. The accuracy of SVM classifiers to classify the sentences that contain the date of occurrence and the number of victims are 96.41% and 93.38%, respectively.**

***Index Terms*—Entity Extraction from the Web, Support Vector Machine, Classification**

I. Introduction

Many tropical disease cases in Indonesia such as lymphatic filariasis, dengue fever, leprosy and malaria are reported every year. Between 2000 and 2009, a total of 11,914 chronic lymphatic filariasis cases have been reported nationally [1]. More than 17 provinces are reported to have malaria transmission with average transmission rate across the country around 5:1,000 population per annum [2]. Kompas online, one of the national Indonesian online newspapers headquartered in Jakarta, wrote that about 14,016 people were infected by *mycobacterium leprae* in 2001 and increased to 19,695 people (40.52%) in 2005 [3].

Due to the proliferation of internet technology, a large number of online news portals report the tropical disease incidence in Indonesia. If a keyword *kasus demam berdarah* (*dengue fever cases*) is searched on google.co.id, about 1,120,000 relevant results were returned.

Web pages, written in hypertext format and in a loosely structured text, are great sources of information in the modern age of internet-based technology today. Anyone can create web pages, and therefore, the size of the Web continues to grow. According to worldwidewebsize.com, the total number of web pages indexed by Google in June 2013 has reached approximately 47 billion pages [4]. A large number of web pages are being added to the Web every day, and thus, classifying web pages into interesting categories is an essential step and it is often treated as an initial step of mining the Web [5].

Classifying a large numbers of web pages into interesting classes is the goal of web classification. Web classification has been studied extensively and many research works in this field have been done, such as classifying web pages without negative examples which eliminates the requirement to manually collect negative training samples that tends to be biased [5], evaluating the capabilities of Bayesian algorithm for web classification and comparing its performance for both binary and multi-classification [6], surveying prominent web page classification methods [7], building SVM web classifiers and selecting web features [8], and learning to classify tropical disease web pages in a large Indonesian web documents [9].

To the best of our knowledge, this paper represents the first attempt to automatically recognize the locations where the tropical disease outbreaks occurred from Indonesian web pages and to identify the sentences that contain the occurrence date and the number of victims. We introduce a rule-based algorithm that incorporates morphological and contextual components as listed in Table 1 and a database of places [10] to recognize the location entities in the sentences, and then, use SVM classifier to classify the sentences and to determine which of those sentences contain the places where the tropical disease incidence occurred. We also build SVM models to identify the sentences that have occurrence date or the number of victims, or both. Previously classified web pages, described in [9], were used as the data source. A

doi:10.4304/jetwi.5.3.302-309

large number of sentences in the web pages were observed and manually annotated. The sentences that have the occurrence date or the number of victims, or both were labeled {+1}, and those that have no occurrence information or the number of victims were labeled {-1}. We took a portion of those labeled datasets for the training set to build SVM classifier and took the rest of the portion for the testing set. In summary, our contributions are twofold:

1. We introduced a rule-based approach to identify the place entities in the sentences and built an SVM classifier [11] to identify which of the sentences contain the place entities of tropical disease incidence.
2. We built SVM classifiers and selected the best classifiers to determine the sentences that contain the occurrence date and to identify the number of victims.
3. We organized the extracted entities and sentences into Keyhole Markup Language (KML) format and integrated them into Google Earth application.

The paper is organized as follows: Section 2 discusses related work. Section 3 describes the proposed approaches, including the contextual and morphological components, the methodology to remove the conflicting words in dictionaries, the construction of features, and the evaluation metrics to measure the accuracy of the SVM classifiers. Section 4 reports the results, and finally, Section 5 concludes our discussion of the automatic extraction of place entities and targeted sentences.

II. Related Work

Research on named entity recognition (NER) that aims at recognizing person, place, organization, time, and numerical expressions from text corpus has become an interesting study since the last two decades. Zhao [12] proposed a Hierarchical Hidden Markov Model to automatically identify product named entity in Chinese text. The entity was constructed using word forms and part-of-speech (POS) features. The findings concluded that the proposed methods outperformed the cascaded maximum entropy model and worked well for electronic and cell phone products. Sari et al. [13] used part-of-speech (POS) and syntactical structure, combined with semi-supervised learning method, to recognize and categorize named entity. They used Natural Language Processing (NLP) software to produce syntactic structure and used Stanford tagger to get POS tags of the sentences. They introduced a new method to automatically extract the date and location patterns from the sentences which have been labeled as prepositional phrase and from the sentences which have not been labeled by the tagger as prepositional phrase. They claimed that the performance of their proposed NER system is in the range of 50-70%.

Chanlekha [14] proposed a methodology to recover the most specific location where the outbreak of infectious disease occurred. They incorporated various features for recognizing spatial attributes into the models and trained the models using machine learning techniques such as Conditional Random Fields (CFP), SVM, and Decision Tree. In that work, Chanlekha considered events as the expression of phrases or grammatical constituents. The drawback of Chanlekha's approach is that the expression of phrases must be entirely defined to ensure that all events reported in the news articles can be recovered.

While research on named entity recognition of location has been intensively studied in English domain, far less attention has been paid to NER of location in Indonesian domain. This paper represents the first attempt to automatically recognize the locations where the tropical disease outbreaks occurred and to identify the sentences that contain the occurrence date and the number of victims.

III. Proposed Approaches

We propose the following approaches: 1) A rule-based approach to identify whether place entities are found in sentences by incorporating contextual and morphological components, and a database of places. The sentences that contain place entities, then, are classified using SVM classifier to ensure that the place entities are the locations where the tropical disease occurred. If the classifier categorizes a sentence as {+1}, then the place entities are extracted; 2) Develop SVM models to categorize sentences containing the date and the number of victims of tropical disease incidence; and 3) Organize the extracted entities into KML to integrate them into Google Earth application. We will discuss the proposed approaches in the following sections.

A. Contextual and Morphological Components

Contextual is a reference component that forms a place entity or negates it. In a sentence, contextual component is commonly written adjacent to a place entity that can be a single-word term, a two-word term, or a three-word term positioned consecutively.

Morphology is a major component in the grammar and it is primarily concerned with the rules of the word formation [15]. For place entities, the words are formally written in title case or uppercase, e.g. Bali or BALI. For date entities, they are usually written as a combination of digits and strings in specific formats such as *dd/dd/dddd*, *dd-dd-dddd*, *dd/dd/dd*, *dd-dd-dd*, *d name-of-month dddd*, and several other forms. The morphology for the number of victims is formally written as a combination of digits and contextual words, such as the word *korban (victim)* or *meninggal (dead)*.

Contextual components such as location prefix (LPRE) [16], popular town (PT), sign of location (SILO), preposition followed by a location (LOPP), and sign of address (SIAD) help us identify a place entity in a sentence, whereas location leader (LLDR) assists us that after the LLDR, the following phrase must not be a place entity. It is a place where the leader leads, instead. Table 1 lists the contextual components for place entity. Let's

discuss a few examples:

Wabah malaria terjadi di Kota Jakarta Utara

(Malaria outbreaks in the City of North Jakarta)

The word *Kota (City)* in that sentence is a location prefix, labeled as LPRE in Table 1. LPRE is a contextual component that gives us a clue that the next adjacent words, *Jakarta Utara*, written in title case, positioned consecutively, found in the database of zip codes, and morphologically true for a location is a place entity.

TABLE 1
CONTEXTUAL COMPONENTS FOR PLACE ENTITY

Label	Description	Examples
LPRE	Location prefix	Kota (city), desa (village), wilayah (region), …
LLDR	Location leader	Gubernur (governor), walikota (mayor), …
GOAG	Government agency	Polda (police), pemda (state government), …
LOGA	Leader of a government agency	Kapolda (chief of a state police), kepala (head of a unit), …
PT	Popular town	Jakarta, Denpasar, Surabaya, Banda Aceh, ...
LOPP	Preposition followed by a location	Di (at), dari (from), …
SILO	Sign of location	Lokasi (location), kawasan (region), …
SIAD	Sign of address	Jl, jln, jalan (street), …
PEOP	Public place	Hotel, taman (park), gedung (building), …
RELO	Religious location	Mesjid (mosque), wihara (temple), …
DAY	Name of day	Senin (Monday), Selasa (Tuesday), …
MONTH	Name of month	Januari (January), Maret (March), …
OPRE	Organization prefix	Universitas (university), institut (institute), ...
OPOS	Position in an organization	Direktur (director), rektor (rector), …
APRO	Abbreviation of a province	Sumut (North Sumatera), Jatim (East Java), …

Gubernur Aceh memberikan bantuan kepada para korban deman berdarah

(Governor of Aceh provides assistance to the victims of dengue fever)

The word *Gubernur* in that sentence is a location leader, labeled as LLDR in Table 1. LLDR is a contextual component that gives us a hint that the next word, *Aceh*, should not be considered as a place entity even though morphologically the first letter of the word is in uppercase and the word is found in the database of places. The word *Aceh* after *Gubernur* in that sentence is actually the name of the province where the governor governs. The two examples discussed here illustrate the roles of contextual and morphological components in assisting the rule-based algorithm to identify the place entities in a sentence.

B. Removing Conflicting Words in Dictionaries

One of important steps in our proposed approach is to construct bag-of-words (*dictionaries*) for each class. We used three different datasets in this research, i.e. place of incidence dataset, date of occurrence dataset, and the number of victim dataset as listed in Table 2. The dictionaries consist of weighted one-gram, bi-gram, and three-gram words extracted from the sentences in class {+1} and {-1}. The dictionaries are used to construct numerical features of each sentence.

To avoid over fitting, which generally gives poor predictive performance, the dictionaries consist of n-gram words were only extracted from the sentences in the training sets. We discovered that many similar n-gram words (one-gram, bi-gram, or three-gram) are found in the dictionary of class {+1} and {-1}, for instance, the word *penyakit (disease)*. The weight of the word in the dictionary of class {+1} is 0.563 and weight of the same word in the dictionary of class {-1} is 0.538. The weights are normalized by dividing the frequency of the word over the maximum frequency of the word in the dictionary.

To remove the words with high or low weight in both dictionaries, the weight ratio is calculated by taking the larger weight as the denominator. If the ratio is greater than a given threshold, 0.5 for this work, then the word with a smaller weight is removed from the dictionary. However, if the ratio is smaller than a given threshold, then the words will be removed from both dictionaries. For the word *penyakit (disease)*, it was removed from both dictionaries, i.e. the dictionaries of class {+1} and {-1}, because the ratio is 0.955 (0.538/0.563), which is greater than a given threshold.

Another example is the bi-gram words *wabah lepra (leprosy outbreak)*. The weights in {+1} and {-1} class dictionaries are 0.023 and 0.003 respectively. Hence, by taking the larger weight as the denominator, the ratio is 0.13. Because the ratio is smaller than 0.5, then the bi-gram words *wabah lepra (leprosy outbreak)* will be removed from the dictionary of class {-1} only, i.e. the class in which the ratio is smaller, while for the dictionary of class {+1}, the bi-gram words are retained.

The process of removing conflicting words in the dictionaries is also done for the date of occurrence and number of victim datasets. The words in the dictionaries play an important role in constructing numerical features of a sentence and achieving good classification accuracy.

C. Constructing Numerical Features for SVM Model

The numerical features of a sentence are the ratio of 1-gram, 2-gram, and 3-gram words in that sentence. There are 3 features for each class, and therefore, a total

of 6 features will be constructed for each sentence. The dictionaries that have been created beforehand are used to construct the numerical features of all sentences.

Mathematically, the ratio of *k*-gram words in class *p*, denoted as $F_{k\text{-}gram,\ p}$, is equal to the number of *k*-gram words in the sentence that are found in the dictionary of class C_p, denoted as $Dic(C_p)$, divided by the total number of *k*-gram words in the sentence.

$$F_{k\text{-}gram,p} = \frac{\sum_{i=1}^{n} t_{k\text{-}gram,i} \in Dic(C_p)}{\sum_{i=1}^{n} t_{k\text{-}gram,i}} \quad (1)$$

where i=1,2,...,n; k=1,2,3; p=1,2 and n is the number of words in the sentence.

In this work, SVM is used as the classification method. SVM has shown high performance in solving classification problems [11], [17]. Its performance results usually outperform other classifiers [9]. SVM is basically a linear classifier, however, by using an appropriate kernel trick, such as linear, polynomial, or radial, SVM also works well on non-linear cases.

Let $x_i \in R^d$ be the data and $y_i \in \{-1,+1\}$ denotes the classes for $i = 1,2,...,l$ where l is the cardinality. The separation of class {-1} and {+1} in the d dimensions is defined as $w \cdot x + b = 0$. A new data x_i will be in the class {-1} if the inequality $w \cdot x + b \leq -1$ is true and x_i will be in the class {+1} if the inequality $w \cdot x + b \geq +1$ is true. The maximum hyper plane is achieved by optimizing the distance between the hyper plane and the support vectors from the two classes, i.e. |1|/w. The flow of SVM models construction is depicted in Figure 1.

D. Evaluation Metrics

The performance of SVM model is usually assessed using testing sets. This performance evaluation has been widely used to avoid potential bias of the result due to over fitting of the model to training set [17]. Precision, recall, and F-measure are used to measure the classification accuracy. Precision (*P*) is the number of correct assignments (true positives) divided by the number of all returned results (true positives + false positives), while recall (*R*) is the number of correct assignments (true positives) divided by the number of correct assignments that should have been returned (the actual number of sentences belong to that class). Recall is similar to the sensitivity (TPR) in ROC analysis [17]. F-measure extends the accuracy metric that just measures the ratio of the correct results and acts as the harmonic mean of the precision and recall. F-measure has a value in the range of 0 to 1, where 0 is the worst and 1 is the best

$$p = \frac{TP}{TP + FP} \qquad r = \frac{TP}{TP + FN} \qquad F = \frac{2pr}{p + r} \quad (2)$$

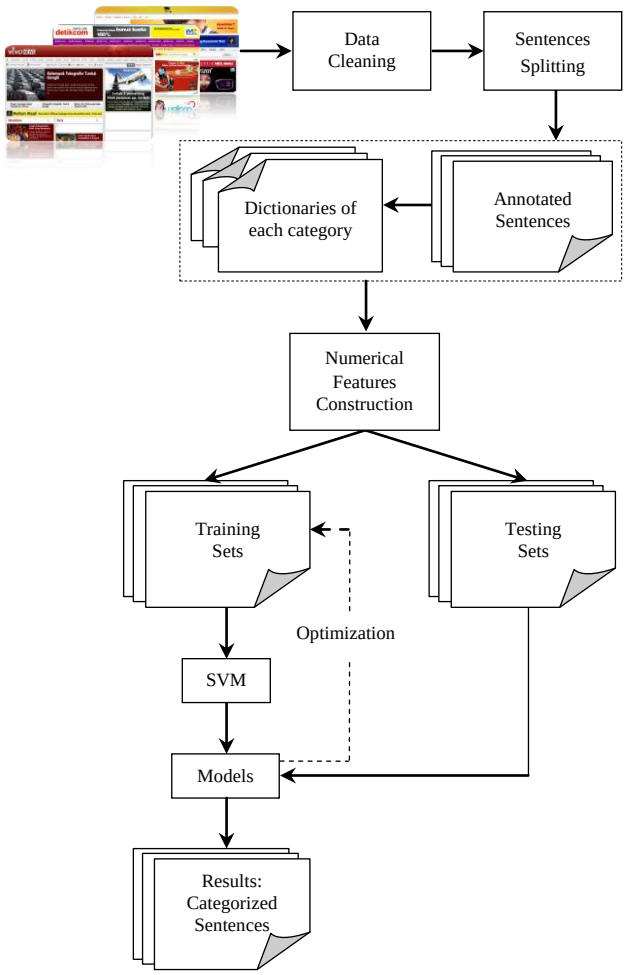


Figure 1. The flow of SVM models construction.

IV. Experimental Results

A. Datasets

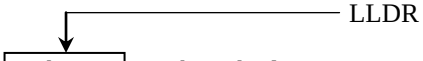
The datasets are sentences collected from 1,863 manually annotated web pages categorized as tropical disease [9]. The sentences were separated into two categories: (a) the sentences that contain the location of the incidence, the date, or the number of victims, labeled as {+1}; and (b) the sentences that contain no incidence information, labeled as {-1}. Table 2 shows the distribution of the sentences in each dataset. We divided the datasets into training and testing sets, and randomly selected 20-40% of the datasets for testing sets. The training set was used to construct SVM models while the testing set was used to evaluate their performance.

B. Results in Identifying Place Entities

We conducted an analysis to find all possible patterns to identify the location entities in the sentences. The patterns can be grouped into 4 cases:

Case 1: The words are possible to be the place entities, however, prior to them, the determinant contextual components are found and negate them as place entities.

The determinant contextual components are LLDR (location leader), GOAG (government agency), LOGA (leader of a government agency), POPL (public place), RELO (religious location), OPRE (organization prefix), SIAD (sign of address), and OPOS (position in an organization). We checked that the rule can handle this case very well as long as the determinant components are complete defined. Let's discuss an example of this case:



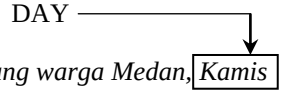
Governors of Aceh, Bali, and West Papua meet to discuss about malaria outbreak

The words *Aceh*, *Bali*, and *Papua Barat* are initially recognized as location entities. However, because prior to them an LLDR component is found, then all of them are canceled out.

TABLE 2
DATASET DISTRIBUTION BY CLASS LABELS

Dataset	Class	Number of Sentences	
		Training Set	**Testing Set**
Place of Incidence	+1	340	147
	-1	441	190
Total		**781**	**337**
Date of Occurance	+1	100	71
	-1	200	76
Total		**300**	**147**
Number of Victim	+1	300	72
	-1	100	38
Total		**400**	**110**

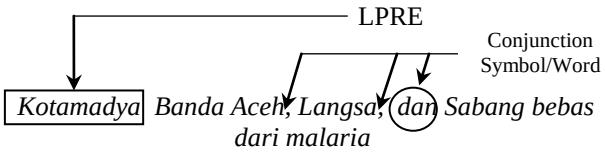
Case 2: The words satisfy the morphology rules, i.e. they are written in title case or uppercase, however if the words are labeled as DAY or MONTH, then they will not be considered as place entities. Let's see an example:



Dengue fever attacks the residents of Medan, Thursday

Kamis (Thursday) is the name of day of the week. Although the word is written in title case and satisfies the morphology rule, however, because it is a DAY, then the word will not be considered as a place entity. In the above example, only *Medan* is identified as a location entity. The same rule is applied if a word is a MONTH.

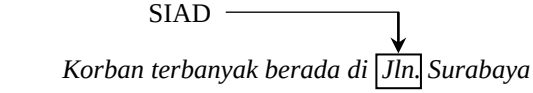
Case 3: The word satisfy the morphology rules, i.e. they are written in title case or uppercase, and prior to them, an LPRE (location prefix) or conjunction symbol is found. If that is the case, then the words will be considered as place entities. Let's discuss this example:



The city of Banda Aceh, Langsa, and Sabang are free from malaria

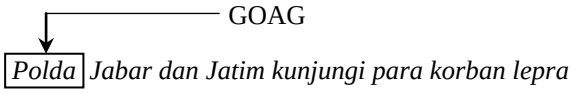
The two-word term *Banda Aceh* satisfies the morphology rule, exist in the database of places, and prior to it, a location prefix *Kotamadya* is found. Thus, *Banda Aceh* is recognized as a place entity. After the words *Banda Aceh*, a comma is found, and the next word after the comma, *Langsa*, satisfies the morphology rule and is found in the database of places, and the word itself is not labeled as LPRE, then the word *Langsa* is also identified as a location entity. The same rule is also true for the word *Sabang*. Hence, for the above example, the words *Banda Aceh*, *Langsa*, and *Sabang* are identified as place entities.

Case 4: The words satisfy the morphology rules, i.e. they are written in title case or uppercase, and they are tagged as APRO (abbreviation of province) or PT (popular town), but prior to them, the words are not tagged as one of the determinant contextual components mentioned in case 1. If this is the case, then the words will not be considered as place entities. Here are a few examples:



Most victims are found in Surabaya Street

Surabaya is listed in a popular town (PT). However, because prior to it a word *Jln* is a SIAD, then the word *Surabaya* will not be considered as a place entity.



West and East Java Police visit the victims of leprosy

Both *Jabar* and *Jatim* are the abbreviation of province (APRO). However, because prior to those words a word *Polda*, tagged as GOAG, is found then both *Jabar* and *Jatim* will not be considered as location entities.

Our empirical results show that from 1,328 location entities, 1,322 location entities were correctly identified by our algorithm and only 6 entities were incorrectly identified. The errors are due to misspelling and the use of capital letters for all the words in the sentences. In other words, the sentences are written in capital letters. The accuracy of our rule-based algorithm to identify the location entities, scored by F-measure, is 99.8%.

The sentences, which have been identified by the rule-based algorithm contain at least one place entity, are further classified by SVM classifier to determine which of those sentences contain the location of the tropical disease incidence. The values of F-measure of all SVM kernels, evaluated on training set, are shown in Table 3.

Empirically, polynomial is the best kernel of SVM for this purpose, i.e. 95.73%. The classification accuracy, evaluated on testing set consists of 337 sentences, is 82%. Table 4 summarizes the results on testing set.

TABLE 3
F-MEASURES OF ALL KERNELS EVALUATED ON TRAINING SET

SVM Kernel	F-Measure (%)
Linear	95.25
Polynomial	95.73
Radial	95.43

TABLE 4
CLASSIFICATION RESULTS TO DETERMINE WHETHER THE SENTENCES CONTAIN THE LOCATION OF TROPICAL DISEASE INCIDENCE

Dataset	Class	Sentences Classified as	
		+1	-1
Place of Incidence	+1	91	30
	-1	10	206
Total		101	236

$$\text{Precision} = \frac{91}{91+10} = 0.90, \text{Recall} = \frac{91}{91+30} = 0.75$$

$$\text{F-measure} = \frac{2 \cdot 0.90 \cdot 0.75}{0.90 + 0.75} = 0.82$$

The results are very conclusive. The accuracy of SVM classifier reaches 82%, and the SVM classifier yields recall and precision up to 75% and 90%, respectively.

C. Experimental Results in Identifying the Sentences that Contain the Occurrence Date

We also built an SVM classifier to determine whether a sentence contains information about the occurrence date of the tropical disease incidence. The occurrence date or time is usually written in a specific format as described in Section III. The numerical features of each sentence, used for SVM classifier, are also constructed using formula (1), i.e. estimating the ratio of 1-gram, 2-gram, and 3-gram words in the sentence and the dictionaries. If the sentences contain continuous time series information, they will be converted into ratio values based on n-grams. For the training set, the numbers of sentence in class {+1} and {-1} are 100 and 200, respectively. For the testing set, the numbers of sentence in class {+1} and {-1} are 71 and 76, respectively. Table 5 shows the values of F-measure for all SVM kernels, evaluated on testing set. The experimental results show that polynomial is also the best SVM kernel for this purpose. The evaluation on testing set demonstrates that the accuracy to classify the sentences that contain the occurrence date of tropical disease incidence is very convincing, i.e. up to 96.41%.

TABLE 5
F-MEASURES OF ALL KERNELS EVALUATED ON TESTING SET

SVM Kernel	F-Measure (%)
Linear	96.25
Polynomial	96.41
Radial	96.25

D. Experimental Results in Identifying the Sentences that Contain the Number of Victims

An SVM classifier is also trained and learned to effectively classify the sentences that have the number of victims of tropical disease in them. In order to complete this task, the numerical features of each sentence are constructed using formula (1). For the training set, the numbers of sentences in class {+1} and {-1} are 300 and 100, respectively. For the testing set, the numbers of sentences in class {+1} and {-1} are 72 and 38, respectively. Table 6 lists the values of F-measure for all SVM kernels, evaluated on testing set. Similar to the previous SVM, polynomial is also the best SVM kernel for this purpose. The evaluation results on testing set show that the accuracy to classify the sentences that contain the number of victims in them is also very promising, i.e. up to 93.38%.

TABLE 6
F-MEASURES OF ALL KERNELS EVALUATED ON TESTING SET

SVM Kernel	F-Measure (%)
Linear	92.73
Polynomial	93.38
Radial	93.21

E. Organizing Extracted Entities into KML to Integrate with Google Earth Application

After the place entities and the sentences that contain the occurrence date and the number of victims are extracted, they are organized into a standard KML file so that the locations of the tropical disease incidence can be viewed geographically in Google Earth application. The information that can be viewed, besides the locations, are the occurrence date of the event and the number of victims. KML is a scheme to describe and define geographic information on Google Earth software. It is a standard XML (eXtensible Markup Language) format containing specific elements and attributes [17].

A *placemark* tag is used to define a location on earth based on longitude and latitude coordinate values. The tag is symbolized by a yellow push pins in Google Earth application. A *point tag* is used to define the coordinates of an object, while a description tag is used to show additional information in a popup window. Figure 2

depicts the integration result in Google Earth.

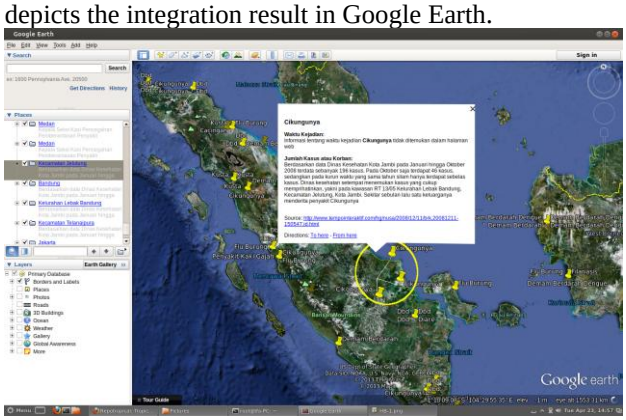


Figure 2. The extracted entities viewed in Google Earth.

V. Conclusion

Many tropical disease incidences in Indonesia are reported online in numerous news portals. News portals are valuable online data sources for creating a tropical disease repository if the locations of the tropical disease incidence, the date of occurrence, and the number of victims can be automatically extracted. In this paper, a rule-based algorithm to automatically identify the locations of tropical disease incidence from the web is proposed. The rule-based algorithm incorporates the database of places and the contextual and morphology components. The accuracy to identify the location entities is very conclusive, i.e. 99.8%. The accuracy of SVM classifier to determine the sentences that contain one or more locations of tropical disease incidence is 82%. The accuracy of SVM classifiers to classify the sentences that contain the date of occurrence and the number of victims are 96.41% and 93.38%, respectively. We believe that if the automatic extraction of location entities and sentences containing the date of occurrence and the number of victims can be scheduled, a tropical disease repository with a frequently updated data can be created.

VI. Acknowledgment

This work was supported by the Directorate General of Higher Education, Ministry of Education and Culture, Indonesia through Hibah Bersaing Grant 2012, 141/UN11/A.01/APBN-P2T/2012. We would to thank Teuku Ardiansyah and Rahmad Dimyati, the members of Data Mining and IR Research Group, Department of Informatics, for their valuable insights and help.

References

[1] Indonesian Ministry of Health, "Neglected Tropical Diseases in Indonesia: An Integrated Plan of Action", 2011.

[2] Behrens, et al., "The Incidence of Malaria in Travellers to South-East Asia: Is Local Malaria Transmission a Useful Risk Indicator?" Malaria Journal, vol. 9, no. 1, p. 266, 2010.

[3] Kompas Cetak, "Penyakit Tropis Tidak Teratasi", cetak.kompas.com/read/xml/2008/08/11/00563886/ penyakit.tropis.tidak.teratasi), cited on May 1, 2011.

[4] Worldwidewebsize.com, http://worldwideweb.com.

[5] H. Yu, J. Han, and K. Chang, "PEBL: Web Page Classification without Negative Examples", Journal of IEEE TKDE, vol. 16, no. 1, pp. 70–81, 2004.

[6] R. Bie, Z. Fu, Q. Sun, and C. Chen, "A Comparison Study of Bayesian Classifiers on Web Pages Classification", New Generation Computing, Ohmsha and Springer, vol. 28, pp. 161–168, 2010.

[7] X. Qi and B. Davison, "Web Page Classification: Features and Algorithms", ACM Computing Surveys Journal, vol. 41, no. 2, 2009.

[8] T. Abidin, A. Misbullah, and M. Subianto, "Determining Features of Web Documents and Building a Web Classifier using SVM", AISS (Advance in Information Sciences and Service Sciences: An International Journal of Research and Innovation), vol. 3, no. 10, pp. 401–408, 2011.

[9] T. Abidin, R. Ferdhiana, and H. Kamil, "Learning to Classify Tropical Disease Web Pages from Large Indonesian Web Documents", In Proc. of the 4th International Conference on Computer and Electrical Engineering, pp. 14–15, 2011.

[10] Pos Indonesia, "List of Places and Zip Codes in Indonesia", http://kodepos.nomor.net, cited on January, 2012.

[11] Joachims, "Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning", B. Scholkopf, C. Burges and A. Smola (ed.), MIT Press, 1999.

[12] J. Zhao and F. Liu, "Product Named Entity Recognition in Chinese Text", Journal of Language Resources and Evaluation, vol. 42, no. 2, pp. 197–217, 2008.

[13] Y. Sari, M. Hassan, and N. Zamin, "Rule-based Pattern Extractor and Named Entity Recognition: A Hybrid Approach", Information Technology, IEEE, vol. 2, pp. 563–568, 2010.

[14] H. Chanlekha and N. Collier, "Analysis of Syntactic and Semantic Features for Fine-Grained Event-Spatial Understanding in Outbreak News Reports", Journal of Biomedical Semantics, vol. 1, no. 3, pp. 1–11, 2010.

[15] A. Spencer, Morphological Theory: an Introduction to Word Structure in Generative Grammar. Oxford & Cambridge, 1991, pp. xviii + 512.

[16] I. Budi, S. Bressan, G. Wahyudi, and Z. Hasibuan, "Named Entity Recognition for the Indonesian Language: Combining Contextual, Morphological, and Part-of-Speech Features into a Knowledge", pp. 57–69, 2005.

[17] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. Khoury, "Application of SVM Modeling for Prediction of Common Diseases: the Case of Diabetes & Pre-diabetes", Journal of BMC Medical Informatics and Decision Making, vol. 10, no. 16, pp. 1–7, 2010.

[18] Google Developers, "KML Documentation Intro.", code.google.com/apis/kml/documentation/, cited on April 2012.

Taufik F. Abidin is a faculty member at the Department of Informatics, College of Science, Syiah Kuala University, Banda Aceh, Indonesia. He received his B.Sc. from Sepuluh Nopember Institute of Technology, Indonesia in 1993 with predicate Cum Laude. He received his Master Degree in Computing from RMIT University, Melbourne, Australia in 2000, and completed his Ph.D. in Computer Science at North Dakota State University (NDSU), USA in 2006. He received the ND EPSCoR Doctoral Dissertation Award from NDSU in 2005 and has been a Senior Software Engineer at Ask.com in New Jersey, USA to develop algorithms and implement efficient production-level programs to improve web search results. His research interests include Data Mining, Text and Web Mining, Database Systems, Information Retrieval, and ICT for Development.

Dr. Abidin is a member of APTIKOM, International Association of Engineers and Computer Scientist (IAENG), and International Association of Computer Science and Information Technology (IACSIT). He is a Program Committee for the International Conference on Software Engineering and Data Engineering (SEDE) for many years. He has US Patent (7,836,090 B2) on the Method and System for Data Mining of Very Large Spatial Datasets using Vertical Set Inner Products.

Ridha Ferdhiana is a faculty member at the Department of Statistics, College of Science, Syiah Kuala University, Banda Aceh, Indonesia. She completed her Bachelor Degree in Mathematics from Sepuluh Nopember Institute of Technology (ITS), Indonesia in 1997 and completed her M.Sc. in Applied Statistics from North Dakota State University (NDSU), USA in 2006. Her research interests include data mining, nonparametric modelling, statistical computing with R, and analyzing key factors that impacted undergraduate students' GPA.

Mrs. Ferdhiana is an active member of FORSTAT, a statistics forum for statistician and people who interested in statistics. She is also an active member of Indonesian Mathematical Society (IndoMS).

Hajjul Kamil was born in East Aceh, Indonesia and he is a faculty member at the Departement of Nursing, College of Medical, Syiah Kuala University, Banda Aceh, Indonesia. He completed his Bachelor of Nursing from the University of Indonesia, Jakarta in 1999 and received his Master of Nursing from the same university in 2001. He is currently pursuing his doctoral degree program at the University of Gadjah Mada, Yogyakarta, Indonesia. His research interests include health, quality of health service, patients' safety, management of nursing, and public health.

Mr. Kamil is a member of The Association of Indonesian Nurse Education Center (AINEC), The Indonesian National Nurse Association (INNA), and The International Nurses of Council (ICN).

Widespread Mobile Devices in Applications for Real-time Drafting Detection in Triathlons

Iztok Fister[a], Dušan Fister[a], Simon Fong[b], Iztok Fister Jr.[a]
[a] University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia
Email: (iztok.fister, dusan.fister, iztok.fister2)@uni-mb.si
[b] University of Macau, Faculty of Science and Technology, Av. Padre Tomas Pereira, Taipa, Macau SAR
Email: ccfong@umac.mo

***Abstract*— Today, global object-positioning is accomplished very precisely by GPS satellite technology. Access to this information is provided globally by widespread mobile devices with integrated GPS receivers from everywhere. On the other hand, mobile devices are connected to world-wide networks that ensure anytime access to application service (also web service) infrastructure based on application servers. Information everywhere at anytime is a key issue of pervasive computing. As a proof of concept that reflects a power of the pervasive computing, the application for drafting-detection in triathlon competitions was developed. This shows that the widespread mobile devices with GPS feature are appropriate for solving of real problems addressing the precise object-positioning.**

***Index Terms*— pervasive computing, real problem, mobile devices, GPS, web services, application server, triathlon**

I. Introduction

The development of mobile technologies enables users to manage information during their lives, as well as within business environments from *everywhere* on the world. However, this can only be realized by convenient applications. Such applications integrate software, hardware, infrastructure and services [18] and provide an *anytime* access to the information. The paradigm "*information everywhere at anytime*" represents the goal of, e.g., *Pervasive* or *Ubiquitous Computing* [28]. Both terms describe the integration of mobile front-end devices with back-end application infrastructure. Device management and application management are the main issues for the back-end systems. On the other hand, these systems must be prepared for serving the growing demands for network access from everywhere. Furthermore, such systems' services become *context-aware* [17], i.e. the answer to context-aware services depends on the contexts' elements as, for example, who, where, when and why someone demands such service.

One of the featured topics of pervasive computing and context-aware services is *positioning* [17]. Location awareness is a basic requirement for new applications on mobile devices [8]. The first step when positioning of object on Earth is the distance calculations between a mobile device and number of reference points. As a result, the mobile device determines the position of the object, whilst the reference points are implemented as a constellation of GPS satellites around the Earth. Typically, the distance is calculated by means of the *triangulation method* [39].

The triathlon is relatively young sport because its beginning only date back to 1978, when a group of enthusiastic athletes decided to finish three marathons using different disciplines, i.e. swimming, bicycling and running, all in one day. The competition got the name Ironman and today represents one of the greatest challenges for the persistence of human beings. Interestingly, this sport is growing wide-world each day, with more and more devotees. Moreover, the triathlon was integrated into the family of Olympic sports by the International Olympic Committee in 2000.

Firstly, in the triathlon competitors should compete in his own right. However, in order to attain better results some competitors forget about fair-ply. In place of competing alone, he exploits the competitor in front of him. This prohibited support is especially employed in bicycling, where the violating competitor rides his bike directly behind an other competitor. Thereby, the violating competitor saves his power for later efforts and rides his bicycle faster. This phenomenon is known as *drafting* (also *slipstreaming*), and is punished by referees. Typically, the referee on a motorcycle can eliminate the drafting competitor from the competition, for even up to five minutes. The rules for drafting-detection are regulated by the World Triathlon Corporation (WTC), and are discussed later.

In European triathlon competitions especially, drafting has been growing and reflects unfavorably on this sport. Referees try to restrict this phenomenon by punishing the drafting competitors. Its detection without modern technology has become impossible because of the increasing number of competitors (more than 2,000 per triathlon). Despite having the appropriate technology for solving this problem, a concrete solution does not exist on the market today.

This article demonstrates that drafting in triathlon competitions can be detected, in practice. An application for drafting-detection in triathlon competitions has been developed in order to prove this concept. This application

doi:10.4304/jetwi.5.3.310-321

consists of two parts:

- pervasive application on a mobile front-end device and
- context-aware service provided by a back-end system.

This pervasive application acts as a gateway that obtains the position of a competitor when riding his bicycle, and transmits this information to the context-aware service trough a worldwide wireless network. This context-aware service acts as follows. Firstly, it identifies the competitor, then, it compares the position of that competitor with the positions of other competitors within his neighborhood. Note that this neighborhood is determined from the WTC rules. If the identified competitor violating the drafting rules, the referees on the motorcycles are notified. Besides trying to find a solution to this problem, the focus is also on technical issues faced when developing this kind of pervasive application. Firstly, its structure is identified [15], the particular elements are then developed according to the recommendations in [6], [8]. Note that this application can be easily integrated within timing system controlled by the domain-specific language EasyTime, as proposed in [13], [14].

The structure of this article is as follows. In Section 2, the phenomenon of drafting in triathlon competitions is explained in detail. Beforehand, however, the characteristics of triathlon competitions are discussed. Section 3 describes the proposed application for drafting-detection in triathlon competitions. In Section 4 experiments and results are presented. In Conclusions results are summarized and directions for further development are placed.

II. Drafting-detection in triathlon competitions

This Section is divided into two parts. The first describes the main characteristics of triathlon competitions, whilst the latter focuses on the drafting-detection in triathlon competitions. In this sense, the rules of the WTC are presented for detecting and punishing this phenomenon. Although today several kind of triathlon exist, this article concentrates on Ironman. This kind of triathlon is still one the most prominent.

A. Ironman

Ironman (also the long triathlon) is held under the auspices of the WTC Association. It consists of three marathons covering different disciplines, in other words (Fig. 1):

- 3.8 kilometer swim,
- 180 kilometer of bicycling and
- 42.2 kilometer run.

The competitors start with the swim, continue with the bicycling, and finish with the run. All disciplines are performed in continuation. Between particular disciplines, however, competitors need to prepare themselves for the next discipline. This preparation is performed in *transition areas*. Here, two transition areas exist. In the first, the competitor takes off his swim suite and prepares himself for bicycling (TA1 in Fig. 1), whilst in the second he takes off his bicycle gear and prepares himself for the run (TA2 in Fig. 1). As a result, the completed achievement of the competitor consists of finishing all three particular disciplines and the time spent in each transition area.

B. Drafting

The phenomenon of drafting arises when one of the competitors purposely rides a bicycle directly behind an other and, thereby, avoids the persistence of wind. A drafting competitor can increase his average speed when bicycling whilst, at the same time, save energy consumption. Usually, in Ironman not only one single competitor drafts but a whole group of them together. Moreover, by exchanging the leading positions within the group (the leading competitor surrenders his position to a fresh competitor riding behind him and goes to take some rest at the rear of his drafting group), thus an additional speed up is achieving.

However, such a grouping has nothing to do with a time-trial competition, where a single competitor overcomes the course. Moreover, the results of those competitors within that group do not express the powers of individuals, and represent drafting violations that are punishable by referees. In fact, the WTC prescribes the following rules in the official Ironman competitions in order to avoid drafting [38]:

- drafting of another bike or any other vehicle is disallowed,
- competitors must keep 7 meters (4 bike lengths) distance between their bikes, except when overtaking,
- overtaking occurs when the overtaking competitors front wheel passes the leading edge of the competitor being overtaken,
- overtaking-competitors may pass on the left for up to 20 seconds, but must move back to the right-side of the road, after passing,
- overtaken competitors must immediately fall-back 7 meters (4 bicycle lengths), before attempting to regain the lead from a front runner.

As illustrated in Fig. 2, competitor B is violating the drafting condition because he is riding his bicycle 4 meters behind competitor A. Although competitor C is 8 meters behind competitor A he is in the so-called drafting zone of competitor B because he is only 4 meters behind competitor B. Note that the drafting zone is defined as an area that is 7 meters long and 2 meters wide, and is in relation to the leading competitor within the group. For example, each competitor located within the drafting area of competitor A in Fig. 2 forms the drafting neighborhood of competitor A. Whenever a competitor enters this zone for more than 20 seconds a drafting violation occurs.

To date, referees are responsible for drafting-detection in Ironman. They monitor competitors along the bicycle-course, from a motorcycle. Thereby, they try to stay as inconspicuous as possible. The time of drafting, as well as the distances between drafting competitors are estimated

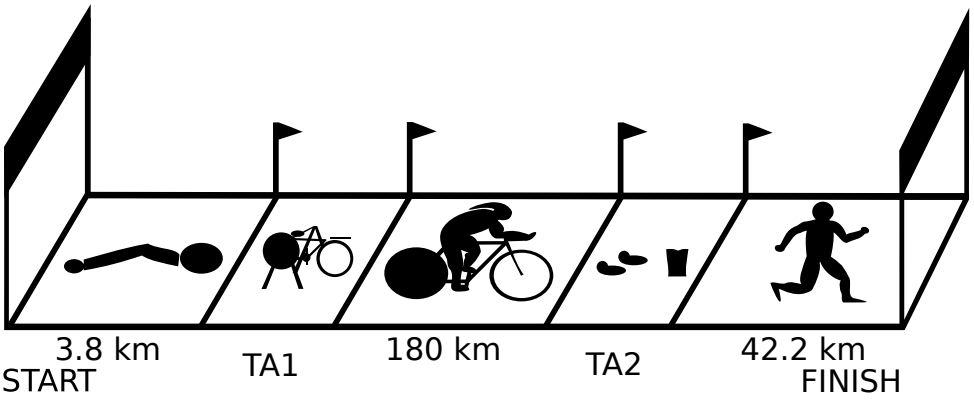


Figure 1. Ironman

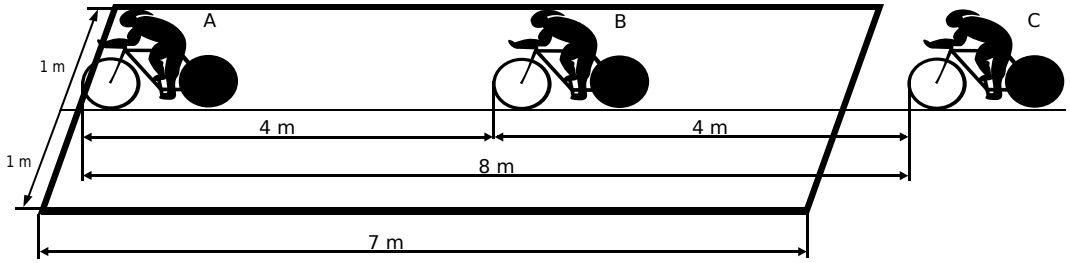


Figure 2. Drafting

by the referees approximately. Furthermore, they could be dealing with a limited number of drafting violations all at the same time. Likewise, they remain powerless when drafting is caused by a group of competitors. As a result, the automaton of drafting-detection is necessary.

III. Design of the application for drafting-detection in Ironman

In order to detect drafting along a bicycle course in Ironman, each competitor needs to be equipped with some mobile device that is capable of positioning his location as precisely as possible, and to transfer this position to a server via some ubiquitous network (Fig. 3). The positioning of particular object on the Earth is today enabled due to satellite technology, i.e. the Global Positioning System (GPS). On the other hand, the Internet is a really truly ubiquitous network today. Widespread use of smart mobile devices can incorporate both demands for drafting-detection: the support of global positioning whilst having access to the Internet, and therefore, appears to be the most suitable for this application. Moreover, referees use the same kind of devices for obtaining information about drafting-competitors.

Smart mobile devices, connected to the Internet due to widespread mobile networks, form complex ecosystem, where all parts work together seamlessly [15]. The mobile ecosystem is divided into the following elements:

- networks,
- devices,
- platforms (operating system, application framework),
- applications and
- web services.

The remainder of this article presents how these elements are addressed when designing this application for drafting-detection in Ironman.

A. Networks

A mobile communication ecosystem is needed in order to make a mobile ecosystem possible for communicating with the Internet. This is comprised of technologies, standards, and networks [31] that have been developing since 1950. A survey of wireless networks from their beginning to recent days, is presented in Table I.

As can be seen from Table I, the evolution of wireless networks can be divided into generations. For example, generation G1 captured analogue mobile telephones, where an user occupies the circuit switched line's whole duration of connections (multiple access to the line is disallowed).

During generation G2, a digital voice transmission via circuit switched networks, i.e. GSM (Global System for Mobile communications), was extended with GPRS (General Packet Radio Services) that allow packet data transfer [26]. The switching between data and voice is conducted by TDMA (Time Division Multiple Access). In addition to GSM, standards, such as USDC (US Digital Cellular) and PDC (Pacific Digital Cellular) are emerged on non-European markets. During the generation 2.5G, GSM was enhanced by EDGE (Enhanced Data rates for Global Evolution) that increase data transmission rates.

The 3G standards are CDMA2000 (Code Division Multiple Access 2000), TD-SCDMA (Time Division - Synchronous CDMA) and UMTS (Universal Mobile Telecommunications Systems). The latter is the successor

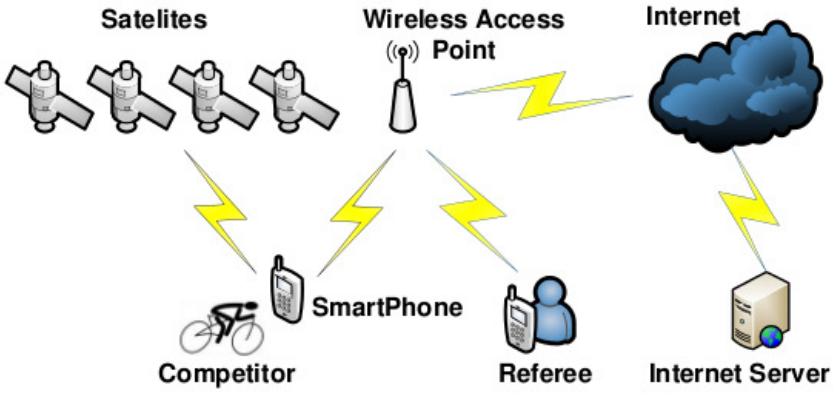


Figure 3. Drafting-detection in Inronman

TABLE I.
SURVEY OF WIRELESS NETWORKS.

Generation	1G	2G	2.5G	3G	3.5G	NGMN
Wireless Technologies	Analog	GPS/GPRS USDC PDC	EDGE	UMTS CDMA2000 TD-SCDMA	HSPA HSDPA HSUPA	EPS WinMAX
Features	No Multiple Access	TDMA	TDMA	CDMA	FDMA	FDMA

of GSM. These standards support the data packet transfer only. Besides higher data transmission rates CDMA was also specified for line-sharing. During generation 3.5G, UMTS was evolved to standards as: HSPA (High Speed Packet Access), HSDPA (High Speed Downlink Packet Access) and HSUPA (High Speed Uplink Packet Access). For line-sharing, the frequency division (FDMA) was introduced that remains the main multiple access mechanism.

The recent state during the evolution of mobile-networks is represented as NGMN (Next Generation Mobile Networks) that embrace two standards: EPS (Evolved Packet System) and WiMAX (Worldwide Interoperability for Microwave Access). The former is an evolution of UMTS systems, whilst the latter is new technology. Both the mentioned technologies can be considered as the last leg to 4G.

B. Devices

Pervasive devices combine the following four paradigms: they are strongly decentralized, diversified, connected, and easy to use [18]. Essentially, the fourth paradigm demands that the complex mobile and Internet technology is hidden behind a friendly user-interface. This interface between the user and machine is at the heart of *human-computer interface* (HCI), i.e. a discipline that has reached its maximum prosperity by the growth of pervasive computing [17].

Pervasive devices are divided into four main categories [18]:

- information access devices,
- intelligent appliances,
- smart controls and
- entertainment systems.

The first category of devices includes pocket-sized smart-phones (iPhones, BlackBerrys, etc.) that allow on-line access to information services (corporate databases, Internet pages, etc.). Intelligent appliances are pervasive devices with specific intelligence, like GPS navigation, industry controller, information car system, smart houses, etc. New kinds of entertainment systems address the world of modern broadcasting, such as interactive digital television, video on demand, etc.

Today, pervasive devices are a combination of more categories. For example, smart-phones incorporate the following features: classical telephone, information access via wireless networks, GPS navigation, IpTV, etc.

1) Global Positioning System: GPS is based on a set of broadcasting satellites that are used as reference points to calculate the position of an object on the Earth. It consists of three segments: space, user and control. The space segment is composed of 24 to 31 satellites orbiting within GPS constellation [2] alligned to the rotation of the Earth, orbiting at an altitude of approximately 20.200 kilometers. The user segment is composed of hand-held receivers (e.g., Polar, SmartPhone, etc.) or devices fixed on a vehicle (e.g., Garmin navigation system). The correct operations of the satellites are provided by the control segment.

The task of a GPS receiver is to identify almost four satellites, to determine the distances to each one, and to use this information for calculating the position of an object on the Earth. This calculation is based upon the mathematical principle of triangulation [39]. Note that a GPS receiver determines a three-dimensional position for the object within geographical coordinate system. Additionally, the Coordinated Universal Time (UTC) is

transmitted by the satellites. In the geographical coordinate system, the position is represented by its longitude, latitude, and altitude.

GPS obtains two levels of services: Standard Positioning Service (SPS) and Precise Positioning Service (PPS). The former can position the object with a precision of less than 20 meters [23], and is devoted to world-wide usage. The latter is more precise (e.g., up to 10 centimeters), therefore, it is deployed for military purposes. However, cost of SPS is much less than the cost of PPS. The mentioned precision is valid when the position is determined by four active satellites only. In practice, the number of active satellites can be increased and, consequently, the position of the object can be better calculated.

On the other hand, the absolute position of a competitor in Ironman is less important than the relative distance to the other competitor using drafting-detection. As a result, this can additionally increase the precision of distance calculation.

2) Differential Global Positioning System: The classical GPS (also stand-alone GPS) cannot be used for the precise positioning of an object on the Earth. This is due to a variety of risks that influence on the GPS performance. These risks relate to the effects of the ionosphere and troposphere, satellite maintenance, unscheduled satellite failures, satellite unavailability due to scheduled maintenance, repairs, repositioning and testing [29]. These anomalies may result in an unpredictable range of errors above the operational tolerances of GPS, which cause degraded availability, reliability, accuracy and safety (integrity monitoring).

Therefore, a supplementary navigation method, named *differential* GPS (DGPS) is used to significantly improve the accuracy and integrity of the stand-alone GPS [7], [11], [16], [20], [21], [25], [30], [35]–[37].

C. Platform

The platform denotes a core programming language in which all the application software is written [15]. Usually, the platform includes the hardware architecture, operating system and application framework (programming languages, run-time libraries or graphic user interface). Each mobile device running an operating system is treated as smart-phone. The operating system performs core services or tools that enable applications to talk to each other and share data or services. The application framework enables the development of a new application. It runs on top of operating system and provides support for sharing core services, e.g., communication, messaging, graphics, positioning, security, etc. Table II displays a review of the most significant mobile platforms with associated operating systems and application frameworks.

Note that all platforms are split into three categories: licensed (e.g., BREW, Windows Mobile), proprietary (e.g., Palm, BlackBerry, iPhone, Nokia) and open-source (Android). The development of the application for drafting-detection in Ironman was performed on Android.

1) Android: In order to write well-formed Android applications, a good understanding of Android's key concepts (e.g., Linux kernel, OpenGL, SQL database, etc.) is necessary. The overall system architecture is illustrated in Fig. 4.

As can be seen from this figure, the Android system architecture is divided into five layers as follows [6]:

- Linux kernel,
- libraries,
- Android runtime,
- application framework and
- applications and widgets.

Each layer depends on the services provided by the layers below it. Android is built on top of a Linux kernel. This is a stable and proven foundation that supplements Android with many operating system services, such as: memory management, process management, networking, etc. An Android developer never use Linux directly but over its utilities.

Android libraries are shared between applications. These are written in C or C++, and compiled for the particular hardware architecture. The most important libraries implement function, such as: surface manager, 2D and 3D graphics, media codecs, and browser engine. Note that these libraries do not represent applications. Conversely, they are used by higher-level applications to call the lower-level services.

Android runtime consists of a Dalvik virtual machine, and the core Java libraries. The Dalvik virtual machine is Google's implementation of Java, optimized for mobile devices. The Java core libraries are also adapted for this virtual machine. That is, all application code are written in Java, compiled from traditional *.java* and *.jar* files to *.dex*, and executed on a Dalvik virtual machine.

The application framework provides high-level building blocks that help the Android developers to create any new application. The framework consists of several managers for the handling of: activities (Android's synonym for process), contents (sharing data between applications), resources (text strings, bitmaps, etc.), positioning (GPS devices), and notifications (messages, appointments, proximity alerts, etc.). The framework is pre-installed as part of Android (Android SDK). However, it can also be extended with new components, as necessary.

The applications and widgets layer present the higher level of the Android architecture. This level is only visible by the end-user. In Android, the end-user interacts with the application over the whole screen. On the other hand, widgets (also gadgets) only operate within a small rectangle of the main screen.

D. Application for drafting-detection in Ironman

Application for drafting-detection in Ironman was written with regard to guidance found in publications, such as [6], [8]. It is a graphical front-end for the Android operating system (Fig. 5). This application supports the following functions:

TABLE II.
REVIEW OF MOBILE PLATFORMS.

Platform	Operating System	Application Framework
BREW	BREW OS	BREW
Windows Mobile	Windows Mobile	Windows Mobile
Palm	Palm OS	WebKit
BlackBerry	BlackBerry OS	Java API
iPhone	Mac OS X	Cocoa Touch
Nokia	Symbian	Qt
Open Source	Android	Android SDK

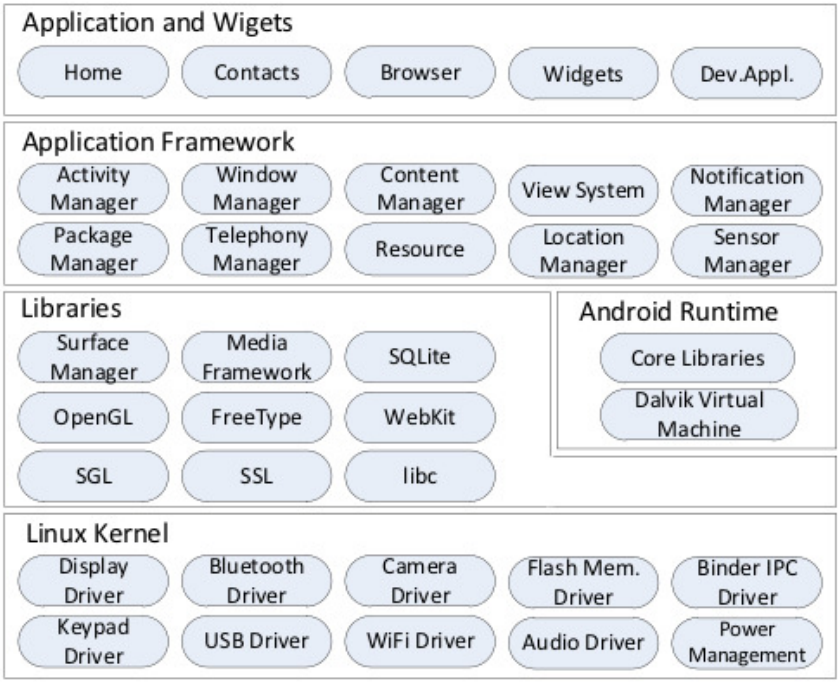


Figure 4. Android Architecture

- about-box,
- setup,
- start GPS agent,
- stop GPS agent and
- exit.

All functions can be activated by pressing the corresponding button. The about-box function displays a current version and the copyright information for the application. Setup enables the end-user (e.g., the competitor) to enter the context-aware and control variables, i.e. the starting number, the URL address of the web service, and the timer interval that determines a frequency of transferring the GPS information to a server. These variables are stored within the application preferences area and implemented by the Android's *PreferenceActivity* class [6]. Obviously, these variables are shared between other applications' activities. The start GPS agent function represents the main part of the application, i.e. the *Send* class. The execution logic of this class is illustrated in Algorithm 1. The stop GPS agent function ends the GPS agent, whilst the exit function finishes the application.

Note that because of this article's limitations, *Send* class (Algorithm 1) is not presented in details. Some variables are omitted but their omission is denoted by dots. Additionally, only the more important functions are presented here.

Send Java class (Fig. 1) extends the *LocationListener* activity class that implements the GPS listener. This class includes several global variables. The more important are the variables denoting the following classes:

- *locationManager* of class *LocationManager*,
- *envelope* of class *SoapSerializationEnvelope*,
- *androidHttpTransport* of class *HttpTransportSE* and
- *myTimer* of class *Timer*.

The *LocationManager* class implements interactions with a GPS device. The next two classes, i.e. *SoapSerializationEnvelope* and *HttpTransportSE*, are devoted to communication with web service provider, whilst the *Timer* class initiates communication with the web service provider.

The function *onCreate()* is called when the *Send* class is created. Firstly, three preference variables, i.e. *str_number*, *tim_tick*, and *URL*, are initialized. In order to create a *LocationManager*, connection with the GPS device is established, whilst *SoapSerializationEnvelope* establishes a connection with the web service provider,

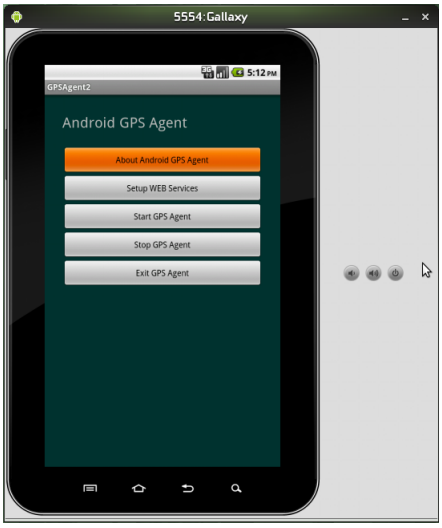


Figure 5. Android Application for Drafting Detection

as determined by the *URL* preference variable. Moreover, a GPS timer is activated by a *Timer* class. Each *tim_tick* seconds, this timer calls a method *TimerMethod* that calls a *postData(longitude, latitude, altitude, utm)* function. The function parameters represent the position of GPS device at *utm* time and are defined globally. This position is obtained by *onLocationChanged()*, which is called when the position of GPS device is changed.

The *postData()* function implements a transfer of positioning data to the web service provider. Firstly, a new SOAP request is created addressing the correct web service namespace (*NAMESPACE*) and method name (*METHOD_NAME*). Then, this request is filled with the GPS position, serialized by the *Ksoap2* library and transfers the SOAP message to the web service directly [22]. Note that all wrap and unwrap SOAP messages are performed by the *Ksoap2* library automatically and, therefore, these messages are transparent for the programmer.

SOAP (Service Oriented Architecture Protocol) is a default message transfer protocol in SOA (Service Oriented Architecture) [10]. SOAP messages are used by wrapping application specific XML messages within a plained XML based envelope structure [4].

E. Web services

The term web services describes a set of open standards that enables web based applications to communicate over the Internet with each other and with clients [3]. This set consists of protocols, such as: Extensible Markup Language (XML), Service Oriented Architecture Protocol (SOAP), Web Service Description Language (WDSL) and Universal Description Discovery and Integration (UDDI), where XML is intended for tagging the data, SOAP is a message transfer protocol, WDSL is used for describing the available services, and UDDI is for finding the services over the Internet. Furthermore, web services allow organizations to communicate transparently with other organizations.

System Oriented Architecture (SOA) is a de-facto standard for web service message exchange [34]. SOA is based on the principle of distributed application services that communicate over the Internet with each other through messages.

Because of many standards, the development of web services is difficult. In order to simplify the development, Apache has prepared an Axis2 engine [27] that allows developers to develop web services on a higher level. This engine can be used for developing the drafting-detection web service, as well. Additionally, the developing tool Eclipse [1] was employed.

The web service for drafting-detection in Ironman is context-aware because each SOAP message containing the position of a GPS device in geographical coordinates is identified by the starting number of competitor. In fact, the enacting of a web service can be divided into three tasks:

- a transformation of geographic coordinates to UTM,
- a distance calculation and
- a drafting-detection.

These tasks are described in detail in the rest of the article.

1) Transformation of geographical coordinates to UTM: Usually, a geographical coordinate system is used by GPS, where a position is represented by:

- a latitude and
- a longitude.

Note that GPS devices also provides an altitude. The longitude represents the angle from the center to a particular *parallel* on the surface of the Earth (direction East-West). Longitude is an angle from the center to a particular *meridian* on the surface of the Earth (direction North-South). Both values can be represented as degrees in the form of decimal number or in discrete form: degree, minutes and seconds (DMS).

The Earth is divided by the equator into Northern and Southern hemispheres, whilst the Prime Meridian divides

Algorithm 1 Android Java Class Implementing GPS Listener.

```
1: public class Send extends Activity implements LocationListener {
2:   ...
3:   private LocationManager locationManager;
4:   private SoapSerializationEnvelope envelope;
5:   private HttpTransportSE androidHttpTransport;
6:   private Timer myTimer;
7:   ...
8:   public void onCreate(Bundle savedInstanceState) {
9:     str_number = Prefs.getStartNumber(getApplicationContext());
10:    tim_tick = Prefs.getTimerTick(getApplicationContext());
11:    URL = Prefs.getURL(getApplicationContext());
12:          locationManager = (LocationManager) getSystemSer-
   vice(Context.LOCATION_SERVICE);
13:    androidHttpTransport = new HttpTransportSE(URL);
14:
15:    myTimer = new Timer(); // activate the GPS timer
16:    myTimer.schedule(new TimerTask() {
17:      @Override
18:      public void run() { TimerMethod(); }
19:    }, 0, tim_tick*1000);
20:  };
21:
22:  private void TimerMethod() {
23:    this.runOnUiThread(Timer_Tick);
24:  };
25:
26:  private Runnable Timer_Tick = new Runnable() {
27:    public void run() {
28:      if(gps_status == GPS_ENABLED)
29:        postData(longitude, latitude, altitude, utm);
30:    }
31:  };
32:
33:  public void postData(double lon, double lat, double alt, long unt) {
34:   SoapObject request = new SoapObject(NAMESPACE, METHOD_NAME);
35:    PackRequest(request, str_num, lon, lat, alt, unt);
36:    envelope = new SoapSerializationEnvelope(SoapEnvelope.VER11);
37:    envelope.setOutputSoapObject(request);
38:    try {
39:      androidHttpTransport.call(SOAP_ACTION, envelope);
40:    } catch(Exception e) {
41:      Error.setText(e.getLocalizedMessage());
42:    }
43:  };
44:
45:  public void onLocationChanged(Location location) {
46:    ...
47:    longitude = location.getLongitude();
48:    latitude = location.getLatitude();
49:    ...
50:    gps_status = GPS_ENABLED;
51:  };
52: }
```

it into Eastern and Western hemispheres. Latitude captures the values from $0°$ to $90°$ in the Northern and the values from $0°$ to $-90°$ in the Southern hemispheres. On the other hand, longitude captures the values from $0°$ to $180°$ in the Eastern and the values from $0°$ to $-180°$ in the Western hemispheres.

With geographic coordinates it is not easy to calculate. Therefore, these need to be transformed into the metric 3-dimensional coordinate system UTM (Universal Transverse Mercator system). This system represents a Mercator projection of the Earth to a plane and is divided into 60 longitude and 30 latitude zones. Each position in this coordinate system is presented as quadruple $\langle lon\_zone, lat\_zone, east, north\rangle$, where lon_zone and lat_zone are the numbers of the longitude and the latitude zone, whilst $east$ is the projected distance from the Prime Meridian, and the $north$ the projected distance from the equator. Both distances are defined in meters.

Although the basis of the geographical coordinates transformation into coordinate system UTM represents a basic trigonometric and algebraic functions the transformation formulas are complex [24]. Therefore, for the necessity of this proof of concept we decided to use the existing implementation of author [32] in Java.

2) Distance calculation: However, the traditional Euclidian distance is used for distance calculation. Let us suppose that the positions of the two competitors $A = (x_1, y_1)$ and $B = (x_2, y_2)$ are given. Then, the distance between both is expressed in 2-dimensional space, as follows:

$$dist2(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (1)$$

However, in 3-dimensional space the Euclidian distance is expressed as:

$$dist3(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \quad (2)$$

if we suppose that the positions of competitors are given as $A = (x_1, y_1, z_1)$ and $B = (x_2, y_2, z_2)$.

3) Drafting-detection: In order to handle a lot of requests addressed by competitors' GPS positions (usually, one position per second) and to make code for serving these requests as simple as possible, an efficient data structure is necessary for the web service. The *map* table is placed on the top of the data structure hierarchy. This maps the starting number of competitor to his current place in the race. In fact, each competitor is represented by the 6-tuple $\langle i, x, y, z, t, l\rangle$ (denoted as $item$ data structure), where i denotes the starting number of the competitor, (x, y, z) the UTM position, t the time of event registration and l the calculated number of kilometers covered by the i-th competitor, i.e. his traveled path length. However, the calculated path length l is obtained by projection of the competitor's current position to the line connecting the points that are sampled the bicycle course with the precise GPS device in small intervals of time (e.g., 1 second).

However, the path length l has an impact on the competitor's current placing. The higher the path length the better his current place. Essentially, this place is gained by a sorting of the map array with regard to the descending number of kilometers covered. Because at one time only one request is handled (serialization), a small number of exchanges is necessary by this sorting algorithm.

An algorithm for drafting-detection in Ironman (Algorithm 2) is an implementation of WTC rules, as described in Sect. II-B.

Note that, in Algorithm 2, an additional 2-dimensional array $viol[i][j]$ is used that contains the starting time of drafting violation between competitors i and j in seconds. However, if the drafting-detection is occurring, i.e. the Euclidian distance $dist2(i, i-1)$ between competitor i and its predecessor $i-1$ amounts to less than 7 meters, for more than 20 seconds, a drafting violation is announced. Note that the Euclidian distance $dist2()$ in 2-dimensional space is used here. However, if the time of sampling

Algorithm 2 Algorithm for the drafting-detection in Ironman.

```
1: if(dist2(item[map[i]].l, item[map[i-1]].l) < 7) { // for i > 1
2:     if(viol[map[i]][map[i-1]] == 0) {
3:         viol[map[i]][map[i-1]] = item[map[i]].t;
4:     } else if((item[map[i]].t-viol[map[i]][map[i-1]]) > 20) {
5:         System.out.println( "Competitor " + i + " drafts the competitor " + i-1);
6:     }
7: } else
8:     viol[map[i]][map[i-1]] = 0;
9: }
```

the competitor's position is relatively small, the third dimension can be neglected.

IV. Experiments and results

The goal of experiments was to show that widespread mobile devices can be used in real-time applications for precise object positioning. In this sense, three experiments were conducted:

- comparing the precision of various GPS devices by positioning of a reference point on the Earth,
- comparing the reference distances on the Earth with the distances that were calculated using data measured by GPS devices statically,
- comparing the reference distances on the Earth with the distances that were calculated using data measured by GPS devices dynamically and
- simulation of drafting-detection.

In the first experiment, a reference point on the Earth was selected and its absolute position was measured by four various GPS devices, as follows:

- differential GPS logger Sanav ML-7,
- differential GPS logger Garmin Etrex-H,
- smart-phone Samsung Gallaxy and
- smart-phone HTC Wildfire.

The main characteristic of the GPS logger is that it can log the current position of a GPS device at a predefined time interval into an internal storage. These positions are copied into a personal computer for additional analysis. Typically, data are saved in GPGGA records [19]. Additionally, these loggers are able to provide differential GPS correction.

The results of the experiment are illustrated in Table III. Note that the data in this table were obtained within intervals of one second. In fact, the average measurements of latitude, longitude, altitude, and distance, calculated according to Equation (1) are presented. However, the average point (line Total in the table) was taken as reference point to calculate the distance. Furthermore, the standard deviations of the average measurements (Stdev1, Stdev2, Stdev3 and Stdev4) are also presented in the table.

It can be seen from Table III that the position of the selected point was measured differently by each device. On average, the position was measured within an accuracy of 1.32 meters. According to altitude, the most accurate was the Garmin Etrex-H device because the measurements were performed at an altitude of 190 meters.

In the second experiment, 14 co-linear points were selected within a plain on the Earth. These points were arranged at distances of one meter between each other. Then, a walk across these was initiated and the appropriate distances from the starting point were calculated using the GPS positions. At each reference point, a halt of 10 seconds was taken. Because this time was enough for the GPS devices to precisely calculate the current positions this experiment was identified as a statical measuring of distances between reference points. Here, the same types of GPS devices were employed as in the first experiment. The average results of calculating the distances after 10 walks, are presented in Picture 6, where the line *Distance* denotes the real reference points.

As can be seen from Picture 6, the logger differential GPS devices (Sanav ML-7 and Garmin Etrex-H) measured the GPS distances of the reference points less precisely than the smart-phones (Samsung Gallaxy and HTC Wildfire), in reality.

The third experiment was performed similarly to the second. However, no halt was taken between walking. Thereby, as the devices have insufficient time for determining the current position precisely, the experiment was also identified as dynamically measuring the distances between the reference points. In this experiments, the uniform movement of GPS devices across reference points was observed. The moving speed of the walk across reference points was $1\ m/sec$. Obviously, this movement is much closer to reality than the movement in the second experiment. The results from calculating the distances obtained from the GPS positions, are presented in Picture 7. Note that the line *Distance* denotes the real reference points, whilst the other lines were calculated from the reported GPS positions. From Picture 7, it can be observed that all the GPS devices used in this experiment followed the real *Distance* line closely except for the smart-phone Samsung Gallaxy.

Finally, the drafting-detection was simulated. The simulation was performed as follows. A competitor A competes with competitor B over a bicycle course of length 3.332 kilometers. Note that the course was flat and only one lap was ridden. Competitor B started one minute after competitor A. Each competitor was equipped with a HTC Wildfire smart-phone. Data about the sports activity were transmitted to the Internet and at the same time, logged into internal storage, while the simulation can be tracked on the Internet using Google Maps online (Fig. 8). Further, the logged data can be downloaded on a personal computer for further analysis. From researchers point of view, however, an offline analysis of logged data was interested in order to explore if these data were accurate enough that the algorithm for drafting detection could be convinced that the drafting condition was reliable arisen.

The results of the simulation are presented in Fig. 9 that the drafting-violation of competitor B by competitor A at 1.591 kilometers was detected, i.e. after 4:52 minutes of the race. Competitor B remained within the drafting zone of competitor A for 2:46 minutes (or the whole 733 meters). After 7:28 minutes (at 2,324 meters) competitor B overtook competitor A and completed the course in

TABLE III.
PRECISIONS OF VARIOUS GPS DEVICES BY POSITIONING A REFERENCE POINT ON THE EARTH.

Device	Latitude	Stdev1	Longitude	Stdev2	Altitude	Stdev3	Dist	Stdev4
ML-7	46.6159988	2.30E-05	16.1487849	9.30E-06	218.65	1.12	1.33	1.57E+00
Etrex-H	46.6160046	9.31E-06	16.1487547	1.05E-05	186.29	0.22	1.21	0.22E+00
Gallaxy	46.6160096	3.27E-06	16.1487881	4.48E-06	235.08	1.04	1.42	4.98E-06
Wildfire	46.6160095	4.99E-14	16.1487544	2.14E-14	238.00	0.00	1.31	0.00E+00
Total	46.6160056	8.90E-06	16.1487705	6.08E-06	219.51	0.59	1.32	0.45E+00

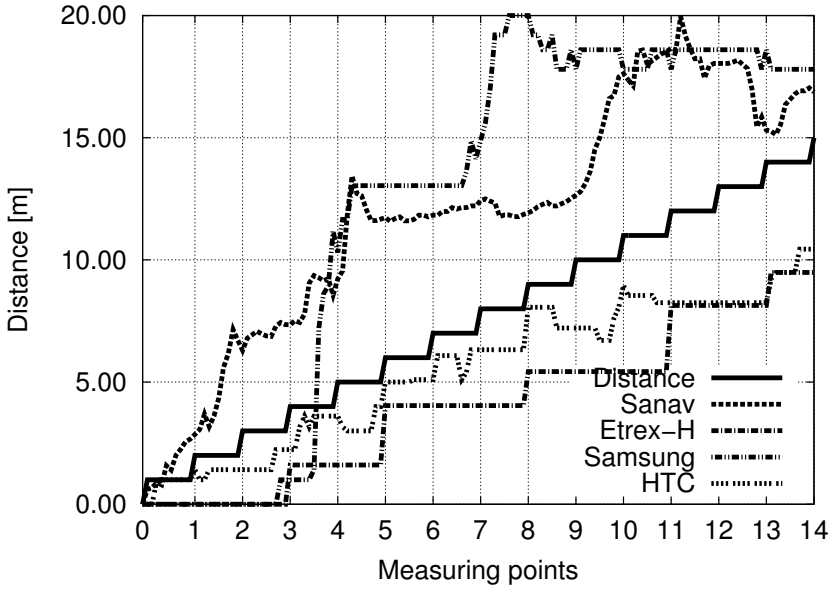


Figure 6. Comparison between the calculated distances obtained by the GPS devices statically

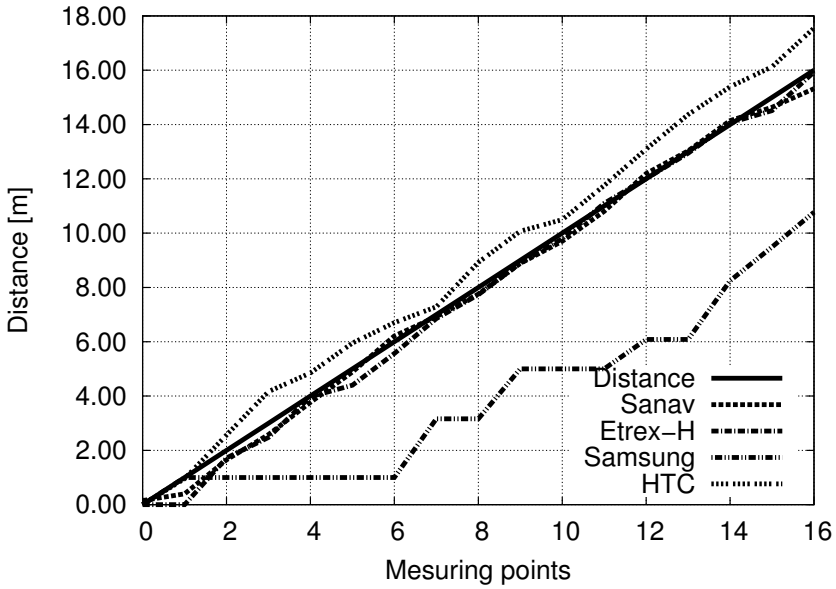


Figure 7. Comparison between the calculated distances obtained by GPS devices dynamically

9:38 minutes, whilst competitor B finished in 10:59 minutes. That is, competitor B overtook competitor A for 1:21 minutes. This simulation showed that the drafting condition could be successfully detected using the mobile smart-phones.

In summary, HTC Wildfire shows that it can be a good candidate for usage in real-time application for drafting-detection in Ironman because it includes a very precise GPS receiver and the reliable UMTS transmitter for connection to the Internet. However, to use this smart-phone on a bicycle would be awkward because of too much size and weight. Competitors in bicycle races are very sensitive to any excessive weight loaded on the bicycle. Furthermore, an additional problem represents the

Figure 8. Simulation of drafting (Powered by Google Maps)

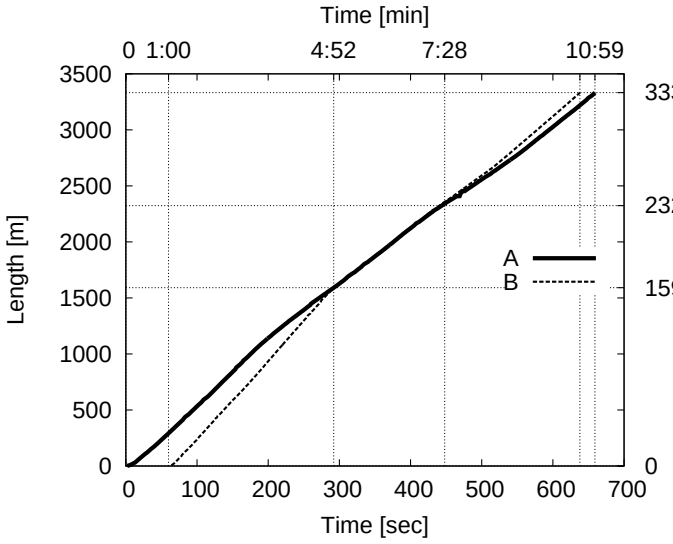


Figure 9. Simulation of drafting-detection

power consumption of smart-phones, which is too high when these are fully-operational.

Smart-phones are dedicated to general usage and support various applications that additionally spend the power consumption. Therefore, only specialized hardware devices with precise GPS, efficient HSPA, and low power consumption, can solve this problem as a whole in the future.

V. Conclusions

In this proof of concept, we have shown that drafting-detection in Ironman is not an illusion and could be used in the near future in practice. This speculation is confirmed by the following facts. The Galileo GPS navigation system that will improve the precision of differential GPS is approaching the end of its construction. An evolution of the mobile network is converging into the fifth generation 5G. The explosion of ubiquitous computation is leading to the creation of specialized hardware devices with integrated GPS and HSPA features, and low-power consumption. These devices are more suitable for use in the application for drafting-detection in Ironman than widespread mobile devices, e.g., smart-phones. In fact, the similar technology is used today by TV transmissions of the greatest bicycle races in the world, e.g., Tour de France, Giro d'Italia, Vuelta a Espana. There, some competitors bear mobile devices that reflect they positions in the race on a graphic illustrating the race course and broadcast this graphic to televisions around the world.

Although we have focused in Ironman, however, this application can be employed without any changes in other triathlons as well. In future work, this real-time application would be integrated into the domain-specific language EasyTime that controls timing systems for measuring time in various sporting competitions.

References

[1] Abbott, D.: Embedded Linux Development using Eclipse. Elsevier Inc., Burlington, US (2009).

[2] Agnew, D. C., Larson, K. M.: Finding the repeat times of the GPS constellation. GPS Solutions, Springer Verlag, Berlin, vol. 11, no. 1, p. 71-76 (2007).

[3] Alonso, G., Casati, F., Kuno, H., Machiraju, V.: Web Services: Concepts, Architectures and Applications. Springer Verlag, Berlin (2010).

[4] Bosak, J., Bray, T.: XML and the Second-Generation Web. Scientific American, vol. 280, no. 5, p. 89-93 (1999).

[5] Braasch, M. S., Van Dierendonck, A. J.: GPS receiver architectures and measurements. Proceedings of the IEEE, p. 48–64 (1999).

[6] Burnette, E.: Hello, Android: Introducing Google's Mobile Development Platform. Pragmatic Programmers LLT., New York, US (2010).

[7] Chen, G., Grejner-Brzezinska, D. A.: Land-Vehicle Navigation Using Multiple Model Carrier Phase DGPS/INS, Proceeding of the 2001 American Control Conference, Arlington, VA, vol. 3, p. 2327–2332 (2001).

[8] Darcey, L., Conder, S.: Android: Wireless Application Development. Addison Wesley, Upper Saddle River NJ, US (2011).

[9] Dierendonck, A. J.: GPS Receivers. In: B. W. Parkinson, and J. J. Spilker (Eds.): Global Positioning System: Theory and Applications. American Institute of Aeronautics and Astronautics, Inc, vol. 1 (1996).

[10] Erl, T.: SOA: Principles of Service Design. Prentice Hall Inc. Upper Saddle River, NJ, US (2009).

[11] Farrell, J., Givargis, T.: Differential GPS Reference Station Algorithm-Design and Analysis. IEEE Transactions on Control Systems Technology, vol. 8, no. 3, p. 519–531 (2000).

[12] Fister, I. Jr., Fister, I., Mernik, M., Brest, J.: Design and implementation of domain-specific language EasyTime. Computer Languages, Systems & Structures, Arcticle in press, doi: 10.1016/j.cl.2011.04.001 (2011).

[13] I. Jr. Fister, I. Fister, M. Mernik, J. Brest.: Design and implementation of domain-specific language Easytime. Computer Languages, Systems and Structures, 2011, 37(4), 276–304.

[14] I. Jr. Fister, M. Mernik, I. Fister, D. Hrnčič.: Implementation of EasyTime Formal Semantics using a LISA Compiler Generator. Computer Science and Information Systems, 2012, 9(3): 1019–1044.

[15] Fling, B.: Mobile Design and Development. O'Really Media Inc., Sebastopol, CA, (2009).

[16] Gehue, H., Hewerdine, W.: Use of DGPS Corrections with Low Power GPS Receivers in a Post SA Environment. Proceedings of IEEE Aerospace Conference, Big Sky, MT, vol. 3, p. 1303–1308 (2001).
[17] Genco, A., Sorce, S.: Pervasive Systems and Ubiquitous Computing. WIT Press, Southampton, UK, (2010).
[18] Hansmann, U., Merk, L., Nicklous, M. S., Stober T.: Pervasive Computing. Springer-Verlag, Berlin, Germany, (2003).
[19] Ibrahim, D.: Design of a GPS data logger device with street-level map interface. Advances in Engineering Software, Elsevier Science Ltd., Oxford, UK, vol. 41, no. 6, p. 859–864 (2010).
[20] Ito, M., Kobayashi, K., Watanabe, K.: Study on the Method to Control the Autonomous Vechile by Using DGPS. Proceeding of the 41st SICE Annual Conference, Vol. 4, Osaka, Japan, p. 2382–2384 (2002).
[21] Kaplan, E. D.: Understanding GPS: Principles and Applications. Artech House, Boston, US (1996).
[22] Knutsen, J.: Web Service Clients on Mobile Android Devices. MSC, Norwegian University of Science and Technology, Department of Computer and Information Science, Trondheim, (2009).
[23] Misra, P., Enge, P.: Global Positioning System: Signals, Measurements, and Performance. Ganga-Jamuna Press, Massachusetts, US (2010).
[24] TM8358.2: The Universal Grids: Universal Transverse Mercator (UTM) and Universal Polar Stereographic (UPS). Defense Mapping Agency, USA (1989).
[25] Parkinson, B. W., Spilker Jr., J. J. (eds.): Global Positioning System: Theory and Applications, Progress in Astronautics and Aeronautics. American Institute of Aeronautics and Astronautics, Inc, vol. 164 (1996).
[26] Patil B., Saifullah Y., Faccin S., Sreemanthula S., Aravamudhan L., S.Sharma, Mononen R.: IP in Wireless Networks. Prentice Hall Inc., Upper Saddle River, NJ, US (2003).
[27] Perera, S., Herath, C., Ekanayake, J., Chinthaka, E., Ranabahu, A., Jayasinghe, D., Weerawarana, S., Daniels, G.: Axis2, Middleware for Next Generation Web Services. Proceedings of the IEEE International Conference on Web Services, IEEE Computer Society, Washington, DC, USA, p. 833–840 (2006).
[28] Poslad, S.: Ubiquitous Computing: Smart Device, Environment, and Interactions. John Willey & Sons Ltd., Chichester, UK (2009).
[29] Prasad, R., Ruggieri, M.: Applied Satellite Navigation Using GPS, GALILEO, and Augmentation Systems. Artech House, Boston, US (2005).
[30] Raquet, J. F.: Multiple GPS Receiver Multipath Mitigation Technique. IEEE Proceedings of Radar, Sonar and Navigation, vol. 149, pp. 195–201 (2002).
[31] Saad, A. Z.: Next Generation Mobile Communications Ecosystem. John Willey & Sons Ltd., Chichester, UK (2011).
[32] Coordinate conversions made easy: http://www.ibm.com/developerworks/ java/library/j-coordconvert (2011).
[33] Schmid, A., Neubauer, A., Ehm, H., Weigel, R., Lemke, N., Heinrichs, G., Winkel, J., Avila-Rodriguez, J. A., Kaniuth, R., Pany, T., Eisfeller, B., Rohmer, G., Niemann, B., Overbeck, M.: Enabling Location-Based Services with a Combined Galileo/GPS Receiver Architecture, Proceedings of the ION GNSS Conference, p. 1–12 (2004).
[34] Schroth, C., Janner, T.: Web 2.0 and SOA: Converging Concepts Enabling the Internet of Services. IEEE Computer Society, IT Professional, vol. 9, no. 3, p. 36-41 (2007).
[35] Shuxin, C., Yongsheng, W., Fei, C.: A Study of Differential GPS Positioning Accuracy. 3th International Conference on Microwave and Millimeter Wave Technology 2002, p. 361–364 (2002).
[36] Soares, M. G., Malheiro, B., Restivo, F. J.: An Internet DGPS Service for Precise Outdoor Navigation. IEEE Conference Emerging Technologies and Factory Automation, vol. 1, p. 512–518 (2003).
[37] Vickery, J. L., King, R. L.: An Intelligent Differencing GPS Algorithm and Method for Remote Sensing. IEEE International Geoscience and Remote Sensing Symposium, vol. 2, p. 1281–1283, (2002).
[38] World Triathlon Corporation: IRONMAN Rules. WTC Technical Report (2010).
[39] Žalik, B.: An efficient sweep-line Delaunay triangulation algorithm. Computer-Aided Design, vol. 37, no. 10, p. 1027-1038 (2005).

Solving Problems of Imperfect Data Streams by Incremetnal Decision Trees

Hang Yang
Department of Computer and Information Science
University of Macau
Macau SAR, China
henry.yh@gmail.com

***Abstract*—Big data is a popular topic that attracts highly attentions of researchers from all over the world. How to mine valuable information from such huge volumes of data remains an open problem. Although fast development of hardware is capable of handling much larger volume of data than ever before, in the author's opinion, a well-designed algorithm is crucial in solving the problems associated with big data. Data stream mining methodologies propose one-pass algorithms that discover knowledge hidden behind massive and continuously moving data. These provide a good solution for such big data problems, even for potentially infinite volumes of data. In this paper, we investigate these problems and propose an algorithm of incremental decision tree as the solution.**

***Index Terms*—Data stream Mining; Big data; Decision Trees; Classification Algorithms.**

I. Introduction

Big data has become a hot research topic, and how to mine valuable information from such huge volumes of data remains an open problem. Many research institutes worldwide have dedicated themselves to solving this problem. The solutions differ from traditional methods, where learning process must be efficient and incremental.

Processing big data presents a challenge to existing computation platforms and hardware. However, according to Moore's Law, CPU hardware may no longer present a bottleneck in mining big data due to the rapid development of the integrated circuit industry. Then, what is the key point of big data mining?

In author's opinion, a well-designed mining algorithm is crucial in solving the problems associated with massive data. The methodology shall efficiently discover the hidden information behind massive data and then present the real-time findings in a user-friendly way.

One on hand, amongst those methods of data mining, fortunately, the decision tree is a non-linear supervised-learning model, which classifies data into different categories and makes a good prediction for unseen data. The decision model is into a set of if-then-else rules within a tree-like graph. The high-degree comprehension of tree-like model makes it easy to understand the discovered knowledge from massive and big data, for both human and machine. Based on data stream mining, incremental decision tree has become a popular research topic.

On one hand, however, imperfect data problem is a barrier of the mining process. Missing data, either value- or case-based, will increase difficulties to data mining process. Noisy data are usually the culprits when contradicting samples appear. Bias data causes an irregular class distribution that will influence the reliability of evaluating model. On the other hand, decision tree model will face tree size explosion and detrimental accuracy problems when including imperfect data. In the past decade, incremental decision trees algorithms [1,2,3,4] apply the Hoeffding bound with a tie-breaking threshold, for dealing with the problem of tree-size explosion. This threshold is a fixed user-defined value. We do not know what the best configuration is unless all possibilities have been tried, but undesirable in practical. Although the pre-processing technique is to handle these imperfections, it may not be possible because of the nature of incremental access to the constantly incoming data streams. In addition, concept-drift problem is a characteristic of time-changing data, referring to that the most types of an attribute remain the same while only particular type changes with time. This problem will reduce the utility of a decision model that increases the difficulties of data mining.

II. Imperfect Data Streams

A. Nosiy Data

A significant advantage of decision tree classification is that the tree-like graph has a higher degree of interpretability. Ideally we want a compact decision tree model that possesses just sufficient rules for classification and prediction with certain accuracy and interpretability. One culprit that leads to tree size explosion is noisy data, a well know phenomenon is called over-fitting in decision trees. Noise data in data samples are considered as a type of irrelevant or meaningless data, which do not typically reflect the main trends but makes the identification of these trends more difficult. However, prior to the start of the decision tree induction, we do not know which samples are noise data; filtering noise is thus difficult.

Noise data is considered a type of irrelevant or meaningless data that does not typically reflect the main trends but makes the identification of these trends more difficult. Non-informative variables may be potentially

doi:10.4304/jetwi.5.3.322-331

random noise in the data stream. It is an idealized but useful model, in which such noise variables present no information-bearing pattern of regular variation. Tree size explosion problem, not only exists in incremental trees [1,2], but also in traditional trees [5,6,7]. However, data stream mining cannot eliminate those non-informative candidates in preprocessing before starting classification mining, because the concept-drift problem may also bring non-informative variables into informative candidates.

On one hand, a previous study [8] reenacts this phenomenon that the inclusion of noise data reduces the accuracy and increasing model size. This consequence is undesirable in the decision tree classification. There has been an attempt to reduce the effect of noise by using supplementary classifiers for predicting missing values in real-time and minimizing noise in the important attributes [9]. Such methods still demand extra resources in computation.

B. Missing Data

It is known that a major cause of over-fitting in a decision tree is the inclusion of contradicting samples in the learning process. Noisy data and missing values are usually the culprits when contradicting samples appear. Unfortunately, such samples are inevitable in distributed communication environments such as wireless sensor network (WSN). Two measures are commonly employed to define the extent of values missing from a set of data [10]: the percentage of predictor values missing from the dataset (the value-wise missing rate) and the percentage of observation records that contain missing values (the case-wise missing rate). A single value missing from the data usually indicates transmission loss or malfunctioning of a single sensor. A missing data value record may result from a broken link between sensors. In WSN, can distinguish the missing data to two categories:

- Incomplete data with lost values: Because of an accidence of sensor itself, like a crash or a reboot, the instant data of the last event before the accidence will be lost. Hence, such kind of missing values is permanent, which is lost forever.
- Unstable data with late arrival: Because of temporal disconnection or network delay, data stream capture faces asynchronous issues. The missing value caused by asynchronous is not permanent, which is temporally lost and will arrive in a short while.

C. Bias Data

Bias data is also called imbalanced distribution data. The term "imbalanced" refers to irregular class distributions in a dataset. For example, a large percentage of training samples may be biased toward class A, leaving few samples that describe class B. Imbalanced classification is a common problem. This problem occurs when the classifier algorithm is trained with a dataset, in which one class has only a few samples, and there are a disproportionally large number of samples in the other classes. Imbalanced data causes classifiers to be over-fitted (i.e., produce redundant rules that describe duplicate or meaningless concepts), and, as a result, perform poorly, particularly in the identification of the minority class.

Most of the standard classification algorithms assume that training examples are evenly distributed among different classes. In practical applications where this was known to be untrue, researchers addressed the problem by either manipulating the training data or adjusting the misclassification costs. Resizing training data sets is a common strategy that attempts to downsize the majority class and over-samples the minority class. Many variants of this strategy have been proposed [10,11,12]. A second strategy is to adjust the costs of misclassification errors to be biased against or in favor of the majority and minority classes, respectively. Using the feedback from the altered error information, researchers then fine-tune their cost-sensitive classifiers and post-prune the decision trees in the hope of establishing a balanced treatment of each class in the new imbalanced data collected by the network [12,13]. However, they are not suitable for data stream mining because of the nature of incremental access to the constantly incoming streams.

D. Concept-drift Data

Data stream is also an infinite big data scenario that the underlying data distribution of newly arrival data may be appeared differently from the old one in the real world, so called concept-drift problem. For example, click-streams of user's navigating e-commerce website may reflect the preferences of purchase through the analysis systems. When people's preferences of product change, however, the old user's behavior model is not applicable any more that the drifting of concepts appears.

The hidden changes in the attributes of data streams will cause a drift of target concept. In terms of the occurring frequency, commonly it can be distinguished in two kinds: abrupt drift and gradual drift. For data streams, the data arrive continuously that the concept-drift is local, for instance, only particular types of attribute may change with time while the others remain the same.

III. INCREMENTAL DECISION TREE ALGORITHMS

A. Decision Tree Learning using Hoeffding Bound

A decision-tree classification problem is defined as follows: N is the number of examples in a dataset with a form (X, y), where X is a vector of I attributes and y is a discrete class label. I is the number of attributes in X. k is the index of class label. Suppose a class label with the k^{th} discrete value is y_k. Attribute X_i is the i^{th} attribute in X, and is assigned a value of x_{i1}, x_{i2}… x_{iJ}, where $1 \le i \le I$ and J is the number of different values of X_i. The classification goal is to produce a decision tree model from N examples, which predicts the classes of y in future examples with high accuracy. In stream mining, the example size is very large or unlimited that $N \rightarrow \infty$.

VFDT [1] constructs an incremental decision tree by using constant memory and constant time-per-sample. It is a pioneering predictive technique that utilizes the Hoeffding bound (HB) that $HB = \sqrt{R^2 \ln\left(\frac{1}{\delta}\right)/2n}$, where R is the range of classes distribution and n is the number of instances which have fallen into a leaf. Sufficient

statistics is used to record the counts of each value x_{ij} of attribute X_i belonging to class y_k. The solution, which doesn't requiring the full historical data, is a node-splitting criterion using a HB. To evaluate a splitting-value for attribute X_i, it chooses the best two values. Suppose x_{ia} is the best value of $H(.)$ where $x_{ia} = \arg\max H(x_{ij})$; suppose x_{ib} is the second best value where $x_{ib} = \arg\max H(x_{ij}), \forall j \neq a$; suppose $\Delta H(X_i)$ is the difference of the best two values for attribute X_i, where $\Delta H(X_i) = \Delta H(x_{ia}) - \Delta H(x_{ib})$. Let n be the observed number of instances, HB is used to compute high confidence intervals for the true mean r_{true} of attribute x_{ij} to class y_k that $r - HB \leq r_{true} < r + HB$ where $r = (1/n)\sum_i^n r_i$. If after observing n_{min} examples, the inequality $r + HB < 1$ holds, then $r_{true} < 1$, meaning that the best attribute x_{ia} observed over a portion of the stream is truly the best attribute over entire stream. Hence, a splitting-value x_{ij} of attribute X_i can be found without full attribute-values, even when we don't know all values of X_i (from x_{i1} to x_{iJ}).

When data contains imperfect values, it may confuse the values of heuristic function. The difference of the best two heuristic evaluation for attribute X_i, where $\Delta\bar{H}(X_i) = H(x_{ia}) - H(x_{ib})$, may be negligible. To solve this problem, a fixed tie-breaking τ, which is a user pre-defined threshold for incremental learning decision tree, is proposed as a pre-pruning mechanism to control the tree growth speed [2]. This threshold constrains the node-splitting condition that $\Delta\bar{H}(X_i) \leq HB < \tau$. An efficient τ guarantees a minimum tree growth in case of tree-size explosion problem. τ must be set before a new learning starts, however, so far there has no a unique τ suitable for all problems. In other words, there is not a single value that works well in all tasks. The choice of τ hence depends on the data and their nature.

B. Evoluation in the Past Decade

According to node-splitting process of a decision tree, we can distinguish it into two categories: singletree algorithm and multi-tree algorithm. Singletree is a decision model that only builds one tree in the tree-building approach while does not require any optional branches or alternative trees. Multi-tree builds a decision tree model dependent on many other trees at the same time. The advantage of singletree is lightweight favored for data streams environment and easy to implement, although in some cases, multi-tree may bring a higher accuracy.

VFDT is the pioneer singletree of using HB to construct incremental decision tree for high-speed data streams, but it can't handle concept drift. Functional tree leaf is originally proposed to integrate to incremental decision tree [3]. Consequently, Naïve Bayes classifier on the tree leaf has improved classification accuracy. The functional tree leaf is able to handle both continuous and discrete values in data streams. OcVFDT [14] provides a solution to deal with unlabeled samples based on VFDT and POSC4.5. The experiment shows four fifths of samples are unlabeled, while the performance still gets close to VFDT of fully labeled streams. OcVFDT is a one-class classification that classifiers are trained to distinguish only a class of objects from all other objects. FlexDT [15] proposes a Sigmoid function to handle noisy data and missing values. Sigmoid function is used to decide what true node-splitting value, but sacrificing algorithm speed.

For handling concept-drift problem, CVFDT [2] proposed a fixed size of sliding-window that integrated to VFDT. It constructs an alternative tree in the tree growing. When tree model is out-of-date within a window, the alternative branch will replace the old one so that it adapts to concept-drift data. HOT [16] proposes an algorithm producing some optional tree branches at the same time, replacing those rules with lower accuracy by optional ones. The classification accuracy has been improved significantly while learning speed is slowed because of the construction of optional tree branches. ASHT [4] is derived from VFDT adding a maximum number of split nodes. ASHT has a maximum number of split nodes. After one node splits, if the number of split nodes is higher than the maximum value, then it deletes some nodes to reduce its size.

IV. Hypothesis and Motivation

A. Hypothesis

The research is on the basic of the following assumptions:

One-pass Process The feature of proposed method implement as a one-pass approach, which requires loading and computing the data records only one time. Therefore, this is potentially applicable for big data, even unbounded data problem.

Data Volume The data is multi-dimensional, with bounded and constant values of attributes. The data is also labeled. A data record is called the instance. The data has a large scale of instances, even infinite. The algorithm builds an incremental decision tree, in which suppose there enough instances for the node splitting using the HB.

Imperfect Data The imperfect data include: the noisy data, the data with missing values, the data with imbalanced class distribution, as well as the data with concept-drift.

Performance Measures Accuracy is the number of correctly classified instances divided by the number of total instances. Tree size is the number of the rules in a decision tree. This also equals to the number of leaves in the tree model. Learning speed is the time to construct the decision tree. It is an immediate time in the incremental learning process. Memory cost is the memory size used to build the tree model.

Classifier Due to the one-pass process, the incremental decision tree implements a test-then-train process. When a new instance arrives, it will traverse from the root to a leaf according to the tree model. This is also a testing process. During the traversing, the node splitting is triggered so that tree model is trained incrementally. Besides, the post-pruning mechanism is infeasible since the nature of fast-moving data scenario. No extra time is allowed to stop tree building and prune tree structure.

Application The result of the proposed methodology is a decision tree model, which presents rules from the root to the leaves. The tree-like structure shows a collection of complex rules intuitively in terms of IF-THEN-ELSE rules. Both human and machine can understand this rules easily.

B. Motivation

In this paper, we propose an incremental decision tree learning method that is suitable for big data analysis. What is the difference between traditional decision tree learning and incremental decision tree learning?

In Figure 1, we provide an example of traditional decision tree learning. The criteria of splitting-node selection is based on heuristic function. For example, ID3 algorithm uses the information entropy while C4.5 applies the information gain as the heuristic function. In general, the traditional tree learning requires loading the full data and analyzes the whole training data to build a decision tree. The splitting criteria is according to the heuristic result, splitting from the attribute with the larger heuristic value, until all candidates become internal nodes.

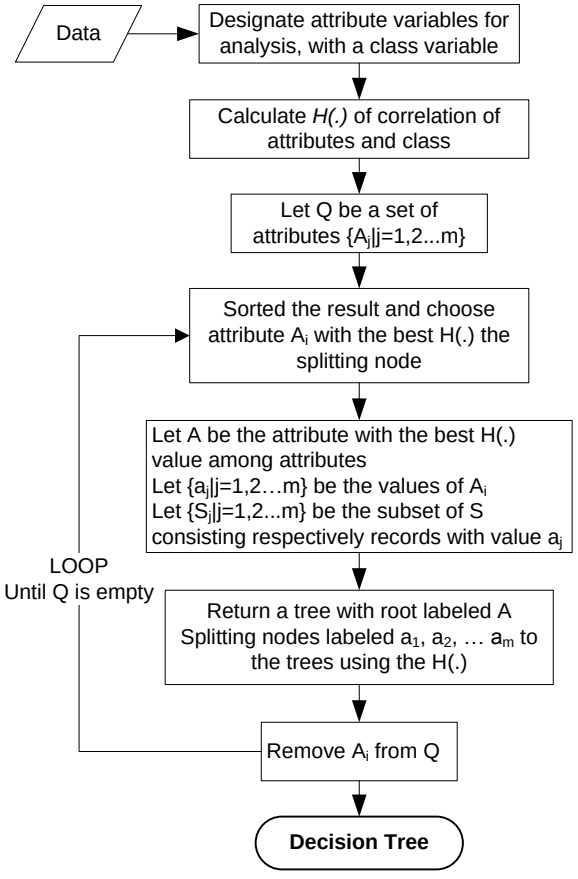


Figure 1. Workflow of Buiding A Traditional Decision Tree.

Differently in Figure 2, incremental learning process using Hoeffding bound in the splitting criteria. It does not require loading the full data, instead, it only needs a part of data to train decision tree model. When new data arrives, the sufficient statistics are updated. If checking condition satisfied, it will compare splitting candidates with the best and the second best heuristic result. In this case, the tree model is updated incrementally, with newly arrival data.

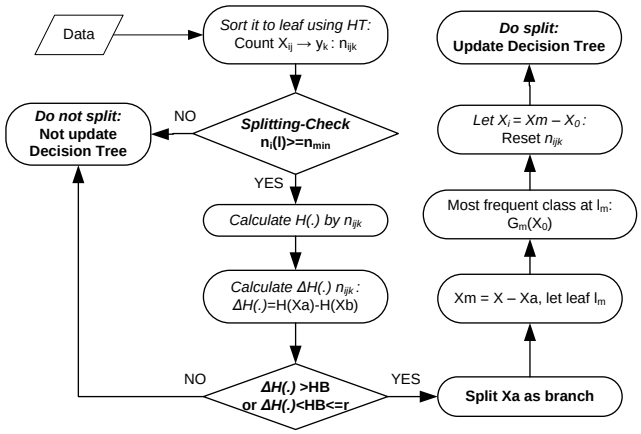


Figure 2. Workflow of Buiding An Incremental Decision Tree.

From the comparison above, obviously, the traditional method is not suitable for big data scenario, because loading full data is inapplicable in practical. That is why we propose an incremental method to deal with big data. The incremental process is applicable for continuously arrival data, even infinite data scenario.

V. METHODOLOGY DESIGN

A. Overall Workflow

The proposed methodology, which inherits the use of HB, implements on a test-then-train approach (Figure 3) for classifying continuously arriving data streams, even for infinite data streams. The whole test-then-train process is synchronized such that when the data stream arrives, one segment at a time, the decision tree is being tested first for prediction output and training (which is also known as model updating) of the decision tree then occurs incrementally.

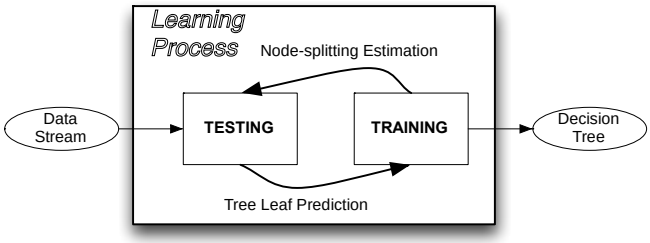


Figure 3. Test-then-train Workflow.

B. Auxiliary Reconciliation Control

The Auxiliary Reconciliation Control (ARC) is a set of data pre-processing functions used to solve the problem of missing data streams. The ARC can be programmed as a standalone program that may run in parallel and in synchronization with the test-and-train operation. Synchronization is facilitated by using a sliding window that allows one segment of data to arrive at a time at regular intervals. When no data arrive, the ARC simply

stands still without any action. The operational rate of the sliding window should be no greater than the speed at which the decision tree building is operated and faster than the speed at which the sensors transmit data.

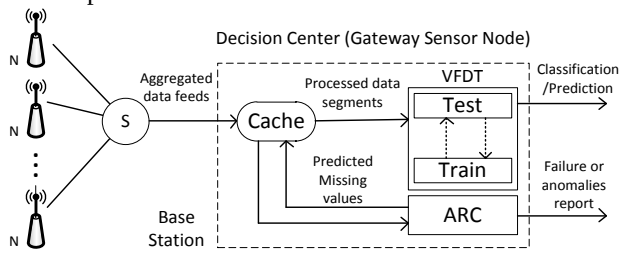


Figure 4. The workflow of ARC in a gateway sensor node.

To tackle the problem of missing values in a data stream, a number of prediction algorithms are commonly used to guess approximate values based on past data. Although many algorithms can be used in the ARC, that deployed should ideally achieve the highest level of accuracy while consuming the least computational resources and time. Some popular choices we use here for simulation experiments include, but are not limited to, mean, naïve Bayesian, and C4.5 decision tree algorithms for nominal data, and mean mode, linear regression, discretized naïve Bayesian and M5P algorithms for numeric data. Missing value estimation algorithms require a substantial amount of past data to function. For example, before using a C4.5 decision tree algorithm as a predictor for missing values, a classifier must be built using statistics from a sample of sufficient size.

C. Functional Tree Leaf

Functional tree leaf [3], can further enhance the prediction accuracy via the embedded Naïve Bayes classifier.. In this paper, we embed the functional tree leaf to improve the performance of prediction by *HT* model. When these two extensions – an optimized node-splitting condition ($\Delta\bar{H} > HB$ or $\mathrm{Opt}.\Phi(HT_x{}^*) > Max.\Phi(HT_x)$ or $\mathrm{Opt}.\Phi(HT_x{}^*) < Min.\Phi(HT_x)$) and a refined prediction using the functional tree leaf – are used together, the new decision tree model is able to achieve unprecedentedly good performance, although the data streams are perturbed by noise and imbalanced class distribution.

For the actual classification, iOVFDT uses a decision tree model HT_F to predict the class label $\widehat{y_k}$ with functional tree leaf F when a new sample (*X*, *y*) arrives, defined as $HT_F(X) \rightarrow \widehat{y_k}$. The predictions are made according to the observed class distribution (OCD) in the leaves called functional tree leaf *F*. Originally in VFDT, the prediction uses only the majority class F^{MC}. The majority class only considers the counts of the class distribution, but not the decisions based on attribute combinations. The naïve Bayes F^{NB} computes the conditional probabilities of the attribute-values given a class at the tree leaves by naïve Bayes network. As a result, the prediction at the leaf is refined by the consideration of each attribute's probabilities. To handle the imbalanced class distribution in a data stream, a weighted naïve Bayes F^{WNB} and an error-adaptive $F^{Adaptive}$ are proposed in this paper. These four types of functional tree leaves are discussed in following paragraphs.

Let Sufficient statistics n_{ijk} be an incremental count number stored in each node in the iOVFDT. Suppose that a node $Node_{ij}$ in *HT* is an internal node labeled with attribute x_{ij} and *k* is the number of classes distributed in the training data, where $k \geq 2$. A vector V_{ij} can be constructed from the sufficient statistics n_{ijk} in $Node_{ij}$, such that $V_{ij} = \{n_{ij1}, n_{ij\,2} \ldots n_{ij\,k}\}$. V_{ij} is the OCD vector of $Node_{ij}$. OCD is used to store the distributed class count at each tree node in iOVFDT to keep track of the occurrences of the instances of each attribute.

***Majority Class Functional Tree Leaf*:** In the OCD vector, the majority class F^{MC} chooses the class with the maximum distribution as the predictive class in a leaf, where F^{MC}: arg max $r = \{n_{i,j,1}, n_{i,j,2} \ldots n_{i,j,r} \ldots n_{i,j,k}\}$, and where $0<r<k$.

***Naïve Bayes Functional Tree Leaf*:** In the OCD vector $V_{i,j} = \{n_{i,j,1}, n_{i,j,2} \ldots n_{i,j,r} \ldots n_{i,j,k}\}$, where *r* is the number of observed classes and $0<r<k$, the naïve Bayes F^{NB} chooses the class with the maximum possibility, as computed by the naïve Bayes, as the predictive class in a leaf. $n_{ij,r}$ is updated to $n'_{i,j,r}$ by the naïve Bayes function such that $n'_{i,j,r} = \mathrm{P}(X|C_f) \cdot \mathrm{P}(C_f) \,/\, \mathrm{P}(X)$, where *X* is the new arrival instance. Hence, the prediction class is F^{NB}: arg max $r = \{ n'_{i,j,1}, n'_{i,j,2} \ldots n'_{i,j,r} \ldots n'_{i,j,k} \}$.

***Weighted Naïve Bayes Functional Tree Leaf*:** In the OCD vector $V_{i,j} = \{n_{i,j,1}, n_{i,j,2} \ldots n_{i,j,r} \ldots n_{i,j,k}\}$, where *k* is the number of observed classes and $0<r<k$, the weighted naïve Bayes F^{WNB} chooses the class with the maximum possibility, as computed by the weighted naïve Bayes, as the predictive class in a leaf. $n_{i,j,r}$ is updated to $n'_{i,j,r}$ by the weighted naïve Bayes function such that $n'_{i,j,r} = \omega_r \cdot \mathrm{P}(X|C_f) \cdot \mathrm{P}(C_f) \,/\, \mathrm{P}(X)$, where *X* is the latest received instance and the weight is the probability of class *i* distribution among all the observed samples, such that $\omega_r = \prod_{r=1}^{k} (v_r / \sum_{r=1}^{k} v_r)$, where $n_{i,j,r}$ is the count of class *r*. Hence, the prediction class is F^{WNB}: arg max $r = \{ n'_{i,j,1}, n'_{i,j,2} \ldots n'_{i,j,r} \ldots n'_{i,j,k} \}$.

***Adaptive Functional Tree Leaf*:** In a leaf, suppose that $V_F{}^{MC}$ is the OCD with the majority class F^{MC}; suppose $V_F{}^{NB}$ is the OCD with the naïve Bayes F^{NB} and suppose that $V_F{}^{WNB}$ is the OCD with the weighted naïve Bayes F^{WNB}. Suppose that *y* is the true class of a new instance *X* and E_F is the prediction error rate using a *F*. E_F is calculated by the average $E=error_i/n$, where *n* is the number of examples and $error_i$ is the number of examples mis-predicted using *F*. The adaptive Functional Tree Leaf chooses the class with the minimum error rate predicted by the other three strategies, where $F^{Adaptive}$: arg min $F = \{E_F{}^{MC}, E_F{}^{NB}, E_F{}^{WNB}\}$.

D. Incremental Optimization

The model is growing incrementally so as to update an optimal decision tree under continuously arriving data. Suppose that a decision tree optimization problem Π is defined as a tuple (X, HT, Φ). The set *X* is a collection of objects to be optimized and the feasible Hoeffding tree

HT solutions are subsets of *X* that collectively achieve a certain optimization goal. The set of all feasible solutions is $HT \subseteq 2^X$ and $\Phi: HT \to \mathbb{R}$ is a cost function of these solutions. The optimal decision tree *HT** exists if *X* and Φ are known, and the subset S is the set of solutions meets the objective function where *HT** is the optimum in this set. Therefore, the incremental optimization functions can be expressed as a sum of several sub-objective cost functions: $\Phi(HT_x) = \bigcup_{D=1}^{M} \Phi_D(HT_x)$, where $\Phi_m : HT \to \mathbb{R}$ is a continuously differentiable function and *M* is the number of objects in the optimization problem. The optimization goal: $minimize\ \Phi(HT_x)\ subject\ to\ HT_x \in X$. $HT(X) \to \hat{y}$ is used to predict the class when a new data sample (*X*, *y*) arrives. So far timestamp *t*, the prediction accuracy defined as: $accu_t = \frac{\sum_{i=1}^{t} Predict(D_i)}{|D_t|}, Predict(D_i) = \begin{cases} 1, if\ \widehat{y_k} = y_k \\ 0, if\ \widehat{y_k} \neq y_k \end{cases}$.

To measure the utility of the three dimensions via the minimizing function, the measure of prediction accuracy is reflected by the prediction error in: $\Phi_1 = 1 - accu_t$.

The new methodology is building a desirable tree model by combining with an incremental optimization mechanism and seeking a compact tree model that balances the objects of tree size, prediction accuracy and learning time. The proposed method finds an optimization function $\Phi(HT_x)$, where $M = 3$. When a new data arrive, it will be sorted from the root to a leaf in terms of the existing *HT* model. When a leaf is being generated, the tree size grows. A new leaf is created when the tree model grows incrementally in terms of newly arrival data. Therefore, up to timestamp *t* the tree size is:

$$\Phi_2 = \begin{cases} size_{t-1} + 1 & , if\ \Delta\bar{H} > HB \\ size_{t-1} & , otherwise \end{cases}.$$

It is a one-pass algorithm that builds a decision model using a single scan over the training data. The *sufficient statistics* that count the number of examples passed to an internal node are the only updated elements in the one-pass algorithm. The calculation is an incremental process, which tree size is "plus-one" a new splitting-attribute appears. It consumes little computational resources. Hence, the computation speed of this "plus one" operation for a new example passing is supposed as a constant value *R* in the learning process. The number of examples that have passed within an interval period of in node splitting control determines the learning time that $\Phi_3 = R \times (n_{y_k} - n_{min})$. n_{min} is a fixed value for controlling interval of node splitting.

Suppose that n_{y_k} is the number of examples seen at a leaf y_k and the condition that checks node-splitting is $n_{y_k} mod\ n_{min} = 0$. The learning time of each node splitting is the interval period – the time defined as Φ_3 – during which a certain number of examples have passed up to timestamp *t*.

Returning to the incremental optimization problem, the optimum tree model is the HT_x structure with the minimum $\phi(x)$. A triangle model is provided to illustrate the relationship amongst the three dimensions – the prediction accuracy, the tree size and the learning time. The three dimensions construct a triangle utility function in Figure 5. A utility function computes the area of triangle, reflecting a relationship amongst the three objects in:

$$\Phi(HT_x) = \frac{\sqrt{3}}{4} \cdot (\Phi_1 \times \Phi_2 + \Phi_1 \times \Phi_3 + \Phi_2 \times \Phi_3)$$

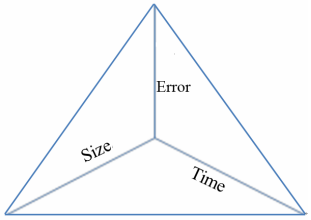


Figure 5. Three-obective Optimization.

The area of this triangle $\Phi(HT_x)$ changes when node splitting happens and the *HT* updates. A min-max constraint of the optimization goal in (4) controls the node splitting, which ensures that the new tree model keeps a $\Phi(HT_x)$ within a considerable range. Suppose that $Max.\Phi(HT_x)$ is a *HT* model with the maximum utility so far and $Min.\Phi(HT_x)$ is a *HT* model with the minimum utility. The optimum model should be within this min-max range, near $Mean.\Phi(HT_x)$:

$$Mean.\Phi(HT_x) = \frac{\text{Max.}\Phi(HT_x) - Min.\Phi(HT_x)}{2}$$

According to the Chernoff bound, we know:

$$|\text{Opt.}\Phi(HT_x{}^*) - Mean.\Phi(HT_x)| \leq \sqrt{\frac{\ln(1/\delta)}{2n}}$$

where the range of $\Phi_x(HT_x)$ is within the min-max model $\text{Min.}\Phi(HT_x) < \text{Opt.}\Phi(HT_x{}^*) < \text{Max.}\Phi(HT_x)$. Therefore, if $\Phi(HT_x)$ goes beyond this constraint, the existing *HT* is not suitable to embrace the new data input and the tree model should not be updated. Node-splitting condition is:

$$\Delta\bar{H} > HB,$$
$$\text{or Opt.}\Phi(HT_x{}^*) > Max.\Phi(HT_x),$$
$$\text{or Opt.}\Phi(HT_x{}^*) < Min.\Phi(HT_x).$$

,

VI. Evaluation

A. Synthetic Data Streams

Hyper-plane data is another typical data streams for concept-drift study [4,17]. We use MOA hyper-plane data generator to simulate the data streams without noise-included (10 attributes and 2 classes, 2 of 10 attributes are randomly drifting). The performance measurement is Interval Test-then-train Evaluation in MOA. The aforementioned contents have verified that Error-adaptive is the best strategy of functional tree leaf, hence, it is applied in this test.

The synthetic streams are marked when attributes drifting. A piece of streams is visualized (50 instances included) in Figure 6. Similar result appears that iOVFDT

outperforms the other two algorithms. In addition, it is obvious that: when a drift occurs, the accuracy is declining consequently. This test shows iOVFDT has a good performance dealing with attributes drifting.

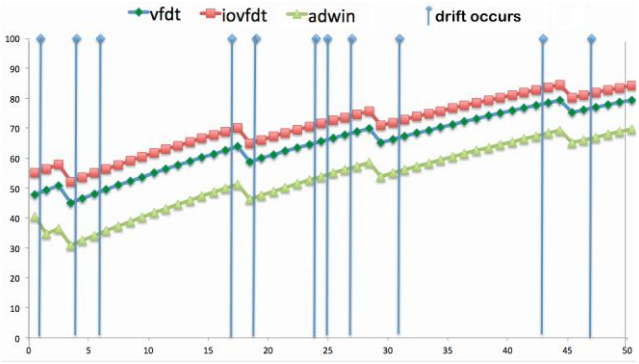


Figure 6. Concept-drift evaluation for hyper-plane data streams.

B. Sensor Data with Missing Values

The complex nature of incomplete and infinite streaming data in WSNs has escalated the challenges faced in data mining applications concerning knowledge induction and time-critical decision-making. Traditional data mining models employed in WSNs work, which are mainly on the basis of relatively structured and stationary historical data, and may have to be updated periodically in batch mode. The retraining process consumes time as it requires repeated archiving and scanning of the whole database. Data stream mining is a process that can be undertaken at the front line in a manner that embraces incoming data streams.

To the best of the author's knowledge, no prior study has investigated the impact of imperfect data streams or solutions related to data stream mining in WSNs, although the pre-processing of missing values is a well-known step in the traditional knowledge discovery process. We propose a holistic model for handling imperfect data streams based on four features that riddle data transmitted among WSNs: missing values, noise, delayed data arrival and data fluctuations. The model has a missing value predicting mechanism called the auxiliary reconciliation control (ARC). A bucket concept is also proposed to smooth traffic fluctuations and minimize the impact caused by late arriving data. Together with the VFDT, the ARC-cache facilitates data stream mining in the presence of noise and missing values. To prove the efficacy of the new model, a simulation prototype is implemented based on ARC-cache and VFDT theories by using a JAVA platform. Experimental results unanimously indicate that the ARC-cache and VFDT method yield better accuracy in mining data streams in the presence of missing values than VFDT only. One reason for this improved performance is ascribed to the improved predictive power of the ARC in comparison with other statistical counting methods for handling missing values, as the ARC computes the information gains of almost all other attributes with non-missing data. In future research, we will continue to investigate the impact of noisy or corrupted data and irregular data stream patterns on data stream mining.

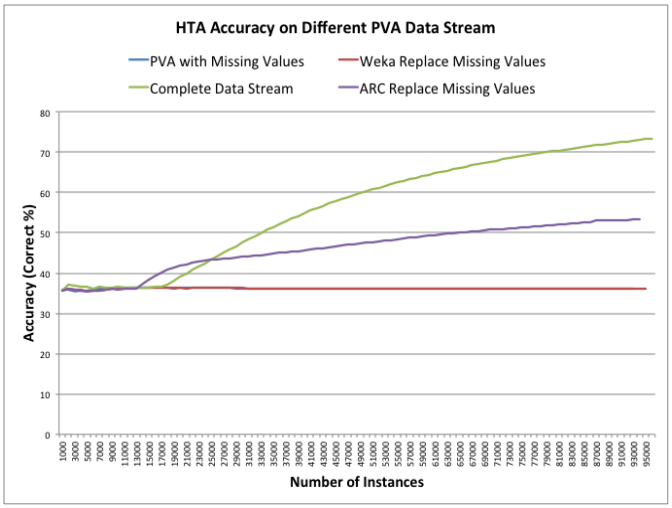


Figure 7. Performance of ARC-cache missing values replacement

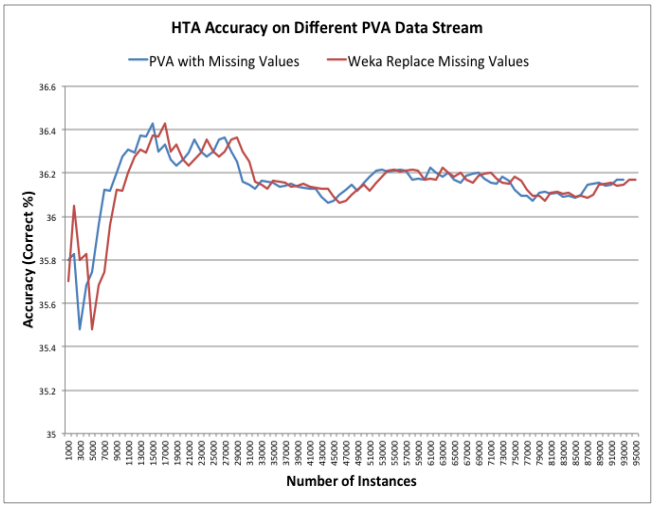


Figure 8. Magnified version of the diagram

In this part, we use a set of real-world data streams downloaded from the 1998 KDD Cup competition provided by Paralyzed Veterans of America (KDD Cup, 1998). The data comprise information concerning human localities and activities measured by monitoring sensors attached to patients. We use the learning dataset (127MB in size) with 481 attributes originally in both numeric and nominal form. Of the total number of 95,412 instances, more than 70% contain missing values.

In common with the previous experiment, we compare the ARC-Cache and VFDT method with the standard missing values replacement method found in WEKA using means. The results of the comparison are shown in Figure 7 and 8. Considering the number of attributes is very large, we apply a moderate window size (W = 100) for the ARC to operate. A complete dataset given by PVA is used to test the ARC-Cache (115MB). The experiment results demonstrate that using WEKA mean values to replace missing data yields the worst level of VFDT classification accuracy. Although using the ARC-Cache to deal with missing values in the dataset

does not yield results as accurate as the complete dataset without any missing values, ARC-Cache performance is much better than that achieved using WEKA means to replace missing values. The enlarged chart shows the WEKA replacement approach has very little effect in maintaining the level of performance because of the very high percentage of missing data (70%) in this extreme example.

C. UCI Data Streams

Dynamic data dominate many modern computer applications nowadays. They are characterized to be vast in size, fast moving in speed and consist of many attributes, which do not make sense individually, but they describe some behavioral patterns when analyzed together over some time. Traditional data mining algorithms are designed to load a full archive of data, and then build a decision model. New data that arrive would have to be accumulated with the historical dataset, and together they would be scanned again for rebuilding an up-to-date decision tree.

TABLE I.

DESCRIPTION OF DYNAMIC DATA

Name	Abbr.	#Ins	#Attr	#Cls	#Nom	#Num
iPod Sales	IP	7882	29	3	16	12
Internet Usage	IU	10104	72	5	71	0
Network Attack	NA	494021	42	23	3	38
Cover Type	CT	581012	55	7	42	12
Robot Sensor	RS	5456	25	4	0	24
Person Activity	PA	164860	6	11	2	3

Six scenarios of dynamic data are tested in the experiment, shown in Table 1. Each type of dynamic data represents typical decision-making problems on the topics of web applications, real-time security surveillance and activities monitoring. The data of web applications are Internet Usage (IU) data and iPod Sales on eBay (IP) data, which are generated from the recording of user's click-streams on the websites. The data of real-time security surveillance are Network Attack (NA) and Cover Type (CT) data. The data of activities monitoring are Robot Sensor (RS) and Person Activity (PA) data. The datasets are extracted from real-world applications that are available for download from UCI Machine Learning Repository.

TABLE II.

ACCURACY ANAYLSIS OF DYNAMIC DATA

Method\Data	RS	IP	IU	CT	PA	NA
C4.5 Pruned	99.45	99.82	82.34	91.01	75.20	99.94
C4.5 Unpruned	99.65	99.70	81.50	92.77	74.10	99.95
Incre.NB	~~55.35~~	~~89.90~~	~~75.29~~	~~60.52~~	~~49.28~~	~~96.55~~
VFDT	~~40.21~~	~~90.7~~	~~79.0~~	~~67.45~~	~~43.75~~	~~98.2~~
		~~1~~	~~6~~			~~7~~
VFDT_NB	~~55.35~~	99.07	82.03	~~77.16~~	~~61.03~~	99.68
VFDT_ADP	~~55.35~~	99.21	82.31	~~77.77~~	~~61.01~~	99.79
iOVFDT_MC	~~71.92~~	~~81.79~~	~~78.24~~	~~70.52~~	~~59.15~~	~~99.23~~
iOVFDT_NB	81.60	98.78	~~78.65~~	90.66	73.45	99.69
iOVFDT_WNB	81.91	98.13	~~78.95~~	90.51	72.35	99.69
iOVFDT_ADP	83.32	98.92	79.84	90.59	73.52	99.85
Standard Deviation.	20.21	6.09	2.31	11.80	11.28	1.08
Variance	408.54	37.14	5.31	139.17	127.13	1.16
Average	72.41	95.61	79.71	80.90	64.28	99.26

From Table 2, in general, it is observed that C4.5 had better accuracy than the other methods in all tested datasets because it built its decision model from the full dataset. Therefore it can attain a globally best solution by going through all the training data at one time. The other methods are incremental learning process that obtained a locally optimum solution in each pass of data stream. The strikethroughs indicate those accuracies that are below the average. Obviously, one can see that only C4.5 and iOVFDT_ADP (iOVFDT with Error-adaptive functional tree leaf) are able to achieve a 'full win' of satisfactory accuracies over the average across all the datasets. Fig. 8.1 shows a graphical representation of the accuracies in the form of a stacked bar chart – despite C4.5, the iOVFDT family of algorithms (except MC) obtains pretty good accuracies. Therefore, when batch learning such as C4.5 is not feasible or available in scenarios of dynamic data stream mining, iOVFDT_ADP would be a good candidate.

Table 3 shows the model size (the number of nodes / the number of leaves) which is calculated as the number of leaves over the number of nodes for different datasets. For all dataset, C4.5 built the decision model requiring largest tree size. Naïve Bayes does its prediction by using distribution probabilities, so that the decision model does not exhibit a tree-like structure. Although smaller tie-breaking threshold might bring respectively smaller tree size for VFDT, the accuracy is obviously worse than iOVFDT. It is interesting to see that the size of a globally best model (C4.5) is not much bigger than a locally optimum model (iOVFDT) because the latter algorithm allows tree to grow incrementally over time.

TABLE III.

MODEL SIZE ANAYLSIS OF DYNAMIC DATA

	RS	IP	IU	CT	PA	NA
C4.5 Pruned	18/35	20/39	847 /911	10149 /20297	13265 /24120	724 /838
C4.5 Unpruned	22/43	24/46	1028 /1267	14903 /29805	6467 /10357	679 /801

IncreNB	N/A	N/A	N/A	N/A	N/A	N/A
VFDT	1/1	5/9	46/47	127/253	167/187	87/94
VFDT_NB	1/1	5/9	46/47	127/253	167/187	87/94
VFDT_ADP	1/1	5/9	46/47	127/253	167/187	87/94
iOVFDT_MC	22/43	6/11	325/329	1280/2559	2211/2500	185/249
iOVFDT_NB	22/43	8/15	325/329	1864/3727	2440/2821	188/255
iOVFDT_WNB	22/43	8/15	325/329	1864/3727	2233/2551	188/255
iOVFDT_ADP	22/43	8/15	325/329	1864/3727	2440/2821	188/255

The speed of learning decision model was reflected by the time in seconds as shown in Table 4. In general, C4.5 has the slowest learning speed for all datasets. Comparing the average learning times of VFDT to iOVFDT, our experiment result shows both algorithms have a very similar learning speed. iOVFDT has a learning speed almost as fast as the original VFDT. This implies that the improved version, iOVFDT can achieve smaller tree size, good accuracy without incurring cost of slowing down the learning speed. Fast learning speed is important and applicable to time-critical applications.

TABLE IV.

LEARNING SPEED ANAYLSIS OF DYNAMIC DATA

Methods\Data	RS	IP	IU	CT	PA	NA
C4.5 Pruned	0.30	0.22	0.40	931.34	180.88	120.68
C4.5 Unpruned	0.80	0.40	0.39	1717.35	121.62	187.44
IncreNB	0.26	0.10	0.24	11.98	0.95	17.77
VFDT	0.18	0.07	0.19	6.65	0.63	4.50
VFDT_NB	0.13	0.09	0.24	9.88	0.96	6.64
VFDT_ADP	0.14	0.09	0.30	10.18	1.28	7.95
iOVFDT_MC	0.12	0.08	0.20	6.86	0.64	4.36
iOVFDT_NB	0.17	0.09	0.30	8.80	0.97	6.62
iOVFDT_WNB	0.16	0.10	0.29	8.61	0.98	6.41
iOVFDT_ADP	0.13	0.11	0.31	13.09	1.26	6.78
Avg. C4.5	0.55	0.31	0.40	1324.35	151.25	154.06
Avg. Increm.NB	0.26	0.10	0.24	11.98	0.95	17.77
Avg. VFDT	0.15	0.08	0.24	9.57	0.96	6.36
Avg. iOVFDT	0.14	0.10	0.27	8.34	0.96	6.04

D. Real-time Recommendation Data

Recommendation system is an important application of data mining that tries to refer the right products to the right customers in the right time. We use some real-life online recommendation data from the GroupLens Research:

MovieLens www.grouplens.org/node/73

Book-cross www.informatik.uni-freiburg.de/~cziegler/BX/

They are the typical dataset for the recommending system. This data is consisted of three files: movie/book information, user information, and rating. The three files are joined together by the user ID and movie/book ID.

After combining the data, MoiveLens includes 1,000,209 instances, 1 numeric attributes, 24 nominal attribute. The target class is the type of movie. There are 18 distinct types. Book-crossing includes 1,316,100 instances, 2 numeric and 5 nominal attributes. The target class is the country where the users are. There are 61 investigated countries. For a recommendation system, the classification model is used to predict what type of the movie does the user like, or which region does the user live in, from the previous rating data. The benchmark algorithms are VFDT, ADWIN and iOVFDT, with Error-adaptive functional tree leaf.

For MovieLens data, after normalized the result, we can see the comparison of these three algorithms in Figure 9. In general, iOVFDT and ADWIN have better accuracy than VFDT, but ADWIN results bigger model size than iOVFDT, as well as the learning time. For Book-crossing data, the accuracy and tree size analysis are shown in Figure 10 and 11 respectively. It reflects that ADWIN still obtains a bigger tree size. iOVFDT outperforms the others in terms of the accuracy and the tree size.

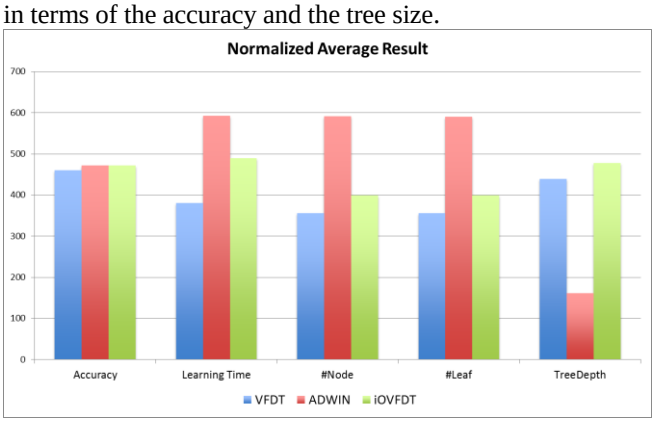


Figure 9. Normalized comparison result of MovieLens data

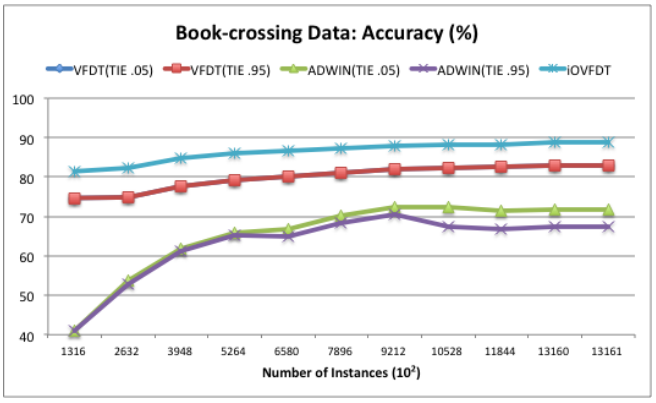


Figure 10. Accuracy of Book-crossing data

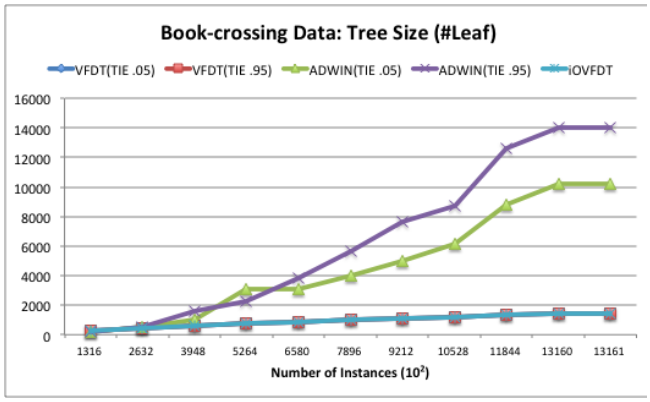


Figure 11. Tree size of Book-crossing data

VII. Coclusions

How to uncover the knowledge hidden within massive and big data efficiently, remains an open question. In the opinion of author, a well-designed algorithm is crucial in solving the problems associated with big data.

A data stream model is usually defined as a model in which data move continuously at high-speed. Most big data can be considered as data streams, in which many new data are generated in a short time, and moving continuously. Data streams contain very large volumes of data, which cannot be stored in either internal or external memory. A one-pass algorithm, therefore, forms the basis of data stream mining, which briefly stores a sufficient statistical matrix when new data passes, but does not require the full dataset being scanned repeatedly. However, imperfect data streams, like missing values, noise, imbalanced distribution and concept-drift, are common in the real world applications. To the best knowledge of the author, no suitable methods have solved all above problems well so far.

The main contributions of this research propose:

- An incremental decision tree algorithm handling imperfect data streams.
- A mechanism so called Auxiliary Reconciliation Control (ARC) is used to handle the missing data.
- An adaptive-tie breaking threshold is robust to the noisy data.
- A new functional tree leaf of weighted Na ïve Bayes is brought forward to deal with imbalanced distributions in data streams.
- A test-then-train learning approach monitors the performance of decision model in real-time so that the model is sensitive to concept-drift occurrence.

Experiment shows the proposed methodology can solve the aforementioned problems as a result.

References

[1] Domingos P., and Hulten G.. 2000. ‘Mining high-speed data streams’, in Proc. of 6th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD’00), ACM, New York, NY, USA, pp. 71-80.

[2] Hulten G., Spencer L., and Domingos P., 2001. ‘Mining time-changing data streams’, in Proc. of 7th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD’01), ACM, New York, NY, USA, pp. 97-106.

[3] Gama.J. Rocha R. and Medas P., 2003.‘Accurate decision trees for mining high-speed data streams’, in Proc. of 9th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD’03), ACM, New York, NY, USA, pp. 523-528.

[4] Bifet A. and Gavalda R. 2007. “Learning from time-changing data with adaptive windowing”. In Proc. of SIAM International Conference on Data Mining, pp. 443–448.

[5] Quinlan R, 1986. Induction of Decision Trees, Machine Learning, 1(1), pp.81-106.

[6] Quinlan R, 1993. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco.

[7] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J., 1984. 'Classification and Regression Trees', in Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

[8] Yang H., and Fong S., 2011. ‘Moderated VFDT in Stream Mining Using Adaptive Tie Threshold and Incremental Pruning’, in Proc. of 13th international conference on Data Warehousing and Knowledge Discovery (DaWak2011), LNCS, Springer Berlin / Heidelberg, pp. 471-483.

[9] Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, *41*(12), 3692-3705.

[10] Ding, Y., & Simonoff, J. S. (2010). An investigation of missing data methods for classification trees applied to binary response data. *The Journal of Machine Learning Research*, *11*, 131-170.

[11] Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*(Vol. 539). New York: Wiley.

[12] Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, *11*(3), 259-275.

[13] Street, W. N., & Kim, Y. (2001, August). A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*(pp. 377-382). ACM.

[14] Li, C., Zhang, Y., & Li, X. (2009, June). OcVFDT: one-class very fast decision tree for one-class classification of data streams. In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*(pp. 79-86). ACM.

[15] Hashemi, S., & Yang, Y. (2009). Flexible decision tree for data stream classification in the presence of concept change, noise and missing values.*Data Mining and Knowledge Discovery*, *19*(1), 95-131.

[16] Pfahringer, B., Holmes, G., & Kirkby, R. (2007). New options for hoeffding trees. In *AI 2007: Advances in Artificial Intelligence* (pp. 90-99). Springer Berlin Heidelberg.

[17] Hoeglinger, S., Pears, R., & Koh, Y. S. (2009). CBDT: A Concept Based Approach to Data Stream Mining. In *Advances in Knowledge Discovery and Data Mining* (pp. 1006-1012). Springer Berlin Heidelberg.

Call for Papers and Special Issues

Aims and Scope

Journal of Emerging Technologies in Web Intelligence (JETWI, ISSN 1798-0461) is a peer reviewed and indexed international journal, aims at gathering the latest advances of various topics in web intelligence and reporting how organizations can gain competitive advantages by applying the different emergent techniques in the real-world scenarios. Papers and studies which couple the intelligence techniques and theories with specific web technology problems are mainly targeted. Survey and tutorial articles that emphasize the research and application of web intelligence in a particular domain are also welcomed. These areas include, but are not limited to, the following:

- Web 3.0
- Enterprise Mashup
- Ambient Intelligence (AmI)
- Situational Applications
- Emerging Web-based Systems
- Ambient Awareness
- Ambient and Ubiquitous Learning
- Ambient Assisted Living
- Telepresence
- Lifelong Integrated Learning
- Smart Environments
- Web 2.0 and Social intelligence
- Context Aware Ubiquitous Computing
- Intelligent Brokers and Mediators
- Web Mining and Farming
- Wisdom Web
- Web Security
- Web Information Filtering and Access Control Models
- Web Services and Semantic Web
- Human-Web Interaction
- Web Technologies and Protocols
- Web Agents and Agent-based Systems
- Agent Self-organization, Learning, and Adaptation
- Agent-based Knowledge Discovery
- Agent-mediated Markets
- Knowledge Grid and Grid intelligence
- Knowledge Management, Networks, and Communities
- Agent Infrastructure and Architecture
- Agent-mediated Markets
- Cooperative Problem Solving
- Distributed Intelligence and Emergent Behavior
- Information Ecology
- Mediators and Middlewares
- Granular Computing for the Web
- Ontology Engineering
- Personalization Techniques
- Semantic Web
- Web based Support Systems
- Web based Information Retrieval Support Systems
- Web Services, Services Discovery & Composition
- Ubiquitous Imaging and Multimedia
- Wearable, Wireless and Mobile e-interfacing
- E-Applications
- Cloud Computing
- Web-Oriented Architectrues

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/jetwi/.

(Contents Continued from Back Cover)

Applying Clustering Approach in Blog Recommendation 296
Zeinab Borhani-Fard, Behrouz Minaei, and Hamid Alinejad-Rokny

Automatic Extraction of Place Entities and Sentences Containing the Date and Number of Victims of Tropical Disease Incidence from the Web 302
Taufik Fuadi Abidin, Ridha Ferdhiana, and Hajjul Kamil

Widespread Mobile Devices in Applications for Real-time Drafting Detection in Triathlons 310
Iztok Fister, Dušan Fister, Simon Fong, and Iztok Fister Jr.

RISING SCHOLAR PAPERS

Solving Problems of Imperfect Data Streams by Incremental Decision Trees 322
Hang Yang