# Journal of Emerging Technologies in Web Intelligence

ISSN 1798-0461

Volume 4, Number 1, February 2012

# Contents

## Special Issue: Intelligent Web Interaction

## Guest Editors: Seiji Yamada, Takashi Onoda, and Yasufumi Takama

RISING SCHOLAR PAPERS

Special Issue on Intelligent Web Interaction

# Guest Editorial

Seiji Yamada
National Institute of Informatics/SOKENDAI, Chiyoda, Tokyo, Japan
Email: seiji@nii.ac.jp

Takashi Onoda
CRIEPI, Komae, Tokyo, Japan
Email: onoda@criepi.denken.or.jp

Yasufumi Takama
Tokyo Metropolitan University, Hino, Tokyo, Japan
Email: ytakama@sd.tmu.ac.jp

Various Web systems/services are currently providing a great deal of benefits for users, and Web interaction to design interaction between users and Web systems is becoming especially important both for research and business. Web interaction has been realized through related technologies including interactive data mining/information retrieval, intelligent systems, personalization, user interfaces and so on. However, each study and development has been done independently in different research fields, which might discourage us from studying Web interaction from unified view of human-system interaction and making Web interaction more intelligent by applying machine learning and soft computing.

We organized the 2011 International Workshop on Intelligent Web Interaction (IWI-2011) at Lyon to bring together a variety of researchers in diverse fields like Web systems, Artificial Intelligence, computational intelligence, human-computer interaction and user interfaces. The workshop was collocated with 2011 IEEE/WIC/ACM International Conference on Web Intelligence (WI-2011). The IWI workshop has been held from 2006 yearly, and has grown to be one of the largest workshops affiliated with the WI conference.

This special issue is consisting of six selected papers from IWI-2011. The purpose of this special issue is to present Intelligent Web Interaction as a new and promising research field. Presenters of the IWI-2011 were encouraged to submit papers to this special issue. All submitted papers are equivalently reviewed in terms of relevance, originality, significance and presentation based on standard review criteria of Journal of Emerging Technologies in Web Intelligence.

The first paper "Co-Authorship Networks Visualization System for Supporting Survey of Researchers' Future Activities" (Takeshi Kurosawa and Yasufumi Takama) describes a visualization system to support users to predict future research activities from current co-authorship networks. This is a quite challenging study because it is strongly concerned to prediction of future dynamic development of human-relational networks. Since collaboration of researchers is essential for researchers' activities, co-authorship network is suitable for predicting future activities. This paper focuses on the task of discriminating growing researchers from supervisors. The effectiveness of the proposed system is evaluated through the detailed analysis of two participants' analyzing process of InfoVis 2004 Contest dataset.

The second paper "Computational Approach to Prediction of Attitude Change Through eWOM Messages Involving Subjective Rank Expressions" (Kazunori Fujimoto) proposes a computational model to predict potency-magnitude relations of electric word-of-mouth messages involving subjective rank expressions. This paper defines three message classes, which are also studied in the areas of opinion mining and sentiment analysis, and investigates mathematically how the potency-magnitude relations change based on the values of the evaluation parameters.

The third paper "Balancing the Trade-Offs Between Diversity and Precision for Web Image Search Using Concept-Based Query Expansion" (Enamul Hoque, Orland Hoeber and Minglun Gong) experimentally investigates trade-off between the promotion of diversity and the precision of the most common sense in diversifying image search results. The image search is done by concept-based query expansion with Wikipedia. As a result of these experiments, an automatic method for tuning the diversification parameter is proposed based on the degree of ambiguity of the original query.

The fourth paper "Which is the best?: Re-ranking Answers Merged from Multiple Web Sources" (Hyo-Jung Oh, Pum-Mo Ryu and Hyunki Kim) proposes a novel method to determine the best answers collected from multiple Web sources. Local optimal answers are selected by several specialized sub-QAs in a distributed QA framework. In order to find global optimal answers, merged candidates are re-ranked by adjusting confidence weights based on the question analysis. The proposed system applies a SVM classification algorithm to adjust confidence weights calculated by own ranking methods in


doi:10.4304/jetwi.4.1.1-2

sub-QAs. The effects of the proposed re-ranking algorithm are evaluated through a series of experiments.

The fifth paper "Graph-cut based Constrained Clustering by Grouping Relational Labels" (Masayuki Okabe and Seiji Yamada) proposes a novel constrained clustering method based on a graph-cut by semi-definite programming. The proposed algorithm begins with a single cluster of a whole dataset and repeatedly divide the larger cluster into two sub-clusters. The division is done by swapping rows and columns of a matrix obtained from a graph-cut problem. Experimental results using datasets from the Open Directory Projects and WebKB corpus support their method is promising for interactive Web clustering.

The last paper "Careful Seeding Method based on Independent Components Analysis for k-means Clustering" (Takashi Onoda, Miho Sakai and Seiji Yamada) applies ICA (Independent Components Analysis) to effective initial seeding for K-means clustering. Although the k-means clustering is a widely used clustering technique for the Web because of its simplicity and efficiency, the clustering results significantly depend on the initial clustering centers. This paper provides a novel seeding method to determine effective clustering centers by selecting the nearest data to independent components obtained by ICA. They evaluate performance of the proposed method by comparing with other seeding methods using various benchmark datasets.

As mentioned at the beginning, Intelligent Web Interaction is a new and promising research field. We strongly hope this special issue will motivate many other researchers to join this growing research field.

**Seiji Yamada** is a professor at the National Institute of Informatics and SOKENDAI. Previously he worked at Tokyo Institute of Technology. He received B.S. (1984), M.S. (1986) and the Ph.D. (1989) degrees in artificial intelligence from Osaka University. His research interests are in the design of intelligent interaction including Human-Agent Interaction, Intelligent Interactive Systems and interactive data mining. He is a member of IEEE, AAAI, ACM, JSAI, IPSJ and HIS.

**Takashi Onoda** graduated from International Christian University, Tokyo, Japan in 1986. He received the M.S. degree in nuclear engineering from Tokyo Institute of Technology, Tokyo, Japan in 1988. He works at Central Research Institute of Electric Power Industry from 1988. He received the Dr. Eng. degree in mathematical engineering from University of Tokyo, Tokyo, Japan in 2000. He worked as a visiting researcher in GMD FIRST in Berlin from September in 1997 to September in 1998. He is a sector leader at Central Research institute of Electric Power Industry and a visiting professor at Tokyo Institute of Technology. His research interests are in statistical learning theory and its applications. He is a member of JSAI.

**Yasufumi Takama,** Japan, 1971. Dr. Eng, University of Tokyo, Tokyo Japan, 1999. He was a JSPS Research Fellow from 1997 to 1999. From 1999 to 2002 he was a Research Associate at Tokyo Institute of Technology in Japan. From 2002 to 2005, he was an Associate Professor at Department of Electronic Systems and Engineering, Tokyo Metropolitan Institute of Technology, Tokyo, Japan. Since 2005, he has been an Associate Professor at Faculty of System Design, Tokyo Metropolitan University, Tokyo, Japan. He also participated in PREST, JST from 2000 to 2003. His current research interest includes Web intelligence, information visualization, and intelligent interaction. He is a member of IEEE, JSAI, IPSJ, IEICE, and SOFT.

# Co-Authorship Networks Visualization System for Supporting Survey of Researchers' Future Activities

Takeshi Kurosawa, Yasufumi Takama
Graduate School of System Design, Tokyo Metropolitan University, Hino, Tokyo, Japan
Email: kurosawa@krectmt3.sd.tmu.ac.jp, ytakama@sd.tmu.ac.jp

*Abstract*— **This paper proposes a visualization system that supports users getting insight into future research activities from co-authorship networks. A bibliographic network such as a co-authorship network and a citation network is important information for researchers when doing a research survey. In particular, there are many requests on research survey that relate with researchers' future activities, such as identification of remarkable researchers including growing researchers and supervisors. Although a citation network has received many attentions from researchers, it is not suitable for such surveys because it reflects researchers' past activities. Since collaboration of researchers is essential for researchers' activities, co-authorship network is supposed to be suitable for predicting future activities. In order to get insights into future research activities by discriminating growing research areas from grown-up areas, the proposed visualization system provides the functions for identifying research areas as well as for identifying time variation of both network structure and keyword distribution. As a basis for getting insights into future research activities, this paper focuses on the task of discriminating growing researchers from supervisors. The effectiveness of the proposed system is evaluated through the detailed analysis of two participants' analyzing process of InfoVis 2004 Contest dataset. It is observed that different analyzing strategies are employed by even the same participant, when available support functions are different. The result indicates participants can successfully utilize the functions in their exploratory analysis process.**

*Index Terms*— **exploratory data analysis; interactive information visualization; temporal trend information; co-authorship networks; graph visualization;**

## I. Introduction

This paper proposes a visualization system for co-authorship networks that has functionalities for supporting the prediction of future research activities. A bibliographic network is important information for researchers when doing a research survey. A bibliographic network is composed of several networks: a co-authorship network, a citation network, and a co-citation network. Today, there are a number of research bibliography sites on the web such as IEEE Xplore[1] and DBLP[2]. These sites provide detailed information about specific authors, papers and journals, which are useful for a research survey. Research surveys are sometimes conducted by researchers who are getting into their unfamiliar research field. In such a case, the purpose of the survey is to grasp an overview of the research field. However, as the structure of bibliographic networks is usually huge and complicated, it is difficult to grasp such an overview using simple interface of existing research bibliography sites. Therefore, information visualization techniques for bibliographic networks have been studied [1]–[6].

In the field of bibliographic network analysis, a citation network has received many attentions from researchers [1]. There also exist many visualization systems aiming at supporting user's research survey based on it [2]. A citation network represents researchers' past activities, from which we can identify important papers by finding the most frequently cited papers.

On the other hand, there are many requests on research survey that relate with researchers' future activities, such as the identification of researchers who will potentially write an interesting paper. As the direction of citation is from new paper to old one, it is difficult to predict such future activities from a citation network.

It is noted that collaboration of researchers is essential for these research surveys. That is, a research paper is the outcome of collaborative activities, and research areas are often emerged from collaboration among researchers working on different research topics. In that sense, a co-authorship network is suitable for predicting researchers' future activities, because it reflects the past and current status of collaboration.

This paper proposes a visualization system for co-authorship networks that supports users getting insight into future research activities. As for remarkable researchers to be identified for a research survey, this paper focuses on two types of researchers: growing researchers and supervisors. It is noted that this paper defines a supervisor as a researcher who has already achieved in his/her research fields. To identify growing researchers and supervisors, the proposed system provides two functions: the function for identifying research areas and that for identifying time variation of both network structure and keyword distribution. To identify research areas, the proposed system renders a researcher as a pie chart, which represents a ratio of keywords assigned to corresponding researcher's papers. By representing a node in the co-authorship network as a pie-chart, keyword distribution

[1] http://ieeexplore.ieee.org/
[2] http://www.informatik.uni-trier.de/~ley/db/


doi:10.4304/jetwi.4.1.3-14

over the network is easily grasped by analysts.

Identifying time variation is composed of three sub functions. First, to determinate whether a researcher published papers in a period or not, the system uses animation. Second, the brightness and saturation of a segment in a pie chart represent the period when corresponding keyword is used. Finally, to determinate whether the researcher's activity is continuously or sporadically, a pie chart of a researcher can be bi-cylindrical style, in which inner and outer ring correspond to the earliest and the most recent year of his/her publications, respectively.

In order to evaluate the effectiveness of the system, the proposed system is applied to exploratory analysis of the co-authorship network extracted from the InfoVis 2004 Contest dataset. Two test participants used the system for identifying growing researchers and supervisors. Results are analyzed from the viewpoint of analyzing strategies they employed in performing tasks. Two types of systems, each of which provides different sets of support functions, are used in a user study. By comparing the results, how available functions affect participants' analyzing strategies is investigated.

The results show the participants who don't have background knowledge about InfoVis can identify growing researchers and supervisors using the system. It is observed that the proposed functions such as highlighting researchers who published in a certain period and visualizing a node as a pie chart of keywords are heavily used.

The content of the paper is extended from our previous works [7], [8]. Main difference from [7] is a result of comparing user behaviors with using two system having different functions. Expansion from [8] has been made according to the discussion at the conference, including more detailed description about the system and discussion about the result of user study.

The remainder of this paper is organized as follows: Section II discusses related works. In Section III we introduce the proposed visualization system. The effectiveness of the proposed system is shown with a user study, which is described in section IV and V.

## II. Related works

### A. Visual Analysis of Bibliographic Networks

Bibliographic information is important for researchers when doing a research survey. A bibliographic network is composed of three types of networks: a co-authorship network, a citation network, and a co-citation network. A co-authorship network represents the relationship between a researcher and his/her collaborators. A citation network represents reference relationship between papers. It is a directed network, in which a paper has a directed link to other paper by citing it. A co-citation network links two nodes (papers) when those appear together (co-cited) in at least one other paper.

A citation network/relationship is suitable for identifying the most cited papers and researchers, which are considered important papers and researchers. PaperLens provides "year by year top 10 cited papers/authors" view to identify the most cited papers and authors in each year. It can filter papers/authors by specifying research area [2].

On the other hand, a co-authorship network tends to be used to analyze the relationship between researchers [9]. Henry et al. identified collaboration patterns of researchers by visualizing a co-authorship network [3].

As such a network structure is complicated, information visualization techniques are usually employed for the analysis. In a node-link diagram which is the most often used representation of a network structure, a researcher or a paper is represented by a node and the relationship is represented by an edge. Ghoniem et al. [10] have shown the readability of node-link diagrams decreases for dense/large networks. To overcome that problem, various techniques have been proposed.

Node clustering/aggregations are commonly used to improve readability of a node-link diagram. Auber et al. [11] have proposed multiscale visualization. In this visualization, a network is divided into clusters, each of which corresponds to a small network and treated as a macro node, and the edges between clusters are clustered.

Ham et al. [12] have proposed an interactive visualization for node clusters to inspect inside of clusters. Holten [13] has proposed hierarchical edge bundles to improve the readability of global relationship trend. TreePlus [14] and Vizster [15] enabled exploratory analysis of local structure of a large network.

### B. InfoVis 2004 Contest

The IEEE InfoVis 2004 Contest [16] provides a dataset which contains bibliographic information of InfoVis papers from 1995 to 2002 and those references. The aim of the contest was "to promote the development of benchmarks for information visualization and establish a forum to advance evaluation methods" [16]. In the contest, three first place winners [1], [4], [5], one student first place winner [6], and eight second winners were selected.

The tasks of the contest are defined as follows.

1) Create a static overview of the 10 years of the InfoVis
2) Characterize the research areas and their evolution
3) Where does a particular author/researcher fit within the research areas defined in task 2?
4) What, if any, are the relationships between two or more or all researchers?

In summary, the first place winner entries tended to use citation networks to identify research areas, their relationships, and/or evolution (task 2). Co-authorship networks tend to be used to identify collaboration relationships of researchers (task 3 and 4).

In task 2, Ke et al. [1] used burst analysis of keywords to identify the research areas and their evolution. They showed the results as tables. Lee et al. [5] clustered papers into research areas using their titles, references, and keywords. Based on the clusters shown as table, they showed the evolution of research areas. Wong et al. [4] identified discriminating research areas by co-occurrence of words appeared in titles and abstracts.

Papers were placed based on thematic similarity. By filtering the papers by years, they showed the evolution of research areas. Ahmed et al. [6] showed that the evolution of research areas and their citation relationship can be identified by using a 3D "worm" representation in task 2. The research areas are identified by clustering papers by the word histogram of titles, abstracts, and keywords using SOM (Self Organizing Map).

In task 3, Ke et al. [1] analyzed the keyword usage of the researchers in non-visual way and showed researcher's interesting areas. In task 4, they used co-authorship networks to identify the collaboration relationship. In task 3 and 4, Lee et al. [5] analyzed a citation network in terms of research area. For each researcher, research areas of his/her papers as well as those citing his/her papers are identified. They also showed collaborators of a researcher using co-authorship relationship. Wong et al. [4] identified research areas of a researcher based on his/her publications. They also defined the thematic similarity between researchers. The results showed that the influence of a researcher can be identified from citation relationships by highlighting papers which cite his/her papers. Ahmed et al. [6] showed the hierarchical relationships of researchers with using a 3D network visualization of co-authorship networks. They classified the researchers in that visualization according to the number of collaboration relationships (degree): 20+ degree, 10-19 degree, and less than 10 degree. They assumed that researchers with high degree are senior researchers and those with low degree are students and younger researchers.

In summary, above-mentioned entries have identified research areas with the process of evolution. The collaboration of researchers has been also identified. However, each of those two findings has been obtained separately despite both are outcomes of the researchers' collaboration.

## III. Visualization System for Co-Authorship Networks

### A. Overview of the system

Fig. 1 shows the screenshot of the proposed system. The system consists of three parts; Network panel (Fig. 1(a)), Node Detail panel (Fig. 1(b)), and Operation panel (Fig. 1(c)).

Network panel shows researchers and their collaborations as a node-link diagram, in which the size of a node indicates the number of his/her collaborators. The thickness of an edge indicates the number of collaborations (i.e., collaborative papers).

Node Detail panel provides detailed information of the node selected in the Network panel. This panel shows an enlarged pie chart of the selected researcher and lists his/her publication list, collaborators, keywords, and publication years. For each item in the lists, more detailed information is provided by a tooltip.

Operation panel allows a user to change the node visualization according to the purpose of analysis. The operations include enabling/disabling the highlighting of nodes and so on.

### B. Research area identification

To identify research areas, the system provides four functions as follows.

1) Listing all keywords which the selected researcher uses.
2) Rendering a researcher as a pie chart (i.e., a node) showing the ratio of his/her keywords usage.
3) Highlighting researchers who use at least one of keywords used by the selected researcher.
4) Aggregating the nodes of researchers if they use exactly the same keywords, for making it clear their collaboration is tighter than usual.

As for (1), a user can check keywords a researcher uses with the Node Detail panel. The Node Detail panel lists all keywords used by the selected researcher.

Regarding (2), to check distribution of the keywords, a keyword can be assigned a color (hue), which becomes a *selected* keyword, and the system renders its usage with a pie chart in the Network panel. A color can be assigned either manually or automatically. A user can manually assign a color to a keyword using the Node Detail panel.

A user can assign a color to a keyword automatically based on one of following indices as well. The colors are assigned from red to blue in descending order of the index value.

- TF (Term Frequency)
- TF/IDF (Inverse Document Frequency)
- Alphabetic

The TF is suitable for identifying frequently used keywords. Two types of TF indices are available. The TF weight $w_k$ of a keyword $k$ is the frequency of $k$ among all papers in the co-authorship network. The $w_{vk}$ of $k$ for a researcher $v$ is the frequency of $k$ among all papers written by $v$.

The TF/IDF is suitable for identifying keywords used by specific researchers particularly. There are two types of indices as well. The TF/IDF weight of a keyword $w'_k$ and $w'_{vk}$ are given by following formulas, where $N$ is the number of all papers over the network, $N_k$ is the number of papers attached the keyword $k$, $N_v$ is the number of papers written by $v$, and $N_{vk}$ is the number of $v$'s papers attached the keyword $k$.

$$w'_k = w_k \times IDF_k, \quad (1)$$

$$IDF_k = \log \frac{N+1}{N_k+1}, \quad (2)$$

$$w'_{vk} = w_{vk} \times IDF_{vk}, \quad (3)$$

$$IDF_{vk} = \log \frac{N_v+1}{N_{vk}+1}. \quad (4)$$

Alphabetic index sorts the keywords alphabetically and assigns color according to the order.

Segments in a pie chart are also ordered according to one of above-mentioned indices for each researcher $v$ (i.e.,

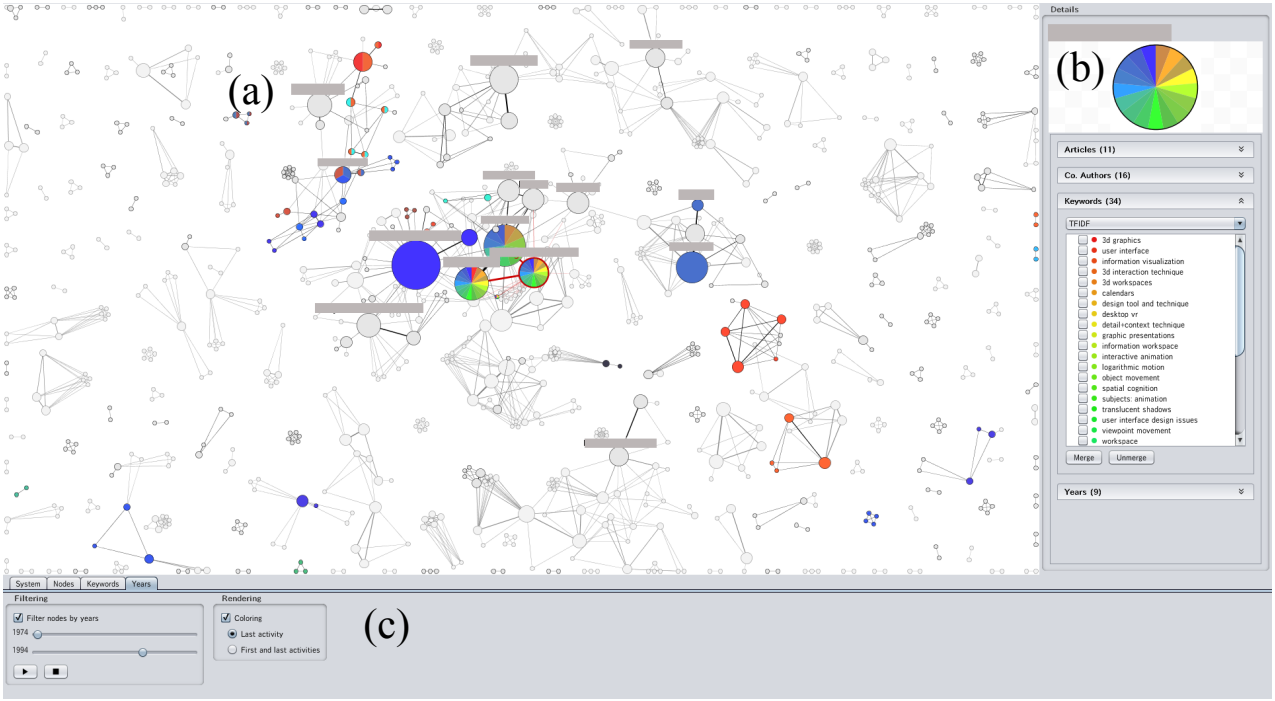



Figure 1. The overview of proposed system; (a) Network panel, (b) Node Detail panel, (c) Operation panel.

$w_{vk}$, $w'_{vk}$ and alphabetic index). This means the positions of a segment in a pie chart can be different between researchers if their keywords usages are different.

If there are keywords that correspond to the same topic, a user can manually group those into one keyword group. A keyword group can be assigned a color in the same way as a single keyword.

The system has two variations of the pie chart representation according to the types of analysis. The first one shows only selected keywords in a pie chart (type *K1*). This is used for checking the distribution of specific keywords over the network. The second one shows all keywords in a pie chart, in which non-selected keywords are rendered in achromatic color (type *K2*). This is used for checking the concordance rate of keywords between researchers.

As for (3), for identifying relations between researchers in terms of keywords, the system can highlight the researchers who use at least one of the keywords used by the selected researcher.

Regarding (4), if the researchers having the collaborative papers use exactly the same keywords, their collaboration is tighter than usual. To show it visually, the system aggregates nodes in the Network panel if the corresponding researchers use exactly the same keywords. This also improves the readability of network view. The left figure in Fig. 2 shows normal node placement. In the right figure in Fig. 2, researchers using exactly the same keywords are aggregated.

### C. Time variation

The system handles two types of time variation; time variation of researcher's collaboration (i.e., publication activities) and that of keywords usage.

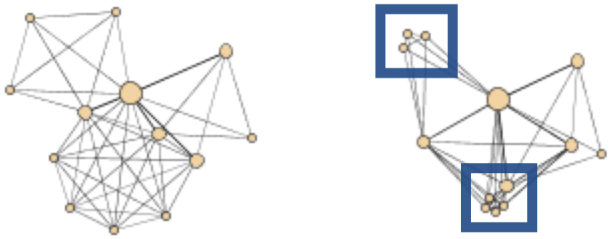


Figure 2. Nodes aggregating based on the keywords usage (left: without aggregation, right: with aggregation). In right figure aggregated nodes are enclosed by rectangles.

This paper divides the task of Identifying time variation of researchers' collaboration into three sub tasks:

- Identifying whether a researcher published papers in a certain period or not.
- Identifying whether a researcher published papers recently or not.
- Identifying whether a researcher published papers continuously or sporadically.

For the first task, a user can highlight the researchers who published papers in a certain period from $y_s$ (year) to $y_e$ (year) in the Network panel. The highlighted area can be varied from $[y_s, y_s]$ to $[y_s, y_e]$ with animation, so that the change of researchers' collaboration can be shown. The system also shows all the publication years of the selected researcher in the Node Detail panel.

For the second task, the brightness and saturation of a node represent the last period when a researcher published papers (type *T1*). Brighter node indicates the corresponding researcher published a paper more recently.

Finally, to identify whether a researcher published papers continuously or sporadically, corresponding node can be represented with bi-cylindrical style (type *T2*). The inner ring represents the earliest year of his/her publication and outer ring represents the last year. In the left figure in Fig. 3, both of inner and outer rings of the node are dark, which indicates the corresponding researcher published papers early in the specified period only. In the center figure in Fig. 3, the inner ring of the node is dark, but the outer ring is bright. This indicates the corresponding researcher published papers continuously. In the right figure in Fig. 3, the both of inner and outer rings of the node are bright, which indicates corresponding researcher published papers late in the period only.

Figure 3. Visualization of time variation of researchers' collaboration (left: a researcher who published papers early in the period only, center: a researcher who published papers continuously, right: a researcher who published papers late in the period only)

Identification of time variation of keyword usage is done by combining {*K1*, *K2*} and {*T1*, *T2*} visualization as shown in Fig. 4. In the same way as type *T1*, the brightness and saturation of a segment in a pie chart represent the period when a certain keyword was used. When the bi-cylindrical style (*T2*) is used, the inner and the outer rings represent the earliest and the most recent year when the corresponding researcher used it, respectively. In Fig. 5(a), the inner ring of the keyword at top right segment is dark and outer ring is bright. This indicates that keyword has been used continuously. On the other hand, the rings of the keyword at middle right (Fig. 5(b)) are colored dark brown. This indicates that keyword was used early in the specified period only. The color of the keyword at the bottom right (Fig. 5(c)) is bright yellow, which indicates that keyword was used in late in the period only. In a pie chart representation, type

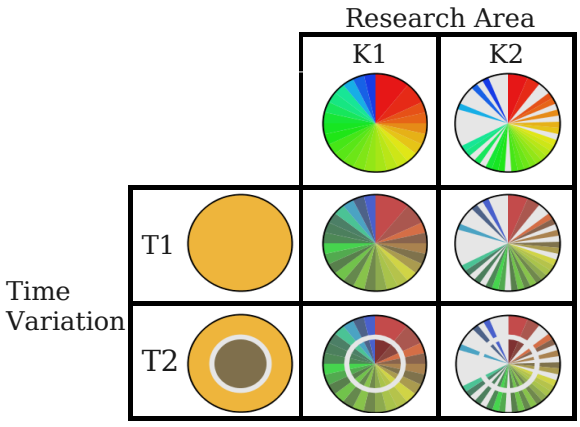



Figure 4. Variations of nodes visualization

*K1* and *K2* as above-mentioned are available also in this case.

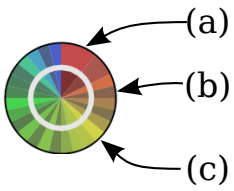



Figure 5. Visualization of time variation of keywords usage; (a) the keyword used continuously, (b) used in early period only, (c) used in later period only

## IV. User Study

It has been shown in [17] that using the proposed functions in combination, authors could find the same insights found by first place winners in InfoVis 2004 Contest. The performed task corresponds to the analysis of past research activities from a co-authorship network.

On the contrary, this paper examines the effectiveness of the proposed system for supporting the analysis of future research activities, and its usability for users unfamiliar with target domain.

Although the comparing with other systems is useful for examining the effectiveness of the proposed system, detailed analysis, such as the contribution of each function, is difficult through the comparison between systems having considerably different interfaces and functions. Therefore, two types of systems are prepared: the system with full functions and that with limited functions. A "limited functions" indicates that test participants cannot use functions for assignment colors to keywords and its visualization, grouping keywords manually, and bi-cylindrical visualization of time variation of a researcher or a keyword. Other than those limitations are as the same as the system with full functions.

Participants are asked to identify researchers of following two types.

Growing researcher
: A researcher who is doing interesting work and would publish research papers actively in future from a period of a dataset.

Supervisor
: A researcher who did or is doing interesting research work but would not publish papers so actively in future from a period of a dataset, because s/he has already achieved in the period of a dataset.

The reason of adopting these kinds of abstract tasks is to let participants perform exploratory data analysis based on various assumptions. That is, when giving such an abstract purpose, participants are supposed to translate it into more concrete assumptions about the conditions the researchers to be identified should satisfy. The participants should perform analysis guided by such assumptions with the combination of functions provided by the system. It is also expected that a participant would

make various assumptions until s/he obtains sufficient results. The purpose of the user study is to examine the relationship between assumptions and used functions.

Two participants took part in the study. Both of them are male graduated students of system design major and aged early 20s. It should be noted we selected test participants who don't have background knowledge about Information Visualization. We think support of research survey by users unfamiliar with target domain is important, because research surveys are inevitable for researchers / companies getting into new domains. In particular, the importance of such surveys is growing for companies trying to adapt to rapidly changing business environment.

A participant first did the task by using the system with limited functions. Before doing the task, participants were lectured about the usage of *T1* for node visualization and keyword list in Node Detail panel without explanation about how to assign color for keyword and group keywords manually. After finishing the task, they are lectured about remaining functions: *T2*, *K1* and *K2* and the manual color assignment to keywords / keyword groups. After the lecture, they did the same task again, with using full functions.

After each task, we had interviews with participants about assumptions and reasons of their analysis procedures.

We use the co-authorship networks extracted from InfoVis 2004 dataset in the experiments. The InfoVis 2004 Contest [16] dataset contains bibliographic information of InfoVis papers from 1995 to 2002 and those references. As some entries of the contest have pointed out that there are duplications or errors in the dataset [1], [6], we cleaned up the dataset based on [1]. The extracted network contains 969 researchers (nodes), 1736 collaborations (edges) and 1777 keywords. We treated the publishing date by year. Published period is from 1974 to 2004, in which there are no missing years.

## V. Results

### A. System with Limited Functions

The analyses of two participants were based on almost the same assumptions:

- A growing researcher has a certain number of papers and also published papers in recent period (2000-2004).
- A supervisor has a certain number of papers but doesn't published papers in recent period (2000-2004).

They used combination of following steps to ensure the assumptions.

- Identify the number of papers which a researcher published by his/her node size
- Highlight researchers who published papers in a certain period

In particular, if corresponding node is relatively large and highlighted in recently period, a researcher is considered as growing researcher. If a node is relatively large but not highlighted, corresponding researcher is considered as supervisor.

They also employed different types of clues to ensure above-mentioned assumptions. Participant A focused on groups of researchers. He looked the co-authorship network as "Map of Researcher Groups," which consists of nodes with various sizes (Fig. 6). By changing time period for highlighting nodes, he identified "Rise and Fall of Research Groups." He considered relatively large nodes highlighted in recent period (2002-2004) as growing researchers and those unhighlighted relatively large nodes as supervisors.

Participant B used *T1* to visualize time variation of node in addition to the function of highlighting researchers. First he highlighted recent 5 years (2000-2004). In this situation, researchers who published papers during 2003 and 2004 become blight nodes (enclosed by red solid rectangle) and those who published during 2000 and 2001 become dark nodes (enclosed by blue dashed rectangles) in *T1* visualization (Fig. 7). From this result he considered blight nodes as growing researchers and dark nodes as supervisors.

It can be said that there is a relationship between their analyzing strategies and the functions provided by the system. That is, as they could not focus on keyword usage, they had to ensure their assumptions only from information about publications.

### B. System with Full Functions

Also in this case, both of participants used almost the same schemes to complete sub-tasks. To identify growing researchers, they used following schemes:

- (Step1) Identify a researcher who published papers in recent 5 years (2000-2004)
- (Step2) Assign colors to keywords according to participant's assumption
- (Step3) Identify keywords usage over the network

For identification of a researcher who published papers in recent period, both participants used the function of highlighting researchers.

On the other hand, in the second step, they selected keywords according to different assumptions. The assumption of participant A is following:

- If a researcher is a growing researcher, some of his/her keywords are used by only his/her collaborators in both recent and early period.

Participant A used the function of automatically assigning colors to selected keywords. As he assigned colors based on TF/IDF value ($w'_{vk}$), in which the colors are assigned from red (high TF/IDF value) to blue (low TF/IDF value), unique keywords used by few other researchers are colored red or yellow (Fig. 8). Therefore he could easily check the usage of such unique keywords over the entire network in recent and early periods. If the red or yellow keywords are concentrated to his/her collaborators in both periods (Fig. 8), he considered the researcher is a growing researcher. In Fig. 8, a researcher enclosed by

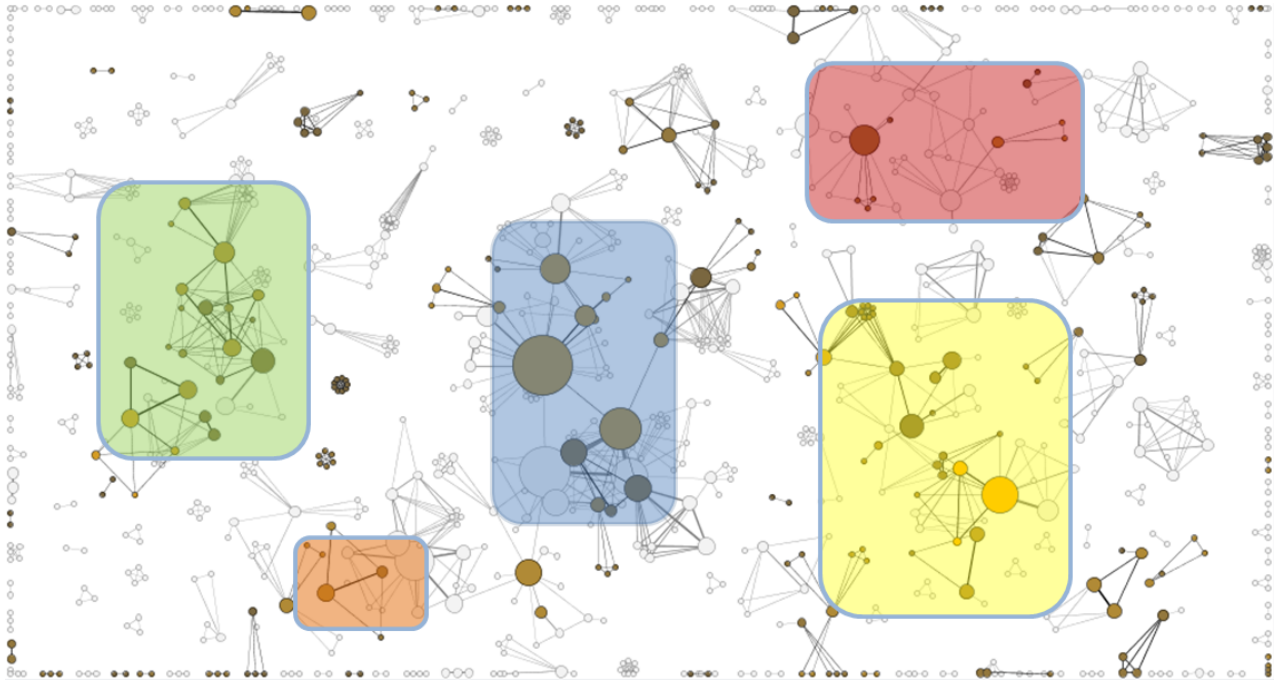


Figure 6. "Map of Researcher Groups" in recent period (2000-2004). Researcher groups are enclosed by rectangles.

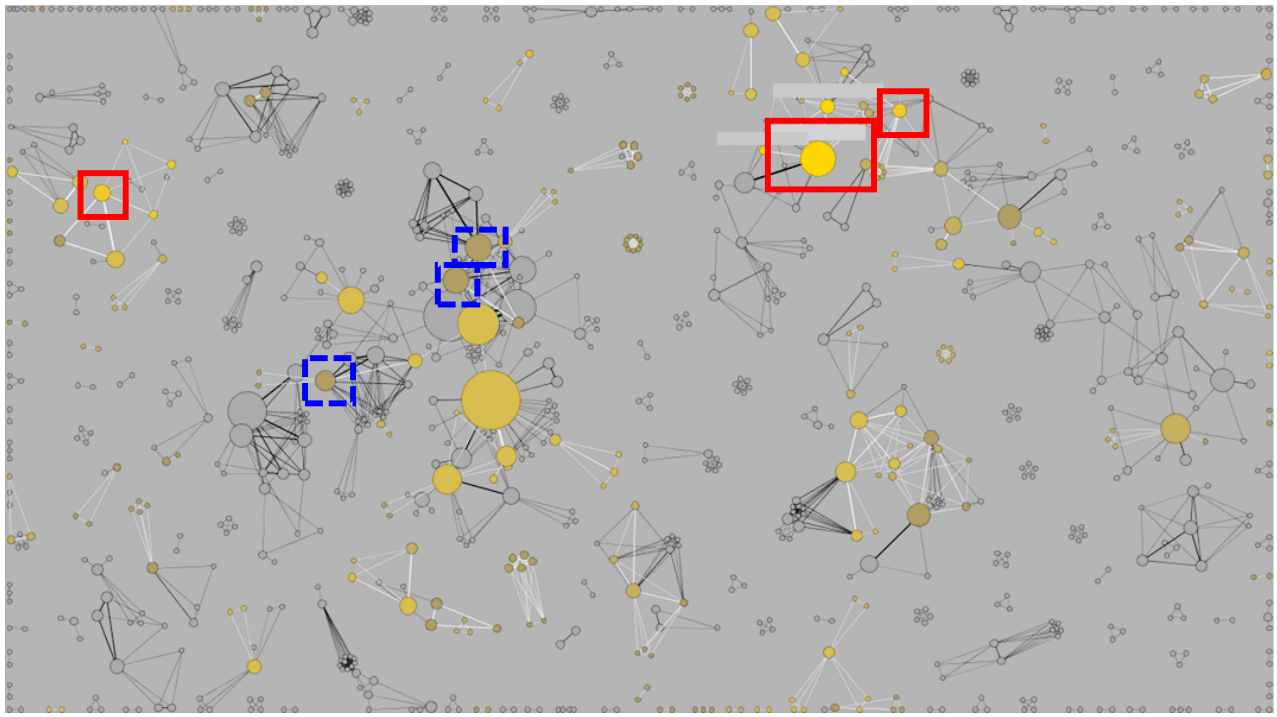


Figure 7. The distribution of last publishing year in recent period (2000-2004).

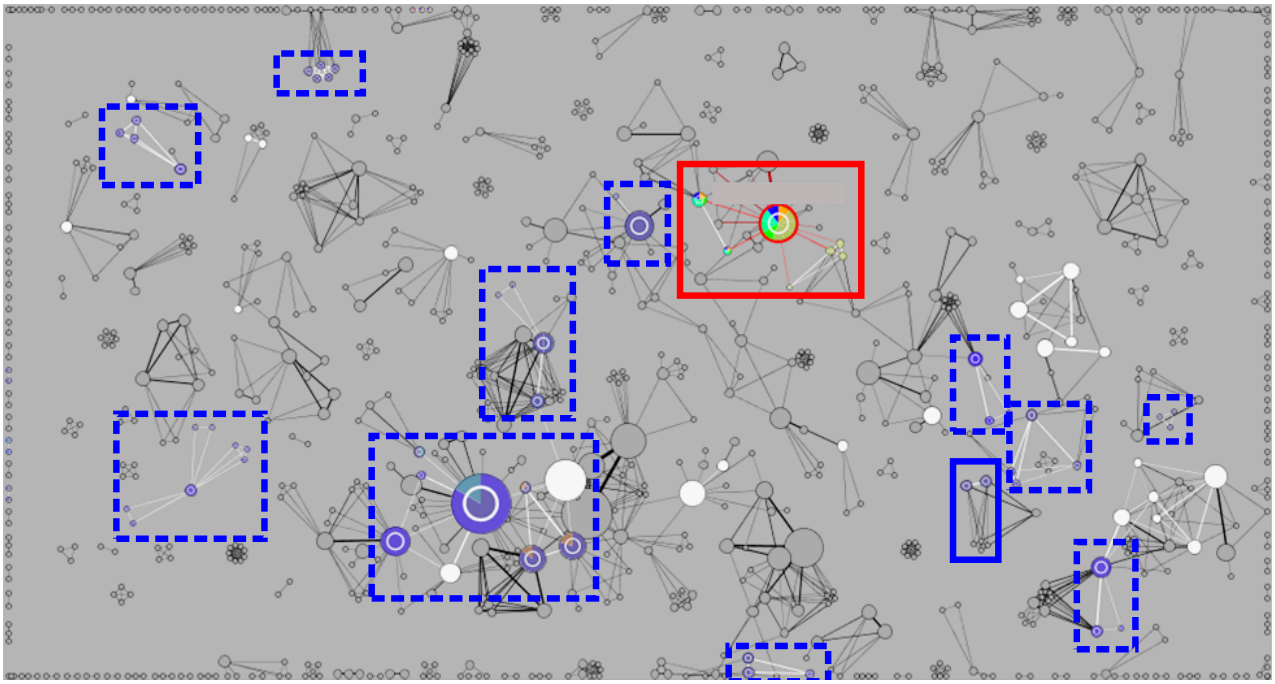


Figure 8. The usage of keywords used by a growing researcher in recent period (2000-2004) (analyzed by participant A)

.

a red solid rectangle is considered a growing researcher. The researchers enclosed by blue dashed rectangles use keywords used by the selected researcher. However they use only non-unique (green or blue) keywords.

Fig. 9 shows the change of his keyword usage from early period to recent period, which satisfies the assumption of participant A.

On the other hand, participant B focused on researcher group, based on the following assumption:

- A growing researcher shares "hot" keywords with several researcher groups in recent period.

In this case, a "hot" keyword means a recently used keyword. He assigned colors to all keywords in the dataset based on TF/IDF value ($w'_k$), then checked whether a researcher group shares keywords with other groups or not. To identify that, he used the function of highlighting researchers who share at least one of keywords used by the selected researcher. Fig. 10 shows the usage of keywords used by a growing researcher in recent period (2000-2004). If the selected researcher and other several groups use the same keywords in recent period, he considered the selected researcher as a growing researcher. In Fig. 10, a researcher group enclosed by a red rectangle contains a growing researcher. The groups enclosed by green or yellow rectangles use at least one of keywords used by the growing researcher.

To identify supervisors, both of participants used similar schemes as used to identify growing researchers.

- (Step1) Identify a researcher who published papers in early period but rarely published papers in recent period (2000-2004)
- (Step2) Assign colors to keywords according to a participant's assumption
- (Step3) Identify keywords usage over the network

In Step2 and 3, they employed the different assumptions each other. The assumption of participant A is following:

- If a researcher is a supervisor, his/her keywords with high TF/IDF value ($w'_{vk}$) are used by researchers other than his/her collaborators in recent and early periods.

As opposite to identification of growing researchers, if unique (red or yellow) keywords are shared without collaborations in early and recent periods, he considered the researcher as a supervisor. In Fig. 11, a supervisor is enclosed by a red solid rectangle. The researchers enclosed by blue dashed rectangles use keywords used by the selected researcher. They use unique (red or yellow) keywords without collaborations.

The assumption of participant B is following:

- If a researcher is a supervisor, keywords which were used by him/her in past period (1974-1999) are used by several other researchers over the network in recent period (2000-2004)

As similar to participant A, participant B assigned colors to keywords based on TF/IDF value ($w'_{vk}$), in which unique keywords used by few other researchers are colored red or yellow. With this visualization, he could easily identify the keywords usage over the network. Interestingly there were two types of supervisors in his results; researchers who used red or yellow keywords (Fig. 12) and those without red or yellow keywords (Fig. 13). In Fig. 12 and 13, a supervisor is enclosed by a red solid rectangle. The researchers enclosed by blue dashed rectangles use keywords used by the supervisor. In Fig.

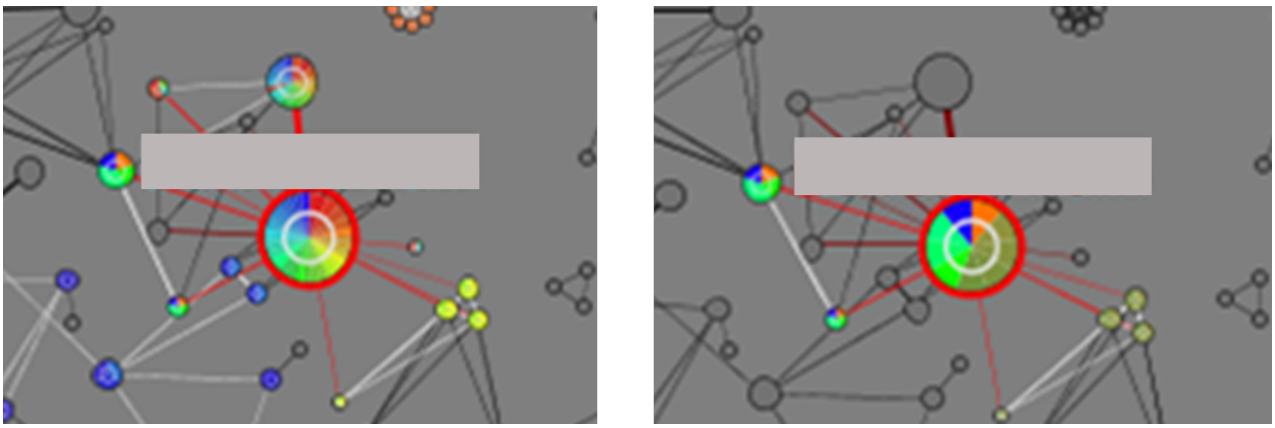


Figure 9. Change of keyword usage: (a) early period (1974-1999), (b) recent period (2000-2004)

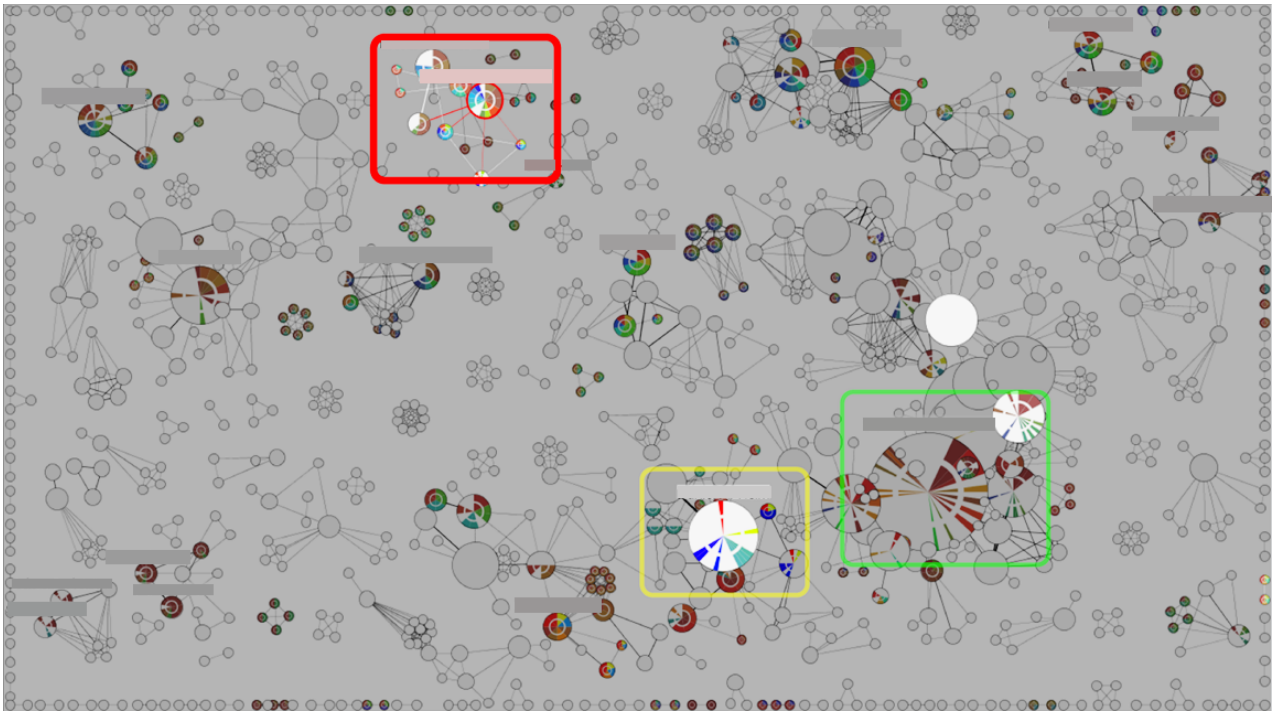


Figure 10. The usage of keywords used by a growing researcher in recent period (2000-2004) (analyzed by participant B)

12, non-unique (green or blue) keywords used by the supervisor are shared by others. In Fig. 13, unique (red or yellow) keywords are shared by those who have no collaboration with the supervisor.

It is not surprising that assumptions of participants are different, as the tasks are abstract and can be performed based on various criteria. It should be noted that both of participants consider keywords usage in a co-authorship networks is important information, in spite of having different assumptions each other. That is, the function is not single-purpose, by which analysis with various assumptions is possible.

Furthermore, the comparison of employed strategies between using the system with limited functions and with full functions shows that the available functions affect their analyzing strategies. By using functions for visualizing keyword usage, they had assumptions regarding research topics.

It is also observed that bi-cylindrical visualization for time variation of node (type *T2*) was not used in their strategies. One possible reason is that it could increase visual complexity. Because there are many nodes and edges in Network panel, participants preferred to the visualization functions which reduce (highlighting) or do not increase (type *T1*) visual complexity. The functions of manual color assignment to keyword and manually grouping keyword were not used as well. Although the functions are useful for analyzing time variation of a few keywords [10], it is supposed that manual operations are not suitable for a large number of keywords.

It is noted that no predetermined correct answers (ground truth) is given in the user study, because our aim

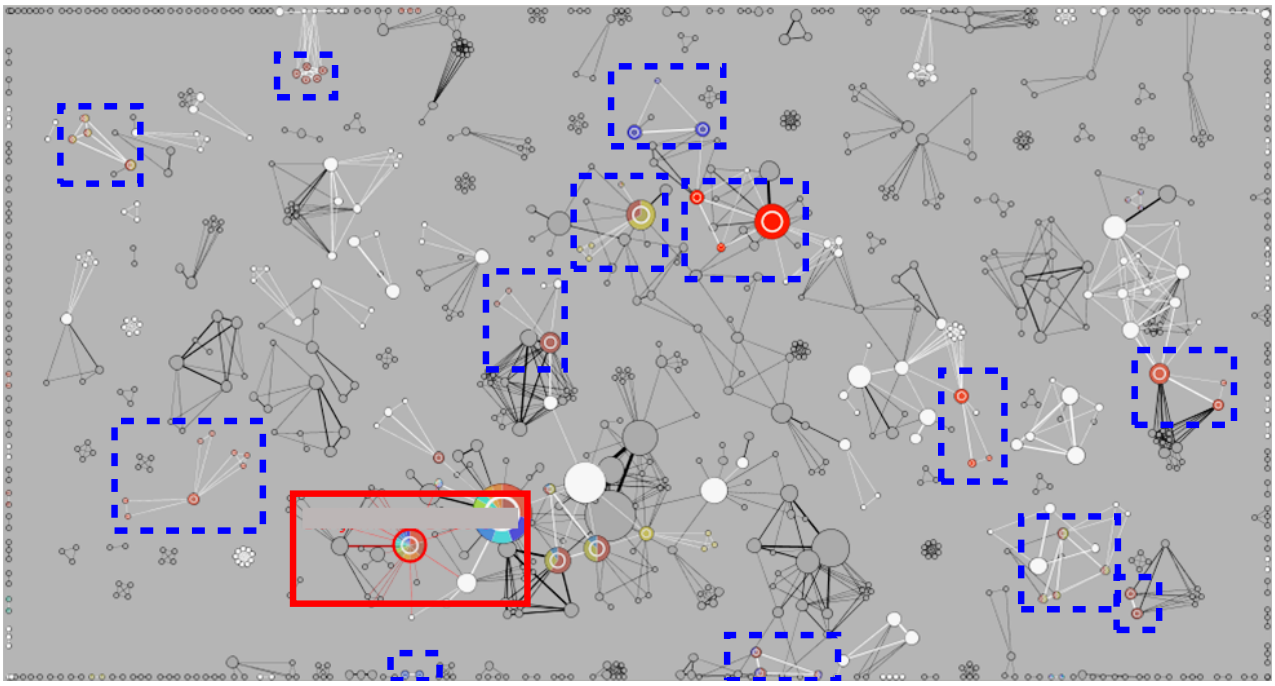


Figure 11. The usage of keywords used by a supervisor in recent period (2000-2004) (analyzed by participant A).

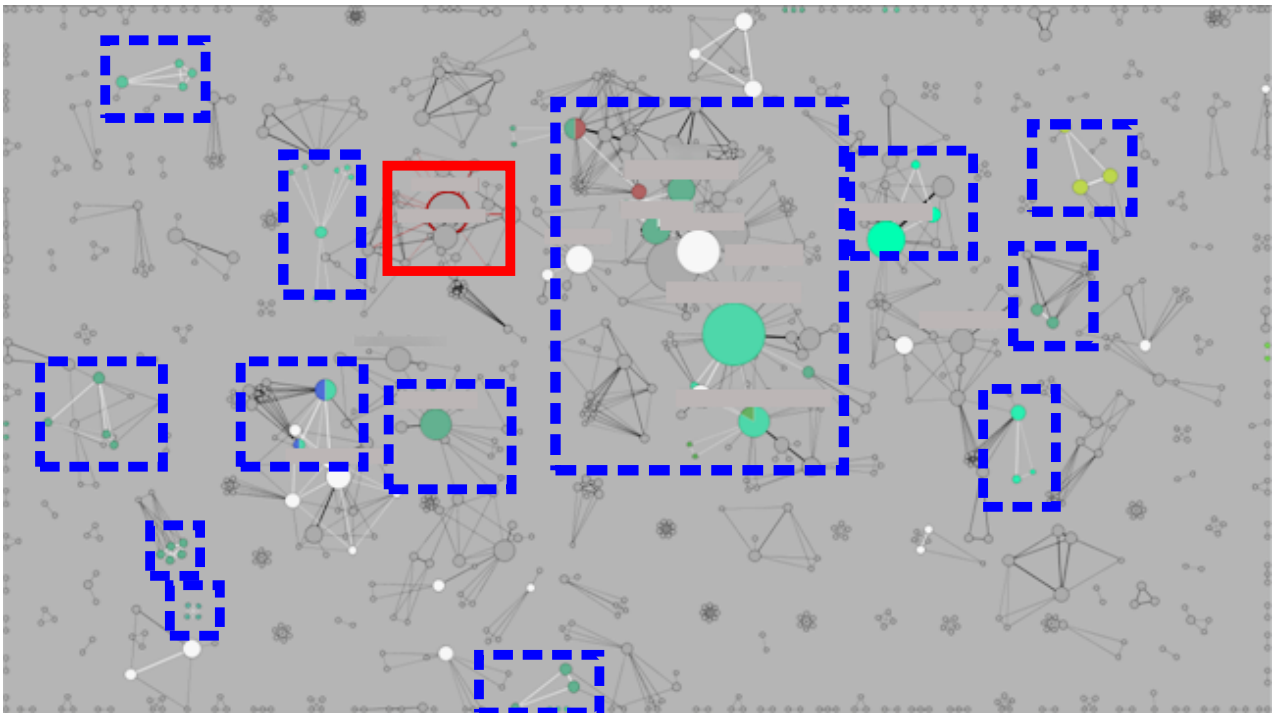


Figure 12. The usage of non-unique (green / blue) keywords used by a supervisor in recent period (2000-2004) (analyzed by participant B).

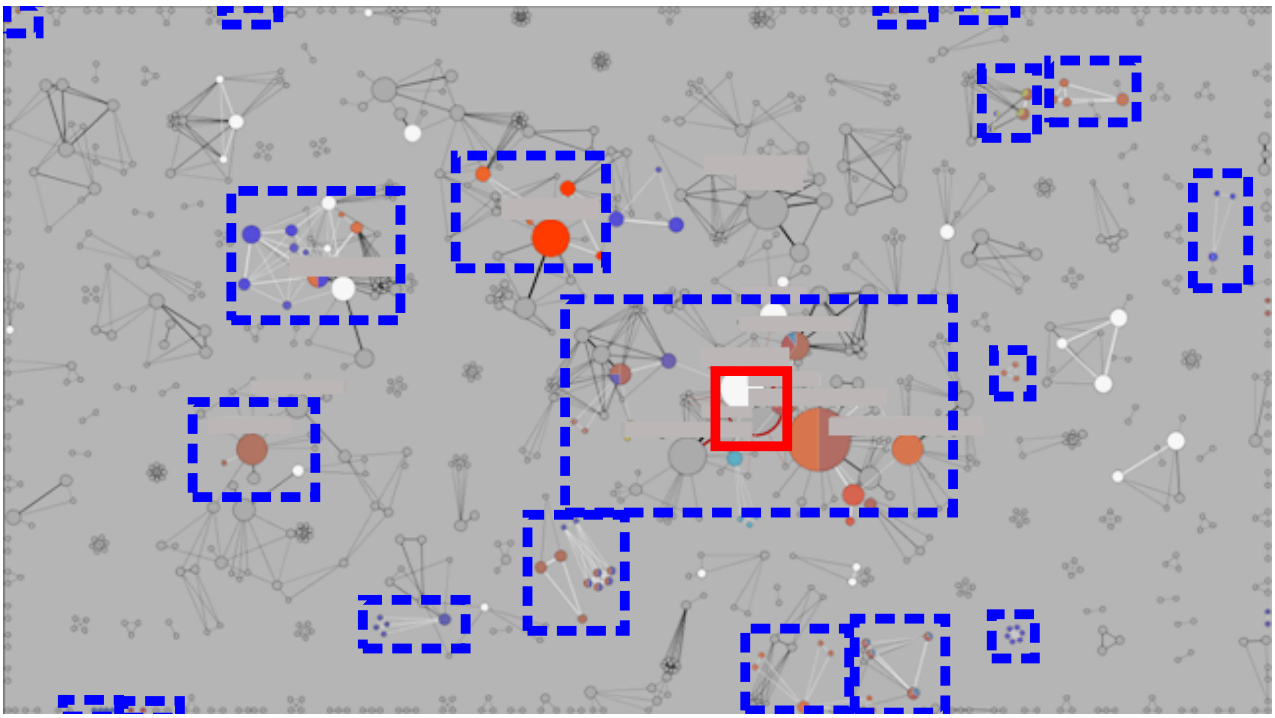


Figure 13. The usage of unique (red / yellow) keywords used by a supervisor in recent period (2000-2004) (analyzed by participant B).

is not to support accurate analysis but to encourage users to perform exploratory data analysis, based on their own assumptions. However, it is also important to confirm that users can analyze data reasonably. Therefore, instead of using ground truth, we asked the participants to conduct an additional task. After finishing all tasks, participants investigated the number of research papers from 2004 to 2007 published by their identified researchers using DBLP. They found that most of growing researchers they identified published 10 or more papers per year and most of supervisors published less than 5 papers per year in that period. These facts show their predictions using the system are reasonable.

## VI. Conclusions

This paper proposes the visualization system for co-authorship networks. To identify growing researchers and supervisors, the system provides support functions for identifying research areas as well as for examining its time variation.

This paper examines whether the proposed system is able to support users predicting future research activities of their unfamiliar domain. In the user study, test participants unfamiliar with InfoVis performed exploratory analysis of the co-authorship network to identify growing researchers and supervisors. The results show that they could perform the tasks even though they had no background knowledge about InfoVis. During performing tasks, it was observed that they employed different analyzing strategies according to available functions. Among the functions the system provided, the function for highlighting researchers who published papers in a specified period is heavily used to identify researchers who recently published papers. The functions for assigning colors to keywords and visualizing node as pie chart of keywords are also frequently used to identify types of researchers. By combining these functions, they could perform the tasks successfully. As one of future works, it is expected that automated assistance of frequently used assumptions would improve the effectiveness of the system.

In this paper, the result of user study by two participants is shown, as it is important for the study of user interface to analyze specific users' behaviors in detail. On the other hand, an experiment with many participants is also important, which should be conducted as one of future studies. The result of this paper will contribute to an experimental design.

Future works also include the introduction of other information than co-authorship networks. Participants suggested the order of authors in a research paper can be useful to determinate whether a researcher is young researcher or a supervisor, as young researcher tends to be first author and supervisor tends to be the second or later. It was also suggested that collaboration of researchers is clearly appeared when they received research grants. As their relationship is stronger than usual, analyzing grant programs could provide interesting information.

## Acknowledgment

We would like to thank Mr. Ippei Ozawa and Mr. Kenji Takamiya for their participations of our study.

This work was partially supported by a grant from National Institute of Informatics (NII) and Japan-Taiwan Joint Research Program by Interchange Association, Japan.

## References


[1] W. Ke, K. Borner, and L. Viswanath, "Major Information Visualization Authors, Papers and Topics in the ACM Library," in *Proc. IEEE Symp. Information Visualization*, 2004, p. 216.

[2] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson, "Understanding Research Trends in Conferences using PaperLens," in *Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1969–1972.

[3] N. Henry, J.-D. Fekete, and M. J. McGuffin, "NodeTrix: a Hybrid Visualization of Social Networks," *IEEE Trans. Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1302–1309, 2007.

[4] P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas, "IN-SPIRE InfoVis 2004 Contest Entry," in *Proc. IEEE Symp. Information Visualization*, 2004, p. 216.

[5] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson, "Understanding Eight Years of InfoVis Conferences Using PaperLens," in *Proc. IEEE Symp. Information Visualization*, 2004, p. 216.

[6] A. Ahmed, T. Dwyer, C. Murray, L. Song, and Y. X. Wu, "WilmaScope Graph Visualisation," in *Proc. IEEE Symp. Information Visualization*, 2004, p. 216.

[7] T. Kurosawa and Y. Takama, "Visualization-based Support of Hypothesis Verification for Research Survey with Co-Authorship Networks," in *Proc. Int'l Workshop on Intelligent Web Interaction*, 2011, pp. 134–137.

[8] ——, "Predicting Researchers' Future Activities using Visualization System for Co-Authorship Networks," in *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence*, 2011, pp. 332–339.

[9] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.

[10] M. Ghoniem, J.-D. Fekete, and P. Castagliola, "A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations," in *IEEE Symp. Information Visualization*, 2004, pp. 17–24.

[11] D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon, "Multiscale Visualization of Small World Networks," in *Proc. IEEE Symp. Information Visualization*, 2003, pp. 75–81.

[12] F. v. Ham and J. J. v. Wijk, "Interactive Visualization of Small World Graphs," in *Proc. IEEE Symp. Information Visualization*, 2004, pp. 199–206.

[13] D. Holten, "Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data," *IEEE Trans. Visualization and Computer Graphics*, vol. 12, no. 6, pp. 741–748, 2006.

[14] B. Lee, C. S. Parr, C. Plaisant, B. B. Bederson, V. D. Veksler, W. D. Gray, and C. Kotfila, "TreePlus: Interactive Exploration of Networks with Enhanced Tree Layouts," *IEEE Trans. Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1414–1426, 2006.

[15] J. Heer and D. Boyd, "Vizster: Visualizing Online Social Networks," in *Proc. IEEE Symp. Information Visualization*, 2005, p. 5.

[16] J.-D. Fekete, G. Grinstein, and C. Plaisant, "
IEEE InfoVis 2004 Contest, The History of InfoVis,"
http://www.cs.umd.edu/hcil/iv04contest, 2004.

[17] T. Kurosawa and Y. Takama, "Visualization System for Co-authorship Networks to Get Insight into Future Research Activities," in *Proc. Joint Int'l Conf. Soft Computing and Intelligent Systems and Int'l Advanced Intelligent Systems*, 2010, pp. 339–344.


**Takeshi Kurosawa,** Japan, 1987. He received the B.E. in Information Communication Technology from Tokyo Metropolitan University Tokyo, Tokyo Japan in 2010. 2010–present, he is a master degree student in Graduate School of System Design, Tokyo Metropolitan University, Japan.

Mr. Kurosawa is a student member of IEEE.

**Yasufumi Takama,** Japan, 1971. Dr. Eng, University of Tokyo, Tokyo Japan, 1999.

He was a JSPS (Japan Society for the Promotion of Science) Research Fellow from 1997 to 1999. From 1999 to 2002 he was a Research Associate at Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology in Japan. From 2002 to 2005, he was an Associate Professor at Department of Electronic Systems and Engineering, Tokyo Metropolitan Institute of Technology, Tokyo, Japan. Since 2005, he has been an Associate Professor at Faculty of System Design, Tokyo Metropolitan University, Tokyo, Japan. He also participated in PREST (Precursory Research for Embryonic Science and Technology), JST (Japan Science and Technology Corporation) from 2000 to 2003. His current research interest includes Web intelligence, information visualization, and intelligent interaction.

Dr. Takama is a member of IEEE, JSAI (Japanese Society of Artificial Intelligence), IPSJ (Information Processing Society of Japan), IEICE (Institute of Electronics, Information and Communication Engineers), and SOFT (Japan Society for Fuzzy Theory and Intelligent Informatics).

# Computational Approach to Prediction of Attitude Change Through eWOM Messages Involving Subjective Rank Expressions

Kazunori Fujimoto
Faculty of Business Administration, Kinki University, Osaka, Japan
Email: kfujimoto@kindai.ac.jp

***Abstract*— Electronic word-of-mouth (eWOM) is an important information source that influences consumer product evaluations. This paper presents a computational model that predicts the potency-magnitude relations of eWOM messages involving subjective rank expressions, which refer to the linguistic representations related to the attitude-levels of the benefits of the product attributes. The amount of required inference for the message receiver to know the attitude-level through the message is quantified as *inference quantum* by using *inference space*, which is characterized by two evaluation parameters: evaluation target size and evaluation scale size. The computational model incorporates the idea of inference quantum into the cognitive hypotheses that were developed to account for the potency differences with reference to the expertise levels - experts or novices - of the message receiver of the products.**

**By applying the computational model to simple eWOM messages, the potency-magnitude relations were observed to depend critically on the values of the message receiver's evaluation parameters. This paper defines three message-classes, which are also studied in the areas of opinion mining and sentiment analysis, and investigates mathematically how the potency-magnitude relations change based on the values of the evaluation parameters.**

***Index Terms*— cognitive modeling; attitude change; electronic word-of-mouth; ewom; social media**

## I. Introduction

In recent years, there has been a focus on electronic word-of-mouth (eWOM) as the information source that influences consumer product evaluations [1]–[3]. eWOM messages refer to statements that are posted electronically in social media such as bulletin boards on the Web. The content includes other consumers' product evaluations and recommendations based on their own experiences and preferences. What kinds of eWOM messages have large potency on the product evaluations made by the consumer who is exposed to the messages? If we can predict the potency on an individual basis, then it will be possible to create an intelligent agent to selectively provide effective statements to individual consumers from among the huge volumes of diverse eWOM messages on the Web. These kinds of intelligent agents would increase opportunities to use eWOM messages and could be expected to promote interactions between consumers via the Web.

The author previously proposed cognitive hypotheses that account for the potency differences in two types - *comparison* and *degree* - of eWOM messages involving subjective rank expressions [4]. This paper develops a computational model of the hypotheses to apply them to various types of messages obtained using techniques from opinion mining and sentiment analysis [5], [6]. The following are the contributions of this paper:

1) Modeled eWOM messages with reference to comprehensive message typology in the areas of opinion mining and sentiment analysis.
2) Developed a computational model that predicts the potency-magnitude relations between two eWOM messages involving subjective rank expressions.
3) Investigated mathematically how the potency-magnitude relations change based on the values of the message receiver's evaluation parameters.

Although the former two contributions were previously presented [7], this is the first appearance for the last contribution (Section V.). In addition, this paper includes four minor modifications from previous work: (1) the idea of "attitude" [8], [9] is incorporated into the definition of subjective rank expressions to clarify the meaning of the "levels" of consumer evaluations (Section II. A.); (2) detailed descriptions of the research background are given (Sections II. B. and C.); (3) the inference quantum is redefined based on the idea of entropy in the inference spaces (Section IV. B.); and (4) computational examples are revised so that two cases with different values of evaluation parameters can be compared (Section IV. C.).

In the following, Section II describes the background of the current research. Section III develops the models of eWOM messages involving subjective rank expressions. Section IV formalizes the computational model and illustrates the prediction processes with example messages. Section V investigates mathematically how the potency-magnitude relations change based on the values of evaluation parameters. Section VI concludes the paper and describes the future work.

## II. Background

### A. Subjective Rank Expressions

Research on word-of-mouth (WOM) communication, which Arndt defined as the oral person-to-person communication between a receiver and a communicator that the receiver perceives as non-commercial [10], has been conducted for many years [11]. It ranges from the motives


doi:10.4304/jetwi.4.1.15-25

for the communications [12], [13] to the effects on the receivers' purchase decisions [14], [15]. Recently, the widespread penetration of social media has increased interest in eWOM communication researches (e.g., [16], [17]) that differ from traditional WOM researches in that they focus on the more detailed aspects of the information content [1], [3].

Lee et al. introduced a differentiation between objective attributes such as size and weight and subjective attributes such as color and shape [1]. Park et al. divided eWOM messages on product attributes into two types, which are "attribute-centric" and "benefit-centric," and verified the differences in the potency of these two types [3]. In contrast, this paper focuses on *subjective rank expressions*, which are related closely to researches in opinion mining and sentiment analysis. Here, subjective rank expressions refer to the linguistic representations related to the attitude-levels of the benefits of the product attributes [4]. A benefit and product attribute pair to be evaluated is called "target" in this paper. The attitude-levels of a target means its ranks or grades with respect to personal attitudes [8], [9]. Thus, subjective rank expressions focus on the benefits and the product attributes that are used as two basic elements to represent product evaluations based on the perspectives in consumer behavior research [18], [19]. The idea of subjective rank expressions is shown in Fig. 1, including similar content from previous researches.

Regarding the benefits, this research examined two types of subjective rank expressions: comparison and degree [4], where the former describes the results of comparisons with other benefits and the latter directly describes the rank of benefits using adjectives and adverbs. A typical example of the comparison type is a message like "The touch panel LCD of product X is easier to use than that of product Y," which claims that this attribute of X is rated higher than that of Y with respect to the benefit; easy to use. On the other hand, a typical example of the degree type is a message like "The touch panel LCD of product X is incredibly easy to use," which claims the attribute of X is rated high with respect to the benefit. Since both messages are concerned with the attitude-levels of product attributes for a benefit, they involve subjective rank expressions. As shown in the example messages, eWOM messages involving subjective rank expressions contain not only information connecting attributes to benefits but also information related to the authors' attitude-levels of the benefits of attributes.

The "potency" of eWOM messages in this paper relates to the attitude change in the product evaluations when the receiver is exposed to the message. Message $m_1$ has larger potency than $m_2$ when the degree of the attitude change by $m_1$ is larger than $m_2$. The potency depends not only on the message content but also on the characteristics of the message receivers and of the evaluated products, so a different person as well as a different product may give different potency with respect to the same message content [20]. There are two types of potency: positive and negative. The former changes the product evaluations positively and the latter changes them negatively [21], [22]. This paper focuses on positive potency because one aim of this research is to develop intelligent agents that selectively provide eWOM messages to increase consumer purchase intention.

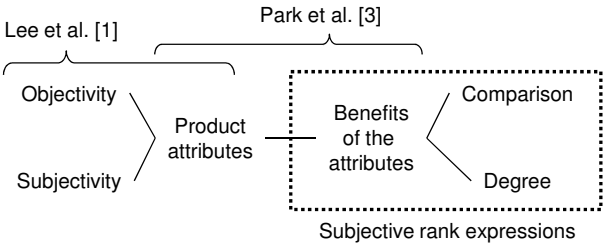



Figure 1. Subjective rank expressions and related work.

Some psychological measurements of attitude change are often used to determine the potency of eWOM messages (e.g., [1], [3]). Such measurements are also used in the area of persuasion research [8], [23], which is closely related to advertising and word-of-mouth researches. In persuasion researches, the term "persuasiveness" is often used instead of potency. The difference between persuasiveness and potency is the presence or absence of a goal and the intention to reach the goal of the message providers; i.e., the term persuasiveness postulates such a goal but the term potency does not.

### *B. Cognitive Hypotheses*

The cognitive hypotheses proposed in [4] focus attention on how much inference is required for the message receiver to know the author's attitude-level of the targets through the message. Since consumers with high expertise in the products are likely to infer based on their own knowledge, they are expected to prefer comparison type in which the attitude-level is not written explictly and leaves room for personal determinations. In contrast, since consumers with low expertise are likely to dislike such inferences, they are expected to prefer the degree type in which the attitude-level has already been determined by the author so that the evaluation can be directly obtained by the message. Thus, the following hypotheses were proposed [4].

Hypothesis A: For consumers with high expertise, comparison type eWOM messages for targets has larger potency on the evaluation of the targets than degree type eWOM.

Hypothesis B: For consumers with low expertise, degree type eWOM messages for targets has larger potency on the evaluation of the targets than comparison type eWOM.

These hypotheses were supported by hypothesis testing on the dataset collected from a questionnaire survey administered to one hundred and fifty two undergraduate students [4].

A theoretical background of the hypotheses is the theory of *implicit conclusions* [24]–[26], which was de-

veloped mainly to account for the persuasiveness of advertising. For example, a typical ad with an explicit conclusion is "Now That You Know the Difference, Shave With Edge – The Disposable Razor That is Best for You." A typical one with an implicit conclusion is "Now That You Have the Facts, Decide for Yourself Which Toothbrush You Should Buy," as introduced in [24]. It states that ads with implicit conclusions are expected to be persuasive when the audience is highly involved in the products instead of being lowly involved. Sawyer et al. explained the persuasiveness as follows [25]: "Perhaps the most important reason is that the absence of any obvious conclusion may lead a motivated audience to try to infer one. ··· Attitudes resulting from effortful self-generated conclusions should be more positive than attitudes resulting from less effortful processing of conclusions explicitly provided in a message and more accessible and persistent over time." The theory of implicit conclusions was empirically supported [25], [26] and extended from wider viewpoints such as attention for visual material [27] and missing attributes [28].

The cognitive hypotheses [4] can be viewed as one application of the theory of implicit conclusions and, in that sense, are characterized from two aspects. First, the hypotheses focus on the inference of the author's attitude-levels toward targets through messages and regard the attitude-levels as "conclusions." Second, the hypotheses incorporate the expertise of message receivers instead of their involvement, which is a motivational parameter used by the message receiver to infer the conclusions. As Chebat et al. suggested, both expertise and involvement should be considered to obtain accurate potency predictions [29]. However, this paper only considers expertise because it is not so difficult to extend the idea with expertise only to the one with both factors by assuming no interaction effect between them.

Expertise of products has various aspects, or dimensions [18], and is measured in various ways. For example, Park et al. used the number of correct responses to questions about the products and performed a median-split technique to divide consumers into experts and novices [3]. As another example, the author defined expertise with respect to having/not having an experience of purchasing products, where expert and novice refer to having and not having [4]. At a more practical setting for intelligent agents, expertise may be determined with keywords or bookmarks used or possessed by users.

### C. Research Purpose

Previous work [4] focused on two message types: comparison and degree; but subjective rank expressions should have a wide variety of message subtypes. For example, gradable comparatives are classified into three subtypes: non-equal gradable, equative, and superlative [30]. The similarity or difference between two objects may generate other subtypes, as shown in [31]. In addition, there may be messages in which two or more types are combined. Such a wide variety of message types requires the cognitive

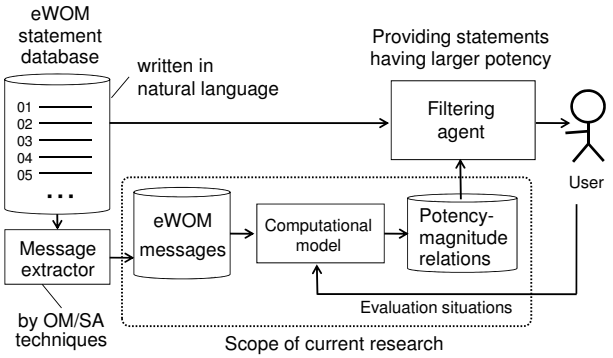



Figure 2. Illustrative application of computational model.

hypotheses to be very generalized. The purpose of the current research is to achieve generalization by developing a computational model that measures the amount of required inference (Q) for such various messages. The generalized hypothesis becomes the following: for any two eWOM messages $m_i$ and $m_j$, if $Q(m_i) > Q(m_j)$ then $m_i$ has larger potency than $m_j$ for experts and, conversely, $m_j$ has larger potency than $m_i$ for novices.

Fig. 2 shows an illustrative application of the computational model and the scope of this research. In the figure, the filtering agent selects the eWOM statements based on the potency-magnitude relations generated from the computational model, which is the main topic of this paper. To obtain the potency-magnitude relations on the eWOM statements written in natural language, the message extractor constructed by opinion mining and sentiment analysis (OM/SA) techniques extracts subjective messages, which are the eWOM messages in the figure, as definite shapes from the natural language statements. Then the computational model generates the potency-magnitude relations on the messages based on the evaluation situations of the user. Although both positive and negative eWOM statements exist when they are gathered through social media, only positively evaluated messages for products are selected and used to promote the purchase intention of system users. As shown in the figure, the scope of this research does not directly include the techniques in the area of OM/SA. However, OM/SA research is closely related to my current research because the formats of the messages should be determined based on the techniques.

## III. Message Modeling

### A. Comparison Type

Two of the most fundamental categories for human opinion are comparative and direct [6]. A direct opinion expresses a subjective idea on a single object, while a comparative opinion expresses a relation of differences or similarities between two or more objects and/or object preferences of the opinion holder. Comparatives are classified into four subtypes: non-equal gradable, equative, superlative, and non-gradable [30].

Based on the subtypes of the comparatives, the eWOM messages in the comparison type of the subjective rank expressions are modeled as follows:

(*target1, target2, type*),

where *target1* and *target2* are the sets of the pairs of a benefit and a product attribute and *type* is one of the three subtypes: non-equal gradable, equative, and superlative. For the non-equal gradable (equative) *type*, the messages insist that the attitude-levels of *target1* are larger than (are equal to) those of *target2*. For the superlative *type*, the messages insist that the attitude-levels of *target1* are the largest among all other targets to be evaluated; *target2* is omitted. The parameter *type* excludes non-gradable from its values because non-gradable does not address the attitude-levels of targets.

The message model proposed here may be obtained by adjusting Jindal's message model using five parameters: *relationWord, features, entityS1, entityS2,* and *type* [30]. Parameters *features, entityS1,* and *entityS2* are related to *target1* and *target2* in the proposed model, while parameter *type* is the same in both models. Parameter *relationWord* takes a keyword such as *-er* or *exceed* that is used to express a comparative relation in a sentence. Although the proposed model does not contain parameter *relationWord*, the benefits in *target1* and *target2* may contain a piece of the parameter information when it has beneficial words such as easier and lightest.

Note that, although the message model proposed here is similar to the Jindal's model in the appearance, they are different in the semantics of the comparison. That is, the message model proposed here compares the attitude-levels of targets whereas the Jindal's model compares certain features like length and size of entities. Therefore, they may generate different structures of the comparative relations. For example, for digital cameras, a message like "The start up time of X is longer than that of Y." constructs the relation $\mathrm{X} > \mathrm{Y}$ with respect to the length of time by the Jindal's model whereas it constructs the relation $\mathrm{X} < \mathrm{Y}$ with respect to the attitude-levels by the proposed model (Shorter is better in this case.). The issue described here is also discussed in [32], [33].

### B. Degree Type

One typical problem in the research area is polarity detection that classifies an online review as positive or negative at a document level [34] or a sentence level [35]. Recently, the rating-inference problem is also studied to classify not into two classes, positive or negative, but into fine-grained rating classes (e.g., one to five "stars") [36]. Rating-inference tasks determine an author's evaluation from the review texts with respect to a multi-point rating scale, which is a kind of ordinal scale. The latent message models that the tasks postulate appear to have three elements: an evaluated object, its rated level, and a multi-point rating scale for the evaluation.

Based on this idea, the eWOM messages in the degree type of subjective rank expressions are modeled as follows:

(*target, level, scaleInfo*),

where *target* is a set of the pairs of a benefit and a product attribute and *level* is the attitude-level based on *scaleInfo*, which is the specifications of the multi-point rating scale used. The specifications include the number of points on the rating scale and, if required, the polarity that each point belongs. Five-point Likert type scales, which are often used in psychological experiments, are one alternative for the rating scales. In the case, the number of points on the rating scales is five and points 1, 2 belong negative, 3 belongs neutral, and 4, 5 belong positive attitude.

The granularity of the multi-point rating scale has variations. Pang et al. discussed a reasonable classification granularity to determine other persons' evaluations by using Internet movie reviews [36]. They examined pairs of reviews extracted from the review set to determine whether the first review in each pair was more positive than, less positive than, or as positive as the second. They concluded that the reasonable scale size, which is the number of points on the rating scale, is not so large and is four or five. As they discussed, much finer-grained may not be reasonable when no information exists to discern such finer-grained levels in the message texts. There may not be enough text samples to create classification rules with finer-grained scales using machine learning techniques. The granularity of the multi-point rating scale is determined practically by considering such properties of the message texts.

## IV. Computational Model

### A. Basic Idea

In the computational model, the amount of required inference is quantified as the *inference quantum*. Messages explicitly containing an attitude-level enable it to be obtained directly, and thus they require no inference; the size of inference quantum is 0. Since messages containing only comparative relations of the attitude-levels require some inference to obtain the levels, the inference quantum is not 0 but has a certain value. The inference quantum does not postulate the levels contained in a single message but postulates the set of the levels of all targets to be evaluated. Therefore, the computational model incorporates the idea of *inference space* that contains all possible attitude-levels inferred by the message receiver.

The dimensions of the inference space correspond to the targets to be evaluated. Fig. 3 shows an example of the inference space where two targets, A and B, are evaluated and a 5-point rating scale ranging from 1 to 5 is used to evaluate the attitude-levels. The horizontal and vertical axes represent the attitude-levels of A and B, which are denoted as $h$(A) and $h$(B), respectively. The inference space consists of 25 points in this case. A certain point in the inference space gives the attitude-levels of all targets, A and B. For example, point $e = (4, 3)$ indicates that the attitude-levels of targets A and B are 4 and 3, respectively.

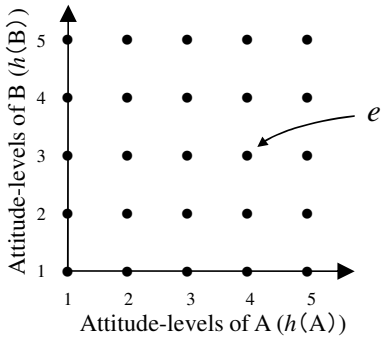



Figure 3. Inference space.

Thus, based on inference space, the determination tasks of the attitude-levels of the targets are regarded as the determination of one point in the inference space. The idea of inference space was inspired in part by distribution hyperspace [37] and its extension [38].

The inference quantum of messages giving stronger constraints in the inference space is considered smaller because such messages limit the inferred space to a narrower region. On the other hand, the inference quantum of messages giving weaker constraints is considered larger because such messages allow the inferred space to be wider. Thus, the inference quantum is expected to be quantified using the size of the compatible regions in the inference space with the message.

### *B. Formalization*

As notations for inference space, the following symbols are used.

- Target set $\Omega$ denotes the set of all targets to be evaluated. Evaluation target size $k$, which is a finite integer greater than or equal to 1, denotes the size of $\Omega$.
- Evaluation scale size $\nu$, which is a finite integer greater than or equal to 2, denotes the number of points on the rating scale used for the attitude-level evaluations. It is also written like $\nu$ point rating scales.
- A pair of $k$ and $\nu$, denoted $\lambda = (k, \nu)$, is called evaluation parameters.
- Inference space $\Theta_\lambda = \{e_1, \ldots, e_{\nu^k}\}$ denotes the set of all possible attitude-levels for all $k$ targets by using the $\nu$ point rating scale. The elements $e_j$, $j = 1, \ldots, \nu^k$ are called points in $\Theta_\lambda$.

Next consider a set of messages $M = \{m_1, \ldots, m_n\}$, each of which is eWOM message involving subjective rank expressions for one or more targets in $\Omega$. The points of $\Theta_\lambda$ compatible with $m_i \in M$ are denoted as $r_i$. Based on the idea of inference space $\Theta_\lambda$ and compatible points $r_i$ of $\Theta_\lambda$, the inference quantum is defined below.

**Definition** (Inference Quantum)

The *inference quantum* $Q$ of a message $m_i \in M$ for a message receiver with evaluation parameters $\lambda$ is defined by

$$Q(m_i) = \log_2 \sum_{e \in \Theta_\lambda} \eta_i(e) \ , \tag{1}$$

where $\eta_i$ is a function:

$$\eta_i(e) = \begin{cases} 1 & \text{if } e \in r_i \\ 0 & \text{if } e \notin r_i \end{cases} . \tag{2}$$

The inference quantum $Q$ takes an integer ranging from 0 to $k \log \nu$. Maximum value $k \log \nu$ of the inference quantum is given to the messages that are compatible with all of the inference space, but minimum value 0 is given to the messages that have only one compatible point in the inference space. The inference quantum is denoted by $Q_\lambda$ when the evaluation parameters should be written explicitly.

Based on the inference quantum, the computational rule for predicting potency-magnitude relations between two eWOM messages is described below.

**Prediction Rule**

If $Q(m_i) > Q(m_j)$, both $m_i$ and $m_j \in M$ give positive support to a target $\in \Omega$,
then $m_i \succ m_j$ for experts and
$m_j \succ m_i$ for novices,
where $m_i \succ m_j$ ($m_j \succ m_i$) denotes $m_i$ ($m_j$) is expected to have larger potency than $m_j$ ($m_i$) with respect to the positive attitude change in the target.

As shown in the prediction rule, potency-magnitude relations are derived by discerning the expertise level of the message receiver of the products. Note that the rating scale for an inference space is determined based on the cognitive perspective of the message receiver's evaluations. Therefore, it may not be compatible with the rating scales for degree type messages because the scales are often determined previously with some practical conditions in the message extraction techniques. However, the identical scale should be used because when a different scale is used, a mapping rule between the scales has to be developed to obtain compatible region $r_i$ with the messages. To conform the scale for the degree type messages to the scale for the inference space, it is necessary to previously prepare degree type messages not in a single type rating scale but in several types and to choose messages in a compatible type when the inference space is determined.

### *C. Example*

This subsection illustrates the prediction processes using two different situations of evaluation parameters, $\lambda_a$ and $\lambda_b$, as shown in Table I(a). The target set on $\lambda_a$ consists of A, B, and C, that is, $k = 3$, whereas that on $\lambda_b$ consists of A, B, C, and D, that is, $k = 4$. The evaluation scale size $\nu$ on $\lambda_a$ and $\lambda_b$ is the same value, 5. The inference space for $\lambda_a$ contains 125 ($= 5^3$) points,

TABLE I.
PARAMETERS AND MESSAGES IN ILLUSTRATIVE EXAMPLE

(a) Evaluation parameters

| | Target set $\Omega$ | Size of $\Omega$ | Scale size |
|---|---|---|---|
| $\lambda_a$ | A, B, C | $k = 3$ | $\nu = 5$ |
| $\lambda_b$ | A, B, C, D | $k = 4$ | |

(b) eWOM messages

| | Representations | Compatible Region |
|---|---|---|
| $m_1$ | (A, B, non-equal gradable) | $h(A) > h(B)$ |
| $m_2$ | (A, , superlative) | $h(A) > h(*)$ |
| $m_2$ | (A, 5, {5-point scale, 1 to 5}) | $h(A) = 5$ |

"*" indicates any other target in $\Omega$.

TABLE II.
COMPUTATIONAL RESULTS

(a) Inference quantum $Q(m_i)$ for each message

| | $m_1$ | $m_2$ | $m_3$ |
|---|---|---|---|
| $\lambda_a$ | 5.64 | 4.91 | 4.64 |
| $\lambda_b$ | 7.97 | 6.64 | 6.97 |

(b) Potency-magnitude relations among messages

| | Experts | Novices |
|---|---|---|
| $\lambda_a$ | $m_1 \succ m_2 \succ m_3$ | $m_3 \succ m_2 \succ m_1$ |
| $\lambda_b$ | $m_1 \succ m_3 \succ m_2$ | $m_2 \succ m_3 \succ m_1$ |

so the inference quantum based on $\lambda_a$ ranges from 0 to $3\log 5$ (=6.97). On the other hand, the inference space for $\lambda_b$ contains 625 ($= 5^4$) points, so the inference quantum based on $\lambda_b$ ranges from 0 to $4\log 5$ (=9.29).

The eWOM messages used in this illustration are shown in Table I(b). Message $m_1$ is the non-equal gradable type and means that the attitude-level of A is larger than that of B. It specifies the region where $h$(A) > $h$(B) in the inference space. Message $m_2$ is the superlative type and means that the attitude-level of A is the largest. It specifies the intersectional region of "$h$(A) > $h$(B)" and "$h$(A) > $h$(C)" for $\lambda_a$ and the intersectional region of "$h$(A) > $h$(B)," "$h$(A) > $h$(C)," and "$h$(A) > $h$(D)" for $\lambda_b$. Message $m_3$ is the degree type and means that the attitude-level of A is 5 on the 5-point rating scale ranging from 1 to 5. It specifies the region where $h$(A) = 5 in the inference space. All these messages positively support target A, so that the prediction rule derives the potency-magnitude relations with respect to the positive attitude change in target A.

Table II(a) shows the calculation results of the inference quantum of the eWOM messages. For $\lambda_a$, the inference quanta of $m_1, m_2$, and $m_3$ are 5.64, 4.91, and 4.64, respectively. For $\lambda_b$, the inference quanta of $m_1, m_2$, and $m_3$ are 7.97, 6.64, and 6.97, respectively. Table II(b) shows the potency-magnitude relations derived from the prediction rule. With respect to $\lambda_a$, relations $m_1 \succ m_2 \succ m_3$ for experts and relations $m_3 \succ m_2 \succ m_1$ for novices are obtained. This suggests that a promising strategy of intelligent filtering agents to promote A is achieved by giving priority to $m_1$ for experts and to $m_3$ for novices. On the other hand, with respect to $\lambda_b$, relations $m_1 \succ m_3 \succ m_2$ for experts and relations $m_2 \succ m_3 \succ m_1$ for novices are obtained. This suggests that a promising strategy of intelligent filtering agents to promote A is achieved by giving priority to $m_1$ for experts and to $m_2$ for novices.

Note that a reversal phenomenon of the potency-magnitude relations between $m_2$ and $m_3$ was observed in the computational results, i.e., the relation $m_2 \succ m_3$ ($m_3 \succ m_2$) in $\lambda_a$ is reversed as $m_3 \succ m_2$ ($m_2 \succ m_3$) in $\lambda_b$ for experts (for novices). This observation suggests that the message receiver's situation of evaluation parameters may change the potency-magnitude relations. In other words, accurate prediction of the potency-magnitude relations can not be achieved without considering the values of evaluation parameters where the message receiver evaluates products with eWOM messages.

## V. MATHEMATICAL PROPERTIES

This section defines three message-classes and mathematically investigates how the potency-magnitude relations change based on the values of evaluation parameters $k$ and $\nu$. The mathematical investigations construct *Q-magnitude Relation Map (Q-Map)*, which (1) partitions the space spanned by $k$ and $\nu$ into disjoint regions such that different regions give different magnitude-relations of inference quanta and (2) labels the regions to give the same label for the regions where the same magnitude-relation holds. This section also derives *Priority Message-Class Map (P-Map)* for experts and novices by applying the prediction rule to the Q-Map. The P-Map contributes to develop eWOM message filtering strategies. The assumptions used in this section are summarized below:

- The prediction rule is applied to an evaluation situation where a message receiver evaluates products with eWOM messages. Therefore, to derive the magnitude relations of the inference quanta, they are compared on the same evaluation parameter values. This means that when we say $Q(m_i) > Q(m_j)$, terms $Q(m_i)$ and $Q(m_j)$ are calculated by the same $k$ and by the same $\nu$.
- Each receiver has target set $\Omega$ and uses messages for the targets in $\Omega$ to make decisions. Therefore, all targets in all the messages to be compared are contained in target set $\Omega$. This means that the evaluation target size $k$ is larger than or equal to 2 on the premise of the non-equal gradable type, which contains two different targets at least.

### A. Calculating Formula of Inference Quantum

Messages $m_1, m_2$ and $m_3$ used in the Section IV example are generalized by using message-classes $M^{(1,1)}_{type1}, M^{(1)}_{type2}$, and $M^{(1)}_{type3}$, respectively:

$M^{(1,1)}_{type1}$ : The set of all non-equal gradable type messages that insist the attitude-level of a target ($\in \Omega$) is larger than that of another target ($\in \Omega$).

$M^{(1)}_{type2}$ : The set of all superlative type messages that insist the attitude-level of a target ($\in \Omega$) is larger than those of all other targets ($\in \Omega$).

$M^{(1)}_{type3}$ : The set of all degree type messages that insist the attitude-level of a target ($\in \Omega$) is a certain value on a $\nu$ point rating scale.

As shown in these definitions, the message-classes neither depend on the target names nor on particular attitude-levels in the degree type messages. This section does not use $m_1, m_2$, and $m_3$ directly, but instead uses messages $m^{(1,1)}_{type1}, m^{(1)}_{type2}$, and $m^{(1)}_{type3}$, each of which belongs to classes $M^{(1,1)}_{type1}, M^{(1)}_{type2}$, and $M^{(1)}_{type3}$, respectively. For example, the statement "$Q(m^{(1,1)}_{type1}) > Q(m^{(1)}_{type2})$" is used to state "$Q(m_i) > Q(m_j)$ for all $m_i \in M^{(1,1)}_{type1}, m_j \in M^{(1)}_{type2}$."

The inference quanta of $m^{(1,1)}_{type1}$, $m^{(1)}_{type2}$, and $m^{(1)}_{type3}$ are calculated from evaluation parameters $k$ and $\nu$ as the following formulae:

$$Q(m^{(1,1)}_{type1}) = \log \frac{\nu^{k-1}(\nu-1)}{2}. \quad (3)$$

$$Q(m^{(1)}_{type2}) = \log \left( \sum_{i=1}^{\nu} (\nu - i)^{k-1} \right). \quad (4)$$

$$Q(m^{(1)}_{type3}) = \log \nu^{k-1}. \quad (5)$$

The explanations for them are described below:

- Eq. (3): The number of compatible points with message $m^{(1,1)}_{type1}$ in the inference space's subspace spanned by the two targets described in the message becomes $(\nu^2 - \nu)/2$ because it excludes the points where two targets have the same attitude-level and the points where a target has smaller attitude-levels than another target. It is multiplied by $\nu^{k-2}$ to consider other dimensions that correspond to other targets in $\Omega$. Then $\nu^{k-1}(\nu-1)/2$ is obtained.
- Eq. (4): When the attitude-level of the target described in the message takes largest value $\nu$ ($i = 1$), the number of compatible points with message $m^{(1)}_{type2}$ in the inference space becomes $(\nu-1)^{k-1}$ because the attitude-levels of other targets in $\Omega$ can take any level less than or equal to $(\nu-1)$. In the same way, when the attitude-level of the target described in the message takes value $\nu - 1$ ($i = 2$), the number of compatible points in the inference space becomes $(\nu-2)^{k-1}$. Considering from $i = 1$ to $\nu$, $\sum_{i=1}^{\nu}(\nu-i)^{k-1}$ is obtained.
- Eq. (5): The number of compatible points with message $m^{(1)}_{type3}$ in the inference space's subspace spanned by the target described in the message becomes 1 because it specifies a single point in the subspace. In the same way for message $m^{(1,1)}_{type1}$, it is multiplied by $\nu^{k-1}$ to consider other dimensions that correspond to other targets in $\Omega$. Then $\nu^{k-1}$ is obtained.

**Example** Simple numerical examples where $k = 2$ and $\nu = 5$ are presented to illustrate the idea of Eqs. (3)-(5). Three messages $m_1$, $m_2$, and $m_3$ introduced in the previous section (Table I(b)) are used to show the number of compatible points visually.

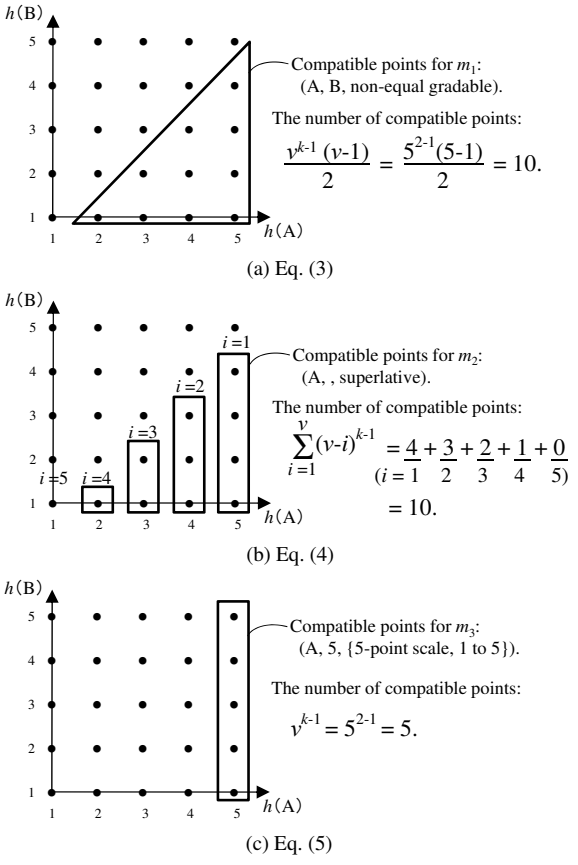



Figure 4. Illustrative examples ($k = 2, \nu = 5$) for Eq. (3), (4) and (5).

(a) From Eq. (3), $Q(m^{(1,1)}_{type1}) = \log \frac{5^{2-1}(5-1)}{2} = \log 10$ is obtained. Fig. 4(a) illustrates the number of compatible points for $m_1 \in M^{(1,1)}_{type1}$.

(b) From Eq. (4), $Q(m^{(1)}_{type2}) = \log \left( \sum_{i=1}^{5} (5-i)^{2-1} \right) = \log 10$ is obtained. Fig. 4(b) illustrates the number of compatible points for $m_2 \in M^{(1)}_{type2}$.

(c) From Eq. (5), $Q(m^{(1)}_{type3}) = \log 5^{2-1} = \log 5$ is obtained. Fig. 4(c) illustrates the number of compatible points for $m_3 \in M^{(1)}_{type3}$.

Some propositions shown in the next section are proved not by using inference quantum $Q$ directly but using function $E$, which is defined without the logarithm function in Eq. (1); that is, $Q(\cdot) = \log_2 E(\cdot)$. The usage of $E$ works with Lemma shown below.

**Lemma**

(a) Suppose two messages, $m_i$ and $m_j$ ($\in M$). If $E(m_i) > E(m_j)$ then $Q(m_i) > Q(m_j)$.

(b) Suppose two messages, $m_i$ and $m_j$ ($\in M$), such that $E(m_j) \neq 0$. If $E(m_i)/E(m_j) > 1$ then $Q(m_i) > Q(m_j)$.

**Proof:** (a) $E(m_i) > E(m_j) \Rightarrow \log E(m_i) > \log E(m_j)$.
(b) $E(m_i)/E(m_j) > 1 \Rightarrow \log\{E(m_i)/E(m_j)\} > \log 1$
$\Rightarrow \log E(m_i) - \log E(m_j) > 0$. □

### B. Mathematical Propositions

Three mathematical propositions for the magnitude relations of the inference quanta between $m_{type1}^{(1,1)}$ and $m_{type2}^{(1)}$, between $m_{type1}^{(1,1)}$ and $m_{type3}^{(1)}$, and between $m_{type2}^{(1)}$ and $m_{type3}^{(1)}$ are presented as follows:

**Proposition 1** *(Between $m_{type1}^{(1,1)}$ and $m_{type2}^{(1)}$)*
(a) When $k = 2$ and $\nu \geq 2$, $Q(m_{type1}^{(1,1)}) = Q(m_{type2}^{(1)})$.
(b) When $k \geq 3$ and $\nu \geq 2$, $Q(m_{type1}^{(1,1)}) > Q(m_{type2}^{(1)})$.

**Proof:** (a) By substituting $k = 2$ for Eqs. (3) and (4), it is easy to confirm $Q(m_{type1}^{(1,1)}) = Q(m_{type2}^{(1)})$ for all $\nu \geq 2$. (b) (i) By substituting $\nu = 2$ for Eqs. (3) and (4), it is easy to confirm $E(m_{type1}^{(1,1)}) > E(m_{type2}^{(1)})$ for all $k \geq 3$. (ii) When $\nu$ becomes $\nu + 1$, the amount of change $\Delta E(m_{type1}^{(1,1)}) > \Delta E(m_{type2}^{(1)})$ for all $k \geq 3$. Thus, by using Lemma (a), $Q(m_{type1}^{(1,1)}) > Q(m_{type2}^{(1)})$ for all $k \geq 3$ and $\nu \geq 2$ follows from the principle of mathematical induction. □

**Proposition 2** *(Between $m_{type1}^{(1,1)}$ and $m_{type3}^{(1)}$)*
(a) When $k \geq 2$ and $\nu = 2$, $Q(m_{type1}^{(1,1)}) < Q(m_{type3}^{(1)})$.
(b) When $k \geq 2$ and $\nu = 3$, $Q(m_{type1}^{(1,1)}) = Q(m_{type3}^{(1)})$.
(c) When $k \geq 2$ and $\nu \geq 4$, $Q(m_{type1}^{(1,1)}) > Q(m_{type3}^{(1)})$.

**Proof:** (a) By substituting $\nu = 2$ for Eqs. (3) and (5), it is easy to confirm $Q(m_{type1}^{(1,1)}) < Q(m_{type3}^{(1)})$ for all $k \geq 2$. (b) By substituting $\nu = 3$ for Eqs. (3) and (5), it is easy to confirm $Q(m_{type1}^{(1,1)}) = Q(m_{type3}^{(1)})$ for all $k \geq 2$. (c) $E(m_{type3}^{(1)}) \neq 0$ allows us to consider ratio $E(m_{type1}^{(1,1)})/E(m_{type3}^{(1)})$. It is easy to confirm that the ratio is greater than 1 for all $k \geq 2,\ \nu \geq 4$. Thus, the statement follows from Lemma (b). □

**Proposition 3** *(Between $m_{type2}^{(1)}$ and $m_{type3}^{(1)}$)*
(a) When $k \geq 2$ and $\nu = 2$, $Q(m_{type2}^{(1)}) < Q(m_{type3}^{(1)})$.
(b) For all $k \geq 2$, there exists $\nu \geq 2$ such that $Q(m_{type2}^{(1)}) > Q(m_{type3}^{(1)})$.
(c) For all $k \geq 2$ and $\nu \geq 2$, $Q(m_{type2}^{(1)}) - Q(m_{type3}^{(1)})$ monotonically increases in $\nu$.

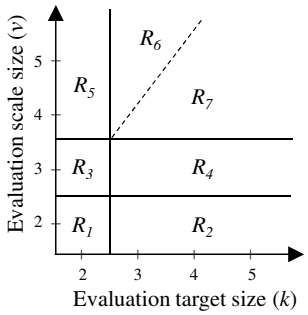
**Proof:** (a) By substituting $\nu = 2$ for Eqs. (4) and (5), it is easy to confirm $Q(m_{type2}^{(1)}) < Q(m_{type3}^{(1)})$ for all $k \geq 2$. (b) $E(m_{type3}^{(1)}) \neq 0$ allows us to consider ratio $E(m_{type2}^{(1)})/E(m_{type3}^{(1)})$ ($\equiv I_\nu$). We can write ratio $I_\nu$ as $(\frac{\nu-1}{\nu})^{k-1} + (\frac{\nu-2}{\nu})^{k-1} + \cdots + (\frac{1}{\nu})^{k-1}$. It is enough to show that the sum of the first two terms of the ratio is larger than 1 because none of the terms of the ratio take negative values. The ratio's first two terms, which are $(\frac{\nu-1}{\nu})^{k-1}$ and $(\frac{\nu-2}{\nu})^{k-1}$, both monotonically increase in $\nu$ and become 1 when $\nu \to \infty$ (They do not become 1 because $\nu$ is finite, but tend to 1 monotonically as g increases without limit.). This holds for all $k \geq 2$. Therefore, by taking a sufficiently large $\nu$, we can find $\nu$ such that $E(m_{type2}^{(1)})/E(m_{type3}^{(1)}) > 1$ for all $k \geq 2$. Thus, the statement follows from Lemma (b). (c) In the same way as (b), $I_\nu$ is used. When $\nu$ becomes $\nu + 1$, the ratio ($I_{\nu+1}$) becomes $(\frac{\nu}{\nu+1})^{k-1} + (\frac{\nu-1}{\nu+1})^{k-1} + \cdots + (\frac{1}{\nu+1})^{k-1}$. The difference $I_{\nu+1} - I_\nu$ is larger than 0 for all $k \geq 2$ and $\nu \geq 2$ because $I_{\nu+1} - I_\nu = \sum_{i=1}^{\nu} \{\frac{(\nu+1-i)^{k-1}}{(\nu+1)^{k-1}} - \frac{(\nu-i)^{k-1}}{\nu^{k-1}}\}$, where $\frac{(\nu+1-i)^{k-1}}{(\nu+1)^{k-1}} > \frac{(\nu-i)^{k-1}}{\nu^{k-1}}$ for all $k \geq 2$ and $\nu \geq 2$. Thus, $Q_{\nu+1}(m_{type2}^{(1)}) - Q_{\nu+1}(m_{type3}^{(1)})$ is larger than $Q_\nu(m_{type2}^{(1)}) - Q_\nu(m_{type3}^{(1)})$ for all $k \geq 2$ and $\nu \geq 2$ follows from Lemma (b). □

### C. Q-magnitude Relation Map (Q-Map)

The three propositions construct a Q-Map for $m_{type1}^{(1,1)}$, $m_{type2}^{(1)}$, and $m_{type3}^{(1)}$. Fig. 5 shows the Q-Map, which consists of seven disjoint regions: $R_1, R_2, \ldots, R_7$.

Regions $R_1, R_2, \ldots, R_5$ are determined by Propositions 1 and 2. For example, Propositions 1(a) and 2(a) specify the magnitude relation of the inference quantum for $R_1$ ($k = 2, \nu = 2$) as $Q(m_{type1}^{(1,1)}) = Q(m_{type2}^{(1)}) < Q(m_{type3}^{(1)})$. For another example, Propositions 1(b) and 2(b) specify the magnitude relation of the inference quantum for $R_4$ ($k \geq 3, \nu = 3$) as $Q(m_{type2}^{(1)}) < Q(m_{type1}^{(1,1)}) = Q(m_{type3}^{(1)})$.

On the other hand, regions $R_6$ and $R_7$ are developed with Proposition 3. In summary, Proposition 3 describes the reversal phenomenon of the magnitude relation between $Q(m_{type2}^{(1)})$ and $Q(m_{type3}^{(1)})$. Specifically, Propositions 3(a) and (b) state that $Q(m_{type2}^{(1)})$ is smaller than $Q(m_{type3}^{(1)})$ when $\nu$ is small ($\nu = 2$), but there exists $\nu$ such that $Q(m_{type2}^{(1)})$ is larger than $Q(m_{type3}^{(1)})$ when $\nu$ becomes large. In addition, Proposition 3(c) states that if once $Q(m_{type2}^{(1)})$ becomes larger than $Q(m_{type3}^{(1)})$ by increasing $\nu$, then $Q(m_{type2}^{(1)})$ never becomes smaller than $Q(m_{type3}^{(1)})$ by additional increases of $\nu$. This allows us to




$R_1$ : $Q(m_{type1}^{(1,1)}) = Q(m_{type2}^{(1)}) < Q(m_{type3}^{(1)})$
$R_2$ : $Q(m_{type2}^{(1)}) < Q(m_{type1}^{(1,1)}) < Q(m_{type3}^{(1)})$
$R_3$ : $Q(m_{type1}^{(1,1)}) = Q(m_{type2}^{(1)}) = Q(m_{type3}^{(1)})$
$R_4$ : $Q(m_{type2}^{(1)}) < Q(m_{type1}^{(1,1)}) = Q(m_{type3}^{(1)})$
$R_5$ : $Q(m_{type1}^{(1,1)}) = Q(m_{type2}^{(1)}) > Q(m_{type3}^{(1)})$
$R_6$ : $Q(m_{type1}^{(1,1)}) > Q(m_{type2}^{(1)}) > Q(m_{type3}^{(1)})$
$R_7$ : $Q(m_{type1}^{(1,1)}) > Q(m_{type3}^{(1)}) > Q(m_{type2}^{(1)})$

Figure 5. Q-magnitude relation map (Q-Map).

divide the region where $k \geq 3$ and $\nu \geq 4$ into two: $R_6$ and $R_7$. Note that the boundary between $R_6$ and $R_7$ (dash-line in the figure) may have another region where $Q(m^{(1)}_{type2})$ equals $Q(m^{(1)}_{type3})$. This indeterminacy disappears if it can be proven that there is no integer $k \geq 2$ and $\nu \geq 2$, except for $k = 2$ and $\nu = 3$, such that $Q(m^{(1)}_{type2}) = Q(m^{(1)}_{type3})$. At this time, it is only confirmed by computer simulation techniques that the condition holds for all $2 \leq k \leq 100$ and $2 \leq \nu \leq 100$.

Thus, the magnitude relations of the inference quanta of the three messages consist of seven patterns, each of which is determined by the region in the evaluation parameter space.

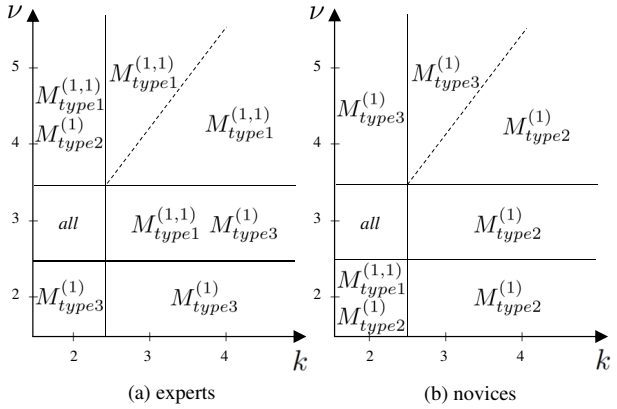



Figure 6. Priority message-class map (P-Map).

### D. Priority Message-class Map (P-Map)

The P-Map for $M^{(1,1)}_{type1}$, $M^{(1)}_{type2}$, and $M^{(1)}_{type3}$ is obtained by applying the prediction rule to the Q-Map with respect to positive attitude changes in a target ($\in \Omega$). Only messages that give positive support to the target are considered when the prediction rule is applied. Figs. 6(a) and (b) show the P-Map for experts and novices, each of which consists of seven disjoint regions, the same as the Q-Map.

The message-classes indicated in each region of the P-Map are the expected classes with the largest potency with respect to the prediction rule. For example, region $k = \nu = 2$ of the P-Map for experts indicates $M^{(1)}_{type3}$. This means, for experts, the potency of $m^{(1)}_{type3}$ exceeds that of $m^{(1,1)}_{type1}$ and of $m^{(1)}_{type2}$. In the same way, the regions indicating two message-classes mean that their potency is the same and larger than the potency of messages in the other class. The regions indicating "*all*," where $k = 2$ and $\nu = 3$ for experts and novices, mean that the potency of messages in the three message-classes is the same.

P-Maps guarantee that no message has larger potency than messages in the message-classes indicated in the region. Thus, it is a rational filtering strategy that gives priority to provide the messages belonging to the message-classes shown in the P-Map's region to which the evaluation parameter belongs. For example, for experts on $k \geq 3$ and $\nu = 2$, message $m^{(1)}_{type3}$ is given priority to provide, in contrast, for experts on $k = 2$ and $\nu \geq 4$, messages $m^{(1,1)}_{type1}$ and $m^{(1)}_{type2}$ are given priority. In the same way, for novices on $k \geq 3$ and $\nu = 2$, message $m^{(1)}_{type2}$ is given priority to provide, in contrast, for novices on $k = 2$ and $\nu \geq 4$, message $m^{(1)}_{type3}$ is given priority.

As illustrated here, message filtering strategies based on P-Maps postulate the values of evaluation parameters, represented by $k$ and $\nu$, where the users evaluate products with eWOM messages. It may be difficult to know previously the values because they depend not only on the type and the number of products evaluated by the user but also on the rating scale used by the user. Fortunately, the P-Map shown in Fig. 6 suggests that there are cases in which the exact values of the evaluation parameters don't have to be known. For example, for experts, not depending on $k$, message $m^{(1,1)}_{type1}$ always belongs to the priority message-classes when $\nu \geq 3$, and message $m^{(1)}_{type3}$ always belongs to the priority message-classes when $\nu \leq 3$. This suggests that we do not have to know $k$ when $\nu$ can be estimated. Such analytical investigations will contribute to reduce the preciseness requirements for $k$ and $\nu$ estimation.

## VI. Conclusion and Future Work

This paper presented a computational model that predicts the potency-magnitude relations of eWOM messages involving subjective rank expressions. This paper defined three message-classes and investigated mathematically how the potency-magnitude relations change based on the values of two evaluation parameters: evaluation target size $k$ and evaluation scale size $\nu$.

The mathematical investigations developed a Q-magnitude Relation Map (Q-Map) and a Priority Message-class Map (P-Map), which are exploited to design eWOM message filtering strategies. Message filtering strategies based on P-Maps postulate the values of evaluation parameters. This paper discussed that some analytical investigations reduce the preciseness requirements for the parameter estimations (Section V. D.). Future work includes further investigations and finding some observable factors for the estimation.

The observable factors for evaluation target size $k$ may be related to the products evaluated by the user. A complex product with various specifications (e.g., digital cameras) will provide larger $k$ than a simple product (e.g., a PC mouse). In addition, the increase of the number of product alternatives to be chosen will increase $k$. According to these clues, the value of $k$ may be estimated roughly. In a practical setting, a key piece of information that enables the estimation is the content in the Web-pages and their number that the user consults for product comparison. For evaluation scale size $\nu$, on the other hand, it may be possible to learn the relationship between the value of $\nu$ and personal characteristics like eWOM involvements and product expertise by doing examinations presented in [36], which is also discussed in Section III. B., on a large scale.

The P-Maps developed in Section V can be regarded as unexplored sub-hypotheses derived from the generalized hypothesis described in Section II. C. Therefore, future work must determine whether the potency-magnitude relations change as the maps predict. The observation of the reversal phenomenon is particularly important; whether the potency-magnitude relation of the two messages is reversed when scale size $\nu$ becomes larger.

To obtain accurate potency predictions, not only the inference quantum proposed in this paper but also many other factors of eWOM messages, such as those discussed in [39]–[41], have to be considered. The combination of factors will determine the potency of the eWOM messages, so these factors should be used properly for practical prediction methods.

## REFERENCES


[1] J. Lee and J.-N. Lee, "Understanding the product information inference process in electronic word-of-mouth: An objectivity-subjectivity dichotomy perspective," *Information and Management*, vol. 46, no. 5, pp. 302–311, 2009.

[2] Y. Chen and J. Xie, "Online consumer review: Word-of-mouth as a new element of marketing communication mix," *Management Science*, vol. 54, no. 3, pp. 477–491, 2008.

[3] D. H. Park and S. Kim, "The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews," *Electronic Commerce Research and Applications*, vol. 7, no. 4, pp. 399–410, 2008.

[4] K. Fujimoto, "An investigation of potency of ewom messages with a focus on subjective rank expressions," in *Proceedings of the International Workshop on Intelligent Web Interaction (WI-IAT10 Workshop)*, 2010, pp. 97–101.

[5] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[6] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, Eds. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2010.

[7] K. Fujimoto, "A computational account of potency differences in ewom messages involving subjective rank expressions," in *Proceedings of the International Workshop on Intelligent Web Interaction (WI-IAT11 Workshop)*, 2011, pp. 138–142.

[8] R. E. Petty and J. T. Cacioppo, *Attitudes and Persuasion: Classic and Contemporary Approaches*. Westview Press, 1996.

[9] A. H. Eagly and S. Chaiken, "Attitude structure and function," in *The Handbook of Social Psychology, 4th ed.*, D. T. Gilbert, S. T. Fiske, and G. Lindzey, Eds. McGraw-Hill, 1998, vol. 1, pp. 269–322.

[10] J. Arndt, *Word of Mouth Advertising: A Review of the Literature*. Advertising Research Foundation, Inc., 1967.

[11] T. M. Y. Lin and C.-W. Liao, "Knowledge dissemination of word-of-mouth research: Citation analysis and social network analysis," *Libri*, vol. 58, no. 4, pp. 212–223, 2008.

[12] E. Dichter, "How word-of-mouth advertising works," *Harvard Business Review*, vol. 44, no. 6, pp. 147–166, 1966.

[13] D. Sundaram, K. Mitra, and C. Webster, "Word of mouth communications: A motivational analysis," *Advances in Consumer Research*, vol. 25, no. 4, pp. 527–531, 1998.

[14] W. R. Wilson and R. A. Peterson, "Some limits on the potency of word-of-mouth information," *Advances in Consumer Research*, vol. 16, no. 1, pp. 23–29, 1989.

[15] P. M. Herr, F. R. Kardes, and J. Kim, "Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective," *Journal of Consumer Research*, vol. 17, no. 4, pp. 454–462, 1991.

[16] D. Godes and D. Mayzlin, "Using online conversations to study word-of-mouth communication," *Marketing Science*, vol. 23, no. 4, pp. 545–560, 2004.

[17] T. Hennig-Thurau, K. P. Gwinner, G. Walsh, and D. D. Gremler, "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet?" *Journal of Interactive Marketing*, vol. 18, no. 1, pp. 38–52, 2004.

[18] J. W. Alba and J. W. Hutchinson, "Dimensions of Consumer Expertise," *Journal of Consumer Research*, vol. 13, no. 4, pp. 411–454, 1987.

[19] D. Maheswaran and B. Sternthal, "The effects of knowledge, motivation, and type of message on ad processing and product judgments," *Journal of Consumer Research*, vol. 17, no. 1, pp. 66–73, 1990.

[20] X. Cheng and M. Zhou, "Study on effect of ewom: A literature review and suggestions for future research," in *Proceedings of the 4th International Conference on Management and Service Science (MASS–10)*, 2010, pp. 1–4.

[21] C. M. Cheung and M. K. Lee, "Online consumer reviews: Does negative electronic word-of-mouth hurt more?" in *Proceedings of the 14th Americas Conference on Information Systems (AMCIS–08)*, 2008, pp. 1–4.

[22] C. Park and T. Lee, "Information direction, website reputation and eWOM effect: A moderating role of product type," *Journal of Business Research*, vol. 62, no. 1, pp. 61–67, 2009.

[23] D. J. O'Keefe, *Persuasion: Theory and Research (2nd edition)*. Sage Publications, Inc, 2002.

[24] A. G. Sawyer, "Can there be effective advertising without explicit conclusions? decide for yourself," in *Nonverbal communication in advertising*, S. Heckler and D. W. Steward, Eds. Lexington, MA: D. C. Heath, 1988, pp. 159–184.

[25] A. G. Sawyer and D. J. Howard, "Effects of Omitting Conclusions in Advertisements to Involved and Uninvolved Audiences," *Journal of Marketing Research*, vol. 28, no. 4, pp. 467–474, 1991.

[26] F. R. Kardes, "Spontaneous Inference Processes in Advertising: The Effects of Conclusion Omission and Involvement on Persuasion," *Journal of Consumer Research*, vol. 15, no. 2, pp. 225–233, 1988.

[27] P. E. Ketelaar, M. S. van Gisbergen, J. A. Bosman, and H. Beentj, "Attention for open and closed advertisements," *Jornal of Current issues and Research in Advertising*, vol. 30, no. 2, pp. 15–25, 2008.

[28] K. Gunasti and W. Ross Jr, "How Inferences about Missing Attributes Decrease the Tendency to Defer Choice and Increase Purchase Probability," *Journal of Consumer Research*, vol. 35, no. 5, pp. 823–837, 2009.

[29] J.-C. Chebat, M. Charlebois, and C. Geinas-Chebat, "What makes open vs. closed conclusion advertisements more persuasive? The moderating role of prior knowledge and involvement," *Journal of Business Research*, vol. 53, no. 2, pp. 93–102, Aug. 2001.

[30] N. Jindal and B. Liu, "Mining comparative sentences and relations," in *proceedings of the 21st national conference on Artificial intelligence (AAAI–06)*, 2006, pp. 1331–1336.

[31] S. Yang and Y. Ko, "Extracting comparative entities and predicates from texts using comparative type classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 1636–1644.

[32] M. Ganapathibhotla and B. Liu, "Mining opinions in comparative sentences," in *Proceedings of the 22nd*

*International Conference on Computational Linguistics (COLING–08)*, 2008, pp. 241–248.

[33] K. Xu, S. S. Liao, J. Li, and Y. Song, "Mining comparative opinions from customer reviews for Competitive Intelligence," *Decision Support Systems*, vol. 50, no. 4, pp. 743–754, 2011.

[34] K. Dave, S. Lawrence, and D. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web (WWW–03)*, 2003, pp. 519 – 528.

[35] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT–05)*, 2005, pp. 347–354.

[36] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 115–124.

[37] M. J. Druzdzel and L. C. van der Gaag, "Elicitation of probabilities for belief networks: Combining qualitative and quantitative information," in *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, 1995, pp. 141–148.

[38] K. Fujimoto, K. Matsuzawa, and H. Kazawa, "An elicitation principle of subject probabilities from statements on the Internet," in *Proceedings of the 3rd International Conference on Knowledge-Based Intelligent Information Engineering Systems (KES–99)*, 1999, pp. 459 – 463.

[39] P. Nelson, "Advertising as Information," *Journal of Political Economy*, vol. 82, no. 4, pp. 729–754, 1974.

[40] R. East, K. Hammond, and W. Lomax, "Measuring the impact of positive and negative word of mouth on brand purchase probability," *International Journal of Research in Marketing*, vol. 25, no. 3, pp. 215–224, 2008.

[41] M. Y. Cheung, C. Luo, C. L. Sia, and H. Chen, "Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations," *International Journal of Electronic Commerce*, vol. 13, no. 4, pp. 9–38, 2009.

**Kazunori Fujimoto** received his Ph.D. degree in informatics from Kyoto University, Japan in 2004, his Master's degree in electrical engineering from Kyoto University in 1992, and his Bachelor's degree in electrical engineering from Doshisha University, Japan in 1989.

From 1992 to 2001 he was with NTT Electrical Communications Laboratories, Tokyo, Japan, and engaged in research on Artificial Intelligence. From 2001 to 2008 he was the President of Fujimoto Research Park Co., Ltd. From 2006 to 2008 he was a Visiting Fellow at Doshisha University Institute for Technology, Enterprise and Competitiveness, Kyoto, Japan. In 2008 he joined the Faculty of Business Administration, Kinki University, Osaka, Japan as an Associate Professor. He is a member of the Program Committee of the International Workshop on Intelligent Web Interaction from 2011. He is a councilor of Japanese Society for Artificial Intelligence from 2010. His current research interests include electronic word-of-mouth, attitude change, decision support, recommender systems, and persuasive technologies.

Dr. Fujimoto is a member of IEEE (Institute of Electrical and Electronics Engineers) and of ACM (Association for Computing Machinery).

# Balancing the Trade-Offs Between Diversity and Precision for Web Image Search Using Concept-Based Query Expansion

Enamul Hoque, Orland Hoeber, and Minglun Gong
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada A1B 3X5
Email: {enamulp, hoeber, gong}@mun.ca

*Abstract*— **Even though Web image search queries are often ambiguous, traditional search engines retrieve and present results solely based on relevance ranking, where only the most common and popular interpretations of the query are considered. Rather than assuming that all users are interested in the most common meaning of the query, a more sensible approach may be to produce a diversified set of images that cover the various aspects of the query, under the expectation that at least one of these interpretations will match the searcher's needs. However, such a promotion of diversity in the search results has the side-effect of decreasing the precision of the most common sense. In this paper, we evaluate these competing factors in the context of a method for explicitly diversifying image search results via concept-based query expansion using Wikipedia. Experiments with controlling the degree of diversification illustrate this trade-off between diversity and precision for both ambiguous and more specific queries. As a result of these experiments, an automatic method for tuning the diversification parameter is proposed based on the degree of ambiguity of the original query.**

## I. Introduction

The primary method for performing image retrieval on the Web is based on document search techniques [1]. Images are indexed based on the text that is related to their use on the Web (keywords, tags, and/or associated descriptions). User-supplied queries are matched to this text to produce a set of images, which are ranked based on the relevance of their associated textual information to the query. This approach can work well when the contents of the images are concisely and accurately described within their source Web pages, and when the searchers' needs are clearly specified.

Recent studies on user behaviour with respect to image search have found that queries are often very short and ambiguous [2]. This ambiguity comes from the difficulty that searchers experience in finding the words to describe an idea or image they have in their mind. From the perspective of image retrieval, the difficulty with ambiguous queries is that they can be open to many different interpretations. It is possible that different searchers may enter the same query, but their intentions and needs may vary significantly from one another. In situations such as this, the matching algorithms used by image search engines promote the interpretation that is most common and popular. However, a more sensible approach may be to produce a diversified set of images that cover the various aspects of the query, under the expectation that at least one of the interpretations matches the searcher's intent. Providing searchers with an overview of the images and allowing them to zoom-in and focus on a particular interpretation [3] may improve the effectiveness and efficiency of the image search process.

Moving beyond traditional relevance ranking, diversification approaches aim to improve the coverage of the search results set with respect to the different senses of the original query. A common method for diversification is query expansion, whereby additional terms are added to the query to generate a collection of new queries that, when taken together, are more broad than the original [4]. However, in doing so there is a danger in broadening the query too much, resulting in a potentially significant decrease in precision. That is, the more broad and diverse the search results are, the less chance that a particular search result will be relevant to the searcher's information need. As such, this trade-off between diversity and precision must be studied in order to understand the situations where more or less diversification is beneficial.

Maintaining a balance between diversity and precision requires an automatic modelling of the searcher's query to determine an appropriate degree of diversification to promote. In many cases, image search queries are inherently ambiguous. For example, "Washington" might be interpreted as "Washington (state)", "Washington D.C.", or "George Washington". Even within a more specific query, searchers might have an interest in seeing images that are related to but not explicitly identified in the query. For example, if a searcher submits a query such as "Hong Kong", they may wish to see some representative images of different landmarks in the Hong Kong area. In other cases, a query might be very specific, and the scope for

This paper is an extended version of "Evaluating the Trade-Offs Between Diversity and Precision for Web Image Search Using Concept-Based Query Expansion," by E. Hoque, O. Hoeber, and M. Gong, which appeared in the Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence Workshops (International Workshop on Intelligent Web Interaction), Lyon, France, August 2011.


doi:10.4304/jetwi.4.1.26-34

broadening the query is limited. For example, queries for a particular landmark within a specific setting like "Eiffel Tower Bastille Day" may leave little room for diversifying the search results.

In previous research, we presented a method for diversifying image search results using concept-based query expansion, which is based on information derived from Wikipedia [3]. In this work, we have modified our approach such that the degree of diversification can be controlled, allowing us to evaluate the trade-offs between diversification and precision among the image search results. Using this knowledge, we have developed a simple method for automatically determining the degree of diversification based on the level of ambiguity of the query, resulting in a balance between diversification and precision.

The remainder of this paper is organized as follows. Section II provides an overview of the existing methods for search results diversification. Section III outlines our method for diversifying image search results, with a specific focus on a parameter that we use to control the degree of diversification. Section IV describes an evaluation that explores how the diversification parameter can be used to find a balance between diversity and precision within the search results, and proposes a formula for automatically determining this parameter. Section V provides a discussion on the outcomes of this study and the implications for diversifying image search results. The paper concludes with a summary of the research contributions in Section VI.

## II. Related Work

The problem of enhancing diversity in search results has been recognized as an important topic in the research community. Diversification can be stated as an optimization problem, where the objective is to maximize the probability of finding a relevant document/image among the top selected results, while minimizing its redundancy with respect to different aspects of the query. The general problem is an instance of maximum coverage problem, which is NP-hard [5].

In the literature of document search techniques, most of the previous approaches to search results diversification can be categorized as either implicit or explicit [6]. Implicit search results diversification approaches perform direct comparison between the retrieved documents, under the assumption that similar documents will cover similar aspects. In an early work, Carbonell et al. [7] proposed a greedy approximation based ranking method called Maximal Marginal Relevance (MMR), which attempted to maximize the relevance of a search result while minimizing its similarity to higher ranked documents. In another implementation of an MMR-based method, Zhai et al. [8] proposed a subtopic search method using statistical language models, which aimed to return results that cover more subtopics.

An alternative approach to diversification is to explicitly utilize different aspects associated with a query by directly modelling these aspects. For instance, Agrawal et al. [5] investigated the diversification problem based on the assumption that a taxonomy is available for both queries and documents. In their work, documents retrieved for a query were considered similar if they were confidently classified into one or more common categories covered by the query. They applied a greedy algorithm that started with an empty list of results and selected the next result with the highest marginal utility until $k$ results were selected. Here, the marginal utility measured the probability that the result satisfied a category the current result set did not yet cover. In a related approach, Radlinski and Dumais [9] proposed to filter the results retrieved for a given query so as to limit the number of those satisfying the same aspect of the query, represented as different query reformulations obtained from a large query log from a commercial search engine. Recently, Santos et al. [10] introduced a probabilistic framework for search results diversification, which explicitly represented query aspects as "sub-queries". They defined diversity based on the estimated relevance of documents to multiple sub-queries and the relative importance of each sub-query.

The diversity problem is more challenging in Web image search for a number of reasons. First of all, image search may involve not only semantic diversity but also visual diversity of the search results [11]–[14]. Although there may be some benefit to promoting both types of diversity, the focus of our research is on the semantic side. Secondly, in the document centric retrieval methods, the documents themselves provide some useful ways to allow direct comparisons and similarity calculations when promoting diversification in the search results. But in image retrieval, the limited amount of textual metadata associated with the retrieved images are not reliable enough nor sufficient to allow for the computation of similarities between their associated images.

A somewhat common approach to dealing with the diversification problem for image search results is to use semantic and/or visual clustering [11]–[13], [15], [16]. Although these approaches differ in the information used and the methods employed for clustering, the end result is a grouping of images that represent the different aspects of the image collection. There are, however, a number of challenges associated with clustering approaches, including determining a suitable similarity measure, the efficiency of clustering algorithms, determining an appropriate number of clusters to use, and deciding how to order or organize the clusters.

Re-ranking methods have also been proposed with an aim to enhance topic coverage in search results. Song et al. [17] addressed the diversity problem using a re-ranking method based on topic richness analysis. The goal here is to enhance topic coverage in the image search results, while maintaining acceptable retrieval performance. A topic richness score was computed by analyzing the degree of mutual topic coverage between an image pair based on the assumption that images are annotated by several words.

Some have also begun to study the trade-off between precision and diversification within the context of image search. For example, van Zwol et al. proposed a method for optimizing this trade-off by estimating a query model from the distribution of tags that favour the dominant sense of an image search query [18].

A promising direction for enhancing the diversity of image search results is the use of query expansion methods based on concepts that are related to the query. Unlike the general domain of document-centric information retrieval, query expansion has been studied in only a few research works in Web image retrieval. Those that have explored such techniques have shown them to be promising [3], [19]. However, one of the challenges associated with query expansion is to find an appropriate source of knowledge required for the expansion process. It has been noted that many image search queries are associated with conceptual domains that include proper nouns (i.e., people's names and locations) [2], [20]. As such, finding a suitable knowledge base that has sufficient coverage of a realistic conceptual domain is very important first step in this approach to query expansion. Wikipedia is a good candidate for such a knowledge base since it includes a large number of articles describing people, places, landmarks, animals, and plants. The challenge in using Wikipedia is to design efficient and effective algorithms that can process the semi-structured knowledge to derive meaningful terms for use in the query expansion process.

## III. Image Search Results Diversification

For the diversification framework used in this research, image search results are explicitly diversified based on concept-based query expansion. For the short and ambiguous queries that are common in image search, query expansion attempts to capture the various aspects of the query. We model the original query by discovering different possible senses, and for each sense a number of concepts pertaining to the query are discovered from within Wikipedia. These concepts are ranked according to the semantic relatedness to the original query, and only the top-$N$ most related concepts are used within the query expansion process to retrieve a range of images that provides a broad view of what is available.

Within this process, the value of $N$ is an explicit indicator of the degree of diversification, which we call the diversification parameter. With a smaller value of $N$, fewer concepts will be used, and the search results will remain more focused. If we increase $N$, then more concepts will be used for query expansion, and in turn the search results will be more diversified covering more aspects associated with the query. The fundamental trade-off between diversification and precision is based on the fact that as we increase $N$, there is a higher chance that a concept will be selected for the query refinement process that is not relevant to the searcher's information needs, resulting in the associated irrelevant images being included in the search results. As such, our goal is to fulfill the diversification objective by setting $N$ sufficiently high to capture the broad range of concepts associated with the query, but not so high as to have a significant adverse affect on the average precision across all of the senses of the query.

In this precision-diversity trade-off context, we can state our diversification objective as follows: *Given a query $Q$, perform query expansion based on $N$ concepts to retrieve a results set $R$, so that it will maximize the value of $N$, and at the same time maximize the precision $P$ with respect to each of the possible senses of the query over the results set $R$.* Note that an inherent feature of such diversification is that it will increase the precision of the search results with respect to the less common senses of the query, at the expense of the most common sense.

To achieve this objective in our diversification method, the possible senses of the query, the number of concepts to be selected for each sense, and the number of images to be retrieved for each concept are determined automatically based on an analysis of the original query and the candidate concepts extracted from Wikipedia. The purpose here is to distribute the concepts and the number of images retrieved for each concept unevenly, so that we can alleviate the problem of harming precision due to the images retrieved for irrelevant concepts that are only loosely related to the senses of the original query.

The process of diversifying the image search results using concept-based query expansion follows three steps: extracting concepts from Wikipedia; ranking the extracted concepts to select the top-$N$ related concepts; and retrieving images based on the expanded queries derived from these concepts. The details for each of these steps are explained in the remainder of this section.

### A. Concept Extraction Using Wikipedia

In order to use Wikipedia as the source knowledge base in this work, a dump of the Wikipedia collection was obtained and preprocessed to support the type of knowledge extraction required for our purposes. Matching a user-supplied query $Q$ to this knowledge base is simply a matter of selecting the best matching article (referred as the home article) using Wikipedia's search feature. In the case where the query is ambiguous and Wikipedia suggests multiple senses of interpretations, the ones with higher commonness values are used as the home articles. Here the commonness value of an article is calculated based on how often it is linked by other articles.

In analyzing Wikipedia, we observed that the in-link articles (ones having links to a home article) and out-link articles (ones to which a home article links) often provide meaningful information that is closely related to the concept of the home article, and hence the user-specified query. Therefore, these linked articles are located and their titles are extracted as candidates for related concepts.

We also found that the captions surrounding the images present within a given article can often provide a valuable perspective on the visual features associated with the article's concept. Due to the importance of this information for the purposes of image retrieval, it is important to

ensure that all relevant concepts associated with the image captions are extracted. We use Wikifier [21] to augment the captions with links to relevant Wikipedia articles that may have been missed by the author of the article, and use these links to extract their associated concepts.

The end result of this process is the selection of a set of home articles $\{h_s|1 \leq s \leq q\}$ (for $q$ senses of given query $Q$), along with a list of all the candidate articles $C_{h_s}$ for each home article $h_s$ that originate from the in-link articles, out-link articles, and the image captions. These concepts provide the basis for the automatic query expansion process.

### B. Ranking the Extracted Concepts

Due to the rich and interconnected nature of Wikipedia, the number of concepts obtained in the process described above may become very large. Thus, a filtering step is necessary to ensure the quality of the concepts that are extracted. Here, our objective is to select the top-$N$ concepts from among all the candidate articles. Considering the difference in the importance of each of the senses, we distribute these top-$N$ concepts among the candidate concepts $C_{h_s}$ of each home article $h_s$. As such, the number of concepts $N_{h_s}$ that are to be selected for a particular home article $h_s$ is determined as follows:

$$N_{h_s} = \frac{|C_{h_s}| \times N}{\sum_{j=1}^{q} |C_{h_j}|}$$

Note that the sum of all $N_{h_s}$ values equals $N$.

To select these $N_{h_s}$ concepts for each home article $h_s$, it is necessary to rank the candidate concepts $C_{h_s}$ based on their relevance to their associated home article. Our approach to this problem is to measure the semantic relatedness between the home article and each of the candidate concepts. A number of different methods have been devised to use Wikipedia for this purpose, including WikiRelate! [22], Explicit Semantic Analysis (ESA) [23], and Wikipedia Link-based Measure (WLM) [24]. Given the computationally efficiency and accuracy of WLM, we use this approach in our work.

For each of the candidate articles $c_i \in C_{h_s}$ extracted from the home article, WLM is applied between the home article $h_s$ and the candidate articles. WLM takes advantage of the hyperlink structure of the associated articles to find out how much they share in common. In order to give preference to the concepts that have been extracted from the image captions within the home article, we use a re-weighting function to determine the relatedness score:

$$r(c_i, h_s) = \min(WLM(c_i, h_s)(1 + \alpha_s), 1)$$

Since WLM provides a value in the [0,1] range, we ensure that the relatedness score remains in this range with the $min$ function. The re-weighting factor $\alpha_s$ is provided according to the following function:

$$\alpha_s = \begin{cases} k\frac{C_{h_s}}{N_{h_s}} & \text{if concept } c_i \text{ originates from a caption} \\ 0 & \text{otherwise} \end{cases}$$

Here, $C_{h_s}$ and $N_{h_s}$ are as defined above, and $k$ is a system parameter that controls the importance of the concepts derived from the captions. In our prototype implementation $k = 0.01$, which results in a 10 - 20% increase in the score for the concepts derived from the captions, with proportionally more importance being given when there are more concepts extracted from the home article.

The outcome of this process is that the top-$N$ concepts are selected from among the candidate articles, such that the ones that came from the image captions are given preference over those from in-link and out-link articles. These concepts are used as the source for the query expansion process. The value of $N$ serves here as an explicit diversification parameter. How it affects the diversification and precision of search results are discuss in Section IV.

### C. Concept-Based Query Expansion and Image Retrieval

In order to ensure that the expanded queries remain focused on the topic of the query itself, the top-$N$ related concepts $\{c_r|0 \leq r \leq N\}$ are prepended with their associated home article $h_r$, resulting in queries of the form $< h_r, c_r >$. We define $c_0$ to be null and $h_0$ to the the original query $Q$, producing the original query plus $N$ expanded queries.

Given that individual expanded queries have differing degrees of relevance to the original query, we dynamically determine how many images to retrieve for each expanded query based on their relatedness score to the home articles. This way we can ensure that more images are retrieved for concepts that are most similar to the sense associated with their home article, even when the original query has multiple meanings. This is done to minimize the number of images retrieved for concepts that are only loosely associated with the sense of the query.

The number of images to retrieve for each expanded query is given by the following formula:

$$I_r = \frac{r(c_r, h_s) \times I_t}{\sum_{k=0}^{N} r(c_k, h_s)}$$

Here, $r$ is the same function used to generate the relatedness score in the concept ranking process, and $I_t$ is the total number of images to be retrieved by all of the queries. We set $I_t = 60$ for the purposes of performing the evaluation within this paper, but it can be set to any reasonable number of images. Since the null expanded query ($c_0$) is the original query, we define $r(c_0, h_s) = 1$ in the above calculation. All of the queries are sent to the Google AJAX Search, and the desired number of images are retrieved. Duplicate images are deleted based on the URL of the source image (as provided by the underlying search engine).

## IV. Evaluation

The goal of this approach is to automatically diversify the images retrieved for a given query, using Wikipedia as the source for a query expansion process. However, it

is unclear to what degree such diversification should be promoted during the search process. In this evaluation of the approach, our goal is to study the inherent trade-off between precision and diversity in detail. In particular, for a set of queries, we explore how the precision changes as diversity in the image search results is promoted. Using this information, we propose a simple approach to automatically determining the degree of diversification based on features of the user-supplied query.

### A. Experimental Setup

For these experiments, we chose 12 query topics, split between those we deemed to be highly ambiguous (having four senses), moderately ambiguous (three senses), slightly ambiguous (two senses), and non-ambiguous (one sense). This distribution of different degrees of ambiguity allowed us to examine the effect of the experimental condition (i.e., the varying of the degree of diversification) in the context of ambiguity. For each of these different degrees of ambiguity we selected two queries, except for the moderate ambiguity, for which we selected six queries. The moderate degree of ambiguity was examined more carefully since it represents the most common case of ambiguity.

To evaluate the effect of diversification on precision, we retrieved the top 60 search results from Google Image Search using our concept-based query expansion method with ten different values of $N$, ranging from 0 to 40 ($N = 0, 2, 4, 6, 8, 10, 15, 20, 30, 40$). Here, $N = 0$ implies that no query expansion has occurred (i.e., the search results are not diversified, and are simply the results provided by the underlying image search engine). At the other extreme, $N = 40$ causes the system to return a highly diversified set of image search results from 40 different associated concepts chosen in the query expansion procedure. Data was collected more frequently in the low end of this range in order to more closely observe the effect of a low degree of diversification. Preliminary experiments illustrated that the effect of the degree of diversification at the higher range became rather stable [25].

For each of the different senses of the query, assessors were asked to judge the relevance of each image. This assessment of relevance provided the ground truth information in the calculation of the precision scores (the ratio of relevant images to the total number of images retrieved). Since there were ten trails (i.e., ten different values of $N$) and 60 images retrieve with each trial, this resulted in the evaluation of a total of 600 images for each test query.

### B. Results

In these experiments we measured the precision for each of the test queries as the diversification parameter $N$ was varied from 0 to 40. Our hypothesis was that as $N$ increased, the distribution of the senses would become more balanced across all of the meanings of the query. This would result in a reduction in the precision for the most common senses of the query, and an increase in the precision for the less common senses. This feature can be readily identified in the graphs in Figures 1, 2, and 3. To further understand this effect, we plotted the average precision (the red lines with the triangle markers) and the total precision (the dark red lines with the x marker) across all of the senses.

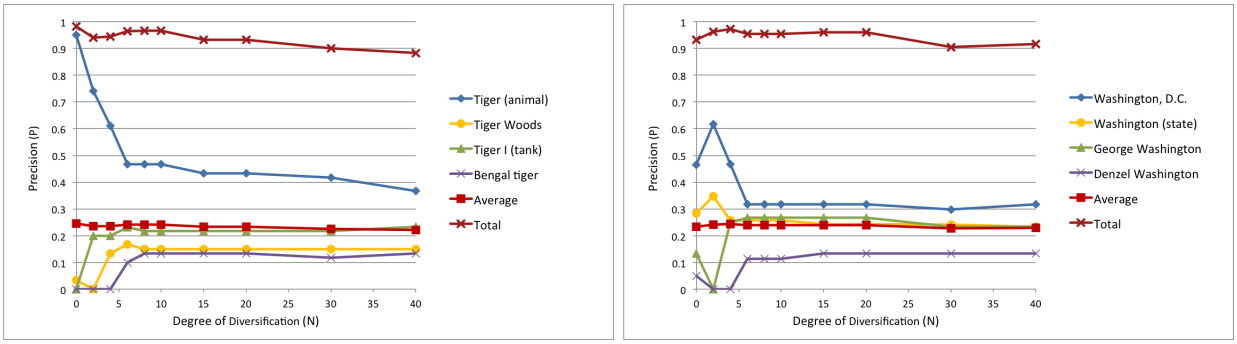
Figures 1 and 2 show the results from the highly and moderately ambiguous queries. The general effect that can be seen here is that the precision for the most common sense (i.e., the blue line in each graph) was automatically reduced as a result of the diversification. In most cases this occurred in a more or less smooth fashion even with very low values of $N$. At the same time, the precision for all other senses increased. In some cases, it was necessary for the value of $N$ to be set higher than six for images from some of the less common senses of the query to be represented in the search results.

Although there was, in some cases, a minor reduction in the total precision for very low values of $N$, this was often accompanied by a subsequent increase in the total precision as more diversification occurred. This effect is a result of the method by which the number of images retrieved for each expanded query is dynamically determined. With very few expanded queries, more images are retrieved for each (which may result in the inclusion of some less relevant images deeper in the search results list). As $N$ is increased and more expanded queries are generated, the images that are retrieved have higher rankings with respect to their source query.

For all of these queries, the precision of the images over each of the senses of the query did not change significantly once the value of $N$ was set beyond six to ten, depending on the specific query. Furthermore, if the value of $N$ was set too large (i.e., 30 or 40), the average and total precision started to decrease, indicating that some non-relevant concepts and their associated images were being included in the search results set due to over-diversification.

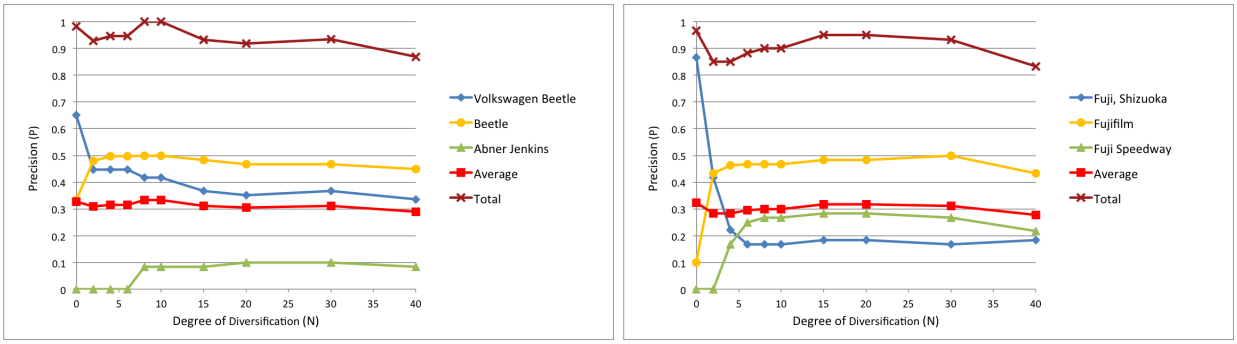
Our expectation when designing these experiments was that for queries that have a high degree of ambiguity, it would be necessary to set the diversification parameter rather high in order to capture enough information on all of the different senses. However, it is clear that even with a diversification parameter set at $N = 10$, the desired effect appears.

In some cases, the most common sense remains the most common regardless of the level of diversification (e.g., Figure 1a and b, and 2d, and e). In the other cases, senses that were less common in the original search results become dominant. The reason for this change is that there may be disagreement between what the underlying search engine assumes is the desired information need (i.e., the most common sense of the query), and the amount of information that can be extracted from Wikipedia on the other senses and their associated concepts.

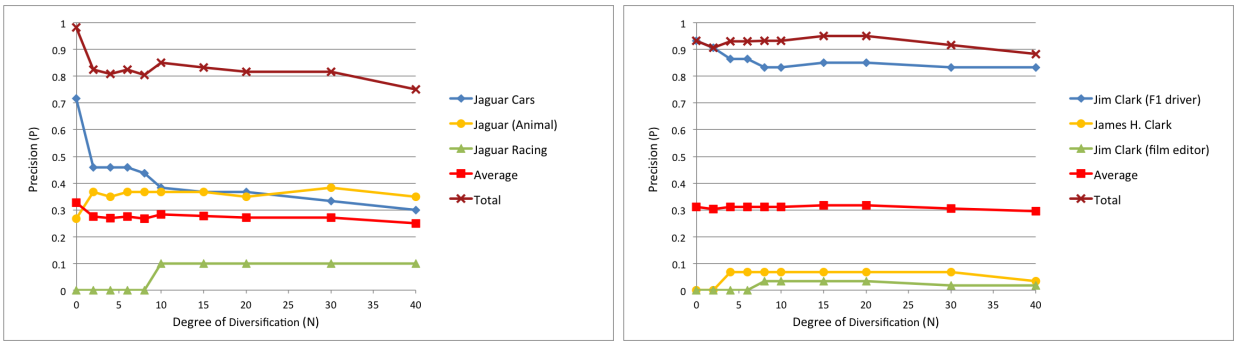



(a) Tiger (b) Washington

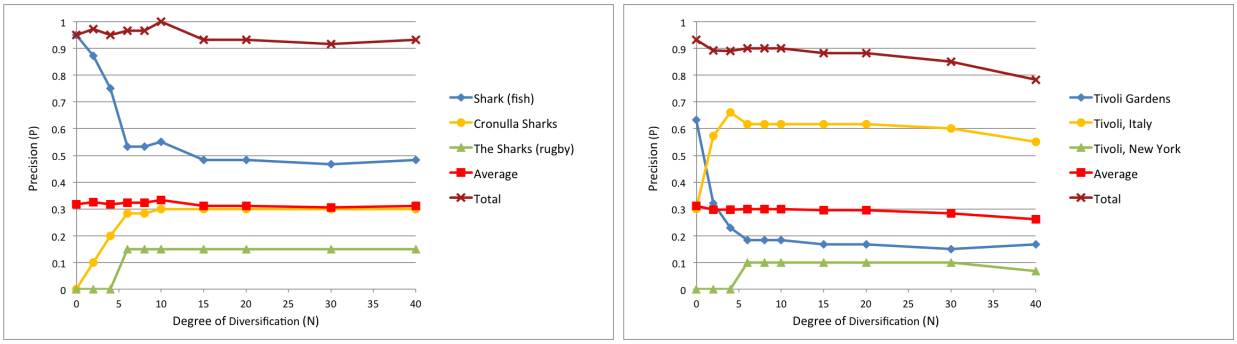
Figure 1. The effect of varying the degree of diversification (N) on precision (P) for highly ambiguous queries with *four* different senses.




(a) Beetle (b) Fuji



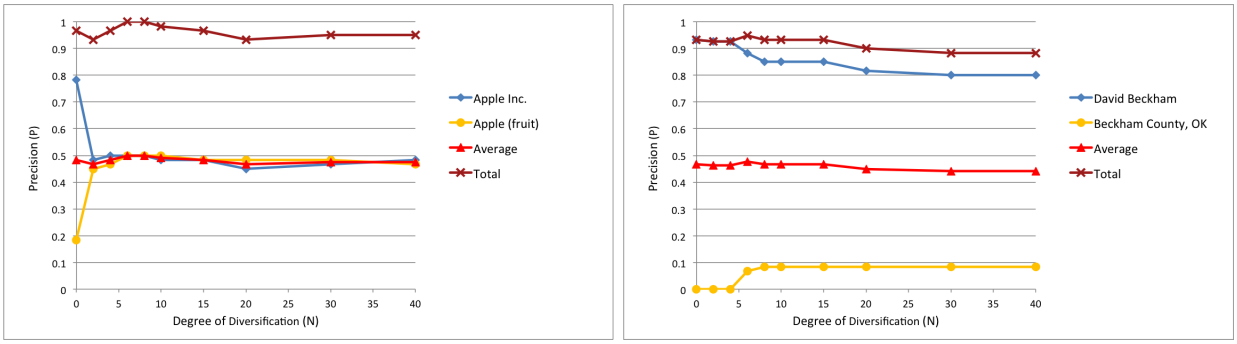

(c) Jaguar (d) Jim Clark




(e) Sharks (f) Tivoli

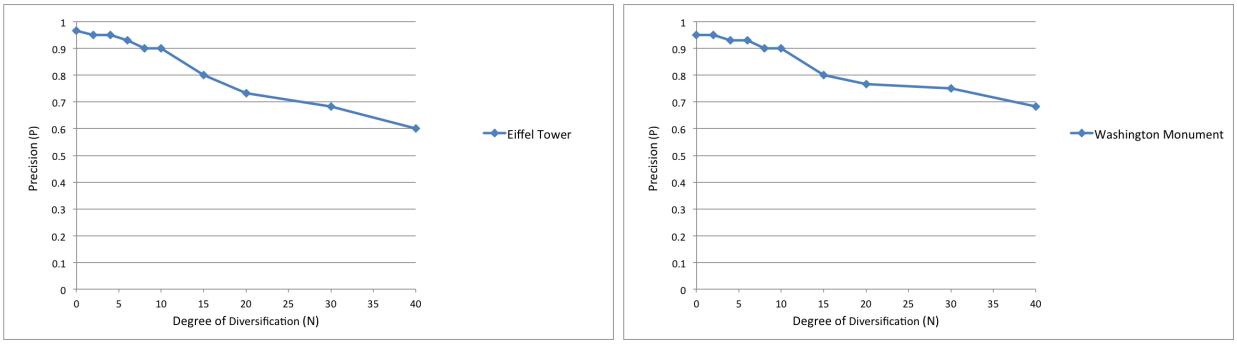
Figure 2. The effect of varying the degree of diversification (N) on precision (P) for moderately ambiguous queries with *three* different senses.

Figure 3 shows the results from the slightly ambiguous queries. As with the highly ambiguous queries, even with a low degree of diversification (from two to six), the outcome of the diversification is a balancing of the precision between the two senses of the query. In addition, the effect of a dropping then increasing total precision over low values of $N$ is present, as is the dropping average and total precision when $N$ is set too large. However, the later effect occurs with lower values of $N$ than with the highly or moderately ambiguous queries (15 - 20 for the




(a) Apple (b) Beckham

Figure 3. The effect of varying the degree of diversification (N) on precision (P) for slightly ambiguous queries with *two* different senses.




(a) Eiffel Tower (b) Washington Monument

Figure 4. The effect of varying the degree of diversification (N) on precision (P) for non-ambiguous queries with only *one* sense.

slightly ambiguous queries).

For the queries where there was only one sense (Figure 4), it is clear that diversifying the search results can very quickly have a negative effect on the precision. However, for small values of $N$ (e.g, 2-4) this effect is negligible. This indicates that even for very specific queries, a small degree of diversification might be tolerable and perhaps even beneficial as highly related images are drawn into the search results set.

As a result of this analysis, we conclude that diversifying the image search results can be very useful for addressing the situation where an ambiguous query has multiple senses. Rather than relying on the search engine to choose the most common sense, we can diversify the image search results and let the user focus on those images that match their needs. The more senses that can be inferred from a query, the more diversification is necessary to sufficiently balance all of these senses in the search results. However, when there are few different senses, the degree of diversification should be limited to avoid including irrelevant concepts and their associated images in the search results.

### C. Automatically Determining the Degree of Diversification

One of our goals in this research is to automatically determine the degree of diversification needed to enhance the searcher's ability to find relevant documents. As illustrated in the experiments, the degree of ambiguity of the query has an impact on the value of diversification. That is, a highly ambiguous query can benefit from more diversification, whereas a very specific query may be harmed by diversifying the search results too much.

Based on this, we propose a linear scaling of the degree of diversification based on the number of senses of the query (as determined by Wikipedia). One such automatic function for determining the how aggressively to diversify the search results is $N = a \times q$ where $q$ is the number of senses for query $Q$ (following the terminology from Section III). Based on the experiments reported in this paper, setting $a = 3$ will produce reasonable results. That is, if a query has one sense, the diversification parameter will be set to 3, performing a relatively low degree of diversification that does not harm the precision of the one sense of the query. For the slightly, moderately, and highly ambiguous queries, this results in the diversification parameter being set to 6, 9, and 12, respectively. By inspecting the graphs in Figures 1 - 4, we can see that this is a near-optimal value for $N$.

This simple approach to determining $N$ could benefit from further research in fine-tuning this formula and its parameters for automatically determining the value of $N$ based on some measure of query ambiguity.

## V. Discussion

Without diversification, the most common sense of an ambiguous query can dominate the image search results. Images that match other senses of the query may not

be very common or even represented at all within the top search results. This is a preferential outcome if the searcher's information needs match the underlying search engine's interpretation of the query. However, when this is not the case, it can be very difficult for the searcher to find images that match their needs. Their only recourse is to attempt to reformulate the query into something more specific. However, studies on Web image search behaviour indicate that this may be a rather difficult task for searchers to perform.

The key benefit of performing image search diversification is that instead of assuming a single interpretation of the query, we retrieve images from different senses discovered using Wikipedia, providing a more balanced distribution of the senses within the set of search results. As a result, this diversified set of search results may be suitable to a wider range of users with a wider range of information needs.

A common problem with query expansion in general, and search result diversification in particular, is that it can reduce the precision by including documents (images) that might not be relevant to all searchers. This remains the case for our work on image search diversification. In particular, we have shown how the precision for the most common sense will invariably be reduced. This is because images from the less common senses are being included in the search results as a result of the diversification process. However, even amid this reduction in the precision of the most common sense, and the increase in the precision of the less common senses, the average precision is not chaining noticeably when the degree of diversification is not set unreasonably large. On the contrary, in some cases the total precision actually increases, when calculated over all senses of the query. This indicates that the proposed approach is effective in keeping the expanded queries focused on the intended senses of the query, even when there are multiple such senses.

Because the process of diversification will commonly result in a mix of relevant and irrelevant images for a given interpretation of an ambiguous query, it is important that an interface to the image search results be used that makes it easy for the searcher to ignore the senses of the query that are not relevant to their needs, and focus on those that are. We have developed such an interface in previous research that organizes the search results according to both visual and conceptual similarities [26], allowing the searcher to filter the search results both visually through a pan-and-zoom operation [27], and based on a hierarchy of the concepts that produced the expanded queries [3]. Using this visual interface to explore the search results, the precision for a particular sense of the query may be improved as irrelevant images are moved out of the field of view or are filtered based concepts that are uninteresting to the searcher. An evaluation of this interface from the perspective of improving the precision of the search results through these interactive operations has shown not only the benefit of the interactive features, but also the benefit of organizing the images based on both visual and conceptual features [28].

It is also worth noting that the benefit that this approach to image search diversification can provide depends greatly on the completeness and interconnected nature of information in Wikipedia. If there is a sense of a given query that is not well represented in Wikipedia, then it will not be well represented within the diversified image search results set. As such, as Wikipedia continues to grow and be enhanced, it will become a better and better tool for providing knowledge for information retrieval research such as what has been discussed in this paper.

## VI. Conclusions

In this paper, we present a novel approach for explicitly diversifying image search results using concept-based query expansion. This approach is based on our previous research, but has been modified such that the degree of diversification can be controlled through the diversification parameter $N$. Using this system, we have evaluated the trade-off between diversification and precision, using a set of test queries of varying ambiguity.

From these experiments, we can see that the degree to which diversification needs to be promoted in order to provide a balance across all of the senses of a query depends on the ambiguity of the query itself. That is, a highly ambiguous query can benefit from a high degree of diversification more so than a very specific query. Based on this, we propose a simple automatic calculation of the diversification parameter based on the level of ambiguity of the query that balances the increased diversification against the decrease in precision.

Instead of only satisfying the information needs for the most common interpretation of the query, our method provides a more balanced view of the different senses of the query, without a significant negative impact on the average precision across all senses. Furthermore, our method of diversification provides an effective means for ensuring that the expanded queries that produce the diversified search results remain focused on at least one sense of the query. Coupled with a visual interface for organizing the image search results based on their visual and conceptual similarity, and interaction methods that support dynamic filtering, this approach to search results diversification can be a very powerful tool for enhancing the image search experience.

## Acknowledgments

This research was supported by Research Development Corporation (RDC) IgniteR&D and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants held by the last two authors.

## References

[1] M. L. Kherfi, D. Ziou, and A. Bernardi, "Image retrieval from the world wide web: Issues, techniques, and systems," *ACM Computing Surveys*, vol. 36, no. 1, pp. 35–67, 2004.

[2] P. André, E. Cutrell, D. S. Tan, and G. Smith, "Designing novel image search interfaces by understanding unique characteristics and usage," in *Proceedings of the IFIP Conference on Human-Computer Interaction*, 2009, pp. 340–353.

[3] E. Hoque, G. Strong, O. Hoeber, and M. Gong, "Conceptual query expansion and visual search results exploration for Web image retrieval," in *In Proceedings of the Atlantic Web Intelligence Conference*, 2011, pp. 73–82.

[4] C. J. van Rijsbergen, *Information Retrieval*. Butterworths, 1979.

[5] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2009, pp. 5–14.

[6] R. Santos, J. Peng, C. Macdonald, and I. Ounis, "Explicit search result diversification through sub-queries," in *Proceedings of the European Conference on Information Retrieval*, 2010, pp. 87–99.

[7] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 335–336.

[8] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 10–17.

[9] F. Radlinski and S. Dumais, "Improving personalized Web search using result diversification," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 691–692.

[10] R. L. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for Web search result diversification," in *Proceedings of the International Conference on the World Wide Web*, 2010, pp. 881–890.

[11] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in *Proceedings of the ACM International Conference on Multimedia*, 2004, pp. 952–959.

[12] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, "Web image clustering by consistent utilization of visual features and surrounding texts," in *Proceedings of the ACM International Conference on Multimedia*, 2005, pp. 112–121.

[13] S. Wang, F. Jing, J. He, Q. Du, and L. Zhang, "IGroup: presenting Web image search results in semantic clusters," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 587–596.

[14] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proceedings of the International Conference on the World Wide Web*, 2009, pp. 341–350.

[15] P.-A. Moëllic, J.-E. Haugeard, and G. Pitel, "Image clustering based on a shared nearest neighbors approach for tagged collections," in *Proceedings of the International Conference on Content-Based Image and Video Retrieval*, 2008, pp. 269–278.

[16] H. Ding, J. Liu, and H. Lu, "Hierarchical clustering-based navigation of image search results," in *Proceedings of the ACM International Conference on Multimedia*, 2008, pp. 741–744.

[17] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," in *Proceedings of the ACM International Conference on Multimedia*, 2006, pp. 707–710.

[18] R. van Zwol, V. Murdock, L. Garcia Pueyo, and G. Ramirez, "Diversifying image search with user generated content," in *Proceeding of the ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 67–74.

[19] D. Myoupo, A. Popescu, H. L. Borgne, and P.-A. Moëllic, "Multimodal image retrieval over a large database," in *Proceedings of the International Conference on Cross-Language Evaluation Forum: Multimedia Experiments*, 2009.

[20] B. J. Jansen, A. Spink, and J. Pedersen, "An analysis of multimedia searching on AltaVista," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2003, pp. 186–192.

[21] D. Milne and I. H. Witten, "Learning to link with Wikipedia," in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2008, pp. 509–518.

[22] M. Strube and S. P. Ponzetto, "WikiRelate! computing semantic relatedness using Wikipedia," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2006, pp. 1419–1424.

[23] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2007, pp. 1606–1611.

[24] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence*, 2008, pp. 25–30.

[25] E. Hoque, O. Hoeber, and M. Gong, "Evaluating the trade-offs between diversity and precision for Web image search using concept-based query expansion," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence - Workshops (International Workshop on Intelligent Web Interaction)*, 2011, pp. 130–133.

[26] G. Strong, E. Hoque, M. Gong, and O. Hoeber, "Organizing and browsing image search results based on conceptual and visual similarities," in *Proceedings of the International Symposium on Visual Computing*, 2010, pp. 481–490.

[27] G. Strong, O. Hoeber, and M. Gong, "Visual image browsing and exploration (Vibe): User evaluations of image search tasks," in *Proceedings of the International Conference on Active Media Technology*, 2010, pp. 424–435.

[28] E. Hoque, O. Hoeber, and M. Gong, "Combining conceptual query expansion and visual search results exploration for Web image retrieval," *Journal of Ambient Intelligence and Humanized Computing*, in press.

**Enamul Hoque** is currently an M.Sc. student in the Department of Computer Science at Memorial University of Newfoundland. His research interests include image retrieval, search interfaces, and information visualization.

**Orland Hoeber** received his Ph.D. from the University of Regina in 2007, and is currently an Assistant Professor in the Department of Computer Science at Memorial University of Newfoundland. His primary research interests include information visualization, Web search interfaces, Web search personalization, human-computer interaction, and geo-visual analytics.

**Minglun Gong** received his Ph.D. from the University of Alberta in 2003, and is currently an Associate Professor in the Department of Computer Science at Memorial University of Newfoundland. His research interests include a variety of topics in computer graphics, computer vision, image processing, pattern recognition, and optimization techniques.

# Which is the best?: Re-ranking Answers Merged from Multiple Web Sources

Hyo-Jung Oh, Pum-Mo Ryu, Hyunki Kim
Knowledge Mining Research Team, Electronics and Telecommunications Research Institute (ETRI)
161 Gajeong-dong, Yuseong-gu, Daejeon, Korea (305-700)
{ohj, pmryu, hkk}@etri.re.kr

***Abstract*—The main motivation of this paper is to devise a way to select the best answers collected from multiple web sources. Depending on questions, we need to combine multiple QA modules. To this end, we analyze real-life questions for their characteristics and classify them into different domains and genres. In the proposed distributed QA framework, *local* optimal answers are selected by several specialized sub-QAs. For fining *global* optimal answers, merged candidates are re-ranked by adjusting confidence weights based on the question analysis. We adopt the idea of the margin separation of SVM classification algorithm to adjust confidence weights calculated by own ranking methods in sub-QAs. We also prove the effects of the proposed re-ranking algorithm based on a series of experiments.**

***Index Terms*—Re-ranking, Multiple Web sources, Question Answering**

## I. Introduction

Depending on questions, various answering methods and answer sources can be used. To find the answer for general questions in a simple factoid-style, many systems of TREC (Text Retrieval Conference, [1]) adopted statistical answering methods [2]. For some questions that ask record information such as "Which is the longest river?", finding the answer in a specific corpus like the Guinness Book is more effective. Otherwise, knowledge bases can be used to answer definition questions such as "Who is J. F. Kennedy?"

One can argue that the same answer from multiple sources would increase the confidence level [3,4]. Depending on the type of question and the nature of Question Answering (QA) module, however, this type of redundancy may not be necessary [5]. For example, a question like, "When was Madam Curie born?", can be answered without ambiguity in an encyclopedia-based QA system, if an answer exists, because it can be handled by a pre-constructed knowledge base. Besides, multiple answers may end up lowering the confidence level of the correct answer if a straightforward merging method is used. We take the position that some redundancy would be useful for answer verification but should be used more judiciously for both efficiency and effectiveness.

More recent research tries to comprehend heterogeneous sources with the aim of improving the performance of QA system. PIQUANT from IBM [6] firstly proved that a multi-source approach to question answering achieve a good correlation of confidence values and correctness. A re-ranker of CHAUCER [7] compiled answers from each of the five answer extraction strategies into a single ranked list and an Answer Selection module identifies the answer which best approximates the semantic content of the original question. However, these researches are focused on the similarity of candidate answer and the given question.

As an advanced research, PowerAqua [8] explores the increasing number of multiple, heterogeneous knowledge sources available on the Web. A major challenge faced by PowerAqua is that answers to a query may need to be derived from different ontological facts and even different semantic sources and domains. To overcome this problem, they presented merging and ranking methods for combining results across ontologies [9]. However, the ontology-based method needs a lot of human efforts.

In Information Retrieval (IR) area, the base technology of QA, Wu and Marian [10] proposed a framework to aggregate query results from different sources. To return the best answers to the users, we assign a score to each individual answer by taking into account the number, relevance and originality of the sources reporting the answer. They took into account the quality of web pages.

In this paper, we build a distributed QA system to handle different types of questions and web sources. Especially, to select the best answer for a given user question, we propose an answer selection algorithm for re-ranking candidate answers from distributed multiple web sources.

To distill characteristics questions and answers, we differentiate varieties of user questions collected from commercial portal services, and distinguish a wide spectrum of potential answers from multiple web sources. Based on these observations, we build vertical sub-QA modules specialized for different domains and genres of web sources. Each sub-QA has own answer extraction methods tailed to various answer types that are identifiable from documents. They built multiple inverted index databases and distributed with Hadoop system [11].

To merge candidate answers from multiple web sources, we develop a special broker to interact with sub-QAs. When a user question is entered, the broker distributes the question over multiple sub-QAs according


doi:10.4304/jetwi.4.1.35-42

to question types. The selected sub-QAs find local optimal candidate answers, however, the weights are calculated by own ranking mechanisms so they have a big diversity.

The merged candidates are re-ranked by adjusting confidence weights based on the question analysis result. The re-ranking algorithm aims to find global optimal answers. We borrow the concept from the margin separation and slack variables of SVM classification algorithm [12, 13], and modify to project confidence weights into the same boundary by training.

In the following, we (1) discuss characteristics of questions and illustrate an overview of our distributed QA model consists of multiple sub-QAs in Section 2, (2) describe how to analyze a given user question and distribute it over corresponding sub-QAs in Section 3, (3) concentrate on the re-ranking algorithm based on the SVM training model in Section 4, (4) analyze the effect of the proposed ranking method with several experiments together with an in-depth analysis of weight distribution in Section 5. Finally we conclude with a suggestion for possible future works in Section 6.

## II. Distributed Question Answering

Our ultimate goal is to build a QA system that can handle a variety of types of questions and answers. In order to make full use of various QA techniques corresponding to different types of questions it is critical to classify user questions in terms of the nature of the answers being sought after. To this end, we collected more than 7,000 questions from the commercial manual QA sites[1]. They were analyzed to characterize the types of questions and answer sources with two points of view: *domains* (or categories) and *genres*.

As shown in Table 1, the most Top-4 domains of user questions are education, game/computer, life (including travel, sports, and local), and entertainment. However, the education domain covers very broad subjects (e.g. mathematics, physics, or chemistry) and over 80% of answers of questions in education could be found in the encyclopedia or Wikipedia [5], so we substituted this domain for the Wikipedia genres. Meanwhile, we also excluded the entertainment domain since many questions belong to this domain are related with gossip or entertainer scandals, so these problems are very difficult to judge that which answers are correct or not. In this paper, four major domains are selected: game/computer, travel, local, and life. The other questions are considered as the "open" domain.

We differentiated answer sources according to genres. Because the corpus such as the Wikipedia contains facts about many different subjects or explains one particular subject in detail, there are many sentences that present definitions such as "X is Y." On the other side, some sentences describe the process of a special event (e.g. World War I), so that they consist of particular syntactic structures (5W1H) similar to those found in news documents. In blogs, there are many personal opinions or know-how for a certain problem. In this paper, major genres of web sources which have many chances to get answers for most of user question are concentrated on four types: news, blogs, Wikipedia, and the others.

TABLE I. Domain Distribution of User Questions

| Domain | | # of Question | Ration |
|---|---|---|---|
| Education | | 1,813 | 25.64% |
| Game/Computer | | 1,218 | 17.23% |
| Life | Local | 891 | 12.60% |
| | Sports | 238 | 3.37% |
| | Travel | 133 | 1.88% |
| Entertainment | | 720 | 10.18% |
| Shopping | | 521 | 7.37% |
| Health | | 516 | 7.30% |
| Economics | | 510 | 7.21% |
| Social/Politics | | 510 | 7.21% |
| Total | | 7,070 | |

Based on overall analysis, we built a distributed QA system to have multiple sub-QA modules as shown in Figure 1. We combined eight sub-QAs: four *domain-QAs* and four *genre-QAs*. While the sub-QA modules are complementary to each other in providing answers of different types, their answer spaces are not completely disjoint.

All sub-QAs except the web-QA have own answering methods tailed to various answer types that are identifiable from documents [5]. Especially, the number of documents indexed in the News-QA and the Blog-QAs are approximately 13,800,000 and 33,600,000, respectively. Thus we needed to distribute local indexing database for efficient. As shown in Figure 1, the News-QA and the Blog-QA consist of 3 and 5 *Hadoop clusters*, respectively. For the Web-QA to complement web documents which we neglected to crawl, we used the Yahoo! API from Yahoo! Search Web Services[2].

To interact with sub-QA modules, we developed special brokers, the B1 and B2 components. The B1 combines multiple indexing blocks for a particular sub-QA and merges candidate answers, whereas the B2 communicates with multiple sub-QAs and ranks their candidates. The B2 has own ranking algorithm to find the local optimal answer for each sub-QA.

The *Answer Manager* component has two roles. The first one is to determine a user question and spread them to appropriate sub-QAs. The other is to re-rank candidate answers merged by B2 and find the best answer as the global optimal solution.

## III. Distributing Questions

Based on question types, the B2 in Figure 1 can determine which sub-QAs should be involved to find the best answer. A user question in the form of natural language is entered into the system and analyzed by the

[1] the *Naver*tm Manual QA Service (http://kin.naver.com). When a user upload a question, then the other users answer the question by manually and get points depending on how satisfied with the answer by the question owner.

[2] http://developer.yahoo.com/search/web/

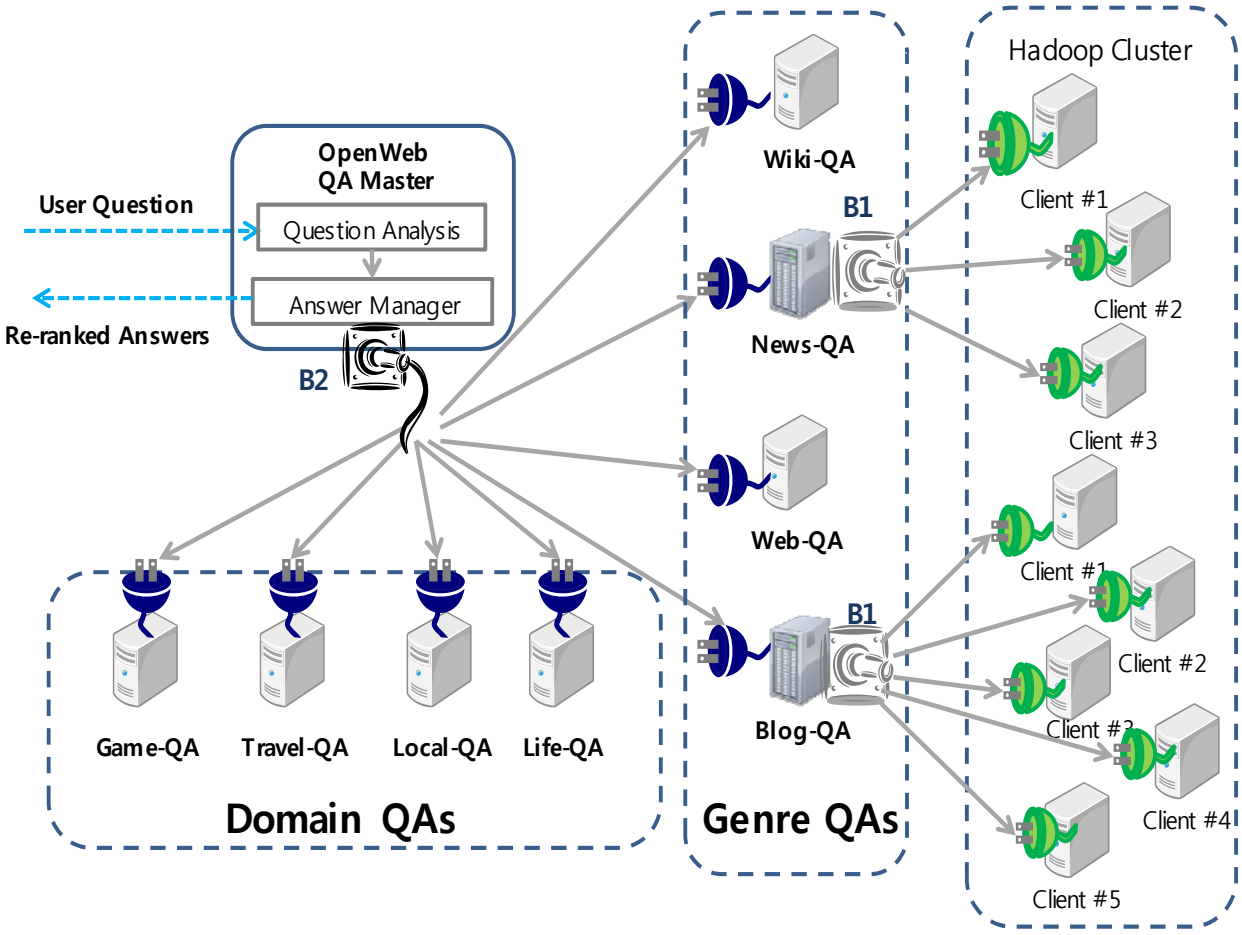



Figure 1. System Overview

*Question Analysis* component that employs various linguistic analysis techniques such as POS tagging, chunking, answer type (AT) tagging [9], and some semantic analysis such as word sense disambiguation [10]. An internal question generated from the Question Analysis component has the following form:

$$Q = <AF, AT, QT, AD> \quad (1)$$

where AF is the expected answer format, AT is the expected answer type, QT is the theme of the question, and AD is the domain related to the expected answer source or sub-QA module from which the answer is to be found.

- The answer format (AF) of a question is determined to be one of these four types: a single, multiple, descriptive, or yes/no question. For example, single is the AF value in the question "*Who killed President Kennedy*?"
- There are 147 fine-grained ATs organized in a hierarchical structure with 15 nodes at the level right below the root, each of which has two to four lower levels [9]. The AT gives information about the type of the entity being sought [11]. The sub-type/super-type relations among the ATs give flexibility in matching. For the example above, the AT would be "people" because of "who", which can be matched with "president" in a passage.
- A question theme (QT) has two parts: a target and a focus. The target of a question is the object or event that the question is about, whereas the focus is the property being sought by the question. In the example above, the target is "J. F. Kennedy" and the focus is "killer".
- The answer domain (AD) of a question indicates the most likely source (sub-QA module) from which an answer can be found, which is determined based on the other traits of the question (AF, AT, and QT). It also contains some detailed information about what should be sought after in the QA module. For example, the answer for the question, "When was Madam Curie born?" might be found in Wikipedia. In contrast, for "*How to play Starcraft*[3] *well using Protos*?", personal blogs or community boards for that game are more suitable.

[3] *StarCraft*[tm] is a military science fiction real-time strategy video game developed by Blizzard Entertainment©.

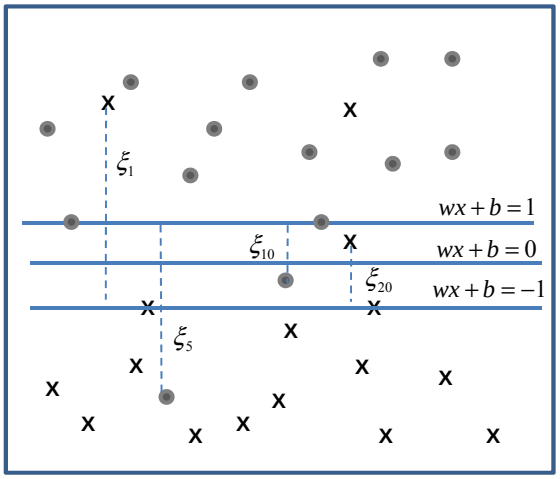



Figure 2. Modeling of SVM

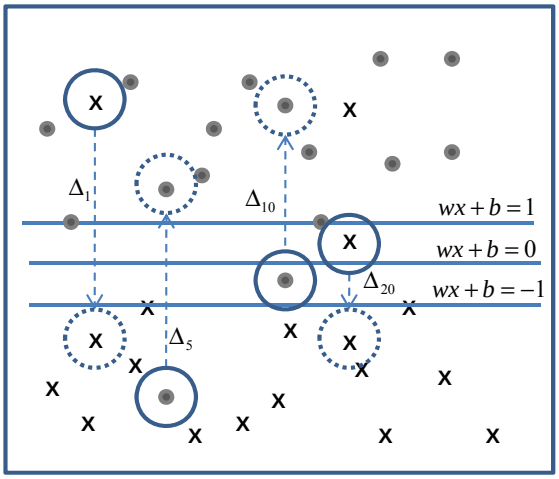



Figure 3. Examples of Adjust Slacks

Based on the question analysis, The B2 in Answer Manager invokes one or more sub-QA modules. For the former example question about Madam Cuire, the B2 might select the wiki-QA at first. If the calculated evidence of answer candidates from the wiki-QA is strong enough, then the B2 provides the answer to the user. If none of the answers from the first module have a confidence weight higher than the threshold, the B2 invokes other sub-QAs to merge candidates.

## IV. Merging and Re-Ranking Answers

The local optimal answers from multiple web sources are collected in the Answer Manager. As mentioned in Section 2, each sub-QA has own ranking mechanism, so that the confidence weights of merged answers are very diverse. For example, the weights from the News-QA are between 0 and 2, while Web-QA weights are between 0 and 0.8 (refer Figure 5). To adjust these variations and to project confidence weights into the same boundary, we devise a new re-ranking algorithm. We borrow the idea from the margin separation of SVM classification algorithm [12, 13], and modify to adjust confidence weights into the same boundary by training.

Figure 2 captures correct answers and uncertain answers in the SVM model and they are marked with "●" and "x", respectively. If the training answer set D is not linearly separable, the standard approach is to allow the fat decision margin to make a few mistakes. We then pay a cost for each misclassified example, which depends on how far it is from meeting the margin requirement given in Equation (2).

$$y_i(\vec{w}^T\vec{x}_i+b)\geq 1 \qquad (2)$$

Asking for small $w\cdot w$ is like "weight decay" in Neural Nets and like Ridge Regression parameters in Linear regression and like the use of Priors in Bayesian Regression-all designed to smooth the function and reduce overfitting [14]. In SVM, slack variable

```
<Original Q> What is the population of the Bahamas?
<Question Analysis>
   <AF> Factoid </AF>
   <AT> QT_NUMBER </AT>
   <QT> target: Bahamas / focus: population </QT>
   <AD> Wikipedia > news > blogs > others </AD>
<Answer rank =1>
   <Ans> 370267</Ans>
   <Ans_sent> The population of The Bahamas on
   January 1st 2010 is approximately 370267.
<Ans_sent>
    <Ans_source> Wiki-QA </Ans_source>
    <org_weight> 0.3 </ org_weight >
</Answer>
<Answer rank =2>
   <Ans> 294,982 </Ans>
   <Ans_sent> The population in the Bahamas is
        currently about 294,982 persons. <Ans_sent>
    <Ans_source> News-QA </Ans_source>
    <org_weight> 0.5 </ org_weight >
</Answer>
…
```

Figure 4. An Example of Evaluation Set.

$\xi_i$ measures by how much example $(\vec{x}_i, y_i)$ fails to archive a target margin of $\delta$. We adjust the weights which are located out of margin area as slacks according to traits of the question analysis result (AF, AT, QT, and AD in Equation (1)) and answer evidences such as snippet, matched keywords, or position in the answer document. After adjusting, the slack weights can be moved the safe boundary, as illustrated in Figure 3.

For training, we built <question, answer> set of various sorts in terms of questions and answer sources. Figure 4 is an example of a <question, answer> pair. We can notice that the weight of wiki-QA should be boosted from 0.3 to 0.6 (or any higher scores than other lower ranked answers), even though the original local confidence weight of the answer from the News-QA (0.5) is higher than wiki-QA's (0.3). On the other hand, some cases should be decreased.

By training, we learned the confidence weight distribution and slack boundaries $\xi_i$ and set boosting ratio $\Delta_i$ in Equation (3):

$$y_i(\vec{w} \cdot \vec{x}_i \Delta_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \qquad (3)$$

According to question types, the boosting ratios are different and they are updated whenever training questions are leaned. We also determined the threshold values for each of sub-QA to avoid superfluous calls for other QA modules.

At the result, when a new user question which is similar with training questions is entered, we can predict which QA modules are likely to find answers and how much we should increase or decrease the confidence weights from multiple sub-QAs.

## V. Evaluation and Analysis

### A. Test Collection and Measure

Among 7,070 questions collected from real users, 577 questions are selected with considering various question/answer types. We used 260 pairs of training and 317 pairs of testing, which is part of the entire set of 577 <question, answer> pairs.

For effectiveness comparisons, we employ precision, recall, and F-score, sometimes with the mean reciprocal rank (MRR) [15] and the well-known "top-n" measure that considers whether a correct nugget is found in the top n-th answers. Because of the large number of comparisons to be made for different cases, we use F-scores for summary tables.

### B. Experimental Background

As described in Section 2, our distributed QA model contains four different genre-specific QAs (News-QA, Blog-QA, Wiki-QA, and Web-QA) and four domain-specific QAs (Game-QA, Travel-QA, Local-QA, and Life-QA), which are covered most frequently asked questions.

Before the main experiment, we have to announce that characteristics of our distributed QA. Table 2 summarizes performance of individual Genre-QAs and Figure 5 shows weight distribution of four Genre-QAs.

As shown in Table 2, the News-QA answered for 244 questions among 317 of entire set. Out of 244 candidate answers, 139 answers are correct so the News-QA precision is 0.570. The best performed Genre-QA is Web-QA since Yahoo web search is covered all kind of web documents, whereas other genre-QAs are focused on specific genres.

Figure 5 depicts original confidence weights of the first answers (Top-1) merged from each four genre-QAs for 317 testing questions. If a sub-QA cannot find appropriate answers for a given question, then the weight regards as 0. In general, weights from the Blog-QA (marked with "■") are higher than others, while the Web-QA weights (marked with "x") are lower. The scores from the Wiki-QA (marked with "▲") are between 0.6 and 0.98, but the Blog-QA and the News-QA (marked with "♦") show inconsistent values from 0 to even 8. We had pruned weights over 2 to avoid overffiting to extremely higher values.

To prove reliability of our proposed QA model and re-ranking algorithm, we measured the lower and upper boundaries.

The accuracy of lower bound can be estimated when selecting the top-ranked answer among the merged candidates from multiple sub-QAs without any adjustment. As shown in the right side of Table 2, the accuracy of simply merging and selecting the top-ranked answer shows just 0.568, even though micro-average precision of all genre-QAs is 0.722. That is poor than the case of performing only one sub-QAs. Because the News-QA and the Blog-QA usually are usually higher than others as shown in Figure 5, the final selected weight depends on them. This result supports that re-ranking is very important, which is the main focus in this paper.

Under assumption that the correct answer for a given question might be exist at least among all local optimal candidate answers generated from multiple sub-QAs, we evaluated all Top-3 ranked answers for each sub-QA. The number of candidates in answer pool can be from 0 to 24 at most (3 answer x 8 sub-QAs). We regarded this result as the upper boundary. Out of 317 questions, user can get correct answers for 273 questions. The fact that the recall is quit high (0.861) prove our assumption is reasonable. Our upper boundary precision is 0.907 and MRR is 0.849.

### C. Analysis of Experimental Results

This paper had pursued adjusting diverse confidence weights from multiple sub-QAs as shown in Figure 5. For a detail example, Figure 6 shows the original weights from the Blog-QA. They are divided into correct and incorrect answers which are indicated with "●" and "x", respectively. Because some fake answers have very high scores as shown in Figure 6, they lead to ignore the other sub-QAs results and ultimately cause false alarm.

After the training process which is described in Section 4, we learned which sub-QAs are more relevant and how to boost their weights depending on question types. As a result, we observed that the Blog-QA is suitable to answer for factoid questions about game/computer or personal life domains. In contrast, answers for questions looking for interesting information about a particular person or thing such as "*Who is Vlad the Impaler?*" or "*What is a golden parachute?*" are founded in the Wiki-QA. Figure 7 presents the adjustment result for the Blog-QA. While some wrong answers are located in higher position, we can cut-off candidates less than 0.2.

Table 3 and 4 summarize the accuracy of our distributed QA adopted the proposed re-ranking algorithm. As mentioned in Section 5.1, multiple sub-QAs answered for 301 questions and could not find any candidate for 16 questions, out of 317 questions. We merged the Top-3 candidates for each sub-QA by the B2 broker, and adjusted their confidence weight by the trained model.

TABLE II. PERFORMANCE OF INDIVIDUAL GENRE-QAS AND BOUNDARIES

| | Individual Genre-QAs (Top-3) | | | | *Lower bound* | *Upper bound* |
|---|---|---|---|---|---|---|
| | News-QA | Blog-QA | Wiki-QA | Web-QA | *Selecting the top-ranked* | *All Top-3 ranked answers* |
| #of answered Q. | 244 | 272 | 135 | 228 | 301 | 301 |
| # of missed Q | 73 | 45 | 182 | 89 | 16 | 16 |
| # of corrected Q. | 139 | 204 | 103 | 184 | 130 | 273 |
| Precision | 0.570 | 0.750 | 0.763 | 0.807 | 0.568 | 0.907 |
| Recall | 0.541 | 0.794 | 0.401 | 0.716 | 0.539 | 0.861 |
| F-score | 0.555 | 0.771 | 0.526 | 0.759 | 0.553 | 0.883 |
| MRR | 0.488 | 0.645 | 0.740 | 0.711 | 0.568 | 0.841 |
| Micro-Average | Precision = 0.722 / Recall = 0.613 | | | | | |

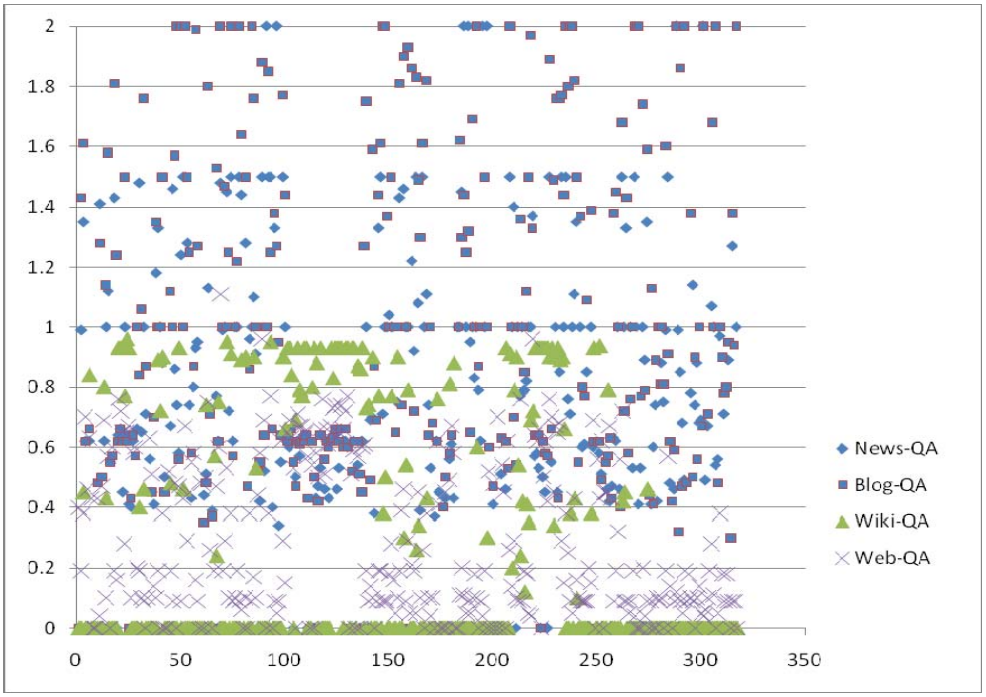



Figure 5. Weight Distribution of four Genre-QAs

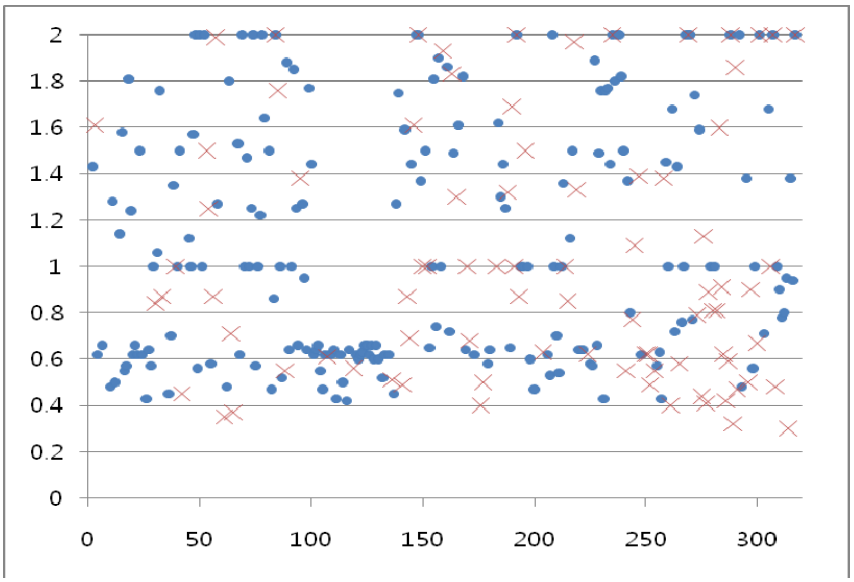



Figure 6. The original weights from the Blog-QAs

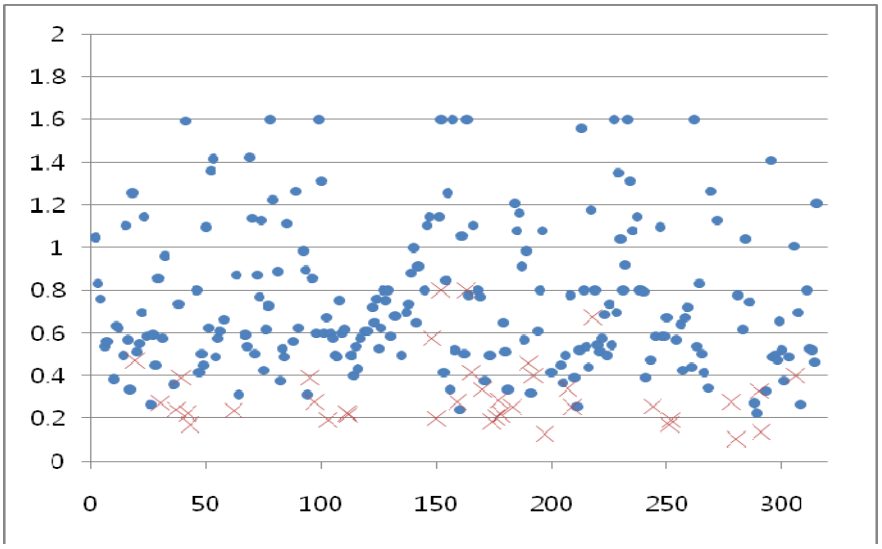



Figure 7. Adjusted weights of the Blog-QAs

TABLE II. EVALUATION RESULT OF THE TOP-1

| | Lower B. | **Top-1** |
|---|---|---|
| # of corrected | 130 | 202 *(+72)* |
| Precision | 0.568 | 0.671 *(+18.13%)* |
| Recall | 0.539 | 0.637 *(18.18%)* |
| F-score | 0.553 | 0.654 *(18.26%)* |

TABLE II. EVALUATION RESULT OF THE TOP-3

| | **Top-3** | Upper B. |
|---|---|---|
| # of corrected | 248 *(-25)* | 273 |
| Precision | 0.824 *(-9.15%)* | 0.907 |
| Recall | 0.782 *(-9.18%)* | 0.861 |
| F-score | 0.803 *(-9.06%)* | 0.883 |
| MRR | 0.801 *(-4.76%)* | 0.841 |

To compare with the lower boundary, we evaluated only the top-ranked answers for 301 questions. As shown as Table 3, we obtained an increased accuracy by about 18% of precision, recall and F-score. 72 (130 to 202) answers are adjusted by our re-ranking algorithm. In Top-1 evaluation, MRR is same with precision.

Compensating the factor that the upper boundary considered all candidates in the Top-3 for each sub-QA, we also evaluated the Top-3 of re-ranked answers, but not all answer pool. As in Table 4, we missed 25 questions' answers while we handled just three candidates on top. We got the loss about 9% (-9.06%) of F-score and 5% (-4.76%) of MRR. The fact that the gap between MRRs (0.801 vs. 0.841) is smaller than other measures indicates the answers of our method are located in higher position. That is by our re-ranking algorithm; we can not only suppress erroneous answers but also save time.

## VI. CONCLUSION

The main motivation behind this work was to devise a way to combine multiple QA modules to answer various user questions. To this end, we analyzed real-life questions for their characteristics and classified them into different domains and genres. In the proposed distributed QA framework, 8 specialized sub-QAs are combined and an advanced re-ranking algorithm are adopted to adjust confidence weights calculated by own ranking methods in sub-QAs.

We ran a series of experiments to see the effects of the proposed re-ranking algorithm against two different cases: (1) the lower boundary when considering only the first answers from sub-QAs, (2) the upper boundary when evaluating all local optimal candidate answers. The result based on 317 questions show that our re-ranking method outperforms the lower boundary by about 18%. Compared with the upper case, the loss is narrowly about 5% in MRR.

Based on the result of question analysis, The B2 in Answer Manager determined invocation of sub-QAs. In particular, the expected answer type and answer domain analysis for a question presents a critical problem because it influences the re-ranking process. We plan to improve upon the answer type classification and domain expectation modules by expanding training corpus toward including various question types.

## ACKNOWLEDGMENT

This work was supported in part by the Korea Ministry of Knowledge Economy (MKE) under Grant No. 2011-SW-10039158.

## REFERENCES

[1] M. Voorhees, "The TREC-8 Question Answering Track Report." Proc. of the 8th Text REtrieval Conference (TREC-8), 1999, pp. 77-82.

[2] Harabagiu, D. Moldovan, C. Clack, et. al., "Answer mining by combining extraction techniques with abductive reasoning" Proc. of the 12th Text REtreival Conference(TREC-12) , 2003, pp. 375-382.

[3] J. Chu-Carroll et al., "A Multi-strategy and Multi-source Approach to Question Answering," Proc. of the 11th Text REtrieval Conference (TREC-11), 2002, pp. 281-288
[4] B. Katz et al., "Answering Multiple Questions on a Topic from Heterogeneous Resources," Proc. of the 13th 11th Text REtrieval Conference (TREC 2004).
[5] H-J. Oh, S. H. Myaeng, and M. G. Jang, "Enhancing Performance with a Learnable Strategy for Multiple Question Answering Modules." ETRI Journal, vol.31, no.4. 2009, pp. 419-428
[6] J. Chu-Carroll, J. Prager, C. Welty et al, "A Multi-Strategy and Multi-Source Approach to Question Answering," Proc. of the 11th Text REtrieval Conference (TREC-11), 2002, pp.281-288.
[7] A. Hickl, J. Williams, J. Bensley et al, "Question Answering with LCC's CHAUCER at TREC 2006," Proc. of the 15th Text REtrieval Conference (TREC 2006).
[8] V. Lopez, M. Sabou, V. Uren, and E. Motta, "Cross-Ontology Question Answering on the Semantic Web –an initial evaluation", Proc. of the Knowledge Capture Conference, 2009
[9] V. Lopez, A. Nikolov, M. Fernandez, et al., "Merging and Ranking Answers in the Semantic Web: The Wisdom of Crowds", Proc. of the ASWC 2009, LNCS 5926, 2009, pp. 135–152.
[10] M. Wu and A. Marian, "A Framework Corroboration Answers from Multiple Web Sources". Information Systems, 2010, doi:10.1016/j.is.2010.08.008, online press.
[11] Dhruba Borthaku. "The Hadoop Distributed File System: Architecture and Design", http://hadoop.apache.org/common/docs/r0.18.0/hdfs_design.pdf , 2007.
[12] C.K. Lee and M.G. Jang, "A Prior Model of Structural SVMs for Domain Adaptation," ETRI Journal, vol. 33, no. 5, 2011, pp. 712-719
[13] H-J, C.K. Lee and C-H Lee, "Analysis of the Empirical Effects of Contextual Matching Advertising for Online News", ETRI Journal, vol. 34, no. 2, April 2012 (to be appeared)
[14] A. W. Moore, tutorial of "Support Vector Machine", http://www.cs.cmu.edu/~awm/tutorials
[15] E. M. Voorhees and M. T. Dawn. "Building a question answering test collection", Proc. of the 23rd Annual International ACM SIGIR, 2000, pp. 200-207

**Hyo-Jung Oh** received the B.S. and the M.S. degrees in computer science from Chungnam National University, Daejeon, South Korea, in 1998 and 2000, respectively. She received the Ph.D. degree in computer engineering from KAIST, Daejeon, South Korea, in 2008. Currently she is a senior researcher in Electronics and Telecommunications Research Institute (ETRI), Deajeon, South Korea. Her research interests include machine learning, question answering, and listening platform for business intelligence.

**Pum-Mo Ryu** received the B.S. degree in computer engineering from Kyungpook National University, Daegu, South Korea in 1995 and the M.S. degree in computer engineering from POSTECH, Pohang, South Korea, in 1997. He received the Ph.D. degree in computer science from KAIST, Daejeon, South Korea, in 2009. Currently he is a senior researcher in Electronics and Telecommunications Research Institute (ETRI), Deajeon, South Korea. His research interests include natural language processing, text mining, knowledge engineering and question answering.

**Hyunki Kim** received the B.S. and the M.S. degrees in computer science from Chunbuk National University, Daejeon, South Korea, in 1994 and 1996, respectively. He received the Ph.D. degree in computer science from University of Florida, Gainesville, USA, in 2005. Currently, He is a principal researcher in Electronics and Telecommunications Research Institute (ETRI), Deajeon, South Korea. His research interests include natural language processing, machine learning, question answering, and listening platform for business intelligence.

# Graph-cut based Constrained Clustering by Grouping Relational Labels

Masayuki Okabe
Toyohashi University of Technology
Tempaku 1-1, Toyohashi, Aichi, Japan
okabe@imc.tut.ac.jp

Seiji Yamada
National Institute of Informatics
Chiyoda, Tokyo, Japan
seiji@nii.ac.jp

*Abstract*— **This paper proposes a novel constrained clustering method that is based on a graph-cut problem formalized by SDP (Semi-Definite Programming). Our SDP approach has the advantage of convenient constraint utilization compared with conventional spectral clustering methods. The algorithm starts from a single cluster of a whole dataset and repeatedly selects the largest cluster, which it then divides into two clusters by swapping rows and columns of a relational label matrix obtained by solving the maximum graph-cut problem. This swapping procedure is effective because we can create clusters without any computationally heavy matrix decomposition process to obtain a cluster label for each data. The results of experiments using datasets from the ODP and WebKB corpus demonstrated that our method outperformed other conventional and the state of the art clustering methods in many cases. In particular, we discuss the difference between our approach and another similar one that uses the same SDP formalization as ours. Since the number of constraints used in the experiments is relatively small and can be practical for human feedback, we consider our clustering provides a promising basic method to interactive Web clustering.**

## I. Introduction

Clustering has long been one of the most essential and popular techniques in data mining [1]. It is used not only for visualization of huge data sets but also for image segmentation [2], medical applications, recommendation systems, and so on.

Constrained clustering is a semi-supervised learning approach that utilizes pre-given knowledge about data pairs to improve normal clustering accuracy [3], [4]. The knowledge used is generally of two simple types: a constraint about data pairs that must be in the same cluster, and a constraint about data pairs that must be in a different cluster. These are usually called *must-link* and *cannot-link*, respectively.

Recent research about distance metric learning interprets the constraint information as the distance or kernel value of data pairs and tries to produce a new distance measure or kernel matrix for a whole dataset to ensure the distance of must-link is small and the distance of cannot-link is large [5], [6], [7]. In this research, we do not interpret the constraint as a distance or kernel value but rather as a relational label that indicates whether data pairs should be in the same cluster or not. Our objective is to predict the correct label for each data pair (not for each individual piece of data) by using sample labels converted from pre-given constraint information according to the framework of the transductive learning.

Our method is based on the graph-cut problem. Although graph-cut based clustering (e.g., spectral clustering) is a well known approach and many methods have been proposed so far [8], [9], their solutions are mostly based on the graph spectrum obtained by eigen decomposition, which requires complicated processes to add in the constraint information. Our approach is to solve it as a semi-definite programming (SDP) problem. The advantage of SDP is that we can naturally incorporate constraints without any complicated processing and do not need any specific objective functions (e.g., normalized cut) to avoid a trivial solution (as is the case with many other spectral clustering methods).

In terms of formalization, our problem is the same as Li's [10] or Hoi's [11], although the introduction is completely different. The most critical difference is the interpretation of the SDP solution. They interpret the solution as a kernel matrix and use it for multi-class clustering, while we interpret it as a label matrix (as described above) and use it for two-class clustering. As we will show in the experiments, our two-class clustering approach performs better than the multi-class clustering approach. Our approach is based on the divide and conquer algorithm. It starts from a single cluster of a complete dataset and repeatedly selects the largest cluster, which it then divides into two clusters until we get the target numbers of clusters. In each iteration, we obtain relational labels for all data pairs from the solution of the SDP problem. We then use the label matrix to create clusters by swapping rows and columns to reduce the clusters' label distribution entropies. This swapping procedure is very effective because we can create clusters without any computationally heavy matrix decomposition processing.

In summary, we propose a constrained clustering method that has the following features.

- Clustering is performed based on the relational labels of all data pairs that are obtained by solving a graph-cut problem formalized by semi-definite programming. Our SDP approach has the advantage convenient constraint utilization compared with conventional spectral clustering methods.
- The interpretation of the obtained matrix is different from Li and Hoi's approaches, although the problem


doi:10.4304/jetwi.4.1.43-50

formalization is similar. They use the matrix as a kernel matrix for one-time multi-class clustering, while we use it as binary label matrix for divide and conquer-based two-class clustering.

These advantages make constrained clustering more efficient, especially in the case of small number of constraints such as interactive web clustering like [12].

The rest of the paper is organized as follows. First we explain the standard maximum graph cut problem and its solution by semi-definite programming relaxation in Section II. Next, we describe our clustering algorithm, which is based on the approximate solution of the SDP. We describe the entire clustering procedure including binarizing and swapping of the solution matrix in Section III. Section IV shows the results of experiments performed using datasets from the ODP and WebKB corpus. We discuss our methods in Section V, and finally we conclude our work in Section VI.

## II. Constrained Graph Cut Problem

Graph-cut formalization is a powerful clustering approach that many algorithms have adopted. In this section, we first formalize the maximum cut problem and then introduce a solution by semi-definite programming.

### A. Maximum Cut Problem

The objective of the problem is to divide a graph into two parts as its cut amount reaches the maximum. More formally, consider a graph $G = (V, E)$, where $V$ is a set of vertices and $E$ is a set of edges. The problem is to find partitioning $(V_1, V_2)$ such as $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \phi$ and a maximum cut amount of $\sum_{i \in V_1, j \in V_2} w_{ij}$. Here, $w_{ij}$ is the weight of an edge between data $i \in V_1$ and data $j \in V_2$. We decide on a "maximum" cut when $w_{ij}$ is defined by some distance (e.g., Euclid distance). In contrast, if $w_{ij}$ is defined by some similarity (e.g., Gauss kernel), the "minimum" cut is appropriate.

By introducing a cluster label variable $u_i$ for each vertex, we can formalize the maximum cut problem as follows.

Maximum Cut Problem

$$\begin{aligned} \text{maximize} \quad & \frac{1}{4} \sum_{i \in V_1} \sum_{j \in V_2} w_{ij}(1 - u_i u_j) \\ \text{subject to} \quad & u_k^2 = 1 \quad (k \in V) \\ & u_k = \begin{cases} +1 & (k \in V_1) \\ -1 & (k \in V_2) \end{cases} \end{aligned}$$

According to the standard method of spectral clustering or segmentation by the random walk model, we can solve this problem with the method of Lagrange multipliers. The $u_i$ labels are obtained as eigen vectors corresponding to the second largest eigen value.

Our aim is to incorporate given constraints into the above problem and find a method to solve constrained maximum cut problems. While there are constrained versions of spectral clustering methods, we adopt a different approach, solution by semi-definite programming (SDP), which is practically easier to use because it can handle constraints intrinsically.

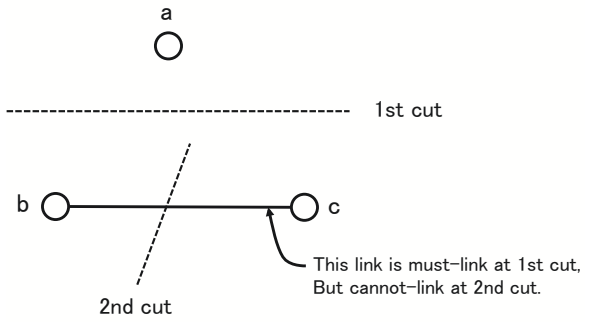



Figure 1. Cannot-link is not applicable in divide and conquer approach

### B. Formalization by SDP

Semi-definite programming is a kind of convex optimization that is used to relax several optimization problems such as combinatorial optimization, 0-1 integer programming, and non-convex quadratic programming. Since the maximum cut problem is an example of 0-1 integer programming, SDP can also relax it.

For the standard formalization of SDP, we transform the above objective function into a matrix representation with a weight matrix $W$ and a matrix $X$ whose element is the product of $u_i$ and $u_j$.

$$\begin{aligned} \sum_{i \in V_1} \sum_{j \in V_2} w_{ij}(1 - u_i u_j) &= (diag(W\mathbf{e}) - W) \bullet X \\ &= L \bullet X \end{aligned}$$

$$X = \mathbf{u}^T \mathbf{u}$$

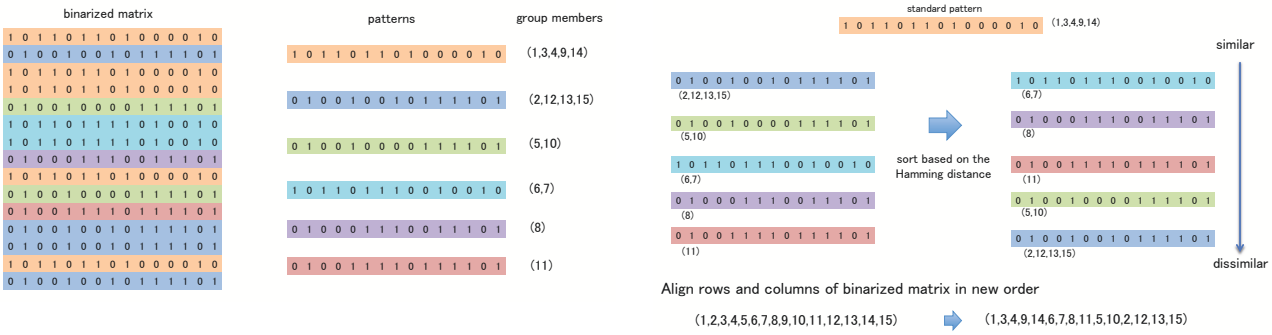
$$\mathbf{u} = (u_1, u_2, ..., u_n), \quad n = |V|$$

$L$ is the graph Laplacian matrix and $\mathbf{e}$ is a vector whose elements are all one. As a final step, we add *must-link* constraints to formalize the constrained maximum cut problem as follows.

Maximum Cut Problem with SDP Relaxation

$$\begin{aligned} \text{maximize} \quad & L \bullet X \\ \text{subject to} \quad & E_{ii} \bullet X = 1, \quad (i = 1 \sim n) \\ & E_{ij} \bullet X = 1, \quad (i, j) \in M \\ & X \succeq O \end{aligned}$$

$E_{ij}$ is an $n \times n$ matrix in which only the $(i, j)$ element is 1, and all others are 0. $M$ is a set of *must-link*.




(a) Grouping patterns

(b) Sort pattern groups

Figure 2. Clustering process using label matrix

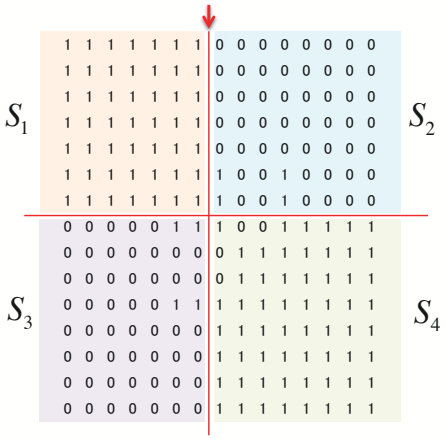



Figure 3. Clustering Boundary

Although available constraints are not limited to must-link, meaning we can also use *cannot-link*, it is very difficult to select applicable cannot-links during multi-class clustering because the partitioning order determined in the graph cut process is usually unpredictable in advance. Figure 1 gives an idea of the difficulty of using cannot-link. There are three data in the figure - $a, b, c$ - and cannot-link is applicable only at the second cut. It cannot be used at the first cut because $b$ and $c$ are in the same cluster at that time.

There are several freely available SDP solvers that we can use to obtain an approximate solution $\tilde{X}$. Although we need to decompose $\tilde{X}$ to obtain partitioning label **u**, we found a way to complete partitioning by using only $\tilde{X}$. We explain this method in the next section.

## III. Clustering through Swapping Rows and Columns in a Label Matrix

In this section, we describe a concrete partitioning procedure using matrix $\tilde{X}$, which is given as the SDP solution, and then an entire clustering with an iterative dual-partitioning process.

As described in the previous section, we solve the maximum cut problem with relaxed SDP, so the elements

**Algorithm 1** Constrained Iterative Graph Cut Clustering

```
Input: D_0      // Dataset
       W        // Weight Matrix
       M        // Must-Link Set
       K        // Number to be Clustered
Output: C = {D_1, D_2, ..., D_K}    // K clusters

Let C = D_0
for i = 1 to K-1 do
    Select the largest cluster D_t^max
    Extract a subset of must-link constraints M_t^sub related to D_t^max
    Input D_t^max and M_t^sub to SDP solver and get divided clusters {D_t^1, D_t^2}
    Subtract D_t^max from C
    Add {D_t^1, D_t^2} to C
end for
```

of $\tilde{X}$ are assigned a real value ranging from -1 to 1. We therefore decided on a different approach in which we first binarize $\tilde{X}$ with 0-1 values, then swap rows and columns to maximize the evaluation measure, and finally determine the partitioning border.

The concrete procedures are as follows.

1) Each element of $\tilde{X}$ is binarized as follows.

$$\tilde{X}_{ij} = \begin{cases} 1, & \text{if } \tilde{X}_{ij} \geq 0 \\ 0, & \text{if } \tilde{X}_{ij} < 0 \end{cases}$$

The value does not matter because we treat 0 and 1 as a character in the following steps.

2) For each row, treat the column's value as a character (0 or 1) and make a string (or pattern) by concatenating each character in the original order. Next, make groups of the same string (select a representative of each kind of string). Figure 2(a) illustrates this procedure.

3) Determine the most frequent string $s_0$, and calculate the Hamiltonian distance between $s_0$ and the other strings. Next, align other strings in descending order of the Hamming distance. String $s_1$ is the most

TABLE I.
DIRECTORIES USED IN ODP CORPUS

| No. | Directory | #data |
|---|---|---|
| 1 | Anomalies and Alternative Science | 47 |
| 2 | Science in Society | 45 |
| 3 | Environment/Water Resources | 42 |
| 4 | Astronomy | 24 |
| 5 | Technology/Structural Engineering | 24 |
| 6 | Agriculture | 19 |
| 7 | Biology/Genetics | 16 |
| 8 | Social Sciences/Linguistics | 15 |
| 9 | Physics | 15 |
| 10 | Earth Sciences | 11 |
| 11 | Math | 10 |
| 12 | Chemistry | 6 |

TABLE II.
DATASETS IN ODP CORPUS

| Dataset | #data | #cluster |
|---|---|---|
| odp_odd | 154 | 6 |
| odp_even | 120 | 6 |
| odp_small | 92 | 7 |
| odp_all | 274 | 12 |

similar to $s_0$. Figure 2(b) illustrates this procedure.

4) Determine the partitioning boundary according to the following measure.

$$F(i,j) = \sum_{k=1}^{4} -p^{S_i}\log(p^{S_i}) - (1-p^{S_i})\log(1-p^{S_i})$$

$(i,j)$ represents the partitioning boundary. If we partition the above aligned matrix in the boundary between $i$'s row and $j$'s row (i.e., also in the boundary of $i$'s column and $j$'s column ), there are four partitioned areas in the matrix. $p_0^k$ and $p_1^k$ are the probabilities of 0 and 1, respectively, that appears in the area $k$. Thus, $F(i,j)$ is the sum of the entropy in four areas when partitioned between $i$th and $j$th row (and column). The more clearly partitioned, $F(i,j)$ becomes lower.

5) Determine the boundary at the lowest $F(i,j)$.

This is a heuristic method because the original problem is a combinatorial one and practically intractable. There is no guarantee for obtaining global optimum. However, experimentally it works well, as described in the next section.

We describe an entire procedure of our clustering algorithm in **Algorithm 1**.

## IV. EXPERIMENTS

### A. Datasets

We evaluated our proposed method on two Web page corpora. One is a ODP corpus, a set of web pages extracted from the Open Directory Project (ODP)[1] by ourselves. We selected 12 subdirectories from the the "Science" top directory, and downloaded top pages of the Web sites listed in each directory. We removed tags and stopwords from the pages, and stemmed each word. The summary of each directory is listed in Table I.

We treated each directory as a target cluster, and made four datasets using those clusters. One is a dataset (odp_all) using all the directories in the corpus. The other two (odp_odd and odp_even) are half size of odp_all. The directories of odp_odd and odp_even are selected from the odd and even number ones in Table I, respectively. The final one (odp_small) is a set of small directories (No.6∼12) that include under 20 data. We summarize about datasets in Table II.

We also used the WebKB corpus [2]. This corpus consists of seven datasets, and each one has fixed five clusters named "Cornel", "Texas", "Washington", "Wisconsin" and "misc", respectively. Since the last "misc" cluster consists of Web pages from miscellaneous universities and lacks unity, we removed it from each dataset. We also removed department dataset because it has only one page for each cluster. The datasets are summarized in Table III. We applied the same preprocessing as ODP corpus and evaluated each method on all the datasets.

TABLE III.
WEBKB CORPUS

| Dataset | #data | #cluster |
|---|---|---|
| student | 558 | 4 |
| faculty | 153 | 4 |
| staff | 46 | 4 |
| course | 244 | 4 |
| project | 86 | 4 |
| other | 3033 | 4 |

### B. Compared methods

We compared our proposed method with other three methods. The notation and brief introduction of each method is listed below.

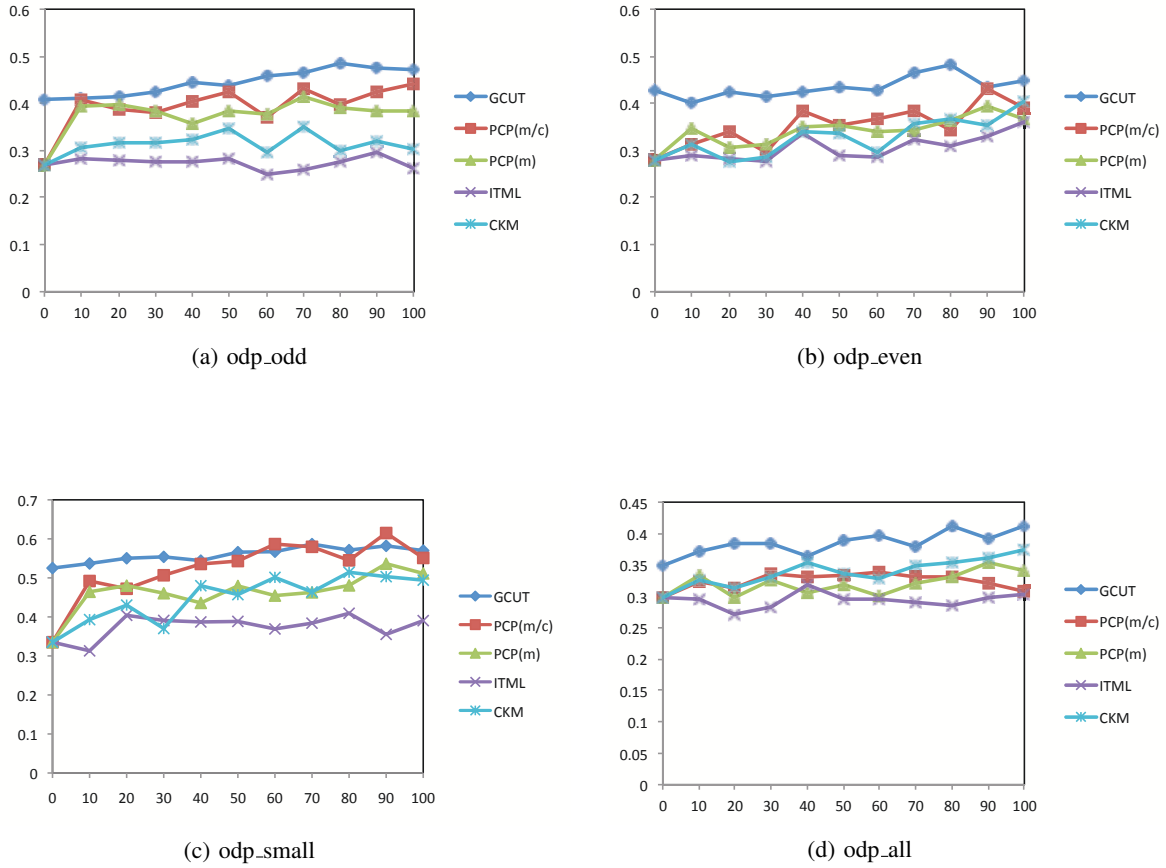
**GCUT** This is our proposed method. We calculated the Euclid distance for the weight of the graph edge, which is indicated as $w_{ij}$ in the maximum graph cut problem in Section II. We used the SDPT3 package [3] to solve SDP. The default parameters are used for each run.

**PCP** PCP is one of the state of the art distance metric learning methods proposed by Li [10]. It learns a kernel matrix using the same SDP formulation with ours. The difference between two methods lies in the usage of the solution matrix. PCP uses it as a kernel matrix for kernel k-means while ours uses it for iterative graph cut. The weight of the graph edge and the parameters for SDPT3 are the same with GCUT. Since this method can use both must/cannot-link constraints, we conducted two trials. One is the trial using only must-link constraints, the other uses both must/cannot-link constraints. the

[1] http://www.dmoz.org/

[2] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[3] http://www.math.nus.edu.sg/ mattohkc/sdpt3.html




(a) odp_odd (b) odp_even

(c) odp_small (d) odp_all

Figure 4. Results of ODP corpus (horizontal axis is the number of constraints and vertical axis is NMI)

former is denoted by PCP(m) and the latter is PCP(m/c).

**ITML** ITML is an another distance metric learning method proposed by Jain [13]. It learns a transform matrix to calculate desired distance. Since this method follows online learning process, we tuned its learning parameter $\eta$ by selecting the best value from 0.1 to 1.0 with 0.1 steps. Clustering is done by normal k-means with learned Euclid distance. Since cannot-link on this method showed terrible performance deterioration, we applied only must-link for it.

**CKM** CKM is the constrained k-means clustering algorithm (called COP-Kmeans) proposed by Wagstaff [14]. Since cannot-link often caused deadlock and stopped all the algorithm procedure, we applied must-link only as well as ITML.

### C. Other settings

We use normalized mutual information (NMI) to measure the clustering accuracy. NMI is calculated by the following formula.

$$\mathrm{NMI}(C,T) = \frac{I(C,T)}{\sqrt{H(C)H(T)}}$$

where $C$ is the set of clusters returned by each algorithm and $T$ is the set of true clusters. $I(C,T)$ is the mutual information between $C$ and $T$, and $H(C)$ and $H(T)$ are the entropies.

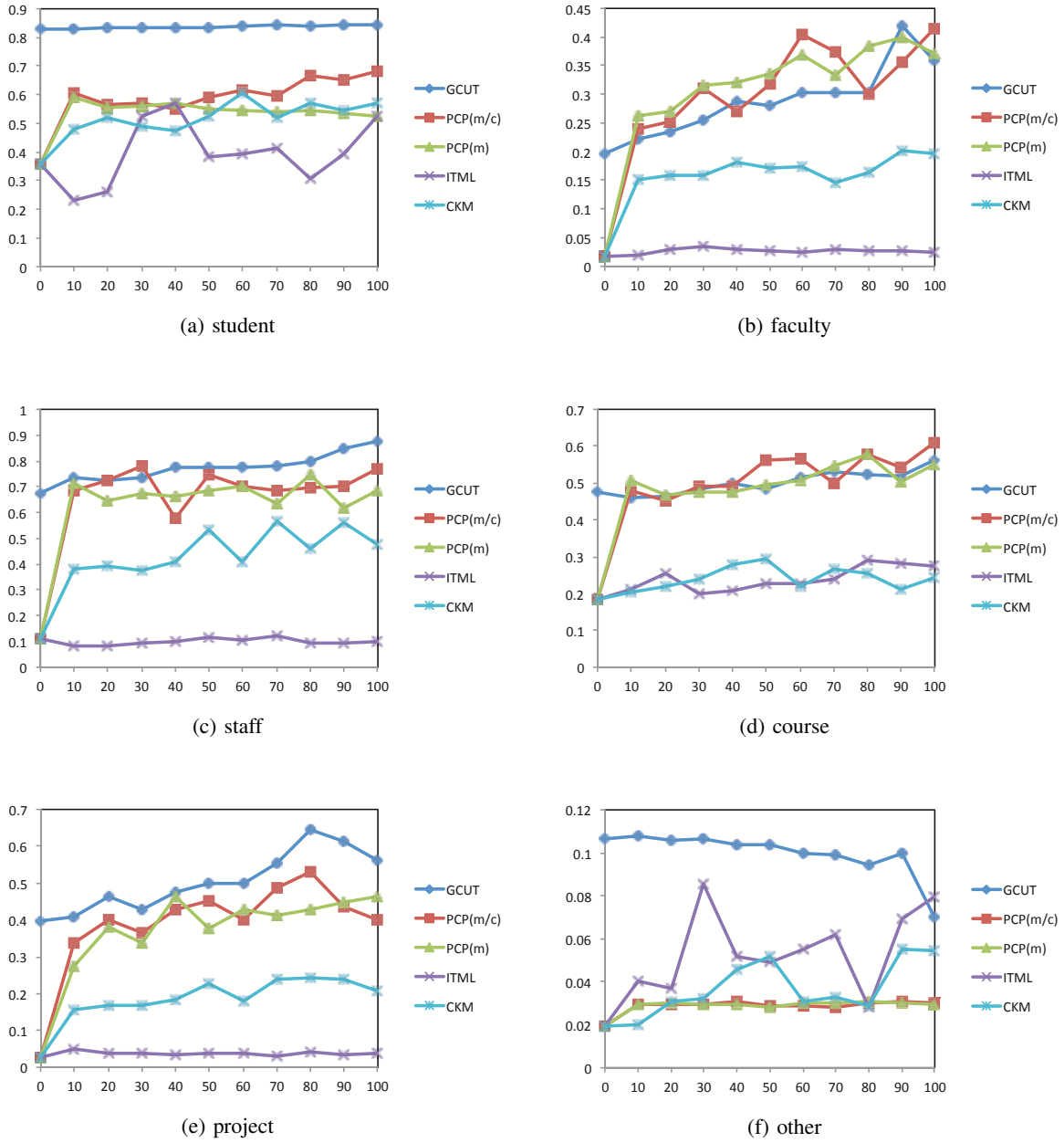
Constraints are selected randomly. We changed the number of constrains from 0 to 100 with 10 steps. For each number of constraints, we selected 10 different sets of constraints and used the same sets for each method. The NMI is calculated as the average value of those 10 sets.

### D. Results

Figure 4 is the results of the ODP corpus. The horizontal axis is the number of constraints and the vertical axis is the value of normalized mutual information (NMI).

In this corpus, GCUT outperformed other methods in all the datasets, especially at the points where the number of constraints is small.

PCP(m/c) showed comparable or slightly better performance at some points in the odp_odd, the odp_even and




(a) student (b) faculty

(c) staff (d) course

(e) project (f) other

Figure 5. Results of the WebKB corpus (horizontal axis is the number of constraints and vertical axis is NMI)

odp_small datasets though it remained the second best or worse in other points. PCP(m) is slightly worse than PCP. CKM outperformed other methods without GCUT in the odp_all dataset. ITML did not show the good results in this corpus.

Figure 5 shows the results of the WebKB corpus. The meaning of the horizontal and vertical axis is the same as the ODP corpus. In this corpus, GCUT indicated the best results in four datasets. The results in the other two datasets is comparable and slightly worse than the other methods. GCUT showed clearly better performance especially in the student and "other" datasets whose number of data is relatively larger than others. The performance of all methods in "other" dataset is low because the topic of this dataset is diverse and does not have unity as a cluster. GCUT also showed better performance than other methods in the staff and project datasets whose number of data is relatively small. In those datasets, performance gap between GCUT and others widened as the number of constraints increased. PCP showed comparable and slightly better performance in the course and faculty datasets. We did not observe significant difference between PCP(m/c) and PCP(m) in this corpus. The performance of CKM and ITML remained worse than GCUT and PCP in all the datasets.

The performance of GCUT as well as all the other

methods does not grow monotonically and sometimes drops despite the increase of constraints. This is because the effectiveness of constraints are generally quite uneven. Some sets of constraints may work better for some methods. On the other hand some sets may bring negative effect (e.g. because of must-link constraints related outliers).

## V. Discussions

Our proposed clustering method (GCUT) repeatedly divides the largest cluster into two sub-clusters until the given number of clusters is obtained. This procedure delivers a good property compared to a one-time multi-class clustering approach represented by PCP as shown in the experiments. The results indicate that we can obtain more accurate clusters if we interpret the SDP problem described in Section II as a constrained graph cut problem and use its solution for iterative two-class clustering.

As described in Section I, our method uses the same SDP formulation as PCP. However the purpose, derivation and interpretation of the solution for our method is different from PCP. In this section, we explain a brief introduction of PCP and clarify the difference from our method, and then discuss the cause of the experimental results in the previous section.

PCP is a method to produce a kernel matrix of a dataset by projecting original feature vectors to a higher dimensional space. Constraints are used as desired inner product values for some selected data pairs in the projected feature space. In order to propagate the effect of the constraints to other data pairs, Li et al. formulated an optimization problem according to a well-known regularization in spectral graph theory.

Optimization Problem derived in PCP

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\sum_{i,j=1}^{n} w_{i,j}\left|\left|\frac{\phi(\mathbf{x}_i)}{\sqrt{d_{ii}}}-\frac{\phi(\mathbf{x}_j)}{\sqrt{d_{jj}}}\right|\right|_F^2 \\
\text{subject to} \quad & <\phi(\mathbf{x}_i),\phi(\mathbf{x}_i)>_F = 1 \quad i=1,2,...,n \\
& <\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)>_F = 1 \quad {}^{\forall}(\mathbf{x}_i,\mathbf{x}_j)\in M \\
& <\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)>_F = 0 \quad {}^{\forall}(\mathbf{x}_i,\mathbf{x}_j)\in C
\end{aligned}
$$

Here, $\phi(\mathbf{x}_i)$ is a feature vector in the projected space, $w_{ij}$ is an initial similarity between data $i$ and $j$, $d_{ii}$ is $\sum_k w_{ik}$, and $<\cdot,\cdot>_F$ indicates inner product calculation in the projected space. The objective function of the above problem can be finally transformed into the formula $L\bullet X$ that is the same one as described in Section II.

In this way, GCUT and PCP solve the same SDP problem except for the use of cannot-link. However their derivations of the problem are different from each other like following.

1) PCP places the SDP problem as a basis to define a proximity measure in a dataset. Thus it solves a SDP problem only once for a clustering trial. The proximity measure is actually got as a kernel matrix.
2) The SDP problem in GCUT is a constrained graph cut clustering problem itself. Thus it iterates to solve SDP $k-1$ times. $k$ is a target number of clusters.

Though the reason for the performance increase of GCUT compared with PCP superficially seems to be multiple SDP executions, it is more important that the derivations of the SDP problem in both methods are different. GCUT derives from a constrained graph cut problem while PCP derives from a problem to define a desired proximity measure.

## VI. Conclusions

In this paper, we proposed a constrained clustering method that is based on a graph-cut problem formalized by semi-definite programming and deterministic iterative two-class partitioning approach. While graph-cut based clustering is a particularly promising way to improve conventional techniques like k-means method, few methods have been proposed, which can naturally incorporate constraint like must-link.

Our method has the advantages of more convenient constraint incorporation compared to other graph-cut based method such as spectral clustering and can utilize SDP's solution matrix more appropriately compared with other SDP-based methods. Since our method adopts deterministic clustering approach unlike k-means using random seeds, the performance is stable and robust to outliers. These advantages were clearly demonstrated through experiments using many datasets from two Web corpus. Results showed that our proposed clustering method constantly outperformed conventional methods in many cases and utilized constraints effectively.

A few problems still remain in this work. First, we need to investigate how the constraint quality influences the clustering. This is an active learning problem and in the future we aim to develop a new active-learning technique by utilizing the properties of our clustering methods. Second, we need to develop a constraint propagation method to improve clustering accuracy even if the number of constraint is small. There are already various propagation methods in place [15], but we need something more powerful to improve the effectiveness of very small constraints.

The advantages of our proposed clustering is efficiency, especially in the case of small number of constraints. Thus we are planning to apply this clustering method to interactive (Web) clustering with GUI [12].

## References

[1] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons, 2003.
[2] Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1101–1113, 1993.

[3] S. Basu, A. Banjeree, E. Mooney, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *In Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04*, 2004, pp. 333–344.
[4] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*. The MIT Press, 2006.
[5] S. Shwartz, Y. Singer, and A. Y. Ng, "Online and batch learning of pseudo-metrics," in *Proceedings of the 21st International Conference on Machine Learning*, 2004, pp. 94–101.
[6] W. Tang, H. Xiong, S. Zhong, and J. Wu, "Enhancing semi-supervised clustering: A feature projection perspective," in *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 707–716.
[7] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, June 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1577069.1577078
[8] S. X. Yu and J. Shi, "Segmentation given partial grouping constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 173–183, 2004.
[9] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Neural Information Processing Systems*, no. 14, 2001, pp. 849–856.
[10] Z. Li, J. Liu, and X. Tang, "Pairwise constraint propagation by semidefinite programming for semi-supervised classification," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 576–583.
[11] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Learning nonparametric kernel matrices from pairwise constraints," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 361–368.
[12] M. Okabe and S. Yamada, "An interactive tool for human active learning in constrained clustering," *Journal of Emerging Technologies in Web Intelligence*, vol. 3, no. 1, pp. 20–27, 2011.
[13] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman, "Online metric learning and fast similarity search," in *Proceedings of Neural Information Processing Systems*, 2008, pp. 761–768.
[14] K. Wagstaff and S. Roger, "Constrained k-means clustering with background knowledge," in *Proceedings of the 18th International Conference on Machine Learning*, 2001, pp. 577–584.
[15] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," in *Proceedings of the 19th International Conference on Machine Learnning*, 2002, pp. 307–314.

**Masayuki Okabe** is an assistant professor at Toyohashi University of Technology. He received B.S. (1996) degree from Soka University and M.S. (1998) and the Ph.D. (2001) degrees from Tokyo Institute of Technology. His research interests include information retrieval, machine learning and data mining. He is a member of The Japanese Society for Artificial Intelligence.

**Seiji Yamada** is a professor at the National Institute of Informatics. Previously he worked at Tokyo Institute of Technology. He received B.S. (1984), M.S. (1986) and the Ph.D. (1989) degrees in artificial intelligence from Osaka University. His research interests are in the design of intelligent interaction including Human-Agent Interaction, intelligent Web interaction and interactive machine learning. He is a member of IEEE, AAAI, ACM, JSAI, IPSJ and HIS.

# Careful Seeding Method based on Independent Components Analysis for k-means Clustering

Takashi Onoda
System Engineering Lab., CRIEPI, Tokyo, JAPAN. Email: onoda@criepi.denken.or.jp
Miho Sakai
Tokyo Institute of Technology, Yokohama, JAPAN. Email: sakai@ntt.dis.titech.ac.jp
Seiji Yamada
National Institute of Informatics/SOKENDAI/Tokyo Institute of Technology, Tokyo, JAPAN
Email: seiji@nii.ac.jp

***Abstract*— The k-means clustering method is a widely used clustering technique for the Web because of its simplicity and speed. However, the clustering result depends heavily on the chosen initial clustering centers, which are uniformly chosen at random from the data points. We propose a seeding method that is based on the independent component analysis for the k-means clustering method. We evaluate the performance of our proposed method and compare it with other seeding methods by using benchmark datasets. We also applied our proposed method to a Web corpus, which was provided by ODP, and the CLUTO datasets. The results from the experiments showed that the normalized mutual information of our proposed method is better than the normalized mutual information of the k-means clustering method, the KKZ method, and the k-means++ clustering method.**

***Index Terms*— k-means clustering method, KKZ method, k-means++ clustering method, independent components analysis, seeding**

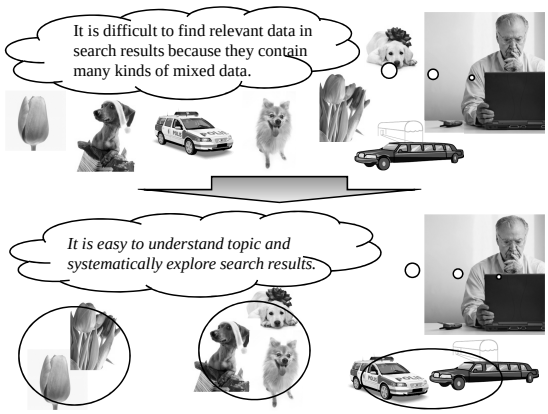



Figure 1. Effect of clustering the results of Web search.

## I. Introduction

Clustering is one of the most useful unsupervised learning in data mining[1][2]. It has been applied to various fields and used widespread both in research and business[3]. We are interested in application of clustering to the Web clustering. The Web clustering[4] has a very wide ranges including clustering searched results[5], [6], [7], [8], clustering Web pages/sites[9], [10], [11], clustering Web multimedia[12] and so on. Especially, we focus on clustering of Web searched results because our final research objective is to build IWI (Intelligent Web Interaction) systems. While Web search engines are definitely good for certain search tasks such as finding an organization's Web page, they may be less effective at satisfying ambiguous queries. The results on different subtopics or meanings of the input query also will come together in a hit list, thus implying that the user may have to sift through a large number of irrelevant items to locate those of interest. On the other hand, there is no way to estimate what is relevant to the user given that the queries are usually very short and their interpretation is inherently ambiguous in the absence of context.

An approach for clustering the results of Web search[5], [6], [7], [8] is different from one for retrieving information from the Web. This clustering approach shows the results, which are manually or automatically associated with clusters that consist of similar items (Figure 1).

We consider this clustering of Web searched results should be quick and accurate because a user never wait for the clustered results so long. In particular, we are interested in the simplest and quickest clustering method. Therefore, we deal with the k-means clustering method in our research. We particularly discuss how to solve the problem of "seeding" in the k-means clustering method[13], [14], [15].

The rest of this paper is organized as follows. Section II discusses the related work and the k-means clustering, the KKZ clustering, and the k-means++ clustering methods. Section III discusses the problem with these clustering methods and introduces our proposed method. Section IV presents our experimental results along with comparison of the performance of the proposed method with those of the k-means clustering, the KKZ clustering, and k-means++ clustering methods. Section V concludes this research.


doi:10.4304/jetwi.4.1.51-59

## II. Related Works

Clustering is a classic problem in machine learning and computational geometry. In the popular k-means formulation, one is given an integer $k$ and a set of $n$ data points $\mathbf{X} \subset \mathbf{R}^m$. $k$ is the number of cluster centers. The goal is to choose $k$ centers $\mathcal{C}$ to minimize the sum of the squared distances between each point and its closest center.

$$\phi = \sum_{\mathbf{x} \in \mathbf{X}} \min_{\mathbf{c} \in \mathcal{C}} \|\mathbf{x} - \mathbf{c}\|^2 \tag{1}$$

Solving this problem is NP-hard, even with just two clusters[16], however Lloyd[17] proposed a local search solution 25 years ago that is still widely used today.

In this section, we formally define the k-means clustering method, the KKZ clustering method and the k-means++ clustering method.

### A. k-means clustering method

The k-means clustering method is simple and fast and locally improves the centers of mass of clusters. It works as follows.

1) Arbitrarily choose $k$ initial centers $\mathcal{C} = \mathbf{c}_1, \cdots, \mathbf{c}_k$,
2) For each $i \in \{1, \ldots, k\}$, set the cluster $\mathbf{c}_i$ to be the set of points in $\mathbf{X}$ that are closer to $\mathbf{c}_i$ than they are to $\mathbf{c}_j$ for all $j \neq i$.
3) For each $i \in \{1, \ldots, k\}$, set $\mathbf{c}_i$ to be the center of the mass of all the points in a set $C_i$ of cluster $i$: $\mathbf{c}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$.
4) Repeat steps 2) and 3) until $\mathbf{c}_i$ no longer changes.

It is standard practice to uniformly choose the initial centers at random from $\mathbf{X}$. For Step 2), the ties may be arbitrarily broken, as long as the method is consistent. Steps 2) and 3) are both guaranteed to decrease $\phi$; therefore, the method makes local improvements to an arbitrary cluster until it is no longer possible to do so.

The k-means method is attractive in practice because it is simple and generally fast. Unfortunately, it is guaranteed only to find a local optimum, which can often be quite poor.

### B. KKZ clustering method

The KKZ method was proposed by Katsavounidis et al. [18]. This method calculates the entire distance among the data and finds the data with a wide distance. The data are selected as the initial cluster centers. At any given time, let $D(\mathbf{x})$ denote the shortest distance from a data point $\mathbf{x}$ to the closest center we have already chosen. Then, the following clustering method is defined as the KKZ clustering method[18].

1a) Choose initial centers $\mathbf{c}_1$ and $\mathbf{c}_2$. The distance between $\mathbf{c}_1$ and $\mathbf{c}_2$ is the widest of all distance between a data point and the other data point (Figure 2).
1b) For all data, $D(\mathbf{x}_j), j \in \{1, \cdots, n\}$ are calculated (Figure 3).

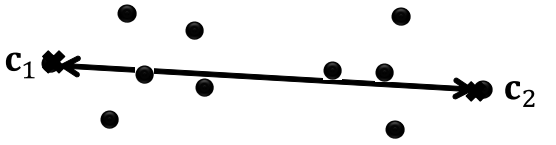



Figure 2. Initial centers of KKZ method

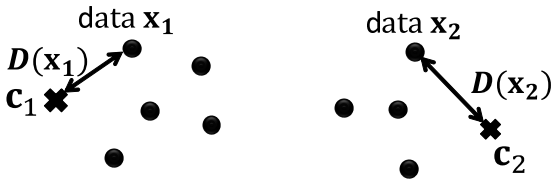



Figure 3. Distance $D(\mathbf{x})$

1c) Choose the next center $\mathbf{c}_i$, selecting $\mathbf{c}_i = \mathbf{x}' \in \mathbf{X}$ with the widest distance $D(\mathbf{x}')$ (Figure 4).
1d) Repeat step 1b) until we have chosen a total of $k$ centers.

Steps 2)-4) proceed just like that for the standard k-means algorithm.

The KKZ method is attractive in practice because it is simple for decision of unique initial centers. However, the KKZ method sometimes find bad clusters because unfortunately it depends on outlier data points.

### C. k-means++ clustering method

The k-means method begins with an arbitrary set of cluster centers. k-means++ clustering proposes specifically choosing these centers. At any given time, let $D(\mathbf{x})$ denote the shortest distance from a data point $\mathbf{x}$ to the closest center we have already chosen. Then, the following clustering method is defined as the k-means++ clustering method[19].

1a) Choose an initial center $\mathbf{c}_1$ uniformly at random from $\mathbf{X}$.
1b) For all data, $D(\mathbf{x}_j); j \in \{1, \cdots, n\}$ are calculated (Figure 3).
1c) Randomly generate a real value $L$ satisfying the following equation.

$$0 < L \leq \sum_{\mathbf{x} \in \mathbf{X}} D(\mathbf{x}_j) \tag{2}$$

1d) Choose the next center $\mathbf{c}_i$, selecting the $\mathbf{c}_i = \mathbf{x}_j$ with satisfying the following equation (Figure 5).

$$\sum_{m=1}^{j-1} D(\mathbf{x}_m) < L \leq \sum_{m=1}^{j} D(\mathbf{x}_m) \tag{3}$$

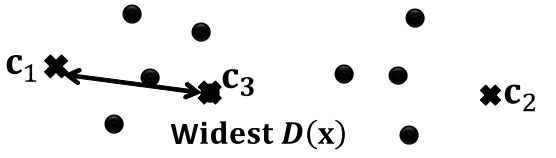



Figure 4. Next center $\mathbf{c}_i$ of KKZ method

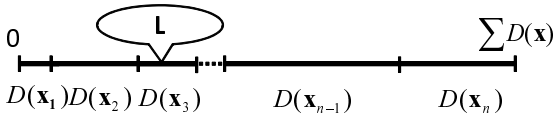



Figure 5. Next center $\mathbf{c}_i$ of k-means++ method

1e) Repeat step 1b) until we have chosen a total of $k$ centers.

Steps 2)-4) proceed in the same way as with the standard k-means clustering method. We call the weighting used in Step 1b) simply "$D^2$ weighting".

## III. Proposed Method

This section describes a problem with the k-means and the k-means++ clustering methods. Then, we propose a k-means combined with an Independent Component Analysis (ICA)[20], [21], [22] based seeding method.

### *A. Problem for k-means and k-means++ clustering methods*

We have six data points, which consist of $\mathbf{x}_i$ ($i$ =1, ..., 6) and these points are divided into two clusters. Figure 6 shows these six data points.

In addition, Figure 7 shows the global optimal clustering result for these six data points. The first cluster consists of $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}$ and the other consists of $\{\mathbf{x}_3, \mathbf{x}_6\}$. We assume that most of clustering methods can find the global optimal clusters. However, the k-means clustering method generates bad clusters if $\mathbf{x}_2$ and $\mathbf{x}_5$ are chosen as the initial $\mathbf{c}_1$ and $\mathbf{c}_2$ cluster centers. Figure 8 shows the local optimal clusters, which are bad clusters. The k-means++ clustering method was developed to avoid this bad clustering.

However, the k-means++ clustering method sometimes generates bad clusters because it depends on the choice of the initial center $\mathbf{c}_1$. The initial center $\mathbf{c}_1$ is chosen uniformly at random from $\mathbf{X}$.

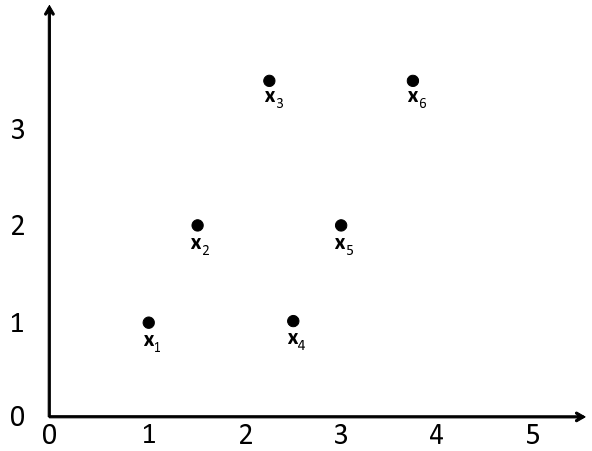



Figure 6. Given Data

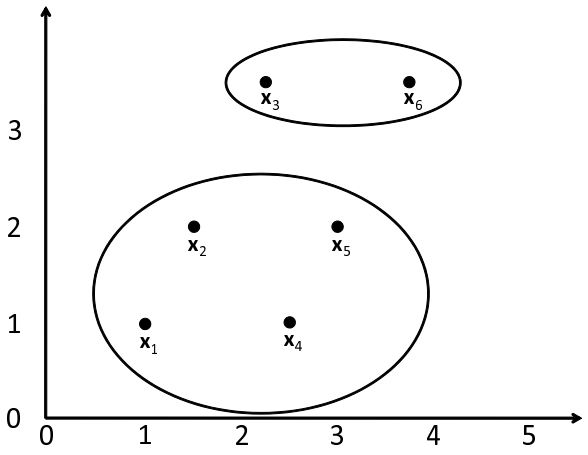



Figure 7. Global Optimal Clustering Case

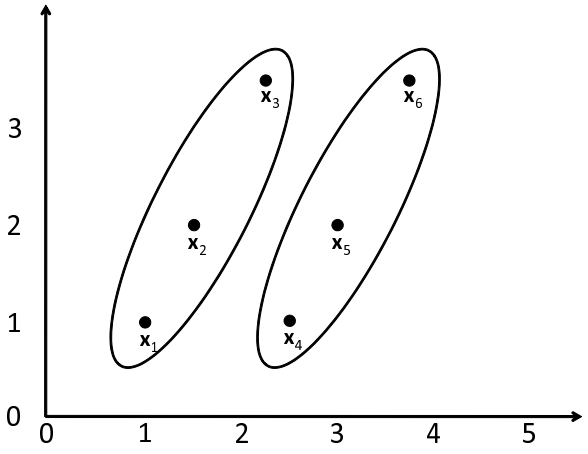



Figure 8. Local Optimal Clustering Case

### *B. k-means combined with ICA based seeding method*

The k-means clustering method begins with an arbitrary set of cluster centers. The k-means++ clustering method begins with a small arbitrary set of cluster centers. As stated above, we propose a method for specifically choosing these centers. At any given time, we can obtain independent components (ICs) from given data $\mathbf{X}$. Then, we define the following seeding method.

1a) Extract $k$ independent components $\mathbf{IC}_1, \cdots, \mathbf{IC}_k$ from given data $\mathbf{X}$ (Figure 9).

1b) Choose $k$ initial centers $\mathbf{c}_i$ ($i = 1, \cdots, k$), selecting $\mathbf{c}_i = \mathbf{x}' \in \mathbf{X}$ with a minimum $\frac{\mathbf{IC}_i \cdot \mathbf{x}'}{|\mathbf{IC}_i||\mathbf{x}'|}$ (Figure 10).

Steps 2)-4) proceed in the same way as with the standard k-means clustering method. Figure 11 shows the concept of the k-means clustering method combined with the ICA based seeding method. In the figure 11, $\mathbf{IC}_1$ and $\mathbf{IC}_2$ denote independent components. The each independent component may become an initial seed to generate the global optimal clustering case.

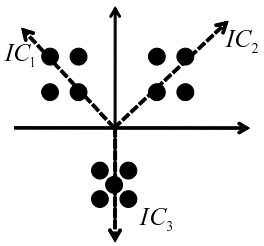



Figure 9. Independent components

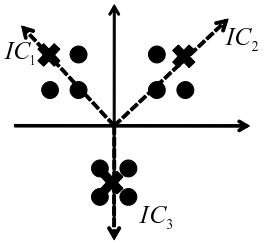



Figure 10. Initial centers $\mathbf{c}_i$ of ICA based method

## IV. Experimental condition

To evaluate the k-means clustering method, KZZ method, k-means++ clustering method and proposed method in practice, we implemented and tested them in matlab. In this section, we briefly explain about the data sets that were used for the experiments, the evaluation metrics in the experiments, some compared seeding methods and the results of the experiments. We found that the k-means clustering method combined with the ICA based seeding method performed well in the experiments.

### A. Data sets

We evaluated the performance of the k-means clustering method, KKZ method, k-means++ clustering method, and the proposed method using on two kinds of data sets. One contained a small amount of data and the other a large data. The small data set consisted of the UCI Machine Learning repository and Open Directory Project(ODP) Web corpus. The large data set consisted of the CLUTO data sets.

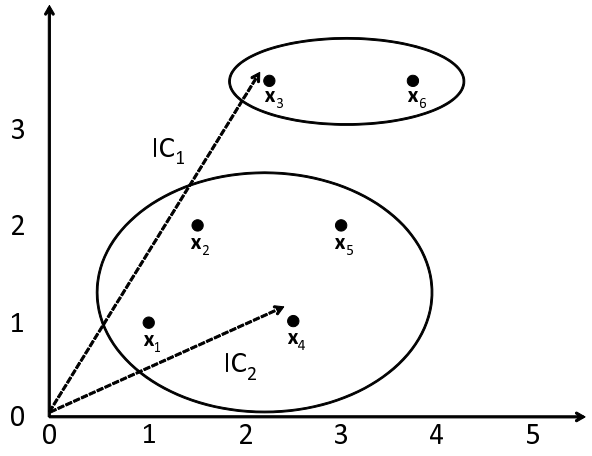



Figure 11. Concept of Our Proposed Method

TABLE I.
No. of clusters, attributes, and samples for UCI repository data sets

| Data set | No. of clusters | No. of attributes | No. of samples |
|---|---|---|---|
| *Iris* | 3 | 4 | 150 |
| *Wine* | 3 | 13 | 178 |
| *Soybean -Small* | 4 | 35 | 47 |
| *Breast -Cancer* | 2 | 9 | 683 |

TABLE II.
No. of directories, attributes and samples for ODP corpus data set

| Data set | No. of clusters | No. of attributes | No. of samples |
|---|---|---|---|
| *ODP* | 4 | 340 | 72 |

*1) UCI Machine Learning repository:* The UCI Machine Learning repository had four data sets in our experiments. The first data set, *iris*, consisted of 50 samples from each of three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). The second data set, *wine*, contained the results of a chemical analysis on wines produced in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wine. The third data set, *soybean-small*, was for diagnosing four soybean diseases. The data set consisted of 47 samples and 35 attributes. The fourth data set, *breast-cancer*, contained diagnosis results of breast cancer. The data set consisted of 683 samples and 9 attributes. Table I lists the numbers of samples, the numbers of attributes and the numbers of clusters in the data sets used in our experiments.

*2) ODP corpus data:* We used the ODP Web corpus data set for our test experiment. The ODP Web corpus was extracted from the Open Directory Project[1] by ourselves. We selected twelve subdirectories from the "Science" top directory, and downloaded top pages of the web sites listed in each directory. We removed tags and stopwords from the pages, and stemmed each word. The summary of each directory is listed in Table I. We treated each directory as a target cluster, and made four datasets using those clusters. Table II lists the number of samples, the number of attributes, and the number of directories of the data sets used in our experiments.

*3) CLUTO data sets:* CLUTO[2] is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of various clusters. In our experiments, seven CLUTO data sets were adopted. The seven CLUTO data sets are *tr11*, *tr12*, *tr31*, *tr41*, *tr45*, *k1b*, and *re1*. Table III describes statistics on CLUTO

[1]http://www.dmoz.org/
[2]http://glaros.dtc.umn.edu/gkhome/views/cluto

TABLE III.
NO. OF CLUSTERS, ATTRIBUTES AND SAMPLES FOR CLUTO DATA SETS

| Data set | No. of clusters | No. of attributes | No. of samples |
|---|---|---|---|
| *tr11* | 9 | 6429 | 414 |
| *tr12* | 8 | 5804 | 313 |
| *tr31* | 7 | 10128 | 927 |
| *tr41* | 10 | 7454 | 878 |
| *tr45* | 6 | 8261 | 690 |
| *k1b* | 6 | 21839 | 2340 |
| *re1* | 25 | 3758 | 1657 |

data sets including the number of samples, the number of attributes and the number of clusters.

### B. Evaluation Metrics

We used normalized mutual information as a metric to evaluate the qualities of the clustering outputs from the different methods. The normalized mutual information measures the consistency of the clustering output compared to the ground truth. It reaches a maximum value of 1 only if the membership $\phi_c$ perfectly matches $\phi_g$ and a minimum of zero if the assignments of $\phi_c$ and $\phi_g$ are independent. The membership function $\phi_c(\mathbf{x})$ is the mapping of a point $\mathbf{x}$ to one of the $k$ clusters. The membership $\phi_g(\mathbf{x})$ represents the true cluster label for $\mathbf{x}$. Formally, the normalized mutual information is derived using the following equation

$$NMI(\phi_g, \phi_c) = \frac{MI(\phi_g, \phi_c)}{\max(H(\phi_g), H(\phi_c))} \tag{4}$$

where $MI(\phi_g, \phi_c)$ denotes the next equation

$$MI(\phi_g, \phi_c) = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{g,c}(i,j) \log \frac{p_{g,c}(i,j)}{p_g(i)p_c(j)} \tag{5}$$

$H(\phi_g)$ comes from the following equation

$$H(\phi_g) = \sum_{i=1}^{k} p_g(i) \log \frac{1}{p_g(i)} \tag{6}$$

and $H(\phi_c)$ denotes the next equation

$$H(\phi_c) = \sum_{j=1}^{k} p_c(j) \log \frac{1}{p_c(j)} \tag{7}$$

The $p_g(i)$ is the percentage of points in cluster $i$ based on the ground truth, i.e.

$$p_g(i) = \frac{\sum_{l=1}^{n} 1(\phi_g(\mathbf{x}_l) - i)}{n}. \tag{8}$$

Similarly, $p_c(j)$ denotes the following equation

$$p_c(j) = \frac{\sum_{l=1}^{n} 1(\phi_c(\mathbf{x}_l) - j)}{n} \tag{9}$$

and $p_{g,c}(i,j)$ is the percentage of points that belong to cluster $i$ in $\phi_g$ and also cluster $j$ in $\phi_c$, i.e.

$$p_{g,c}(i,j) = \frac{\sum_{l=1}^{n} 1(\phi_g(\mathbf{x}_l) - i)1(\phi_c(\mathbf{x}_l) - j)}{n}. \tag{10}$$

The above defined metrics were used to evaluate the accuracy of the k-means clustering method, KKZ method, k-means++ clustering method and the proposed methods.

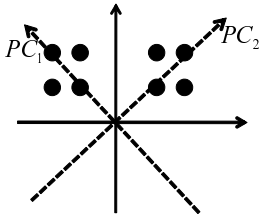



Figure 12. Principal components

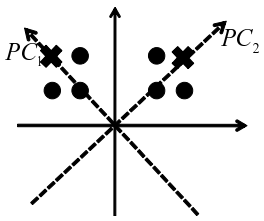



Figure 13. Initial centers $\mathbf{c}_i$ of PCA based method

### C. Compared methods

In our experiments, we tried to compare the performance of our proposed methods with the performance of other methods. The other methods consist of the k-means clustering method, KKZ method, and k-means++ clustering method (See section II). Our proposed methods are based on a k-means combined with an ICA based seeding method and k-means combined with a PCA[23] based seeding method. k-means combined with ICA based seeding method was explained in Section III. Now, we briefly explain the k-means combined with a PCA based seeding method.

At any given time, we can obtain principal components (PCs) from the given data $\mathbf{x}$. Then, we define the following seeding method.

1a) Extract $k$ principal components $\mathbf{PC}_1, \ldots, \mathbf{PC}_k$ from given data $\mathbf{X}$ (Figure 12).
1b) Choose $k$ initial centers $\mathbf{c}_i$ ($i = 1, \cdots, k$), selecting $\mathbf{c}_i = \mathbf{x}' \in \mathbf{X}$ with a minimum $\frac{\mathbf{PC}_i \cdot \mathbf{x}'}{|\mathbf{PC}_i||\mathbf{x}'|}$ (Figure 13).

Steps 2)-4) proceed in the same way as with the standard k-means clustering method. Figure 14 shows the concept of the k-means clustering method combined with a PCA based seeding method.

## V. EXPERIMENTS

This section discusses some of the experimental results under the above experimental condition.

### A. Experimental results for small data sets

The k-means and k-means++ clustering methods were executed 100 times using different initializations over all four data sets from the UCI repository[3]. In our experiments, the Euclid distance was used as a similarity measure when the k-means clustering method was applied to the UCI repository. The KKZ method and the

[3]http://archive.ics.uci.edu/ml/

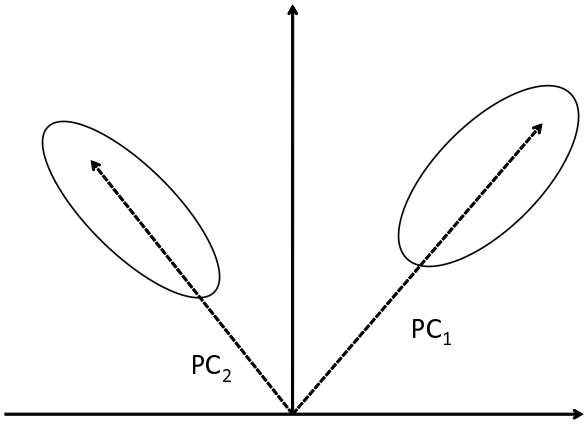



Figure 14. Concept of Our Proposed Method based on PCA

TABLE IV.
EXPERIMENTAL RESULTS FOR *iris* DATA SET

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.751 | 0.751 | 0.532 | 0.703 |
| k-means++ | - | 0.751 | 0.751 | 0.532 | 0.749 |
| KKZ | 0.751 | - | - | - | - |
| PCA | 0.751 | - | - | - | - |
| ICA | 0.751 | - | - | - | - |

proposed method were executed only one time because a unique initial seeding can be set up. Table IV lists the experimental results for the *iris* data set. Table V lists the experimental results for the *wine* data set. Table VI lists the experimental results for the *soybean-small* data set. Table VII lists the experimental results for the *breast-cancer* data set. Tables IV, V, VI, and VII have an averaged $NMI$, a maximum $NMI$, a minimum $NMI$, and a $NMI$ when the clusters achieved minimum variance.

In IV, V, and VII tables, the $NMI$s of our proposed method are the same as the maximum $NMI$s of the k-means clustering method and the k-means++ clustering method. The $NMI$s of our proposed methods are achieved by using only one initial seeding. Therefore, IV, V, and

TABLE V.
EXPERIMENTAL RESULTS FOR *wine* DATA SET

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.429 | 0.429 | 0.387 | 0.418 |
| k-means++ | - | 0.429 | 0.429 | 0.387 | 0.418 |
| KKZ | 0.387 | - | - | - | - |
| PCA | 0.429 | - | - | - | - |
| ICA | 0.429 | - | - | - | - |

TABLE VI.
EXPERIMENTAL RESULTS FOR *soybean-small* DATA SET

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.711 | 1.000 | 0.518 | 0.714 |
| k-means++ | - | 0.711 | 1.000 | 0.711 | 0.806 |
| KKZ | 0.711 | - | - | - | - |
| PCA | 0.711 | - | - | - | - |
| ICA | 0.711 | - | - | - | - |

TABLE VII.
EXPERIMENTAL RESULTS FOR *breast-cancer* DATA SET

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.743 | 0.743 | 0.743 | 0.743 |
| k-means++ | - | 0.743 | 0.743 | 0.743 | 0.743 |
| KKZ | 0.743 | - | - | - | - |
| PCA | 0.743 | - | - | - | - |
| ICA | 0.743 | - | - | - | - |

TABLE VIII.
EXPERIMENTAL RESULTS FOR *ODP Web corpus* DATA SET

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.555 | 0.589 | 0.392 | 0.514 |
| k-means++ | - | 0.555 | 0.589 | 0.425 | 0.525 |
| KKZ | 0.531 | - | - | - | - |
| PCA | 0.500 | - | - | - | - |
| ICA | 0.638 | - | - | - | - |

VII tables show that the proposed method outperforms both the k-means clustering method and the k-means++ clustering method for the *iris*, *wine*, and *breast-cancer* data sets of the UCI repository.

In IV and VII tables, the $NMI$s of the KKZ method are the same as the maximum $NMI$s of the k-means clustering method and the k-means++ clustering method. The $NMI$s of the KKZ method are achieved by using only one initial seeding. Therefore, the IV and VII tables also show that the KKZ method outperforms both the k-means clustering method and the k-means++ clustering method for the *iris* and *breast-cancer* data sets of the UCI repository. However the V table shows that the performance of the KKZ method is the worst among the compared methods in our experiments.

We generally cannot provide true cluster data. Having a $NMI$ with minimum variance is the most important issue for real-world applications. Table VI shows that the $NMI$s of our proposed method are the same as the $NMI$s of the k-means clustering method and k-means++ clustering method when the clusters achieved minimum variance. This situation shows that the performance of our proposed method is the same as the performance of the k-means clustering method and the k-means++ clustering method for the *soybean-small* data set. And the $NMI$ with minimum variance is achieved by using only one initial seeding.

In our experiments, the k-means clustering and k-means++ clustering methods run 100 times using different initializations for the *ODP Web corpus* data set. The proposed method runs only one time because it can set up a unique initial seeding. Table II lists the experimental results of the *ODP Web corpus* data set. When the k-means clustering method was applied to an ODP corpus, the cosine distance was used as a similarity measure in our experiments. The KKZ method and the proposed method were executed only one time because they can set up a unique initial seeding.

Table VIII shows that the $NMI$ of our proposed method is better than the $NMI$ of the k-means clustering and k-

means++ clustering methods when the clusters achieved a minimum variance for the *ODP Web corpus* data set. Table VIII shows that the $NMI$ of the proposed method was 0.638. This value is better than the maximum $NMI$ of the k-means clustering method and the maximum $NMI$ of k-means++ clustering method. In addition, the $NMI$ of the proposed method is better than the $NMI$ of the KKZ method and the $NMI$ of PCA based method. Therefore, Table II shows that the proposed method outperforms the k-means clustering method, k-means++ clustering method, KKZ method and PCA based method for the *ODP Web corpus* data set. In addition, the best $NMI$ is achieved by using only one initial seeding.

### B. Experimental results for large data sets

The k-means and k-means++ clustering methods were executed $100$ times using different initializations for all seven data sets from the CLUTO data sets. In our experiments, the cosine distance was used as a similarity measure when k-means clustering method was applied to CLUTO data sets. KKZ method and the proposed method were executed only one time because they could set up a unique initial seeding. Table IX lists the experimental results for the seven CLUTO data sets.

It is difficult to understand aspects of performance of compared methods from table IX. Therefore, we would now like to introduce the ratio between the number of attributes and the number of samples. We could order the results from the CLUTO data sets by using the value of the ratio. Table X lists the ordered results of CLUTO data sets. The underlined performances indicate the best ones. The ratio denotes the following equation

$$ratio = \frac{\text{No. of attributes}}{\text{No. of samples}}. \tag{11}$$

The $NMI$s of the k-means and k-means++ are $NMI$s with the minimum variance in table IX.

When the ratio is smaller than $10.00$, table X shows that shows the performance the proposed method based on ICA is better than the performance of the k-means clustering method, k-means++ clustering method and KKZ method. In other words, the $NMI$s of the proposed method is better than the $NMI$s of the k-means clustering method, k-means++ clustering method and KKZ method for *k1b*, *tr41* and *re1* CLUTO data sets. However, when the ratio is larger than $10.00$, the performance of the proposed method based on ICA is not better than the performances of the k-means clustering method, k-means++ clustering method and KKZ method in table X. In other words, the $NMI$s of the proposed method are not better than the $NMI$s of k-means clustering method, k-means++ clustering method, and KKZ method for *tr12*, *tr11*, *tr45*, and *tr31* CLUTO data sets. When the ratio is larger than $10.00$, the number of attributes is much larger than the number of samples and it is difficult to find stable independent components. Therefore, the proposed method with ICA may not be able to perform well.

TABLE IX.
EXPERIMENTAL RESULTS FOR *CLUTO* DATA SETS

tr11

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.630 | 0.730 | 0.523 | 0.635 |
| k-means++ | - | 0.669 | 0.717 | 0.545 | 0.632 |
| KKZ | 0.578 | - | - | - | - |
| PCA | 0.619 | - | - | - | - |
| ICA | 0.585 | - | - | - | - |

tr12

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avrg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.664 | 0.752 | 0.521 | 0.664 |
| k-means++ | - | 0.621 | 0.689 | 0.425 | 0.621 |
| KKZ | 0.683 | - | - | - | - |
| PCA | 0.500 | - | - | - | - |
| ICA | 0.638 | - | - | - | - |

tr31

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.579 | 0.676 | 0.377 | 0.503 |
| k-means++ | - | 0.523 | 0.641 | 0.392 | 0.507 |
| KKZ | 0.439 | - | - | - | - |
| PCA | 0.504 | - | - | - | - |
| ICA | 0.438 | - | - | - | - |

tr41

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.623 | 0.699 | 0.533 | 0.611 |
| k-means++ | - | 0.651 | 0.730 | 0.530 | 0.620 |
| KKZ | 0.584 | - | - | - | - |
| PCA | 0.680 | - | - | - | - |
| ICA | 0.667 | - | - | - | - |

tr45

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.769 | 0.787 | 0.594 | 0.696 |
| k-means++ | - | 0.794 | 0.794 | 0.564 | 0.697 |
| KKZ | 0.660 | - | - | - | - |
| PCA | 0.744 | - | - | - | - |
| ICA | 0.722 | - | - | - | - |

k1b

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.537 | 0.649 | 0.421 | 0.518 |
| k-means++ | - | 0.523 | 0.611 | 0.422 | 0.521 |
| KKZ | 0.491 | - | - | - | - |
| PCA | 0.438 | - | - | - | - |
| ICA | 0.587 | - | - | - | - |

re1

| method | *NMI* | *NMI* with min. variance | max. *NMI* | min. *NMI* | avg. *NMI* |
|---|---|---|---|---|---|
| k-means | - | 0.541 | 0.578 | 0.460 | 0.523 |
| k-means++ | - | 0.545 | 0.575 | 0.465 | 0.545 |
| KKZ | 0.557 | - | - | - | - |
| PCA | 0.557 | - | - | - | - |
| ICA | 0.575 | - | - | - | - |

### C. Computational costs

Next, we explain the computational cost of the proposed method from an experimental point of view. In our experiments, we used a Windows Vista 32 bit machine that has an Intel Core 2 Duo E8600 3.34 GHz and 4 GB memory. Table XI lists computational times of four UCI repository data sets.

We can find from this table that the computational time of the proposed method is larger than the computational time of the k-means clustering method for the four UCI repository data sets. The k-means clustering method was

TABLE X.
ORDERED EXPERIMENTAL RESULTS BASED ON RATIO (NO. OF ATTRIBUTES/NO. OF SAMPLES) FOR *CLUTO* DATA SETS

| data set | ratio | proposed | k-means | k-means++ | KKZ |
|---|---|---|---|---|---|
| tr12 | 18.54 | 0.638 | 0.664 | 0.621 | 0.683 |
| tr11 | 15.52 | 0.585 | 0.630 | 0.669 | 0.578 |
| tr45 | 11.97 | 0.722 | 0.769 | 0.794 | 0.660 |
| tr31 | 10.92 | 0.438 | 0.579 | 0.523 | 0.439 |
| k1b | 9.33 | 0.587 | 0.537 | 0.523 | 0.491 |
| tr41 | 8.49 | 0.667 | 0.623 | 0.651 | 0.584 |
| re1 | 2.26 | 0.575 | 0.541 | 0.545 | 0.557 |

TABLE XI.
COMPUTATIONAL TIMES (SEC.) FOR UCI REPOSITORY DATA SETS

| method | *iris* | *wine* | *breast-cancer* | *soybean-small* |
|---|---|---|---|---|
| k-means | 0.0032 | 0.0034 | 0.0059 | 0.0041 |
| proposed | 0.0748 | 0.0948 | 0.0643 | 0.0773 |

executed 100 times using different initializations for all four data sets of the UCI repository.

Table XII lists computational times of the *tr45 CLUTO* data set. We can find from this table that the computational time of the proposed method is smaller than the computational time of the k-means clustering method for a *CLUTO* data set. The k-means clustering method was executed 100 times using different initializations for all four data sets of the UCI repository. In other words, the computational time of the proposed method is smaller than the computational time of the k-means-clustering method for large data sets that contain many attributes. Generally, the Web contains a lot of documents with many attributes. Therefore, the proposed method is useful for the Web.

## VI. CONCLUSION

We proposed a method that combines the k-means clustering method with an Independent Component Analysis based seeding method and a Principal Component Analysis based seeding method, and compared the performances of the proposed method with the performance of the standard k-means clustering method, k-mean++ clustering method, and k-means clustering method with a KKZ seeding method.

From our experimental results for small data sets (UCI repository data sets), our proposed method performed the same as or better than the standard k-means clustering method, k-means++ clustering method, and k-means clustering method with a KKZ seeding method.

From our experimental results for large data sets(CLUTO data sets), our proposed method based on ICA performed better than the standard k-means clustering method, k-means++ clustering method, and k-means clustering method with a KKZ seeding method when the ratio between the number of attributes and the number of samples is smaller than 10.00. When the ratio between the number of attributes and the number of samples is larger than 10.00, our proposed method based on ICA did not perform better than the standard k-means clustering method, k-means++ clustering method, or k-means clustering method with a KKZ seeding method. Generally, the Web has a lot of documents and the ratio between the number of attributes and the number of samples is small. Therefore, the proposed method is useful for the Web.

TABLE XII.
COMPUTATIONAL TIMES (SEC.) FOR *tr45 CLUTO* DATA SET

| method | *tr45* |
|---|---|
| k-means | $90.42 \times 100$ |
| proposed | 52.39+90.42 |

For our future work, we plan to theoretically analyze the computational cost of the proposed method, and to conduct research on how to decide the number of clusters based on the observed data distribution.

## ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 23300063, 2011.

## REFERENCES

[1] S. Basu, I. Davidson, and K. Wagstaff, Eds., *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman and Hall/CRC, 2008.
[2] C. Ding, "A Tutorial on Spectral Clustering Part I: Basic Theory," 2004. [Online]. Available: http://ranger.uta.edu/ chqding/Spectral/
[3] P. Berkhin, "Survey of clustering data mining techniques," Accrue Software, San Jose, CA, Tech. Rep., 2002.
[4] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, "A survey of web clustering engines," *ACM Computing Survey*, vol. 41, pp. 17:1–17:38, 2009.
[5] P. Ferragina and A. Gulli, "A personalized search engine based on web-snippet hierarchical clustering," in *Special interest tracks and posters of the 14th international conference on World Wide Web (WWW'05)*, 2005, pp. 801–810.
[6] R. Navigli and G. Crisafulli, "Inducing word senses to improve web search result clustering," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, 2010, pp. 116–126.
[7] B. Stein, T. Gollub, and D. Hoppe, "Beyond precision@10: clustering the long tail of web search results," in *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM'11)*, 2011, pp. 2141–2144.
[8] Y. Wang and M. Kitsuregawa, "Evaluating contents-link coupled web page clustering for web search results," in *Proceedings of the eleventh international conference on Information and knowledge management (CIKM '02)*, 2002, pp. 499–506.
[9] D. Crabtree, P. Andreae, and X. Gao, "Query directed web page clustering," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*, 2006, pp. 202–210.
[10] P. Li, B. Wang, W. Jin, and Y. Cui, "User-related tag expansion for web document clustering," in *Proceedings of the 33rd European conference on Advances in information retrieval (ECIR'11)*, 2011, pp. 19–31.
[11] C. Lu, X. Chen, and E. K. Park, "Exploit the tripartite network of social tagging for web clustering," in *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*, 2009, pp. 1545–1548.

[12] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, "Web image clustering by consistent utilization of visual features and surrounding texts," in *Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA'05)*, 2005, pp. 112–121.

[13] T. Onoda, M. Sakai, and S. Yamada, "Independent component analysis based seeding method for k-means clustering," in *In Proceedings of the International Workshop on Intelligent Web Interaction 2011 (IWI-2011)*, 2011, pp. 122–125.

[14] T. Onoda, M. Sakai, and S. Yamad, "Careful seeding based on independent component analysis for k-means clustering," in *In Proceedings of the International Workshop on Intelligent Web Interaction 2010 (IWI-2010)*, 2010, pp. 112–115.

[15] T. Onoda, M. Sakai, and S. Yamad, "Seeding method based on independent component analysis for k-means clustering," in *In Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems (SCIS&ISIS-2010)*, 2010, pp. 1306–1309.

[16] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering large graphs via the singular value decomposition," *Maching Learning*, vol. 56, no. 1-3, pp. 9–33, 2004.

[17] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[18] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized lloyd iteration," *IEEE Signal Processing Letters*, vol. 1, no. 10, pp. 144–146, 1994.

[19] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[20] A. Hyvrinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[21] A. Hyvrinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[22] A. Hyvrinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492,, 1997.

[23] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

**Takashi Onoda** graduated from International Christian University, Tokyo, Japan in 1986. He received the M.S. degree in nuclear engineering from Tokyo Institute of Technology, Tokyo, Japan in 1988.

He works at Central Research Institute of Electric Power Industry from 1988. He received the Dr. Eng. degree in mathematical engineering from University of Tokyo, Tokyo, Japan in 2000. He worked as a visiting researcher in GMD FIRST in Berlin from September in 1997 to September in 1998. He is a sector leader at Central Research institute of Electric Power Industry and a visiting professor at Tokyo Institute of Technology.

His research interests are in statistical learning theory and its applications. He is a member of JSAI.

**Miho Sakai** graduated from Musashi Institute of Technology, Tokyo, Japan in 2009. She received her MS degree in engineering from Tokyo Institute of Technology in 2011. Her research interest is in clustering method.

**Seiji Yamada** is a professor at the National Institute of Informatics. Previously he worked at Tokyo Institute of Technology. He received B.S. (1984), M.S. (1986) and the Ph.D. (1989) degrees in artificial intelligence from Osaka University.

His research interests are in the design of intelligent interaction including Human-Agent Interaction, intelligent Web interaction and interactive machine learning. He is a member of IEEE, AAAI, ACM, JSAI, IPSJ and HIS.

# Measuring Emotions from Online News and Evaluating Public Models from Netizens' Comments: A Text Mining Approach

Simon Fong
Department of Computer and Information Science
University of Macau
Taipa, Macau SAR
Email: ccfong@umac.mo

*Abstract*— **Nowadays netizens embark on a prevalent lifestyle to actively voice out their opinions online that includes both forums and social networks (Web 2.0). Their opinions which initially are intended for their groups of friends propagate to attentions of many. This pond of opinions in the forms of forum posts, messages written on micro-blogs, Twitter and Facebook, constitute to online opinions that represent a community of online users. The messages though might seem to be trivial when each of them is viewed singularly; the converged sum of them serves as a potentially useful source of feedbacks to the current affairs after analysis. A local government, for instance, may be interested to know the response of the citizens after a new policy is announced, from their voices collected from the Internet. However, such online messages are unstructured in nature, their contexts vary greatly, and that poses a tremendous difficulty in correctly interpreting them. In this paper we propose an innovative analytical model that evaluates such messages by representing them in different moods. The model comprises of several data analytics such as emotion classification by text mining and hierarchical visualization that reflects public moods over a large repository of online comments.**

*Index Terms*— **Emotion classification; text mining; hierarchical visualization**

## I. Introduction

Netizens nowadays develop a habit of whining out their opinions in the virtual world, through blogs, social networks as well as community forums. Their purpose may be just to share their views, casually or consciously, in response to all kinds of world events and individual topics of interest. From the postings and counter-replies, it has evolved into a trend of social acquaintance in the virtual world [1]. Twitter has more than 180 million unique visitors per month, and a total amount of messages close to a trillion. Facebook also has a population of 166 million active users whose posts amount to an astronomical figure. And these figures are still undergoing some phenomenal growth.

Recently government agencies established their community groups on Facebook. The motive could be in twofold: to disseminate information to the online users, and to probably listen to their opinions. However, to the second motive, assuming that the government agency bothers to pay attention to those opinions, there is an inherent challenge in the format of the data. They are unstructured both in grammar and context. As users are free to post anything under the sun, the format is not in formal writing (unlike official letters); slangs may be used and they differ from culture to culture. On the brighter side, netizens are responsive to new posts and new events. For example, any world news, such as earthquake, terrorist attacks or economic crisis that rock the world would attract them to proactively post and encounter post on each other's messages. They share their views in different emotions, pertaining to the subject that they are commenting about. The online messages come in very different types of wish-making, suggestion, political opinion, critics and praises, or dissatisfaction to share among friends and the rest of the world.

In addition to the obstacles of data formats and contexts, a government or organization may face another challenge due to the dynamic nature of the distributed online comments, which arise both in tremendous quantity and at a very high speed. The contents of the comments may change over time too; for example, an invention of a vaccine for a global epidemic disease may first be cheered as "happy" news. Should it be later found as a hoax, the general comments may gradually switch to mood of "disappointing" or even "hilarious".

Organizations do need some autonomous method to classify the messages into different moods and kinds of opinions, in contrast of the previous works of deciphering their actual meanings. Currently manual work is required by a human user to comprehend the messages by his knowledge background and relate them as opinions being talked about of a particular event. Because there are large diversities of words and vocabularies representing different emotions, it is important to tap on the cultural background knowledge.

Emotion is a complex psycho physiological experience of an individual's state of mind as interacting with biochemical and environmental influences. In humans, emotion fundamentally involves "physiological arousal, expressive behaviors, and conscious experience" [2]. Emotion is associated with mood, temperament, personality and disposition, and motivation. Emotion doesn't exist in computers that are based on logics. Emotion also may not be easily calculated by a formula.


doi:10.4304/jetwi.4.1.60-66

Emotion is a fuzzy state; hence machine learning algorithms that are able to represent non-linear relations between the occurrence of a series of keywords in the text and the predicted class of emotion, are appropriate for handling this type of problems such as Artificial Neural Network and Decision Tree algorithms [3]. Before entering the comments in to an emotion classifier, we need to translate sentences to relative metadata which are represented in abstract levels. To translate sentences to metadata we make use of a linguistic dictionary to categorize, stemming methods that filter out unimportant words, vector space model for establishing the importance of words by measuring their frequencies and group the significant words into meta-data.

"Point of view" is another important factor that contributes to understanding emotions. We utilize information from online news to establish a neutral evaluation standard. Since opinions in newspapers are in journalistic and relatively objective style, we adopt so as a standard for describing neutral opinions. The other usage of newspaper is it may contain the background story of an event. By comparing the evaluation standard and online comments we can have some benchmark for positioning a neutral point in our visualization which shows information in different levels, thus we call it "Hierarchical Visualization". The Hierarchical Visualization can provide the trend of public mood, detail of the range about mood and it can be directly used for government or organization to understand their citizens' or customer's feedbacks. The Hierarchical Visualization reveals the moods of the public in general, instead of displaying a long list of individual comments. Our proposed Hierarchical Visualization is designed for high-level users who often prefer to glimpse at an overall view of the public opinions, without going into details or crunching over the numeric figures. This method proposed in this paper is subtle which doesn't require costly massive scale of survey questionnaires that probe answers directly from citizens.

## II. Our Proposed Model

Before collecting information from the Internet, a specific Research Topic (or topic of interest) to be analyzed should be defined. Research topics could be of any current affair or any latest government policy which netizens are keen to comment about. The next step is to download the data from relevant sources. The easiest way to confirm the date of the event that happened can be referred from the official news. News published on the Internet usually would have highlighted by some keywords that can be extracted from the tags that appear at the bottom of the page – the keywords are useful for us to define the metadata of a research topic. Overall, for each research topic, we use the time, the metadata, as settings of parameters for the web downloading software to congregate Internet comments within a reasonable time range (e.g. 80% of netizens talked about Michael Jackson's death within only 5 months). The information downloaded will be used to build up two kinds of databases. One is a repository of online News that are tagged with date of occurrence, plus the related metadata for ontology [4]; the second one is the postings extracted from some social networks and micro-blogging sites. Twitter and Facebook are used as experiments in this paper. HTML tags are cleansed in the preprocessing step. The information about the poster's information, such as IP (which may not always be available), time of posting, user's background or other will also be collated.

There are several approaches to build a Moods engine which is used to classify the mood of a given online article or piece of text. It was suggested in [5] that a Moods engine to be based on a standard dictionary for embracing the keywords by using an artificial neural network (ANN). The relevant words that are related to different emotions from a well-known dictionary reference are used as training data to build up a number of ANNs, one for each type of emotion so that it can be used subsequently to recognize the perceived emotion out from a testing text. Optionally, one may incorporate MSN-style of acronyms or emoticons to represent emotions [6], e.g. a smiley is a symbol of happiness written as :-) Short-names commonly used as cyber etiquettes like W.T.H/F. (anger plus astonishment) and I.M.H.O (neutral narration) could also be added on. One ANN is to be trained and employed to describe one type of mood. The mood engine is to be fine-tuned with users' subjective experiences for improving the accuracy. So the major function of the mood engine is to distinguish a resultant mood by reading through a pile of text messages.

The evaluated news will also train the ANN of different moods. The ANN of a particular mood is represented by the weights trained by the training data. If a piece of news was marked or flagged as "happiness", the news would be used to train the "happiness" ANN model. Data from the Internet comments databases would be used as testing data to be tested in the ANNs for deciding which mood(s) they belong to.

Alternatively, as proposed in this paper here, a generic text classifier can be trained by some predefined training texts which have already the labeled classes of emotions assigned in the training data. This approach requires pre-assignment of verdict classes (emotions) on the training dataset that is made up of news which we already known the emotion class that they belong to. The training dataset would be processed by text mining techniques that include a sequence of data-preprocessing steps such as stemming, data cleaning and transforming of the keywords to attributes of frequency of occurrences. Text classification is adopted here as a generic approach that could be powered by a range of different underlying algorithms. After a classifier is trained with sufficient records that have the labeled emotions, it could be deployed for automatically classifying the expected emotions from the future texts. It is recommended that the classifier has to be trained first to certain acceptable accuracy from training data taken from online news pertaining to a Research topic or a category of current affairs. Then it would be used for classifying emotions from testing data that are to be scraped from users' posts. The workflow of the model training is shown in Figure 1.

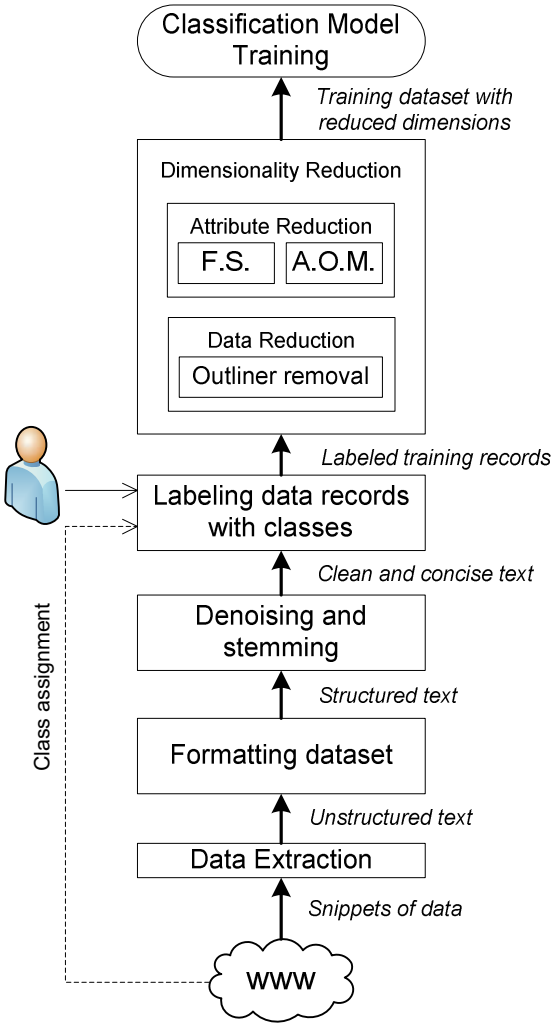



Figure 1. The workflow of training the classifier for recognizing emotions from online news.

Once the emotion classifier is trained and its performance accuracy reaches an acceptable level, the classifier is ready to classify emotions from new data. Posts and messages made by online users are retrieved and formatted in the same way as in the training process for the classifier. Given the new data, the classifier classifies new instances of online users' opinions into emotion groups. The advantage of this text mining approach is the generality that different machine learning algorithms as well as different dimensionality reduction algorithms can be used, even in combination, for the optimal results. Figure 2 shows the subsequent process.

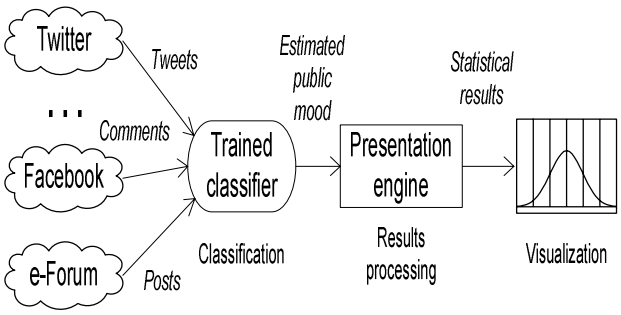



Figure 2. High-level view of the emotion classification process.

The output results will be processed in presentation engine and displayed in the Hierarchical Visualizations. Multiple levels of visualizations are used because the details of the results could be shown in different depths, depending on the choice of the user for the desired resolution. Too much information in visualization is confusing to the users. Users can opt to choose a viewing level interactively in the control panel of the visualization software.

The results are shown in graphical form so that it is easy to captivate the attention of the user, and possibly to spot any special patterns visually. The user can zoom in and out at will, or to display the full details for further analysis when necessary. Colors are used to represent the different emotions respectively. The following example in our experiment is on the topic of "Hengqin Campus Project of University of Macau". We set up a list of colors [7] to represent different moods. The circle represents moods of an event and we use the angles represent the percentage of the moods. In Figure 3, it is easy to get to know the public mood about an event based on the comments collected from the Internet. In the software, we can scale to level 2 of the hierarchical visualization as shown in Figure 4 when we select the two colors in the circle. Since the colors are not fixed, the visualization is interactive with the end-user. The user can focus his attention on which moods that the general public feels about the specific topic. In this zoom-in level of visualization, we may want to analyze about the genders of the users who posted their opinions (and subsequently reflected by their moods), just for example. The wave line in Figure 4 is representing the number of people who are in different moods with different gender, one on each side of the belt. In the next level, we can select a location to be analyzed. The locations, gender and moods relationships are presented in this level. In Figure 5, the chart shows visually that how users in different locations carry certain moods. The level of details can be increased optionally; for example, the locations can further break down to suburbs, streets, etc. Other dimensions can be added or switched too

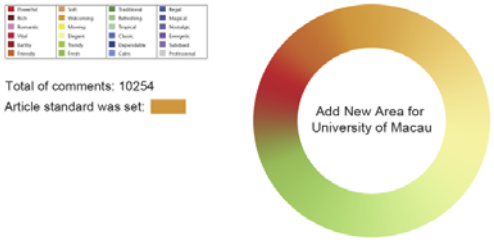



Figure 3. Level 1 of the visualization.

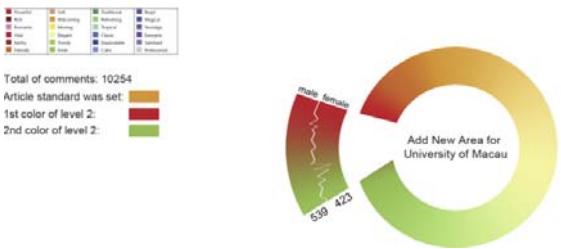



Figure 4. Level 2 of the visualization.

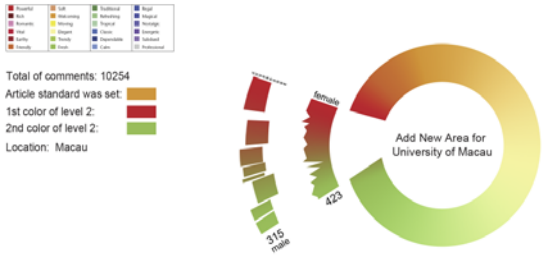



Figure 5. Level 3 of the visualization.

The proposed system can help organizations or government to understand the opinions which are in response to a news event or a policy announcement, without doing meticulous survey to collect public opinions. The system is easy enough to use, for revealing the public moods based on a given event. The system features about consideration of a culture of a country by training the emotion classifier using news samples from the local online news website. For example, CNN for Americans, CBS/BBC for English, ABC for Australian, ChannelNewsAsia for Asians, just to name a few. Cultural perceptions influence on how the citizens express their views, hence the choice of words being said and posted on online forums. For example, the word "cool" may mean a cheerful mood in the Western culture, but otherwise in oriental or other conservative cultures who take the word "cool" literally as "indifferent" in character. However, different versions of the system may be needed to be built for different cultures [8], but the training datasets and hence the classifiers would be unique in each culture. As such, an analyst who uses the system by different cultures can understand where, who, which age group, how many people and what they feel in the visualization, in response to an event, based on the analysis from the Internet comments.

## III. EXPERIMENT

In order to validate the concept of our proposed model, a text mining program is built by using Weka which is an open-source JAVA platform for evaluating machine learning algorithms by University of Waikato. Specifically, text classification is implemented under the text mining domain and training datasets of different emotions are used for the experiment. We aim to test-drive the classifier with different machine learning algorithms and different dimensionality reduction methods. It is a known challenge in text mining that the accuracy is almost directly pegged on how well the dimensionality of the dataset can be tamed. The training data which are obtained from online news platforms are unstructured in nature. In addition to standard data pre-processing techniques like filtering noise and stemming (a process for removing redundant words), dimensionality reduction algorithms for reducing the number of attributes that are used to represent the essence of the text and amount of instance number are applied in our experiment. An outliner removal algorithm is used for trimming off data rows that have exceptionally different values from the norm. For reducing the number of attributes, two standard Feature Selection algorithms (FS) are used, together with a novel approach called Attribute Overlap Minimization (AOM) are applied. Readers who want to have further details about these algorithms should refer to [9].

The training data are excerpted from CNN news website, of the news articles that were released for ten days across the New Year 2012 (one week before and one week after the New Year eve). The news collection has a good mix of political happenings, important world events and lifestyles. One hundred sample news were obtained in total, and they were rated manually according to six basic human psychological emotions, namely, Anger, Fear, Joy, Love, Sadness and Surprise. The data are formatted into ARFF format (as required by Weka), having one news per row in the following structure: <emotion>, <"text of the news"> where the second field has a variable length. The training dataset is then subject to the above-mentioned dimensionality reduction methods for transformation to a concise dataset in which the attributes have substantial predictive powers contributing to recognizing emotions from the text strings.

The Feature Selection algorithms used include, *CfsSubset*, *ChiSquaredAttribute*, *InfoGainAttribute*, *SignificanceAttribute*, and *SymmetricalUncertAttribute*. The full explanation about these algorithms can be found on Weka homepage. As shown in our experimental results in Table 1, *CfsSubset* generally can achieve the best classification accuracy by filtering most but retaining only the minimum set of elite attributes that have most predicting powers. The rest of the FS algorithms produce almost identical results though *ChiSquare* and *InfoGain* are relatively more popularly used in data mining community. Accuracy is defined by the percentage of the number of correctly classified instances over the total number of instances in the training dataset. By applying attribute reduction and data reduction, we can observe that the initial number of attributes have reduced greatly from 8135 to 19. Having a concise and elite amount of attributes is crucial in real-time application, and in text mining, the number of attributes is proportional to the coverage of news articles – the more unique words (vocabularies) that are being covered, the greater the amount of attributes there are. Text classification essentially works on the principle of finding the non-linear relations of co-occurrences of important keywords in a given text, measured by their occurrence frequencies.

It is found from the results in Table 1 that using the FS algorithm *CfsSubset* together with other techniques can effectively achieve a classifier (Decision tree is selected in this example) that has the lowest number of tree size, highest accuracy and shortest training time. A compact tree size in Decision tree algorithm means least consumption of heap memory space that is essential for real-time applications where memory space may be an operational constraint. Short training time implies that the classifier model takes only a short while for updating or even rebuilding the tree model that will be useful for application scenarios where frequent updates may be necessary for fast-changing data inputs.

The experiment is then extended to evaluate the use of machine learning algorithms, with the objective of achieving the highest accuracy. The selection list of the machine learning algorithm used in our experiment here is by no means exhaustive, but will form the basis of a performance comparison which should supposedly cover most of the popular algorithms. The machine learning algorithms are grouped by four main categories, Decision Tree, Rules, Bayes, Meta and Miscellaneous; all of them are known to be effective for data classification in data mining to certain extents. The list of algorithms is shown in Table 1, and their details can be found in [9].

TABLE I. LIST OF CLASSIFICATION ALGORITHMS USED IN THE EXPERIMENT

| Acronym | Algorithm name | Type |
|---|---|---|
| J48 | C4.5 decision tree | Decision tree |
| BFTree | Best-first decision tree classifier | |
| Ftree | Functional trees', which are classification trees that could have logistic regression functions at the inner nodes and/or leaves | |
| NBTree | A decision tree with naive Bayes classifiers at the leaves | |
| LMT | Logistic model trees', which are classification trees with logistic regression functions at the leaves | |
| RandForest | A forest of random trees | |
| RandTree | A tree that considers *K* randomly chosen attributes at each node | |
| REPTree | Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning | |
| DecTable | A simple decision table majority classifier | Rules |
| FURI | Fuzzy Unordered Rule Induction Algorithm | |
| Ripper | A propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) | |
| PART | A PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule | |
| BayesNet | Bayes Network learning using various search algorithms and quality measures | Bayes |
| CompNB | A Complement class Naive Bayes classifier | |
| NB | a Naive Bayes classifier using estimator classes | |
| Bagging | Bagging a classifier to reduce variance | Meta |
| Ensemble | Combines several classifiers using the ensemble selection method | |
| SVM | Support Vector Machine | Misc |
| NN | Backpropagation Neural Network | |

The experiments are conducted according to the workflow depicted in Figure 1. After the training data are cleansed, formatted and labeled, they were subject to the dimensionality reduction algorithms for improving the accuracy of the classifier. Three groups of resultant datasets were text-mined by different classification algorithms that are specified in Table 1. They are the original dataset without any dimensionality reduction, transformed dataset with reduced attributes, and transformed dataset with both attributes reduced and outliners removed. The results in terms of accuracy are shown in Figure 6 and Figure 7. The experiments are repeated for trying two most popular FS algorithms – *CfsSubset* and *InfoGain*. While the former FS algorithm chooses only the minimum number of the attributes that have significant contributing predictive power, the latter algorithm retains most of the attributes that have at least non-zero information gain towards the decision tree induction. Figure 6 shows the classification results from dataset filtered by *CfsSubset*, and Figure 7 shows those filtered by *InfoGain*.

It can be seen that the classifiers performed poorly over the original dataset, but the accuracy greatly improved once the attributes are reduced. A more than 100% gain increase is observed between the original data and the attribute-reduced data. This gain can be observed for most of the classification algorithms except for FURI and RIPPER. It makes little difference between FS algorithms for *CfsSubset* and *InfoGain*, which means *CfsSubset* can be used for minimum number of attributes. From the results of Figures 6 and 7, approximately 4% to 18.3% increases in accuracy are observed between the results obtained from attribute-reduced data and both attribute-reduced and outliner removed data. All the classification algorithms perform consistently well. Naïve Bayes classifier (NB) however, achieves the highest accuracy 86.8% in all cases. It in fact is the only classifier which has no effect by using *CfsSubset* and *InfoGain* feature selection algorithms. NB is independent of the feature selection algorithm and the candidate that yields the highest accuracy when a combination of dimensionality reduction techniques is used. For this reason, mood classification application is suggested to adopt NB for effectively classifying online text messages into a class of one of the six emotions. The following part of the experiment is to classify newly acquired text data by using the trained classification model.

## IV. RELATED WORK

There is a similar project named "Twitter mood maps reveal emotional states of America" in America. It has an idea to present human mood in a timeline superimposed on an America map with different colors. This method takes individual words out of context. If someone tweets "I am not happy", the team's method counts the tweet as positive because of the word "happy". It is based on current state of Twitter user and by possibly keyword matching methods. Unlike our proposed model, it does not shows results from multi-dimensions, and the fuzzy

natures of mood matching and different cultural aspects were not considered.

Another academic research work that is very similar to ours is [10], by the authors Tao et al. Tao regarded that the Web has become an excellent source for gathering and realizing public voice. The paper discusses a method for exploring the public mood levels at the time of posting. Hence the results are presented as two-dimensional graphs with the y-axis being the mood level, and the x-axis as a time-line. Again, the two dimensions of variables for presenting mood levels can be extended to multiple dimensions as proposed in this year. Also the paper [10] used corpus aggregating method for measuring mood level, and the case study was on emergency scenarios, where a flexible classifier that can be chosen from a collection can be used in our model for handling application situations. In addition, our model incorporates with a hierarchical visualization program, and our experiments showed that the prototype can be used in general situations.

## V. CONCLUSION

In this paper we proposed and defined an analytical model for evaluating online users' comments in response to a given event. Our model features a Mood engine made up of a number of dimensionality reduction algorithms, and text classification machine learning algorithms, that can effectively classify a text into one of the six human basic emotions aka moods. The moods can be changed and calibrated according to different cultures by re-training the text classification model. The classifiers can be trained by using standard words or past news with predefined emotions assigned by human experts. The trained classifier is then used to detect types of moods and their intensities from a pool of new messages and postings collected from micro-blogs and social networks that constitute a large online community as a whole. The online comments are inputted to the mood engine and the comments are categorized into types of moods. Aggregating categorized comments and their moods are fed into a hierarchical visualization program that shows interactively different dimensions of the information with respect to the public Mood. A prototype is built and results show that the model is feasible. This paper contributes a text classification model for this job; its efficacy is experimented by considering a wide range of classification algorithms and several feature selection algorithms. It is found that Naïve Bayes classification algorithm outperformed the rest, and it is independent of which feature selection algorithm is being used. The techniques and model presented in this paper are generic which means a specific algorithm can be replaced by a new candidate, and it is believed to work on other similar mood detection scenarios such as analysis of help-desk logs, customers' feedbacks, online reviews etc. The core engine of the model is the Mood Engine which essentially is shown to be possible by implementing it with an appropriate text mining algorithm and a sequence of dimensionality reduction methods. The advantage is a simple and flexible model with reasonable accuracy.

## REFERENCES

[1] Tanase, S., "When web 2.0 sneezes ...everyone gets sick", Engineering & Technology, Volume 5, Issue 5, 27 March-23 April 2010, pp.28-29.

[2] Myers, David G. "Theories of Emotion." Psychology: Seventh Edition, New York, NY: Worth Publishers, 2004, p.500 (Wiki).

[3] Elaleem, O.A.; Elragal, H.M.; Shehata, H.M, "Voice Message Priorities Using Fuzzy Mood Identifier", Proceedings of the Twenty Third National Radio Science Conference, NRSC 2006, pp.1-6.

[4] Seheon Song; Minkoo Kim; Seungmin Rho; Eenjun Hwang, "Music Ontology for Mood and Situation Reasoning to Support Music Retrieval and Recommendation", Third International Conference on Digital Society, ICDS 2009, pp.304-309.

[5] Chan Io Weng, Simon Fong, Suash Deb, "An Analytical model for Evaluating Public Moods Based on Internet Comments", National Conference on Data Mining (NCDM 2011), 19-20 February 2011, Pune, India.

[6] Yamashita, Ryo; Yamaguchi, Sanae; Takami, Kazumasa, "A Method of Inferring the Preferences and Mood of Mobile Phone Users by Analyzing Pictograms and Emoticons Used in their Emails", Third International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services (CENTRIC), 2010, pp.67-72

[7] Chuan-Kai Yang, Li-Kai Peng, "Automatic Mood-Transferring between Color Images", IEEE ComputerGraphics, Issue No.2, March/April 2008, pp.52-61

[8] Zhijun Zhao; Lingyun Xie; Jing Liu; Wen Wu, "The analysis of mood taxonomy comparision between chinese and western music", 2nd International Conference on Signal Processing Systems (ICSPS), 2010, pp.606-610

[9] Simon Fong, "Dimensionality Reduction in Vector Space Model for Multi-class Text Classification", submitted to JIPS-K, 2012.

[10] Tao Xu, Qinke Peng, Chengwei Li, "A Method of Capturing the Public Mood Levels in Emergency Based on Internet Comments", 7th World Congress on Intelligent Control and Automation, WCICA 2008, 25-27 June 2008, pp.3496-3499.

**Simon Fong**

He graduated from La Trobe University in Australia, with a First Class Honours BEng. Computer Systems degree and a PhD. Computer Science degree in 1993 and 1998 respectively. Simon is now working as an Assistant Professor in the Computer and Information Science Department of the University of Macau. He is also one of the founding members of the Data Analytics and Collaborative Computing Research Group in the Faculty of Science and Technology. Before his academic career, Simon took up various managerial and engineering posts, such as being a systems engineer, IT consultant, integrated network specialist, and e-commerce director in Melbourne, Hong Kong. and Singapore. Some companies that he worked at before include Hong Kong Telecom, Singapore Network Services, AES Pro-Data, and the United Overseas Bank in Singapore. Dr. Fong has published over 150 peer-reviewed international conference and journal papers, mostly in the area of E-Commerce and Datamining.

Accuracy

| Classifiers | Original | Attr-Reduced | Attr-Data-Reduced |
|---|---|---|---|
| J48 | 33 | 66 | 75.8242 |
| BFTree | 32 | 64 | 72.5275 |
| Ftree | 36 | 78 | 83.5165 |
| NBTree | 31 | 69 | 76.9231 |
| LMT | 37 | 78 | 83.5165 |
| RandForest | 33 | 75 | 83.5165 |
| RandTree | 33 | 75 | 83.5165 |
| REPTree | 34 | 62 | 71.4286 |
| DecTable | 26 | 60 | 71.4286 |
| FURI | 38 | 55 | 65.9341 |
| Ripper | 32 | 55 | 63.7363 |
| PART | 31 | 65 | 76.9231 |
| BayesNet | 22 | 72 | 79.1209 |
| CompNB | 36 | 68 | 71.4286 |
| NB | 46 | 83 | 86.8132 |
| Bagging | 37 | 71 | 74.7253 |
| Ensemble | 34 | 64 | 69.2308 |
| SVM | 38 | 69 | 67.033 |
| NN | 36 | 75 | 82.4176 |

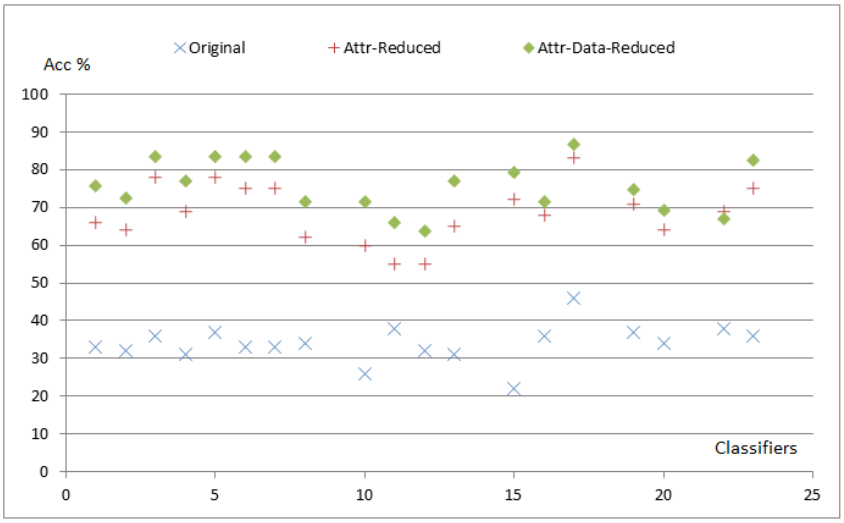



Figure 6. Accuracy results of classifiers in percentage with different types of datasets based on *CfsSubset* FS algorithm.

Accuracy

| Classifiers | Original | Attr-Reduced | Attr-Data-Reduced |
|---|---|---|---|
| J48 | 33 | 69 | 75.8242 |
| BFTree | 32 | 66 | 73.6264 |
| Ftree | 36 | 73 | 79.1209 |
| NBTree | 31 | 69 | 73.6264 |
| LMT | 37 | 73 | 78.022 |
| RandForest | 33 | 76 | 80.2198 |
| RandTree | 33 | 76 | 78.022 |
| REPTree | 34 | 60 | 71.4286 |
| DecTable | 26 | 58 | 70.3297 |
| FURI | 38 | 60 | 62.6374 |
| Ripper | 32 | 56 | 62.6374 |
| PART | 31 | 67 | 75.8242 |
| BayesNet | 22 | 71 | 74.7253 |
| CompNB | 36 | 72 | 73.6264 |
| NB | 46 | 83 | 86.8132 |
| Bagging | 37 | 68 | 74.7253 |
| Ensemble | 34 | 62 | 71.4286 |
| SVM | 38 | 63 | 63.7363 |
| NN | 36 | 75 | 81.3187 |

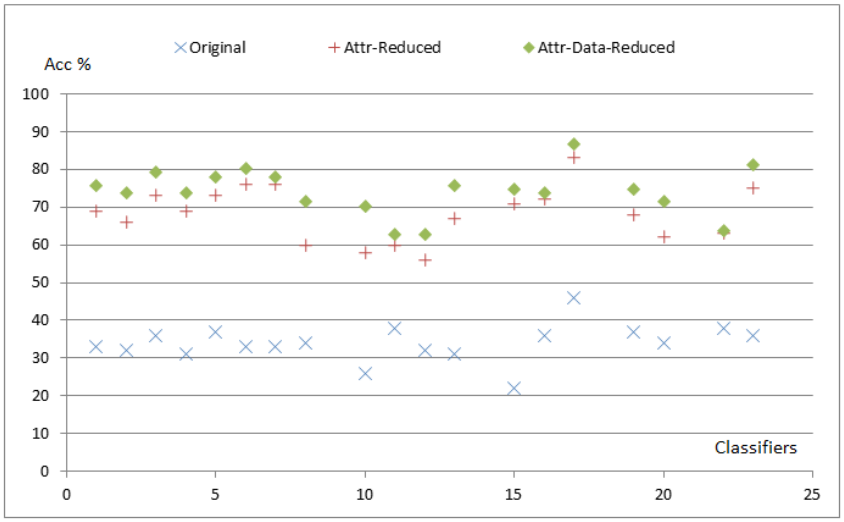



Figure 7. Accuracy results of classifiers in percentage with different types of datasets based on *InfoGain* FS algorithm.

# Distance-Based Scheme for Vertical Handoff in Heterogeneous Wireless Networks

Wail Mardini and Musab Q. Al-Ghadi
Computer Science Department, Jordan University of Science and Technology, Irbid, Jordan
Email: mardini@just.edu.jo

Ismail M. Ababneh
Computer Science Department, Al al-Bayt University, Mafraq, Jordan
Email: ismael@aabu.edu.jo

***Abstract*— Seamless vertical handoff between different access networks in the next generation wireless networks remains a challenging problem. A recent vertical handoff scheme that is based on Signal to Interference and Noise Ratio (SINR) may not be the best scheme for selecting the service access point or base station. Although this SINR-based scheme has higher system throughput and lower disconnection probability as compared with other vertical handoff schemes, we presume that the distance is a good criterion for decreasing service disconnection probability and increasing system throughput. The Distance-based Scheme for Vertical Handoff (DSVH) that we propose in this paper for heterogeneous wireless networks is a reactive vertical handoff scheme. We suggest that vertical handoff be based on the Received Signal Strength (RSS) and the distances to access points or base stations, where the main goal of this scheme is to enhance system performance in terms of reducing service disconnection probability and increasing system throughput. The simulation experiments show that our proposed scheme, DSVH, significantly outperforms the SINR-based scheme. It reduces the number of dropped users by 20%. However, the throughput improvrmrnt is insignificant; it is about 1%.**

***Index Terms*—Vertical handoff, SINR, RSS, Heterogeneous Wireless Networks**

## I INTRODUCTION

Nowadays, mobile users demand to be connected with the Internet while they move freely, and Always Best Connected (ABC) has become a very important service for mobile users so as to get high quality services at high data rates [2].

A state-of-the-art Fourth Generation (4G) wireless network is composed of different wireless subnetworks that complement each other [3]. The integration of such heterogeneous subnetworks should allow mobile stations (MSs) to choose the most appropriate access subnetwork among the available alternatives (these include IEEE 802.11 Wireless Local Area Network (WLAN) and IEEE 802.16 Worldwide interoperability for Microwave Access (WiMAX)), in addition to the traditional cellular networks.

One of the important issues in mobility is mobility management. This is concerned with location management and handoff management [6]. Location management enables the system to track the locations of mobile stations continuously as they move from one location to another; this might be in the same system or to a different one. On the other hand, handoff management aims to maintain an active connection with high quality between the MS and the network during the movement of the MS. This process requires keeping track of the state of the MS, either when it is linked with some Base Station (BS) [7] or when it is moving from one BS to another [8].

Handoff can be either horizontal or vertical [1]. A horizontal handoff takes place when the MS switches between points of attachment supporting the same network technology. For example, between two neighboring BSs of a cellular network. On the other hand, a vertical handoff occurs when the MS switches between points of attachment supporting different network technologies, for example, between a cellular network BS and an IEEE 802.11 AP [9][10]. A handoff process can be divided into three stages: initiation process (radio link transfer), decision process and execution process (channel assignment) [11] [12].

In this paper, we are interested in vertical handoff. Several criteria have been proposed in the literature for use in vertical handoff schemes. The main criteria used are discussed below:

- Received Signal Strength (RSS): is the most widely used criterion to decide which network to use for the handoff from a candidate list of networks. It is easy to measure and it is directly related to service quality. Obviously, RSS depends on the velocity of the MS, and on the distance between the MS and its point of attachmen. When a MS notices that the RSS is gradually decreasing, it can be assumed that the MS is moving away from the AP. When the RSS is increasing, then the MS is moving towards an AP. The speed and direction of a mobile node indicate the length of time the current connection can be maintained. Another factor is the coverage area of the network. A small coverage area would cause excessive handoffs between access points within the same network, which can lead to high packet loss.
- Signal to Interference plus Noise Ratio (SINR): is the


doi:10.4304/jetwi.4.1.67-76

ratio of the power of the received desired signal to the average noise power at the receiver. SINR can be improved by increasing the transmitted power, decreasing the coverage range and using a better Low Noise Amplifier (LNA).

- Available Bandwidth: is the volume of data per unit of time that a transmission medium can handle [1]. It is a good indicator of traffic conditions in the access network. When a high bandwidth network, for example, a WLAN is heavily loaded or congested, the MS can switch to a lower bandwidth connection.

### A. Background on Radio Wave Propagation

Radio wave propagation deals with the properties and behavior of radio waves as they propagate from the sender to the receiver. There are three key factors that may impede the propagation of waves. They are reflection, diffraction and scattering. Reflection occurs when radio waves collide with a very large object, such as a mountain, a hill and a tower. Diffraction occurs due to collision with an aliasing object (i.e., an object that contains wavy edges and many protrusions surfaces). Scattering happens when the radio waves pass through paths that contain a large number of objects with small dimensions compared with the wavelength, such as foliage, herbs and street signs [14].

Path loss indicates the decline in the power of the wave during transmission from the sender to the receiver. In general, the path loss depends on the properties of the environment, the topography of the earth and the propagation medium, and the distance between the sender, the receiver, and the height of the BS/AP [13].

The propagation or path loss from the sender antenna to the receiver antenna is computed using the equation:

$$P_L = 10 \log [ P_t / P_r]$$

Where $P_L$ is the path loss in decibels (db), $P_t$ is the transmitted power in watts, and $P_r$ is the received power in watts.

Several models have been proposed for computing the path loss in various environments. These models are based on experiments in real environments, where all objects that are present in a particular experimental environment are taken into account, whether they are mountains, towers or buildings … etc [15].

Okumura's Model [16] is a well-known model for predicting the value of the path loss in an urban environment. This model is applicable when the frequency of waves is within the 150-1920 MHz range [17], the distance between the sender and receiver is from 1-100 km, and the height of the antenna of base station is from 30-1000 m. To determine the median path loss between the sender and receiver, Okumura developed a set of curves that give the median attenuation ($A_{mu}$) relative to free space in an urban area with BS antenna height of 200 m and mobile station (MS) antenna height of 3 m. [15]. Another model, the Hata model, is an extension of the Okumura model [18]. This model is applicable for the 150-1500 MHz frequency range, the sender-to-receiver distance range of 1-20 km, and base station heights from 30 to 200 meters and MS heights from 1 to 10 meters.

### B. Problem Definition and Motivation

A recent SINR-based vertical handoff scheme [28] [29] has higher system throughput and lower disconnection probability as compared with other vertical handoff schemes**.** However, we presume that the distance is a better criterion for decreasing. service disconnection probability and enhancing system throughput when selecting the best AP or BS. We propose that wave propagation models use the distance to the AP/BS as a main parameter, and propose a reactive distance-based vertical handoff scheme that aims to improve performance as compared with the recent SINR-based vertical handoff scheme. The proposed scheme has been designed and simulated using MATLAB. In order to evaluate the performance of our proposed scheme, we have compared our results to those of the SINR scheme. The reason behind choosing the SINR scheme is that it has high system throughput and low disconnection probability as compared with other vertical handoff schemes.

## II LITERATURE REVIEW

Vertical handoff schemes are essential components of the structural design of the forthcoming 4G heterogeneous wireless networks. These schemes need to be designed to provide the required QoS to a wide range of applications while allowing seamless roaming among a multitude of access network technologies [1].

There are many proposed schemes for vertical handoff in heterogeneous networks, which can be grouped into four categories: RSS-based, bandwidth-based, area-based and fuzzy logic-based vertical handoff scheme.

### A. RSS-based Vertical Handoff Schemes

The idea of the RSS based vertical handoff schemes is to calculate and compare the RSS of the current point of attachment against the others to make handoff decision. A lot of previous work and studies have been conducted in this topic [23] [24] [25] [26]. In this section we discuss three representative RSS based vertical handoff schemes.

Zahran et al [23] [27] proposed an adaptive lifetime-based vertical handoff (ALIVE-HO) scheme which takes into consideration the RSS, handoff latency, application QoS and delay tolerance, by presenting an application-based signal strength threshold (ASST) tuning mechanism to study the performance of vertical handoff between Third Generation (3G) cellular network and WLAN. The ASST have a significant role in future generation wireless networks where access technologies with different characteristics are expected to seamlessly co-exist and efficiently inter-operate. Therefore, the ASST can be optimally tuned for any access network based on practical system characteristics and requirements. In this scheme, the vertical handoff between 3G cellular network and WLAN can be described through two cases; in the first case, when the MS moves towards a WLAN cell. The handoff to the WLAN is trigger if the average RSS measurement of the

WLAN signal is larger than a threshold ($MIT_{WLAN}$) and the available bandwidth of the WLAN meets the bandwidth requirements of the application. While in the second case, when the MS moves away from the coverage area of a WLAN into a 3G cellular network cell, a handoff to the 3G cellular network is initiated under the conditions that the average RSS of the WLAN connection falls below a predefined threshold ($MOT_{WLAN}$), and the expected lifetime is less than or equal to the handoff delay. An analytical framework has been proposed to evaluate the performance of adaptive lifetime-based vertical handoff (ALIVE-HO) scheme, which is validated by computer simulation. This analytical framework proved that by introducing the lifetime metric the algorithm adapts to the application requirements and the MS mobility reducing the number of superfluous handoffs, and there is an improvement on the average throughput that provides for the MS because of the MS's ability to remain connected to the WLAN cell as long as possible.

Another scheme was designed to minimize the probability of unnecessary handoffs and to improve the overall network utilization. Yan et al [25] [26] proposed a vertical handoff decision scheme based on the distance of traveling distance within a WLAN cell (i.e. the time that MS is expected to spend within a WLAN area). The proposed scheme uses two thresholds which are calculated by the MS as it enters the WLAN area: Distance traveling threshold which based on RSS change rate (i.e. time that the MS is expected to spend within the WLAN area) and distance threshold which is calculated based on various network parameters such as handoff failure, handoff probability, radius of the WLAN area and handoff delays. A handoff to a WLAN is initiated if the WLAN coverage area is available and the estimated traveling distance inside the WLAN area is larger than the distance threshold. While a handoff to the cellular network is initiated if the WLAN RSS is continuously, fading and the MS reaches a handoff commencement boundary area based on its speed. The performance analysis showed that the main improvement of this scheme is that it minimizes the probability of handoff failures, unnecessary handoffs and connection breakdowns whenever the predicted traveling distance inside the WLAN cell is smaller than the distance threshold value.

### *B. Bandwidth-based Vertical Handoff Schemes*

Bandwidth based vertical handoff schemes considers the available bandwidth for MS as the main criterion to make handoff decision. A lot of previous work and studies have been conducted in this topic [22] [27] [28] [29]. In this section, three representative bandwidth based vertical handoff schemes are discussed.

Yang et al [22] proposed a bandwidth-based vertical handoff scheme between WLAN and Wideband Code Division Multiple Access (WCDMA) network, using the received Signal to Interference plus Noise Ratio as the handoff criteria. This scheme consider the combined effects of SINR from different access networks; where the SINR value from one network being converted to equivalent SINR value to the target network. A handoff to the network with larger SINR is performed, so the handoff algorithm can provide the knowledge of the achievable bandwidth from both access networks to make handoff decisions with QoS consideration. In addition to that, Yang et al [27] recently propose a Multi-dimensional Adaptive SINR based Vertical Handoff scheme (MASVH) for next generation heterogeneous wireless networks. This scheme uses the combined effects of SINR, MS required bandwidth, MS traffic cost and utilization from participating access networks to provide seamless vertical handoff with multi-attribute QoS support. Simulation results confirm that the new MASVH scheme improves the system performance in terms of higher throughput and lower dropping probability, as well as reduces the MS traffic cost for accessing the integrated wireless networks.

Ayyappan et al [28] [29] proposed SINR based vertical handoff scheme for QoS in heterogeneous wireless networks. In order to provide QoS inside the heterogeneous network, the vertical handoff scheme needs to be QoS aware, which can be achieved by gives the SINR based handoff better than RSS based handoff. This scheme considers the received SINR as a handoff criterion, which can be calculated using the Shannon's capacity theorem as $R = W \log2 (1 + \gamma / \Gamma)$ Where, R is the maximum throughput, W is the carrier bandwidth, $\gamma$ is SINR received at MS, $\Gamma$ is the gap between uncoded quadrature amplitude modulation and channel capacity. The handoff is initiated when the MS receives higher equivalent SINR from another network. In such cases, the MS tries to switch to another network that will satisfy the service QoS attributes. Simulation results prove that the proposed SINR based vertical handoff scheme provides higher overall system throughput as well as minimum number of dropped MS.

Rafiq et al [30] proposed a vertical handoff scheme that takes into account end-to-end QoS in addition to other common parameters. The scheme present an architecture involving an external host based light-weight server, called Access Link Utilization Monitor (ACUM) that disseminates the available end-to-end bandwidth to the mobile node to assist it in making a decision to maintain end-to-end service quality. The authors also describe a fuzzy logic based algorithm that is used in the handoff decision.

### *C. Area-based Vertical Handoff Schemes*

Area-based schemes make use of geographical information that is gathered by either GPS devices or a physical layer support. Such additional coverage area information is exploited in making proper vertical handoff decision. In this section, two representative area-based vertical handoff schemes are discussed.

The mobility models of MS are used as input data for predicting the next served AP. Zhang et al [31] predict a handoff based on the movement of MS using its current location, direction and velocity, to predict the next location L [x, y] after certain period. It finds a serving AP

of the location L and if it different from the current AP, it initiates the handoff to that AP. Distance from AP is another scheme to predict a handoff that based on the current position of the MS. The MS compares the distance from the current associated AP with the distance from the APs of neighbor cells. When MS is moving away from the current AP, it calculates the time when it will get out of the cell. If it determines that it will be out of cell in several scans later, it decides the handoff and searches for the next nearest AP. If there is a nearer AP than the current associated AP, the MS determines the handoff to the nearest AP [31].

*D. Fuzzy Logic-basedVvertical Handoff scheme*

Shih-Jung [33] proposes the Fuzzy Normalization - HandOver Decision strategy algorithm (FUN-HODS), to obtain system-loading balance and to avoid failures caused by mobile node handovers to network access points with lower velocity capabilities and weaker RSS requirements. The characteristics of fuzzy normalization were applied to handover decisions to make vertical handover decisions as simple as the horizontal one. The simulation experiments proved that the handover fail probability in the FUN-HODS algorithm is lower than the handover fail probability for a traditional fuzzy algorithm when each mobile node has random velocity.

## III DISTANCE-BASED SCHEME FOR VERTICAL HANDOFF (DSVH) IN HETEROGENEOUS WIRELESS NETWORKS

*A. Overview*

As discussed before in section two, the radio wave propagation model has been determined for hotspot communication in wireless access technologies like WLAN and WCDMA network. This model is based on extensive experimental data and statistical analysis to compute RSS for WLAN and WCDMA network. The first idea in our scheme is to rely on this radio wave propagation model to calculate the RSS in both WLAN and WCDMA network as well as to study and evaluate the handoff process between WLAN and WCDMA network and vice versa. Our distance-based algorithm is based on RSS, WLAN and WCDMA network vertical handoff threshold value and minimum distance between mobile user and corresponding APs or BSs. The second idea in our scheme is to calculate the mean throughput value for WCDMA and WLAN based on Shannon capacity theorem [22] [28]. We consider WCDMA as an example of 3G networks. Shannon's theorem is based on the average RSS value, available bandwidth and the total noise or interference power over the bandwidth. It uses these parameters in computing channel throughput. In this paper we compare mean throughput values for WLAN and WCDMA network in both SINR and DSVH schemes, this is to see whether the proposed scheme improves the throughput. The Shannon theorem states that the throughput, R, is an upper bound of data rate that can be sent with a given average RSS through an analog communication channel subject to an additive white Gaussian noise of power Γ.

The expression below represents the maximum possible rate of information transmission through a given channel or system. The channel bandwidth, the received signal level, and the noise level set the maximum throughput rate. It is computed as follows:

$$R = W \log2 (1+ \gamma / \Gamma) \quad (1)$$

Where R is the channel capacity (throughput) in bits per second. W is the bandwidth of the channel in hertz. γ is the total received signal power. Γ is the total noise or interference power over the bandwidth, measured in watt or volt. The RSS and throughput for both SINR and DSVH schemes are calculated and explained next.

*B. The SINR-based Scheme for Vertical Handoff in Heterogeneous Wireless Networks*

This section illustrates how we can calculate the RSS and throughput for both WLAN and WCDMA network based on SINR Scheme.

γ that is received at mobile user i when associated with $WCDMA_{BSj}$ [22] [29] can be represented as:

$$\gamma_{BSj,i} = G_{BS}\ P_{BS} / (P_B + \Sigma (G_{BS}\ P_{BS}) - G_{BS}\ P_{BS}) \quad (2)$$

Where $G_{BS}$ is the channel gain power between mobile user i and $BS_j$. $P_{BS}$: is transmitting power of $BS_j$. $P_B$: is the background noise power at mobile user receiver end.

γ that is received at mobile user i when associated with $WLAN_{APj}$ [22] [28] can be represented as follows:

$$\gamma_{APj,i} = G_{AP}\ P_{AP} / (P_B + \Sigma (G_{AP}\ P_{AP})) \quad (3)$$

Where $G_{AP}$ is the channel gain power between mobile user i and $AP_j$. $P_{AP}$: is transmitting power of $AP_i$. $P_B$: is the background noise power at mobile user receiver end.

In SINR scheme, Shannon's capacity formula is applied, as it is, where the value of γ represents the RSS at the mobile user. Therefore, we can use the Shannon's capacity formula (1).

*C. The Distance-based Scheme for Vertical Handoff (DSVH) in Heterogeneous Wireless Networks*

The distance-based scheme for vertical handoff in heterogeneous wireless networks that we propose is a reactive vertical handoff scheme. We suggest that vertical handoff be based on the RSS and the distances to access points or base stations, where it is able to consistently offer the mobile user with maximum available throughput during vertical handoff.

Okumura et al and Bertoni et al have developed various empirical path loss models based on RSS measurements [11]. These models are the best and most-used models for path loss distance in urban areas where there are many urban structures but not many tall buildings.

The Path Loss (PL) in dB for cellular networks (CN) environment is given by:

$$PL = 135.41+12.49 \log(f) - 4.99 \log(h_{bs}) + [46.84 - 2.34 \log(h_{bs})] \log(d) \quad (5)$$

Where d is distance in kilometer and f is the frequency in MHz. $h_{bs}$ is the effective base station antenna height in meters.

Moreover, RSS for cellular networks is expressed in dBm as:

$$P_{CN} = P_t + G_t - PL - A \quad (6)$$

Where $P_{CN}$ is the RSS of CN in dBm. $P_t$ is the

transmitted power in dBm. $G_t$: is transmitted antenna gain in dB. PL: is total path loss in dB. A: is connector and cable loss in dB.

In WLANenvironment, the path loss in dB is given by [11]:

$$PL = L + 10\, n \log (d) + S \quad (7)$$

Where L is constant power loss. n is path loss exponent with values between 2 to 4. d: represents the distance between the MS and WLAN access point. S: represents shadow fading which is modeled as Gaussian with mean $\mu=0$ and standard deviation $\sigma$ with values between 6-12 dB depending on the environment.

Moreover, the RSS for WLAN is expressed in dBm as:

$$PW = Pt - PL \quad (8)$$

Where Pt is the transmitted power and PL is the path loss in dB.

In DSVH, Shannon's capacity formula is applied, where the value of $\Upsilon$ represents the total RSS at the mobile user based on the distance, which it is computed previously in (6) and (8). Therefore, we can rewrite the Shannon's capacity formula, which is used to calculate the throughput in DSVH, as follows:

$$R = W \log2 (1 + RSS / \Gamma) \quad (9)$$

Where R is the channel capacity (throughput) in bits per second. W is the bandwidth of the channel in hertz. RSS is the total received signal power. $\Gamma$ is the total noise over bandwidth measured in watt or volt.

## IV Performance Evaluation and Analysis

In this section, we discuss the performance metrics used to evaluate our scheme. We then present and analyze the results of the simulation experiments that compare SINR with DSVH.

### *D. Performance evaluation metrics for vertical handoff schemes*

For the purpose of evaluating our proposed scheme DSVH and comparing its performance with the performance of the SINR scheme, we examined the overall system throughput and number of dropped mobile users for DSVH and SINR under identical input parameters, such as number of nodes, bandwidth and transmitted power.

Vertical handoff schemes can be quantitatively compared under various usage scenarios by measuring the mean and the maximum handoff delays, the number of handoffs, the number of failed handoffs due to erroneous decisions, and the overall throughput of a session maintained over a typical mobility model.

These metrics are further explained below:

- Number of Handoffs: the movement of MS would cause the change of RSS value received either from AP or BS. Reducing the number of handoffs is usually preferred as frequent handoffs affect the network resources. A handoff is considered dispensable when a handoff back to the original point of attachment is needed within certain time duration [1].
- Throughput: refers to the average data rate of successful packets delivery over a communication channel to all MSs in a network. Handoff to a network candidate with higher throughput is usually desirable. The throughput is usually measured in bits per second (bps), and sometimes in data packets per second or data packets per time slot.

### *E. Simulation Environment*

All simulation experiments that we were carried out on a mobile technology T3400 2.2 GHz laptop that has a dual-core Intel Pentium 64 × 2 CPU and 2 GB RAM. The operating system is the Windows Vista 32-bit operating system. The proposed scheme was added to the MATLAB (matrix laboratory) version 7.7.0.

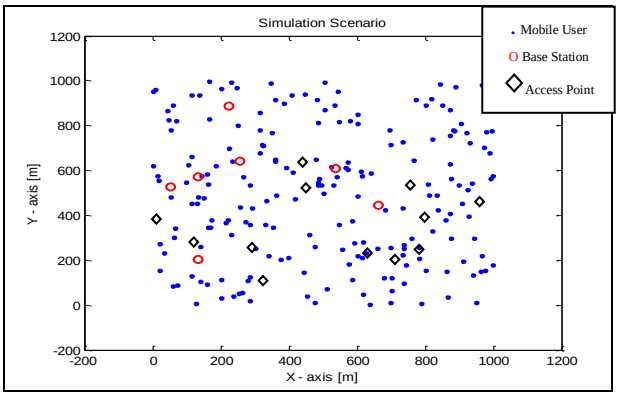



Figure 1. Simulation Scenario

Fig.1 present the simulation scenario of the network that we are based on it to evaluated and analyze the performance of the proposed distance-based scheme comparing with SINR scheme. For the experiments, the simulated network consists of 7 BSs, 12 APs and 200 MS randomly located in a space of 1000 m × 1000 m. Positions changes randomly based on a random way model [34][22].

Table 1 summarizes the different configuration values that were used in the simulations, these values that are used in SINR scheme [22][28] are also used as it in our scheme because our work aims to compare our scheme with the SINR scheme.

Table 1. simulation parameters

| Parameter | Values |
|---|---|
| Simulator | MATLAB version 7.7.0 |
| Simulation area | 1000 × 1000 $m^2$ |
| Simulation time | 200 seconds |
| Number of nodes | 200 nodes |
| Number of access points | 12 |
| Number of base stations | 7 |
| Threshold (cellular network to WLAN) | -80 dBm |
| Threshold (WLAN to cellular network) | -85 dBm |
| Antenna height of base station | 30 m |
| Access point transmitter power | 20 dBm |
| Base station transmitter power | 33 dBm |
| Cable loss | 5 dB |
| Channel gain power | 33 dBm |
| Base station 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 |

### *F. Simulation Process Flow Chart*

The following flowchart represents the process that we have used in our simulation. As we can see in Fig.2, the positions of the mobile users, APs and BSs are determined randomly. The distance between each mobile user and all APs and BSs are computed to connect each mobile user with the nearest AP or BS. Then, Path Loss (PL) and RSS are computed for each mobile user based on the attached AP or BS by using the equations that are illustrated previously in section four. The threshold (-80 or -85 dBm) is used to determine dropped users. Finally, we apply Shannon's capacity theorem to compute the mean throughput for both SINR and DSVH based on the RSS and SINR values that result from the previous steps.

### *G. Number of Mobile Users Factor*

The purpose of the simulation presented in this experiment is to study the effect of varies number of mobile users on the mean value of dropped mobile users in both SINR and DSVH schemes. In all scenarios of this experiment, the number of mobile users ranges from 100 to 300 mobile users with an increment of 50 mobile users. The subsequent sections discusses the effect of varies number of mobile users on the mean value of dropped mobile users in both WLAN and WCDMA network. The simulation results presented in Fig.3 illustrate the mean value of dropped mobile users for both SINR and DSVH schemes in WCDMA network. From this Figure we can show that as the number of mobile users increases in DSVH scheme, the mean value of dropped mobile users not changed, it is equal to 7. The reason is that no equations that we have adopted in the DSVH scheme take into account the number of mobile users as in (5) and (6).

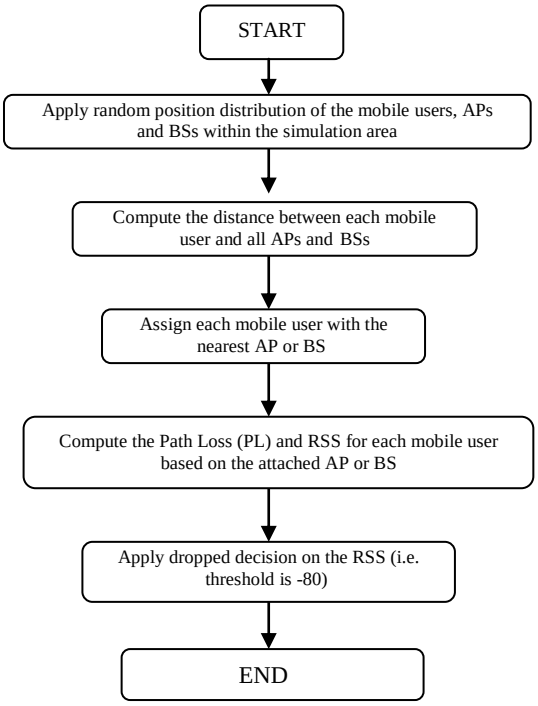



Figure 2. Simulation Process

While in SINR scheme, as the number of mobile users increases the mean value of dropped mobile users becomes decreases from 17 until it reaches to 4 when the number of mobile users becomes 300 MSs. The reason is that the equation that we have adopted in the SINR scheme take into account the channel gain power between each mobile user and it's base station (2). In general, the DSVH scheme outperforms the SINR scheme, if the number of mobile users not exceeded the 250 mobile users, but after that, the SINR scheme will be better than DSVH scheme for handling the vertical handoff process, whereas the mean value of dropped mobile users becomes lower.As a result, DSVH scheme achieves major enhancement in terms of reducing the mean value of dropped mobile users comparing with SINR scheme when the number of MS not exceeded the 250 by 20%. The simulation results presented in Fig.4 illustrate the mean value of dropped mobile users for both SINR and DSVH schemes in WLAN. From this Figure we can show that as the number of mobile users increases in DSVH scheme, the mean value of dropped mobile users is fixed, it is equal to 4. The reason is that no equations that we have adopted in the DSVH scheme take into account the number of mobile users as in (7) and (8). While as the number of mobile users increases in SINR scheme, the mean value of dropped mobile users decreases from 18 until it reaches to 4 when the number of mobile users becomes 300 MSs.

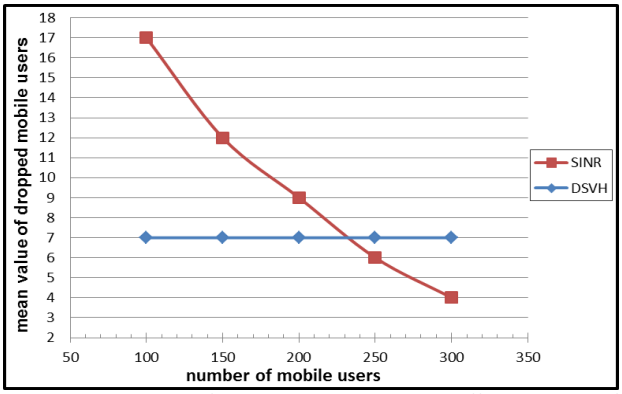



Figure 3. Mean value of dropped mobile users with different number of MSs for WCDMA network in both SINR and DSVH schemes

Thereason behind decreasing the number of dropped mobile users is that the decrease in the number of access points could leads to reduce the interference that may affect on the mobile users, and thereby increase the signal strength and then decrease the number of dropped users (3). In general, the DSVH scheme outperforms the SINR scheme, if the number of mobile users not exceeded the 300 mobile users, but after that when the number of mobile users exceeded the 300 mobile users, the SINR scheme will be better than DSVH scheme for handling the vertical handoff process, whereas the mean value of dropped mobile users becomes lower. As a result, DSVH scheme achieves major enhancement in terms of reducing the mean value of dropped mobile users comparing with SINR scheme by 10%, as the number of users increases until it reaches 300 mobile users, after that the contrast may occurred.

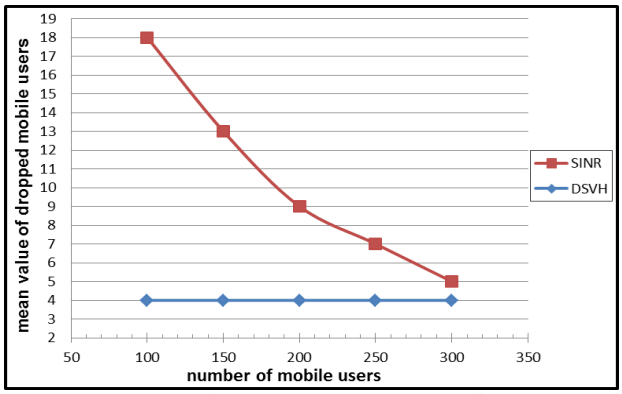


Figure 4. Mean value of dropped mobile users with different number of MSs for WLAN in both SINR and DSVH schemes

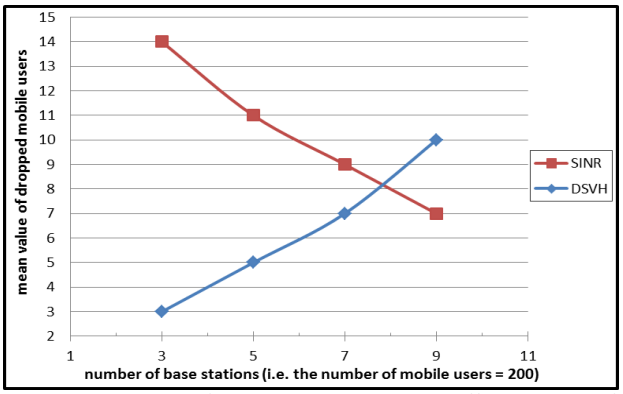


Figure 5. Mean value of dropped mobile users with different number of base stations for WCDMA network in both SINR and DSVH schemes

*H. The Number of Base Station Factor*

The purpose of the simulation presented in this experiment is to study the effect of using different number of base stations on the mean value of dropped mobile users in both SINR and DSVH vertical handoff schemes. In all scenarios of this experiment, the number of base stations varies from 3, 5, 7 and 9 base stations, while the number of mobile users is fixed, it is equal to 200. The simulation results presented in Fig.5 illustrate the mean value of dropped mobile users for both SINR and DSVH schemes in WCDMA network. From this Figure , we can show that the mean value of dropped mobile users in our proposed scheme DSVH is less than the mean value of dropped mobile users in SINR scheme while the number of base station not above 8, but when the number of base station exceeded 8, the SINR scheme becomes better than DSVH scheme for handling the vertical handoff process.

In addition, we can see that as the number of base stations increases in the DSVH scheme, the mean value of dropped mobile users becomes increase from 3 until it reaches to 10. The reason maybe that the increase in the number of base stations may lead to decrease the distance between the base station and the mobile user during his movement, but this decrease will be for a few periods.

Thus, the vertical handoff process for the mobile user to another base station may increase with a decrease of time that may be linked to the mobile user with a previous base station. While as the number of base stations in SINR scheme increases, the mean value of dropped mobile users becomes decreases from 14 until it reaches to 7 when the number of base stations becomes 9 base stations. This is because the SINR scheme influenced by a number of base stations in a positive ratio, so when the number of base stations increased, the percentage of signal strength to noise ratio be higher.

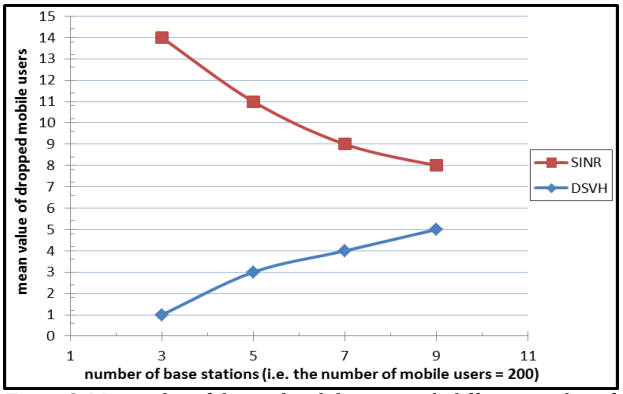


Figure 6. Mean value of dropped mobile users with different number of base stations for WLAN in both SINR and DSVH schemes

On the contrary, from DSVH scheme. In general, DSVH scheme achieve major enhancement in terms of reducing the mean value of dropped mobile users comparing with SINR scheme when the number of base stations not exceeded 8 base stations by 20%.

The simulation results presented in Fig.6 illustrate the mean value of dropped mobile users for both SINR and DSVH schemes in WLAN. From this Figure , we can show that the mean value of dropped mobile users in our proposed scheme DSVH is less than the mean value of dropped mobile users in SINR scheme. In addition, as the number of base stations increases in the DSVH scheme, the mean value of dropped mobile users becomes increase from 1 until it reaches to 5. While in SINR scheme, as the number of base stations increases the mean value of dropped mobile users decreases from 14 until it reaches to 8 when the number of base stations becomes 9 base stations. Of course, in both cases, the reasons are the same as those shown for the WCDMA network. In general, DSVH scheme achieve major enhancement in terms of reducing the mean value of dropped mobile users comparing with SINR scheme when the number of base stations not exceeded 8 base stations by 30%.

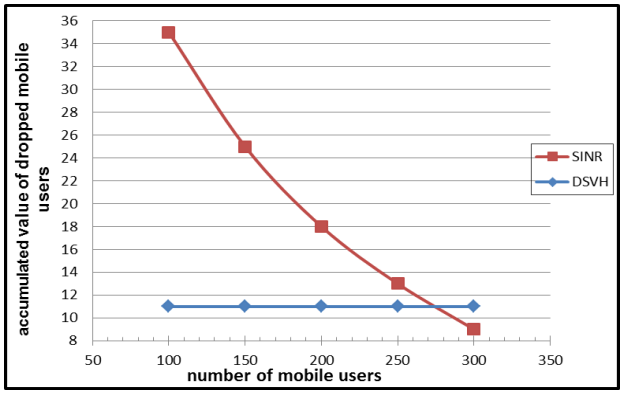



Figure 7. Accumulated value of dropped mobile users in both WLAN and WCDMA network for SINR and DSVH schemes with varies number of mobile users

### *I. Accumulated value of dropped mobile users with varies Number of Mobile Users*

Fig.7 display the accumulated value of dropped mobile users in both WLAN and WCDMA network with varies number of mobile users. From this Figure , we can show that as the number of mobile users increases in DSVH from 100 to 300, the value of dropped mobile users is fixed, it is equal to 11. While in the SINR scheme as the number of mobile users increase the mean value is decreases from 35 to 13 if the number of mobile users is 250. However, when the number of mobile users reaches to 300 the mean value of dropped mobile users in SINR scheme becomes lower than the mean value of dropped mobile users in the DSVH scheme.

In general we can said, if the number of mobile users not exceeded the 250 the enhancement of mean value of dropped mobile users in DSVH outperforms the enhancement of SINR scheme by 20 %. On the other hand, Fig.8 shows the accumulated value of dropped mobile users in both WLAN and WCDMA network with varies number of base stations. From this Figure , we can see that as the number of base stations increases from 3 to 9, the mean value of dropped mobile users in DSVH scheme is increases from 4 to 15 mobile users. While as the number of mobile users increase in the SINR the mean value of dropped users is decreases from 28 to 15 if the number of base station is 9.

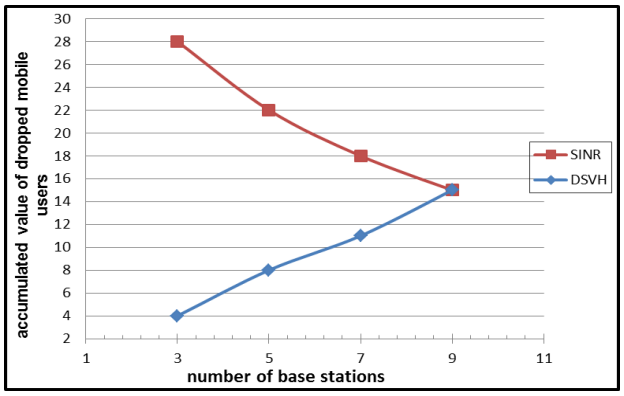



Figure 8. Accumulated value of dropped mobile users in both WLAN and WCDMA network for SINR and DSVH schemes with varies number of base stations

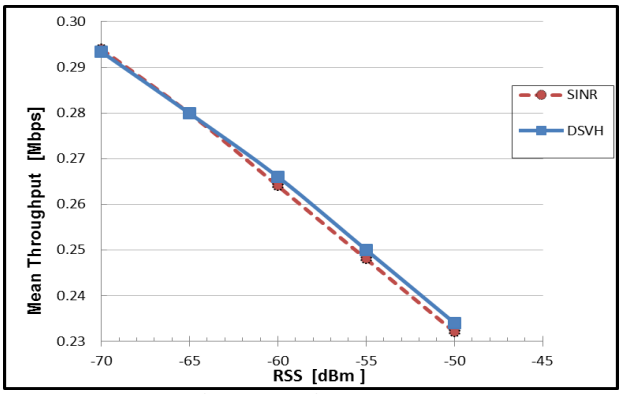



Figure 9. mean value of throughput for WCDMA network in both SINR and DSVH schemes

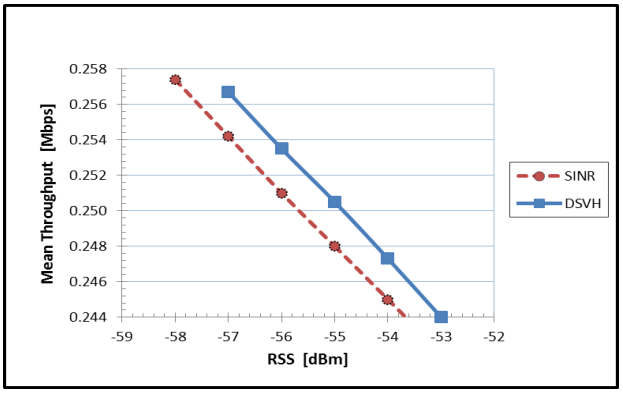



Figure 10. Mean throughput with zoom for WCDMA network in both SINR and DSVH schemes

### *J. Mean Value of Throughput*

Fig.9 illustrates the mean value of throughput for both SINR and DSVH in WCDMA network. As we can see, the DSVH scheme outperforms the mean value of throughput for SINR. Fig.10 represents the same Figure but with a zoom to better illustrate the differences between the two schemes. We can see that the enhancement on the mean throughput is very slight (less than 1%).

Next, we measure the ratio of improvement defined as:

***Improvement* = (*New value* – *Old value*)/*Old value* * 100 %**

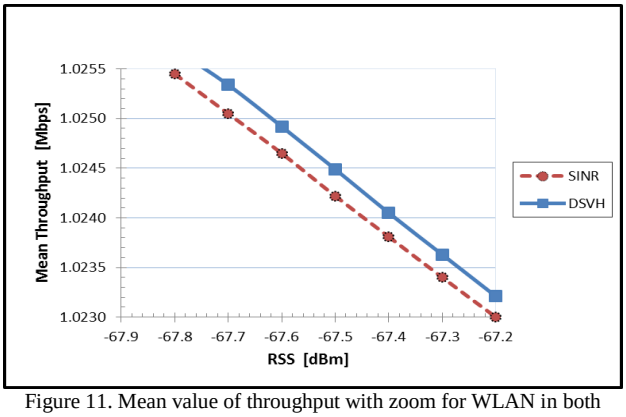



Figure 11. Mean value of throughput with zoom for WLAN in both SINR and DSVH schemes

Fig.11 illustrates the mean values of throughput for both SINR and DSVH in WLAN with a zoom to better

illustrate the differences between the two schemes. As we can see the DSVH scheme outperforms the mean value of throughput for SINR. The enhancement on the mean throughput is very slight (less than 1%).

## V Conclusions and Future Work

In this paper, we present the design and simulation of our distributed distance-based scheme for vertical handoff in heterogeneous wireless networks and provide performance measurements using the MATLAB. The major issues in our paper are presented below.

Our scheme shows that the distance is a better metric for minimizing service disconnection probability and maximizing system throughput when selecting the best AP or BS. The main goal of our scheme has been achieved. It is to enhance and provides higher overall system performance in terms of minimizing service disconnection probability during vertical handoff as compared with the SINR based vertical handoff scheme.The simulation experiments show that our proposed scheme, DSVH, significantly outperform the SINR scheme in terms of reducing the number of dropped users. It reduces this number by 20%. The throughput remains almost the same; its improvement is very slight (i.e. Less than 1%). There are still several research points that can be investigated based on this work. It would be interesting to explore our proposed scheme with a scheme that takes in consideration additional parameters, such as the velocity and mobility direction and the location of mobile users by using GPS technology. Devising an algorithm that is useful in a wide range of conditions and user preferences. Once possible solution would be to implement several vertical handoff algorithms and then adopt adaptive methods that choose an algorithm intelligently based on conditions and user preferences.

## VI References


1. Xiaohuan Y, Y.Ahmet S, Sathya N. a survey of vertical handover decision algorithms in fourth generation heterogeneous wireless networks. The International Journal of Computer and Telecommunications Networking; 2010; 54(11): 1848-1863.
2. Eva G, Annika J. always best connected. IEEE Wireless Communications; 2003; 10(1): 49 – 55.
3. Enrique S, Yuxia L, Vincent W. an MDP-based vertical handoff decision algorithm for heterogeneous wireless networks. In Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2007); 2007 March; 57(2): 1243 – 1254.
4. Sasha D, JP S, Upkar V, Geoffrey K. evolution and emerging issues in mobile wireless networks. Communications of the ACM; 2007; 50(6):38–43.
5. Suk H, Kai Y. challenges in the migration to 4G mobile systems. IEEE Communications Magazine; 2003; 41(12): 54–59.
6. Qing-An Z, Dharma A. handoff in wireless mobile networks. In: Handbook of wireless networks and mobile computing; 2002: 1–25.
7. Hossen A. vertical handover scheme for next generation mobile networks. [dissertation]. Faculty of Engineering at AL-Mergib University; 2007 May.
8. George L, Alexandros K, Nikos P, Lazaros M. handover management architectures in integrated WLAN/cellular networks. Communications Surveys and Tutorials, IEEE Communications Society; 2005; 7(4): 30–44.
9. Janis M, Fang Z. vertical handoffs in fourth-generation multinetwork environments. IEEE Wireless Communications; 2004; 11(3): 8–15.
10. Enrique S-N, Vincent W. comparison between vertical handoff decision algorithms for heterogeneous wireless networks. In Proceedings of the 63rd Vehicular Technology Conference (VTC'06), Melbourne, Australia; 2006 May: 947–951.
11. Ayyappan K, Dananjayan P. RSS measurement for vertical handoff in heterogeneous network. Journal of Theoretical and Applied Information Technology; 2008; 4(10): 989-994.
12. Syuhadal A, Mahamod I, Firuz S. performance evaluation of vertical handoff in fourth generation (4G) networks model. In Proceedings of IEEE 2008 6th National Conference on Telecommunication Technologies and IEEE 2008 2nd Malaysia Conference on Photonics, Putrajaya, Malaysia; 2008 August: 392 – 398.
13. Mian I. Radio Propagation. [Online] [Accessed August 2010]. Available from URL http://web.uettaxila.edu.pk/CMS/teWCbs/notes%5CLec%207%20Radio%20Propagation.pdf.
14. Theodore R. mobile radio propagation: large-scale path lose. In: Wireless Communications Principles and Practice; 2002: 69–138.
15. Tapan S, Zhong Ji, Kyungjung K, Abdellatif M, Magdalena S-P. a survey of various propagation models for mobile communication. IEEE Antennas and Propagation Magazine; 2003 June; 45(3): 5-82.
16. Mike W. mobile systems. [Online] [Accessed August 2010]. Available from URL http://www.mike-willis.com/Tutorial/MP.pdf.
17. Omorogiuwa O, Edeko F. investigation and modeling of power received at 1800 MHz in a mountainous terrain. case study of Lgarra in Edo State, Ajaokuta and Okene in Kogi State, Nigeria. International Journal of Electrical and Power Engineering; 2009; 3(3): 129-135.
18. Hata M. empirical formula for propagation loss in land mobile radio service. IEEE Transactions on Vehicular Technology; 1980; 3: 317-325.
19. Javier D, Rodrigo M. bluepass: an indoor bluetooth-based localization system for mobile applications. Symposium on Computers and Communications (ISCC), 2010 IEEE, Italy; 2010 June: 778 – 783.
20. Alexander W, Johnny K, Pooya S. long distance path-loss estimation for wave propagation through a forested environment. Antennas and Propagation Society International Symposium, IEEE Communications Society; 2004 June; 1: 922-925.
21. Zhu H. wireless communications and networks. [Online] [Accessed July2010]. Available from URL http://www.slidefinder.net/5/552_452_Spring_2008_Wireless/6995630.
22. Kemeng Y, Iqbal G, Bin Q, Laurence D. combined SINR based vertical handoff algorithm for next generation heterogeneous wireless networks. In Proceedings of the 2007 IEEE Global Telecommunications Conference (GLOBECOM '07), Washington, DC, USA; 2007 November: 4483 – 4487.
23. Ahmed Z, Ben L, Aladin S. signal threshold adaptation for vertical handoff in heterogeneous wireless networks. ACM/Springer Mobile Networks and Applications (MONET) journal; 2006; 11(4): 625–640.
24. Ahmed Z, Ben L. performance evaluation framework for vertical handoff algorithms in heterogeneous networks. In Proceedings of the 2005 IEEE International Conference on Communications (ICC'05), Seoul, Korea; 2005 May; 1: 173–178.
25. Xiaohuan Y, Mani P, Sekerciog-lu Y. a traveling distance distance based method to minimize unnecessary handovers from cellular networks to WLANs. IEEE Communications Letters; 2008; 12(1): 14–16.
26. Xiaohuan Y, Sekerciog-lu Y, Mani P. a method for minimizing unnecessary handovers in heterogeneous wireless networks. In Proceedings of the 2008 International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'08), Newport Beach, CA, USA; 2008 June: 1–5.
27. Kemeng Y, Iqbal G, Me, Bin Q. multi-dimensional adaptive SINR based vertical handoff for heterogeneous wireless networks. IEEE COMMUNICATIONS LETTERS; 2008 JUNE; 12(6): 438 – 440.
28. Ayyappan K, Dananjayan P, Narasimman K. SINR based vertical handoff scheme for QoS in heterogeneous wireless networks. In Proceedings of the 2009 International Conference on Future Computer and Communication, IEEE Communications Society; 2009 April: 117 – 121.
29. Ayyappan K, Kumar R. QoS based vertical handoff scheme for heterogeneous wireless networks. In Proceedings of the International Journal of Research and Reviews in Computer Science

(IJRRCS'10); 2010; 1(1): 1-6.
30. Rafiq M, Kumar S, Kammar N, Prasad G, Krishna G. a vertical handoff decision scheme for end-to-end QoS in heterogeneous networks: an implementation on a mobile IP testbed. In Proceedings of the 2011 National Conference on Communications (NCC), IEEE Communications Society; 2011 Jan: 1 - 5.
31. Jie Z, Henry C, Victor L. a location-based vertical handoff decision algorithm for heterogeneous mobile networks. In Proceedings of the 2006 IEEE Global Telecommunications Conference (GLOBECOM '06), San Francisco, CA; 2006 November: 1 - 5
32. Lott M, Siebert M, Bonjour S, Hugo D, Weckerle M. interworking of WLAN and 3G systems, IEEE Communications Society; 2004 October; 151(5): 507-513.
33. Shih-Jung Wu. fuzzy-based handover decision scheme for Next-generation heterogeneous wireless networks. Journal of Convergence Information Technology; 2011 April; 6(4): 285-297.
34. Jain R. art of computer systems performance analysis. In: Wireless Communications Principles and Practice; 1991.

# Framework of Competitor Analysis by Monitoring Information on the Web

Simon Fong
Department of Computer and Information Science
University of Macau
Taipa, Macau SAR
Email: ccfong@umac.mo

***Abstract*—Competitor Analysis (CA) is an important part of the strategic planning process. By spying every move of the competitors offers a company an advantageous position in decision making. With the advent of World-Wide-Web technology, many businesses extend their activities to the Internet platform, from online marketing to customer services, thereby left over traces of valuable information for competitor analysis. The scattered information can be collected from the Web for studying what the competitors are doing and what products and services they offer up-to-date. In the past CA was conducted manually and it was a tedious process. The sources of such Web information are diversified and the contents are updated frequently too. It is therefore desirable to build an efficient and automated tool that gathers competitor information, monitors the updates and formulates them into useful competitor intelligence with minimum human intervention. Although many information retrieval and monitoring technologies have been developed, they are more for generally tracking changes and downloading the whole websites for offline browsing. In this paper, a framework of automated competitor analysis is proposed as a holistic solution. It is a Web monitoring tool that works for both Web 1.0 and Web 2.0. The designs of the components and the operational challenges are discussed respectively.**

***Index Terms*—competitor analysis; Web information; Information system design**

## I. Introduction

World-wide-web (or simply the Web) has grown into a huge virtual world of information by itself, where information are dynamically generated, disseminated, and viewed by millions of users on a global scale. Businesses tapped on the power of the Web for doing online marketing, online customer services and released a significant amount of information about their products, services, promotions and latest news. More and more companies commit to update their websites for showing their latest business information for publicity. At the same time, Web users posted, commented and discussed about the businesses and their products and services online. Out of this tremendous pool of information, some information, either implicit or explicit, which are pertaining to competitors' activities and news, are valuable for doing competitor analysis for a company. From the perspectives of competitor analysis, it is vital for a company to keep informed of what their competitors are doing and what products and services they offer up-to-date; and then absorb feedbacks and opinions from the Web users about their products and services.

By acquiring such information from the Web, the company can gather business intelligence for planning countermeasures and remain competitive. Hence it is crucial for a company to have the right tool to effectively gather such information from the Web. Many information retrieval and monitoring technologies have been proposed in the literature, such as OpenCQ [1], WebCQ [2], CONQUER [3] and Niagara [4]. There are many commercial products too in models of Application Service Provider such as the ones listed in Table 1.

TABLE I. Web Monitoring Systems

| Service/Product | URL |
|---|---|
| WatchThatPage | http://www.watchthatpage.com |
| Wisdomchange | http://www.wisdomchange.com |
| ChangeDetection | http://www.changedetection.com |
| ChangeDetect | http://www.changedetect.com |
| Track Engine | http://www.trackengine.com |
| WebsiteWatcher | http://www.aignes.com |

They are generally content monitoring services that periodically watch Web pages and other Internet resources for keyword related content or changes. However they are more for generally tracking changes and downloading the whole websites for offline browsing than for tactical CA.

In this research, we focus on Web business environment in which knowing one's competitors is of crucial importance to the survival and growth of any business. Before the Internet became popular, it used to be an expensive process in obtaining business intelligence information quite often from printed publication and other media channels. Although nowadays much of the information is freely available from the Web, they are scattered and dynamic in nature. Like searching for a grain in an ocean, acquiring useful information from the competitors' Web sites as well as from other Web news portals is still a tedious and time-consuming task. Furthermore, many companies nowadays extended their e-business activities to Web 2.0 for example Facebook and Twitter, for engaging customers and extending their marketing platform online. Data from Web 2.0 are known to be scattered, distributed and unstructured in contents; the updates of such data of Web 2.0 are more dynamic than that of page contents in traditional website. Web 2.0 certainly imposed extra challenges for the job of


doi:10.4304/jetwi.4.1.77-83

competitor analysis, especially in the information retrieval and updates tracking.

Many companies today opt to invest certain resources in collecting information about their competitors from the Web and other channels. It is a regular routine that they keep track of what their competitors are doing, what products and services they offer and any news that concern about them. This is usually done by manual browsing, by the marketing personnel. From our industrial experiences, the approach of acquiring business intelligence from the Web by manual browsing poses a number of problems, such as:

1. Business websites especially that of the large international cooperates often have a large number of pages in the number of hundreds, which makes it very tedious for manual browsing without any automated assistance. The overwhelming amount is a major cause to human errors in selecting the correct information.
2. Different companies may organize the same information very differently, due to differences in culture and practices. One company may use one format and another may follow a different style. The diverse and unstructured formats make manual browsing a daunting.
3. Beyond the competitors' websites, there are certainly other websites, forums, news portals feature news about the competitors and their products. Searching the World-Wide-Web for all possible sites that might have mentions of the competitors is an extremely tedious task if done manually.
4. The speed of updates on some information portal, such as stock market, headline news, news feed from social networks could be beyond that of a human task that consumes time in continuously searching, downloading, extracting, analyzing and archiving. The balance of completeness and timeliness of downloading online information has been discussed in [5]. This implies some automated process must be implemented to capture the right information over the Web at the right time intervals.

The amount of information within a site and new sites is growing at a phenomenal rate. Also for social networks, new user accounts, blogs or groups that are relevant to the CA concerns, emerge very rapidly; their contents are constantly refreshed and are forever growing. Monitoring such information can no longer be easily done manually.

## II. Our Proposed Framework

In this paper we propose a competitor analysis framework that is based on the design of an automated market monitoring Web agent system, namely Market Watcher Agent, for gathering business information relevant to a company in an automated approach. The technology is designed to assist competitor analysis that has the following important roles in strategic planning:

- *To help management understand their competitive advantages/disadvantages relative to competitors.*
- *To generate understanding of competitors' past, present (and most importantly) future strategies.*
- *To provide an informed basis to develop strategies to achieve competitive advantage in the future.*
- *To help forecast the returns that may be made from future investments (e.g. how will competitors respond to a new product or pricing strategy?)*

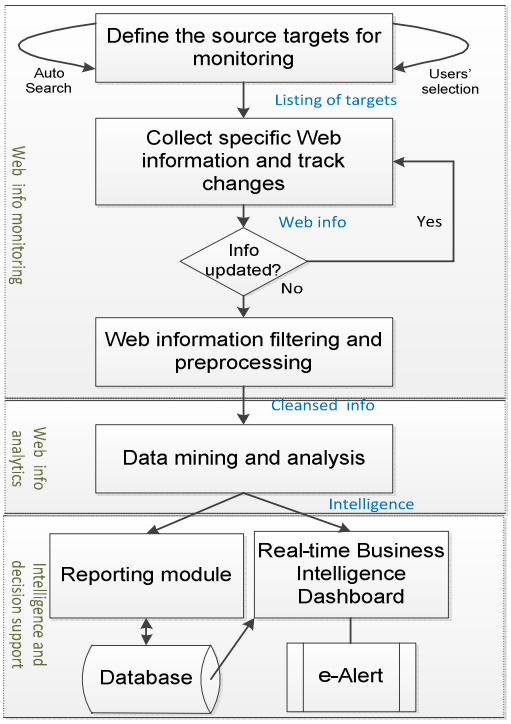



Figure 1. The workflow of the CA framework by monitoring Web info.

The workflow of the CA framework by monitoring Web information is presented in Figure 1. All the steps are supposed to fulfill the roles of CA as an ultimate goal as mentioned above. The workflow can be briefly divided into three phases, namely *Web Information Monitoring*, *Web Information Analytics* and *Intelligence and Decision Support*.

Data mining and decision support techniques are applied that convert collected Web information into meaningful business intelligence. The business intelligences are in turn, presented and used by the management through reporting and dashboard technologies respectively. In real-time, business intelligence of two levels, statistical numeric level and raw messages, and abstract level in which the messages carry semantic meanings can be presented. If needed, users can preset certain rules so that instant alerts can be sent to the managers' mobile phones or PDA for immediate attention.

For example, the decision support module should be able to answer the following questions when undertaking competitor analysis:

- *How did our competitors compete with us?*
- *What threats do they pose?*
- *What are the estimated impacts as a result of their latest competitive move? (forecast and chronicle)*
- *What are the objectives of our competitors? And how far they have achieved or failed them?*
- *What strategies are our competitors pursuing and how successful are these strategies?*
- *What are the strengths and weaknesses of our competitors?*
- *How are our competitors likely to respond to any changes to the way we do business?*

The Market Watcher Agent that is the main power horse in the Web Information Monitoring layer (as in Figure 1) is an autonomous software program that "spies" on the competitors' prices and news information over the Web. The watcher agent consists of mainly two parts, namely, market watcher and price watcher. In this paper, market watcher is described in detail while detailed description on price watcher can be found in [6]. The Market Watcher is an information-collecting tool, which assists the users to monitor the specified Web sites, e.g. competitors' web sites, and to locate the relevant information automatically. The market watcher has two sub-components, namely market monitor and market explorer. The Market Monitor works as an information filter. The objectives of the Market Monitor is to find the news updating on competitors' web sites and to find news articles from some established news portals for particular products, for example, CNN and BBC. The Market Explorer, on the other hand, is an information provider, which is able to get the most updated information worldwide. As the Internet is extremely dynamic, it will never be enough to get news from a fixed number of Web sites, i.e., the competitors' official web sites and certain news portals. With the help from various types of the Internet search engines, worldwide information can be collected by passing users' queries to the search engines and retrieving the top matches. At least two kinds of information can be discovered with the search records from such popular search engines. One is the information about a particular product that cannot be easily found from the competitors' official web sites or news portals: for example, user's feedback or technical web site's product reviews. The other is the product ranking, or how prominent your product can be reached by the Internet users from search engines. For example, if the query "inkjet printer" is given to Google search engine, manufacturers in the top 20 matches will be Epson, Kodak and HP.

In summary, Table 2 lists the information which constitutes to business intelligence, and from where over the Web such information could be obtained.

TABLE II. SUMMARY OF WEB INFORMATION TO BE MONITORED

| Where can be found | Web information as business intelligence |
|---|---|
| Competitors' websites | Competitors price information |
| | Competitors product information |
| | Competitors company information |
| News portals | Product and company news on media |
| Social networks | Product and company news |
| | Customers' feedbacks and community impacts |
| Popular search engines | Product and company news as search results returned from search engine |
| | Search engine rankings |

The rest of the paper is organized as follows. In Section 3, the system architecture is described in detail. Operational processes for market monitor and market explorer are presented in Section 4 and 5 respectively. Section 6 discuses about the performance evaluation. Finally we conclude our work in Section 7.

## III. SYSTEM ARCHITECTURE

As shown in Figure 2, in the Information Retrieval Layer, the URL Retrieval Engine takes two parameters as input, the URL and downloading level. The URL Retrieval Engine issues requests to the corresponding Web server and retrieves Web pages. The Web pages are then stored as data files. On the other hand, the matches from the Internet search engines are also retrieved in retrieval layer. The search queries are from the Market Watcher.

The Compilation Layer is the core of the Watcher Agent, which includes the Price Watcher and Market Watcher. The Price Watcher takes the data files and the list of product names. It then detects the matched product names, and extracts price respectively from the Web pages. The Market Watcher is made up of two sub-components, namely, Market Monitor and Market Explorer. The Market Monitor monitors the Web pages for interesting updates and news information as an information filter. The Market Monitor can be set to work repeatedly on either daily or weekly basis. However, there could possibly be overwhelming amount of news on the Web sites. Hence it provides an instant events section. A small module named Instant News Watcher will be designed as a part of the Market Watcher Agent. The Instant News Watcher can monitor the instant news section from few important Web sites on an hourly base or even every few minutes upon scheduled.

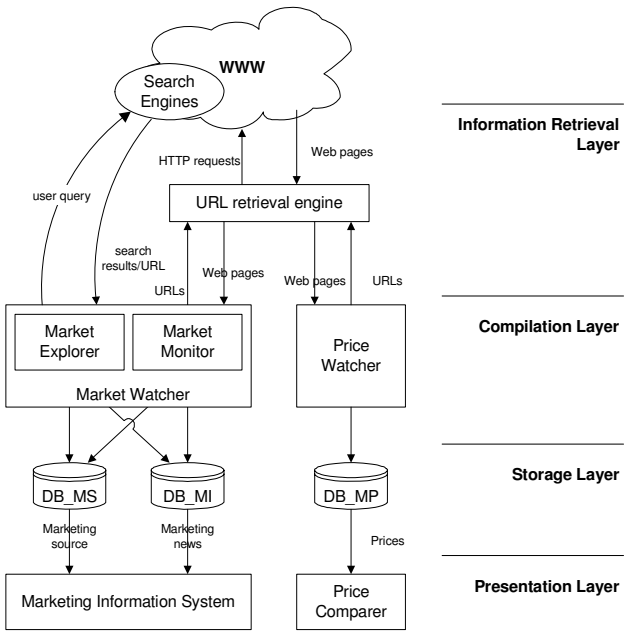



Figure 2. Watcher Agent System Architecture

The Market Explorer is a both retrieval and formatting tool, which passes the user queries to a number of popular search engines available on the Internet and collects the matches from each of the search engines.

The Storage Layer is the local database, which stores price information collected by Price Watcher, the marketing news from Market Watcher, and the searching matches from the Market Explorer.

The Presentation Layer consists of a set of Web pages generated from the local database upon request. Part of the Web interface works for the price comparison of various Web sites on any particular product. For the Market Monitor, a set of market news reports will be generated directly from the database. The instant news report can also be generated upon user's request. The local search engines that directly work on the database are also a part of the presentation layer. These search engines can help user to locate the relevant information from the local database. Since the data is ready to use, local searching will save users' time substantially. For instance, three local search engines can be developed for Market Watcher Database, Price Database, and Market Explorer search history respectively. Search engine for the Market Watcher Database will help user search for the news reports about the query. Search engine for the Price Watcher Database will be able to locate the price of a particular product easily. Users may find relevant information from the search history with the help of a local search engine rather than make a search session over the Internet.

## IV. Market Monitor

### A. Information Extraction

One important operation of market monitor is to extract information from competitors' web sites. Given a list of pre-defined competitors' web sites, information about new product release or similar in other area should be extracted using a full or semi-automatic HTML wrapper [7]. A HTML wrapper is a kind of software that extracts a certain paragraph or a section from HTML pages based on the HTML tags, formatting or structural information. HTML wrappers normally make use of rule-based leaner or machine learning techniques to learn how to extract the desired information accurately based on the given sample pages. With such a HTML wrapper, market monitor scans the web sites (given by the users) based on the schedule setting and extracts the detailed news information.

### B. Information Filtering

Another information resource for the market monitor is from the news portals. Different from the pre-defined the competitor's web sites, majority of the news articles from a new website, say CNN or BBC, are normally not relevant to competitors. Taking an Information filtering approach is necessary to filter out the unrelated information. Filtering is the operational mode in which the queries remain relatively static while new documents come into the system (and leave).

In this filtering task, a user profile describing the user's preferences is constructed. Such a profile is then compared with the incoming documents in an attempt to determine those that might be of interest to this particular user profile. Hence, the filtering approach can be used to select news articles broadcast every day or the newly uploaded Web pages from specific Web sites. One of the difficulties to design such a filtering system is on how to construct the user profile that truly reflects the user's preferences.

Typically, the filtering task simply indicates to the user the documents that might be of interest to him. The task of determining which ones are really relevant is fully reserved to the user. The documents ranking generated by the system may or may not be presented to the user. However an internal ranking is normally computed to determine potential relevancy of documents. For example, the documents with a higher ranking than the predefined threshold may be selected.

### C. User's Profile Construction

Two approaches, which are introduced in [8], can be described as words "static" and "dynamic". The static approach is to simply take a set of keywords from user to construct the user's profile. The keywords then can be directly compared with the documents arriving at the system. A similar way is used by a number of Web sites like Hotmail where a set of choices are listed to be selected by users. The choices may be personal interests such as sports news, music, or computer news. The result of the selection will be used to construct a simple user profile. The dynamic approach works exactly the same as the static one in the beginning. A set of keywords is required from user to construct an initial profile. As new documents arrive, the system uses the initial profile to select the documents of potential interest and present to user. The user will then go through the recommended documents, select the relative ones and pass this information back to the system with a feedback process. The system uses this feedback information to adjust the user profile so that it reflects the new preferences just declared. The dynamic approach keeps catching up with the user's feedback and adjusts the profile to be as close to the user's preferences as possible. This is how the dynamic approach will be employed in the Market Watcher Agent.

Since the Market Watcher is designed to be categorization supported, one user profile is constructed for each category. The user profile or category profile in this case, is the keywords given by both the system users. The keywords are called domain keywords for that category.

The dynamic category profile is achieved with a user feedback session. For each Web page recommended by the Market Watcher Agent, there are three choices: worth reading - user agreed with the recommended Web page, no comments - user did not give any choice, and don't waste my time - user disagreed with the recommended Web page, for the user to feedback. Among the total number of users willing to give feedback, the feedback value is derived with the following formula.

$$\text{feedback value} = \frac{\text{User Agreed - User disagreed}}{\text{Total number of user feedback}}$$

The total number of feedbacks received must be greater than the predefined threshold to make the feedback value valid. In case of the feedback value is

high enough (i.e. greater than the threshold), the top matched keyword and the top index term from the Web page will be added to the category profile. The top matched keyword refers to the one from the Web page and has the most number of matches with any of the keywords from user's query. The top index term is the most frequently used index term from the Web page. As a result, the feedback from users is represented by the weight increase of the corresponding index terms in the user query.

### *D. News article filtering based on document similarity*

The input for the Market Watcher is the category profile and a set of Web pages. The output is the information indicating which pages are relevant to the user's requirement. The information is stored in the local database and news report can therefore be generated. The information includes the Web page title, updated time, URL, similarity level, and a summary. The summary will be the first one or two sentences of the page as most of the news writers give a summary in the beginning of the news article. The similarity is calculated with the following formula:

$$Similarity = Query \text{ Vector} \times \text{Document Vector}$$

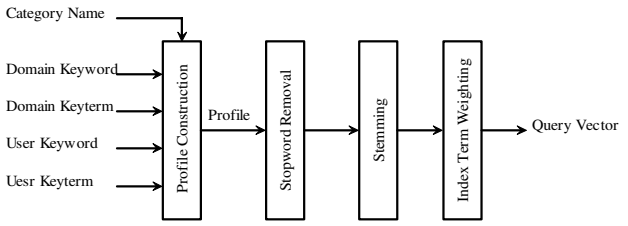



Figure 3. Query Vector Process

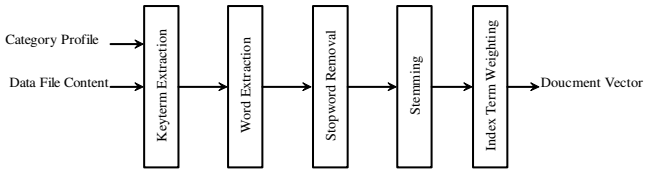



Figure 4. Document Vector Process

### *E. Instant News Watcher*

The Instant News Watcher is developed to retrieve the most updated news from homepages of some Web sites where there are instant event sections. The inputs are category profiles and homepages, and the output will be the updated news extracted.

The Instant News Watcher is a special case of the Market Watcher where more frequent monitoring is required. However, the output is the news instead of the summary of the entire page. The entire structure and most of the contents of the homepage of one Web site will not be updated on daily basis. The frequently updated part is the instant news section or instant event section. Therefore the document-ranking algorithm to calculate the similarity level of the profile and entire Web page cannot be applied to the Instant News Watcher.

In the Instant News Watcher design, the content of the Web page is divided into small sections based on its internal structure with help of the Semi-Data Tree Model that has been used in [9, 10]. Each time, one small section, e.g., one paragraph, one table cell, or one list item, is compared to the category profile. Since the input data is not so much different from the entire Web page, the similarity level given by the document-ranking algorithm will be relative low. For this reason, it's hard to set any default threshold. Therefore, the exact pattern matching is good enough in this case. If one section matches any of the keywords or key-terms from the category profile, the section is considered to be relevant and will then be saved to the database. The duplicated sections will be detected before storing to the database in order to save space.

## V. MARKET EXPLORER

Market Explorer is a part of the Market Information System. The objective of the Market Explorer is to assist users to locate the required information with a number of popular search engines available such as AltaVista, Lycos, Excite, Yahoo, Catcha and InforSeek. Other than news from the competitors' official web sites and popular news portals, information about product review or user feedback cannot be easily extracted from the market monitor. With the help of such engines, these kinds of information can be located with the user keywords queries. Similar to market monitor, two functions have been incorporated into the market explorer.

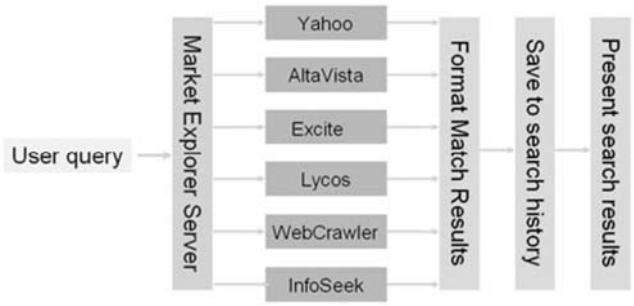



Figure 5. Operational block diagram of Market Explorer

### *A. Information Extraction*

To extract information from the popular search engines, simple keywords queries need to be derived in advance. The keywords can be product names, competitors' names, product brands or a combination of these. From the top matches of the search engines, say top 100, relevant information can be readily extracted. Each record that has been successfully extracted will be associated with a time stamp and stored in the local database for further analysis.

As the matching records from search engines are dynamically generated from databases, the search resultant web pages are normally in the similar format in terms of HTML structures. A HTML wrapper for each search engine is therefore necessary to extract the match records in the search resultant page.

Similar to market monitor, the search queries in market explorer will be relatively static and a scheduled information extraction process needs to be conducted

frequently, say once a day or once a week. Furthermore, a simple local search engine should be developed in order to navigate the extracted information easily as in [11].

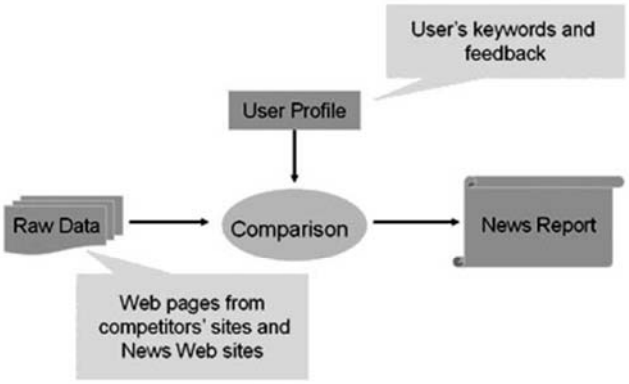



Figure 6. Information filtering process for Market Monitor

### B. Product Ranking

With the popularity of the Internet, the World Wide Web has become one of the important information sources for many users. Browsing and search are the two major information access methods. When user wants to find more information about a particular product, a search with the popular search engine is necessary for him/her. Therefore, get to know how easy their products can be accessed from the search engines is important to managers. In market explorer, we have come out with a way of product ranking.

Given a product name, for example inkjet printer, all top $N$ matches from each popular search engine e in the defined search engine list $E$ will be retrieved. For each retrieved match record $m$, the URL, denoted by $m.url$, the order in the returned search page, $m.o$ can be easily extracted. For any given manufacture's URL, denoted by p.*url*, the rank of the manufacture is defined as:

$$p.rank = \sum_{e \in E} (N - m.o) \text{ where } m.url.domain = p.url.domain$$

With such a rank, how easy the competitor's web pages can be reached from the search engines is clear.

Term Weight $w_{i,j}$ of index term $k_i$ in document $d_j$

$$w_{i,j} = \frac{freq_{i,j}}{\sqrt{\sum_{i=0}^{t} (freq_{i,j})^2}}$$

Similarity of the document $d_j$ and query $q$

$$sim(d_j, q) = \sum_{i=0}^{t} w_{i,j} \times w_{i,q}$$

### C. Performance Evaluation

he Watcher Agent includes the Price Watcher and the Market Watcher. They can be set to collect price information and the marketing news repeatedly on either daily or weekly basis. Then the results are stored in the local database. Since the data is ready to use, the search engine can get the requested information directly on the local database. Thus it highly increases the performance of search engine and reduces the users' waiting time.

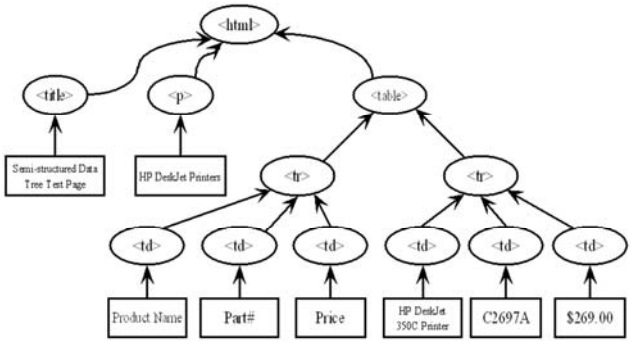



Figure 7. The SDT search steps through HTML files

## VI. Conclusion

Many information retrieval and monitoring technologies have been developed. But they are more for generally tracking changes and downloading the whole websites for offline browsing. This paper is to shed some light on specifically the design of a Web monitoring system for gathering business information relevant to a company, especially those related to competitors. For enabling competitor analysis, we proposed an autonomous software agent called Market Watcher that collects competitors' product prices, news, and information on social networks, and monitors their updates on the Web. Market Watcher is built as a market research tool for the users at the back-end. They both run autonomously as to relieve monitoring tasks over websites and search engines otherwise to be tediously carried out by human. The collected intelligence information is usually supplied to marketing managers for business decision making. So far this project has implemented up to the monitoring functions. It is envisaged that Market Watcher can be scaled up to include data-mining functions on competitors' information and automatic reporting as well, in the near future. We are in the progress of integrating Market Watcher into a full business intelligence infrastructure.

## References

[1] L. Liu, C. Pu, and W. Tang, "Continual queries for internet scale event-driven information delivery", Knowledge and Data Engineering, 11(4):610–628, 1999.

[2] L. Liu, C. Pu, and W. Tang, "WebCQ: Detecting and delivering information changes on the web", In Proceedings of International Conference on Information and Knowledge Management, November 2000.

[3] L. Liu, C. Pu, W. Tang, and W. Han, "CONQUER: A continual query system for update monitoring in the WWW", International Journal of Computer Systems, Science and Engineering, 1999.

[4] J. Naughton, D. DeWitt, D. Maier, A. Aboulnaga, J. Chen, L. Galanis, J. Kang, R. Krishnamurthy, Q. Luo, N. Prakash, R. Ramamurthy, J. Shanmugasundaram, F. Tian, K. Tufte, E. Viglas, Y. Wang, C. Zhang, B. Jackson, A. Gupta, and

R. Chen, "The Niagara internet query system", IEEE Data Engineering Bulletin, 24(2):27–33, 2001.

[5] S. Pandey, K. Dhamdhere, C. Olston, "WIC: A General-Purpose Algorithm for Monitoring Web Information Sources", Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

[6] S. Fong, A. Sun, and K.-K. Wong, "Price Watcher Agent for E-Commerce", in Proc. of the 2nd Asia-Pacific Conf. on Intelligent Agent Technology (IAT-2001), pp. 294--299, Maebashi City, Japan, Oct. 2001.

[7] S. J. Lim and Y. K. Ng, "An automated approach for retrieving hierarchical data from HTML tables", In Proc. of the 8th Inter. Conf. on Information and Knowledge Management, pages 466-474, 1999.

[8] B. Y. Ricardo and R. N. Berthier, Modern Information Retrieval, ACM Press, New York, 1999.

[9] S. Chawathe, A. Rajaraman, H. Garcia-Molina, and J. Widom, "Change Detection in Hierarchically Structured Information", Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp.493-504, Montreal, Quebec, June 1996.

[10] S. Chawathe, S. Abiteboul, J. Widom, "Representing and Querying Changes in Semistructured Data", Proc. of the Int. Conf. on Data Engineering, pp.4-13, Orlando, Florida, February 1998.

[11] F. Douglis, T. Ball, Y. Chen, and E. Koutsofios, "The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web", *World Wide Wide*, Vol.1, Issue 1, pp.27-44, Baltzer Science Publishers, 1998.

**Simon Fong**

He graduated from La Trobe University in Australia, with a First Class Honours BEng. Computer Systems degree and a PhD. Computer Science degree in 1993 and 1998 respectively. Simon is now working as an Assistant Professor in the Computer and Information Science Department of the University of Macau. He is also one of the founding members of the Data Analytics and Collaborative Computing Research Group in the Faculty of Science and Technology. Before his academic career, Simon took up various managerial and engineering posts, such as being a systems engineer, IT consultant, integrated network specialist, and e-commerce director in Melbourne, Hong Kong. and Singapore. Some companies that he worked at before include Hong Kong Telecom, Singapore Network Services, AES Pro-Data, and the United Overseas Bank in Singapore. Dr. Fong has published over 150 peer-reviewed international conference and journal papers, mostly in the area of E-Commerce and Datamining.

# Similar Document Search and Recommendation

Vidhya Govindaraju
HP Labs, Bangalore, India
vidhya.govindaraju@hp.com

Krishnan Ramanathan
HP Labs, Bangalore, India
Krishnan_ramanathan@hp.com

***Abstract* - Query formulation is one of the most difficult aspects of search, especially for a novice user. We propose a new search interaction where the user searches with a reference document and the system learns from the user inputs over a period of time to "push" relevant and new content without additional user interaction. Our method is based on identifying key phrases from the input document. The key phrases are used to query a search engine and the results are evaluated for similarity to the original document. By caching documents received from a user over a period of time, a user profile is built. The profile is then used to provide recommendations to the user.**

**Evaluations show that this method has a good precision in finding documents of interest to the user. Also our key phrase extraction method has good recall in retrieving the input document. Additional experiments reveal that our recommendation system is of help in exploring documents of interest to the user.**

***Index Terms* - Key phrase extraction, recommendation system, similarity search.**

## I. INTRODUCTION

In spite of the ubiquity of search engines, navigating information spaces remains a complex affair. Traditional search operates by matching a search query to (pre-processed) document representations. While current search algorithms perform reasonably well when the goal is navigation and known item search, they are not well suited when the goal is more exploratory and persistent in nature (e.g. the user is looking to learn a new topic). The user's ability in finding relevant information depends on his ability to frame good queries. However, query formulation is harder when the user is unfamiliar with the topic. There is also very little support on the web for stating persistent interest; this is necessary for facilitating ongoing learning. These long term interests are often stated in the form of short queries for which an engine can provide alerts [11], however ongoing maintenance of alerts is a problem.

Often, users have a set of documents obtained through browsing or from their social network (via emails, recommendations etc). These documents could serve as a good starting point for further search and recommendations. These documents are highly reflective of the user interests, yet they are hardly used in fulfilling ongoing information needs. Today, it is the responsibility of the user to identify salient keywords from the specific document of interest and use them in a search query.

In this paper, we propose an interaction paradigm whereby a user can provide to a relevant document and ask the system to retrieve similar documents without having to formulate search queries. For example, we would like the system to take as input the PDF version of John Hopcroft's talk on "Future directions in Computer science"[1] and output Ed Lazowska's talk titled "Computer Science-past, present and future"[2] as a similar document. Since the user is likely to be interested in similar documents that get created at a later point in time, it would be useful for a system to scout for similar documents on a continuous basis and send it to the user whenever they become available.

There are three main goals of the system. They include

1. Fetching documents similar to an input document.
2. Learn user interests periodically and recommend documents covering multiple user interests.
3. Provide enough content exploration via result diversification.

Key phrases are often used as a brief summary of documents. Hence they could prove useful for retrieving and recommending similar documents. Since manual key phrase extraction is time-consuming, automatic extraction becomes an important task. We describe a novel key phrase extraction method to extract key terms in a document and use them in a query to find similar

[1] www.cs.cornell.edu/jeh/**China**%202007.ppt

[2] lazowska.cs.washington.edu/fcrc/Lazowska.FCRC.pdf


doi:10.4304/jetwi.4.1.84-93

documents. The aim of this is to find key phrases that best describe the context of the document.

The amount of content on the web is increasing with new articles, documents, blogs etc being posted every day. Users face the problem of finding documents in their area of interest. There is a lot of support in the web for finding new and relevant information. Web based recommender systems help in choosing the right documents for the user. Often such systems suffer from the problem of data sparsity and hence produce redundant results. We develop a personalized recommendation framework for recommending documents based on the past user requests to the system. A user profile is built from the past requests and then used to source and recommend documents to the user on an ongoing basis. We feel that such a push based document system will be highly valuable in finding content from the web without querying for it.

We extend our solution of finding similar and relevant content to result diversification. A document has multiple modalities and providing the user with similar content in all these dimensions becomes an essential component in this scenario. We study the problem of diversifying search results and present ways to maximize relevance and diversity of search results.

There are a number of applications where this kind of technology can be useful. For example, in e-discovery, a patent attorney looking for relevant documents among millions of documents can identify one relevant document and request the search system for similar ones. In an online video application, a user can mark videos as interesting and request retrieval of similar videos. Finally, in an exploratory search scenario, users might find the results useful even if they are not very similar to the input document. We would like to stress that in this work, our motivation is not to detect duplicate web pages or documents.

The remainder of this paper is organized as follows. In section 2 we survey related work on similar document search, keyword extraction, results diversification and user profiling. In section 3 we describe the architecture of the system. Section 4 discusses the system components and algorithms in greater detail. In section 5, we describe the different fronts on which we evaluated our system. Section 6 concludes the paper.

## II. RELATED WORK

### A. *Similar Document Search*

Similarity search has recently become a field of active research [6] [7]. Despite this, there are very few systems that use similarity search to facilitate user interaction.

One of the earliest approaches was the “Similar pages” or “More like this”[3][4] link provided by search engines for search results. In [8] related article search in Pubmed through citation links in the database is presented. The user study reveals that such a system which helps in exploring new and relevant information is a useful feature and it becomes an integral part of user’s interaction with Pubmed. A direct way of finding similar text that is conceptually related to the input document is presented by Yang et.al. [6]. They have built a system for finding similar articles in BlogScope. They have developed a system of cross-referencing information created by different users. There is a large amount of related work on retrieving similar images. For example, Flickner [9] developed a system for querying with images to get similar images and videos.

### B. *Keyword and Key phrase Extraction*

The simplest approach to key phrase extraction is taking top n most frequent n-grams in a document [4]. A method for extracting keywords based on frequency and co-occurrence phenomenon is presented in [1]. In [6] key phrases are extracted using part of speech tagger. All noun phrases are extracted as key phrases.

Yahoo Phrase Extractor[5] takes a text snippet and returns key terms in the text. In [12], the task of key phrase discovery is accomplished using suffix arrays or suffix tree structures. They have also presented the benefits of using key phrases as a feature in natural language processing. The authors have used key phrases extracted from web pages in clustering web search results.

Kea algorithm [5] uses the Naïve Bayes machine learning algorithm for training a classifier with user generated key phrases for sample documents. The trained set is then used to extract key phrases from other documents.

Extracting key terms from noisy documents by exploiting the graph of semantic relationships between terms in the document is explained in [3]. This method is close to ours except that they exploit the Wikipedia information to filter redundant phrases. In [11], clustering based unsupervised key phrase extraction algorithm is presented.

Since most key phrase extraction methods generate a large number of key phrases, some form of ranking is used to select key phrases [2]. Most key phrase extraction algorithms are based on TFIDF for ranking key phrases [4].

[3] googleguide.com/similar_pages.html

[4] **http://www.google.com/alerts**

[5] http://developer.yahoo.com/search/content/V1/termExtraction.html

### C. *Diversifying search Results*

Documents have multiple themes associated with it. Our hypothesis is that in the process of finding a document that is similar to the input document, users want variety and coverage of different themes that the input document covers. Hence diversifying search results to cover these multiple interpretations becomes important. Diversification can be achieved in two ways: framing multiple queries that are intrinsically diversified or clustering results so as to achieve diversification. Agrawal et.al. [10] employed a greedy algorithm that minimizes user dissatisfaction in a web search scenario. This method considers the popularity of the category while diversification.

Clustering of words [11] will help in framing queries that represent various themes in a document. Clustering words that are semantically similar is done based on known ontologies. Also mutual information between words could be used as a factor to classify words into different clusters.

The traditional method of clustering documents considers a document as a vector of words and distance between two documents is found by taking the cosine similarity between them. This is then used in a hierarchical clustering algorithm to get document clusters. But this method leads to high dimensionality and the computational costs are often huge. Using key phrases as document features for clustering is discussed in [12].

### D. *Recommending documents based on user profile*

Web based recommender systems are primarily based upon Collaborative Filtering (CF) techniques which filters information based on user preferences. It is based on measuring the similarity of users or items or both [18]. Though the user based CF has a lot of commercial applications, when sorted for recommending documents this suffers from the problem of data sparsity and noise. Zhou [14] has used co-citation graph, author-document relationship and document-venue relationship in an item based CF for recommending documents. They implemented a single low dimensional embedding of documents that capture the similarities between them. A semi-supervised learning on this graph was used to develop a recommendation system. In [15], a content based recommendation system for Citeseer database is proposed. They classify the documents in Citeseer into predefined set of concepts which they use to build a user profile and recommend documents accordingly.

Query specific recommendation depending on standing interests is proposed in [16]. Xu et. al [17] proposes a personalized method for recommending documents based on eye tracking of keywords in the document.

## III. SYSTEM OVERVIEW

In this section, we present an overview of our system architecture and designed a solution to the problem of recommending relevant documents based on a set of input documents.

When a user queries our system with a document, our system generates keywords and key phrases from the document and uses them in a search query to get an initial set of documents from the web. It ranks these documents based on the similarity to the input document and diversifies the results so as to cover all the themes in a document. It then presents the top ranked similar documents to the user. It builds a profile with the documents received from a user and exploits them in sending recommendations. An overview of the system architecture is shown in Fig. 1.

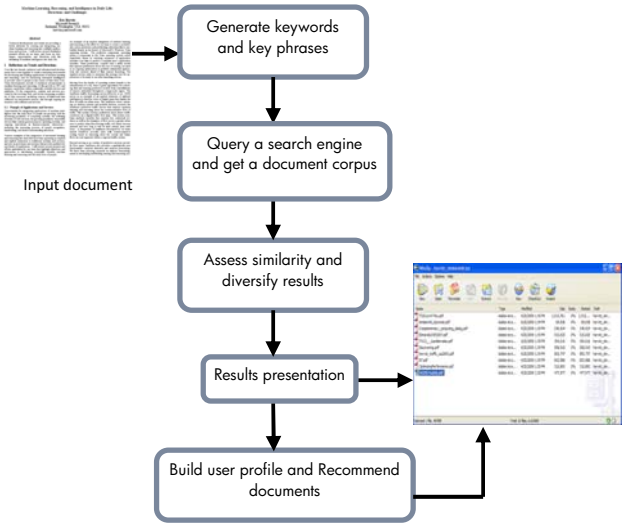



**Figure 1: System Overview**

The first step in our solution is to generate a set of keywords and key phrases from the input document. A document can be viewed as a bag of words. Identifying the most important words can help in framing the right queries for searching similar documents. The keywords and key phrases in the documents appear often in document titles, paragraph titles, and important sentences, often associated with more meaningful terms. We exploit this feature in finding keywords and key phrases in the document. Since key phrases are more descriptive than keywords in explaining the context of the document, we use them in framing a search query. We also add the most important keywords that may not have a co-occurrence feature to the list while framing a search query.

We achieve search result diversification by framing multiple queries representing various dimensions of a document. We cluster the key phrases into different sets

and frame the queries for each cluster using the key phrases in it. We query the search engine (e.g. Google scholar) with these clusters and get an initial set of documents. Since the queries are intrinsically diversified, the initial corpus contains documents diversified on various topics. For each document retrieved, we assess their similarity with the input document. We return most similar documents to the user.

We extend our solution to develop a recommendation framework based on content similarity. We build the user profile with the keywords and key phrases of the documents that are sent to the system by the user. We also use the author information in the input documents to recommend recent documents published by the authors to the user. This is highly useful in a research scenario to find other content that is posted or created by the same author whose documents the users are interested in. Also in a document search application, finding new publications from the top publishers who publish content of interest to the user is also an important feature. Hence we use the profiler data in finding the publishers and extract new and relevant content published by them for recommending to the users. Our evaluations show that such a system is very useful in exploring the web for finding relevant information.

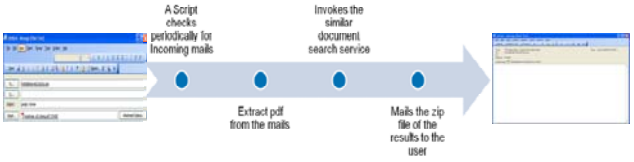



**Figure 2: User Interface Overview**

The service is exposed as a cloud service. On the server side, it is implemented as a shell script in Linux which runs every minute and checks the incoming mails and parses them to get the required information. This script then runs the program for the similar document search and collects the similar documents, compresses them and mails it back to the user (Fig. 2). Currently, ten relevant documents are mailed to the user.

## IV. PROPOSED METHOD

In this section, we describe the algorithms used in our system.

### A. *Key phrase Extraction and Ranking*

In our method, we model the document as a graph of words in which the important words in the document tend to co-occur with other important words. We exploit this feature in finding the keywords from the document. We extend the algorithm in [1] to generate keywords and key phrases for a document.

We first construct a graph where the nodes are high frequency words in the document. The weight of a word is taken as the document frequency of the word (excluding paragraphs and document titles) multiplied by its weight. The weight of a word is measured by the frequency of the word in paragraph and document titles. If it does not appear in paragraph and document titles, the weight is taken as one.

We create edges between those nodes of the graph if words associated with the nodes co-occur in the same sentence. The edge weight is the minimum of the word weights assigned to words in the previous step. We add to the graph those words that co-occur with the high frequency words in any sentence. We then find the maximally connected components in the graph. Each maximally connected component is called a concept. We use the concept graph ($C_G$) for finding the keywords and key phrases.

We find all 2-3 gram words in the document that does not contain a stop word in the middle. For each phrase, we test whether the words in the phrase are part of a concept. We find the rank of the phrase as follows.

$$\text{Rank of a phrase} = \frac{\text{Number of words in phrase which are present in the concept graph} * \text{Frequency of a phrase}}{\text{Number of words in the phrase}} \quad (1)$$

We select phrases with higher rank as key phrases. Sometimes, key phrases generated by our method are permutations of each other. These key phrases are further filtered so as to avoid redundancy. While selecting a new key phrase, only those that contain a new keyword (compared to previous key phrases) are chosen.

From our initial experiments, we found that adding a few keywords that are very important in the document, to the seed query can improve the precision of the results drastically. This is because certain words may not have a co-occurrence feature. So for adding the keywords that are very important to the query we use the weight of the keywords. We sort the keywords in the descending order of their weight. We find the point at which there is steep decrease in weight. All the keywords before this point are taken for query formulation.

### B. *Querying the search engine and assessing similarity*

A query is formulated using the key phrases and important keywords and is used to query the search engine to get a set of similar documents.

The retrieved documents are ranked based on the similarity to the input document. There are a number of ways in which similarity could be assessed. In our method, keywords and key phrases are extracted from the retrieved

document. Jacquard similarity measure is found between the keywords and key phrases of input and retrieved document using the equation

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Here A and B are the key phrases extracted from the input and retrieved documents. Documents with higher similarity score are sent to the user.

C. *Diversifying search results*

A document could be addressing multiple themes. For instance, a research paper on mobile security could be addressing issues with mobility and security. Diversifying search results so as to cover all the major themes of a document will increase user satisfaction. This can be achieved by framing multiple queries one for each theme. The themes are found as discussed below.

*Finding themes in a document:* For finding different themes of a document, the phrases and important keywords extracted above are clustered based on their semantic relationships into different groups. We use the Normalized Google Distance (NGD) [13] as a measure to find semantic relationships between words. NGD uses Google page counts of words and phrases to find the relative distance between them. All the key phrases and important keywords in the input document are clustered based on the NGD between them. The clustering algorithm requires the number of required clusters as the input. Experimentally we found the number of themes (clusters) in research papers to be 3. This can also be set by the user. These clusters are used to frame multiple queries so as to cover all the context of an input document. A sample cluster of key phrases and keywords for the research paper in reference [6] is listed in Fig. 3.

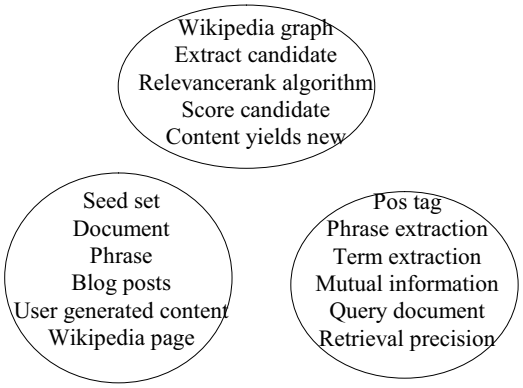



Figure 3. Key phrase clusters for reference [6]

*Algorithm for clustering key phrases:*

Input: Number of required clusters N, Array of key phrases

Output: Clustered key phrases

Steps:

**1.** Find Normalized Google Distance (NGD) between every pair of key phrases and important keywords.

**2.** Use the NGD between key phrases in a hierarchical agglomerative clustering algorithm to cluster the key phrases. Compute the cluster centroid. This step is used to find cluster centriods for seeding the k-means clustering algorithm.

**3.** With these cluster centroids, use the NGD between key phrases in a k-means clustering algorithm to get clusters of key phrases.

D. *Building User Profiles*

A profile is a description of user interests. A push based data delivery system requires a profile to be built for every user interacting with the system. The user interacts with our system by sending a document of significance to him and requesting for similar documents. The input document serves as a tool to predict user interest. We aim at building a user profile by implicitly predicting user interest from user interactions.

We add the keywords and key phrases from the input document to the profile. The profile thus built for a user contains terms that broadly specify his interest. When keywords and key phrases collected over a period of time repeat in the profile or have a close semantic relationship between them, it represents consistent user interests. The recommendation algorithm exploits this feature in getting valuable recommendations for the user.

We also add other metadata of the input document such as author name to the profile. This specifies the list of authors whose documents the user has read. The user profile thus built is used in a recommendation system to push relevant content to the user.

E. *Recommendation System*

The system is used to recommend more relevant and recent documents to the user based on his profile. The recommendation system uses multiple approaches to improve relevance.

a) Recommend documents based on user profile

b) Recommend new documents of favorite authors

c) Recommend new documents from recent conferences or journals.

*Recommend documents based on user profile*: The profile built for a user implicitly using the documents received from the user, is used to select documents for recommendation. The profile has a list of keywords and key phrases from the input documents and author names of the documents. All the key phrases and keywords in the profiler are clustered (using the algorithm in Section

4.3.1.1) into different clusters. Since the clusters are framed from keywords and key phrases taken from multiple documents, they capture interrelationships between various topics. One cluster may contain key phrases from different documents if they have a close semantic distance.

These clusters are then used to frame multiple queries. An initial set of documents are retrieved using Google Scholar as the search engine for each of these queries. For each document retrieved we extract the keywords and key phrases.

Each document is ranked based on three parameters:

$S_P$ - Similarity with the profiler

$D_P$ - Days since the document was published

N - Number of queries that retrieved the same document.

$$\text{Rank} = (S_P * w_1 + (1/ D_P) * w_2) * N \quad (3)$$

The similarity score is the strength of user interest in the document. Similarity with the profiler is computed as follows.

$$\text{Profile Similarity Score, } S_P = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

Here, A is the set of key phrases in the document and B is the set of key phrases in the user profile.

The second parameter is used to filter out old documents and ensure recency. For experimental purposes, we used $w_1$ =1 and $w_2$ =4. The documents are sorted based on the final rank and top ranked documents are recommended to the user.

*Recommend new documents of favorite authors*: The profiler together with the list of keywords and key phrases of the documents received from a user, has a list of authors of each document. We assume that this list of authors as favorite authors for the user. The number of documents of a particular author, that a user has read is taken as the authority of an author. We extract papers for each author by adding the string author:<name> to the query in the search engine. Also we find all the co-authors from the input list and retrieve papers for each double author. We extract keywords and key phrases for each document retrieved.

Each document is thus ranked on four parameters

$S_P$ - Similarity with the profiler

$D_P$ -Days since the document was published

$N_A$ - Number of favorite authors who wrote this document

$A_W$ - Authority of the authors who wrote the document

$$\text{Rank} = S_P * w_3 + (1/ D_P) * w_4 + N_A * A_W * w_5 \quad (5)$$

Author names may be misleading when the name is shared by more the one person who publishes content in different fields. To prevent such errors, we compute the similarity of the document with the profiler. Similarity with the profiler is computed as in Equation (4). The second parameter in the above equation is used to filter old documents. The third parameter increases the weight of a document according to the number of favorite authors who wrote the document. For experimental purposes we used $w_3$ =4, $w_4$ =4 and $w_5$ =2.

*Recommend new documents from recent conferences or journals:* We exploit the profiler key phrases in fetching a list of conference and journal names that publish content of interest to the user. This helps in alerting the user with relevant documents from recent conferences.

All the key phrases and keywords in the user profile are clustered (using the algorithm in Section 4.3.1.1) into different clusters. These clusters are then used to get a list of conference and journal names from the search engine. For each clustered dataset, we get publisher details of documents from the search engine (e.g. in Google Scholar using the Bibtex output) and add them to the publishers list. We consider publishers with higher frequency from the results list and retrieve recent documents published by them. This is done by querying the conference name in the search engine and setting the recent preference to the current year. For each conference, recent documents that were published by them are obtained.

Each document is ranked based on two parameters

$S_P$ - Similarity with the profiler

$D_P$ - Days since the document was published

$$\text{Rank} = S_P * w_6 + (1/ D_P) * w_7 \quad (6)$$

Similarity with the profiler is computed as in Equation (4). The second parameter is used to filter old documents. For experimental purposes, we used $w_6$ =1 and $w_7$ =4. We used these values because more likely the documents retrieved with the conference names tend to be relevant and the major parameter here is the published date (since users desire documents that are recommended to be more recent).

## V. EXPERIMENTAL EVALUATION

Our experimental evaluation is designed to answer the following questions

1. Are the key phrases obtained by our method sufficiently discriminative?

2. Are the similar documents retrieved by the system of good precision?

3. How good are the recommendations made by our recommender?

### A. *Evaluation of Key phrases*

We first evaluated the usefulness of using key phrases in representing the context of a document. This was done because the hypothesis we use for retrieving similar documents is that they are clustered close to the input document. We extracted keywords and key phrases from a document published in the web. We used the keywords and key phrases separately as a query to find the rank at which the input document is retrieved. Results (in Fig. 4) show that key phrases perform better than keywords in retrieving documents at higher rank.

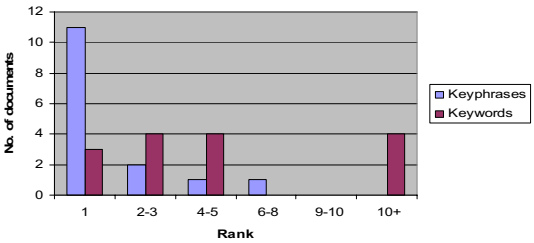



Figure 4. Comparison of keywords vs. key phrases in fetching the input document

We evaluated whether the key phrases extracted for a document is able to retrieve the same document back in top results in Google scholar. An automated test run for 500 documents shows that 80% of the documents were retrieved at top ranks by querying with the key phrases. Results are shown in Fig. 5.

We next compared our method of finding key phrases against existing methods. We compared our key phrase extraction method with the Yahoo Phrase extractor and manually extracted key phrases. Yahoo! Phrase Extractor is on online tool for extracting phrases. It takes a text document as input and returns a list of key phrases for the document.

We conducted a study in which the users were presented a list of key phrases for each document in the test data set. The list comprised of the key phrases extracted by our method and Yahoo Phrase Extractor. The users were asked to rate the relevancy of each key phrase (0, if it is irrelevant and 1, if relevant). We present the comparison using the measure of precision which is calculated as in Equation 7.

$$\text{Precision} = \frac{\text{No of relevant key phrases}}{\text{Total no. of key phrases}} \quad (7)$$

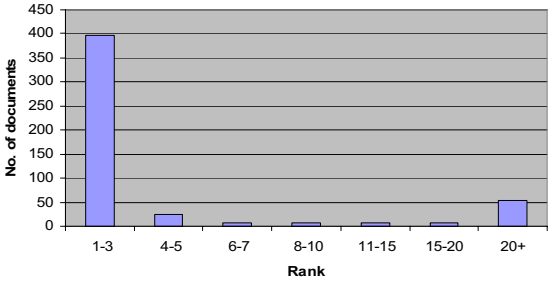



Figure 5. Performance of key phrases in fetching the input document

Table 1 summarizes the performance of our method and Yahoo Phrase Extractor for the test data set.

TABLE I

Performance of different key phrase extraction methods

| Method | Precision |
|---|---|
| Yahoo! Phrase Extractor | 0.478 |
| Our key phrase extraction algorithm | 0.758 |

A study was done in which users were asked to find key phrases in 30 documents. Key phrases were also extracted by our method. Using these phrases a search string was framed and the rank at which the input document is retrieved back is found in Google. Fig. 6 shows that our key phrase extraction algorithm performs better than manually extracted key phrases in retrieving the input document at higher rank.

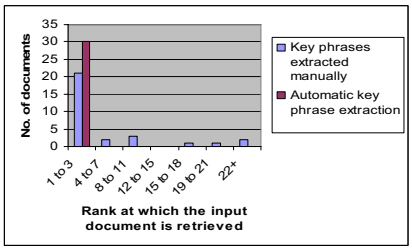



Figure 6. Comparison of automatic key phrase extraction against manually selected key phrases

### B. *Evaluation of Similar Documents*

In this section we evaluate the relevance of similar documents that are retrieved by our method. We conducted an evaluation of the system with 10 users by exposing the solution as a cloud service. The users were primarily research scholars.

To compare the retrieval quality of the system, we computed the following commonly used measure:

1 **Precision at K**: Prec@K(q) is the fraction of relevant documents within the top K results for a query q. This needs a binary classification of documents.

2 **Mean Average Precision** : It returns a single

value for each method of ranking and is computed as follows

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\Pr ecision(R_{jk}) \quad (8)$$

Q is a set of queries, $m_j$ is the number of documents on which precision is computed and is chosen as 20. MAP takes the position of the results in the ranking into account and is based on binary relevance classification.

3 **DCG at K**: The Discounted Cumulative Gain explicitly takes the ranking of first K results into account. Hence it rewards highly relevant results less than less relevant results. It is computed as follows :

$$DCG@K(q) = \sum_{j=1}^{K}(2^{r(j)} - 1)/\log(1+j) \quad (9)$$

Here, r(j) is an integer for the relevance rating given to the result at position j for the query q and is taken on a scale of 1-3.

We did not evaluate recall because of 2 reasons. One, we are using a search engine backend and there is no way to know the set of all relevant documents. Second, for research papers, the users are more interested in top N relevant documents that the entire set of relevant documents.

We first evaluated the efficiency of key phrases in finding similar documents. Here, all the key phrases are given as one long query. In this experiment the user sends a document to the service and is sent a set of documents similar to the input document. The user is asked to rate each document on a scale of 1-5 based on his interest in the document.

We chose ten documents from varied topics for evaluating the relevancy of similar documents. For each input document, we retrieved 20 similar documents. We used the above specified measures for evaluation. Fig. 7 shows the precision graph and Fig. 8 shows the DCG graph. Results are summarized in the Table 2.

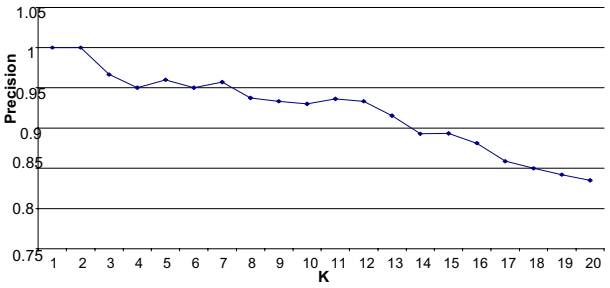



Figure 7. Precision at K

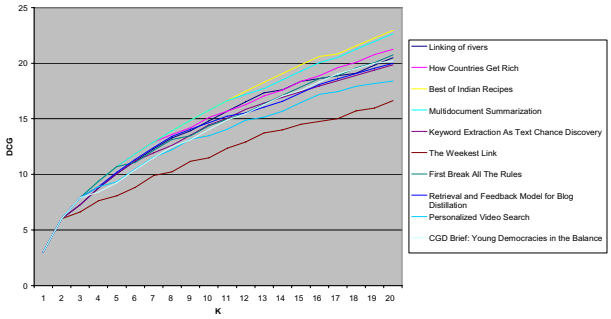



Figure 8. DCG at K

TABLE II

Evaluation of similar document search results

| | |
|---|---|
| Precision @ 10 | 0.93 |
| Precision @ 20 | 0.835 |
| Mean Average Precision (MAP) | 0.90 |

Also, we evaluated the importance of diversification in search results. We will compare the similar documents that are retrieved by the system with and without clustering of key phrases. For this, we took ten documents from the users and sent two lists of 10 documents for each input document. The documents considered were primarily research papers that were much focused. For each document sent, the user rated the relevancy of it on a scale of 1-5. We use the standard measure of precision and DCG for comparison. Fig. 9 shows the DCG graph. Table 3 summarizes the performance of our method.

TABLE III

Precision of similar documents retrieved with and without explicit result diversification

| Method | Precision @10 |
|---|---|
| Similar Documents retrieved with result diversification | 0.64 |
| Similar Documents retrieved without result diversification | 0.65 |

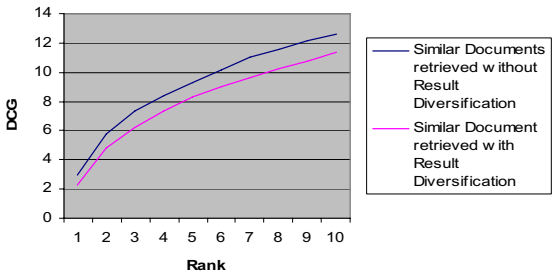



Figure 9. DCG graph of similar documents retrieved with and without explicit result diversification

### C. *Evaluation of Recommended Documents*

For evaluating the recommender system, the users are sent a series of recommended documents based on the documents submitted by them to the system. The user feedbacks are collected for each recommended document on a relevancy scale of 1-5. The DCG graph (Fig. 10) shows that our recommender system based on user profiles performs well in selecting the right documents for recommendation. Table 4 summarizes the performance of our individual recommendation modules. This study was done with 10 users, however we could obtain more than five relevant documents for only 7 users (because of subscription requirements), hence results are reported for only 7 users.

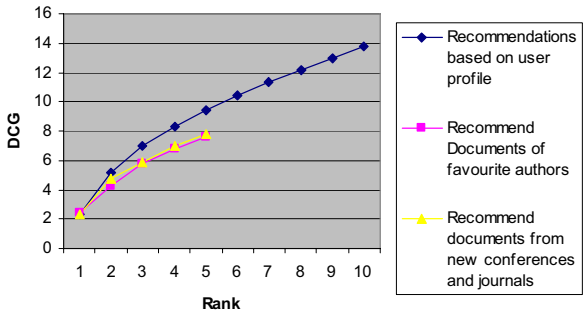



Figure 10. DCG graph for recommended documents

TABLE IV

Performance of Recommender System

| Recommendation Module | Performance | |
|---|---|---|
| Recommend documents based on user profile | Precision @ 10 | 0.89 |
| Recommend new documents of favorite authors | Precision @ 5 | 0.51 |
| Recommend new documents from recent conferences or journals | Precision @ 5 | 0.65 |

### D. *Discussion of Experimental results*

The evaluation of extracted key phrases reveals that the key phrase extraction algorithm performs better than some existing key phrase extraction methods. It is also noted that key phrases performs better than keywords in describing the context of a document. Evaluation of similar documents shows that our method helps in exploring the web in finding documents of interest to the user. However, it is noted that the diversifying search results does not improve the user satisfaction. This could be because the user may be interested in only the broader theme of a document and not all the themes. Clustering has taken the results to a different document space that is more precise to the individual themes in a document.

From the evaluation of our recommender system we infer that such a system can form a useful component of the user's information exploration in the web. Also it is evident that the recommendation based on user profile has fetched documents of interest to the user.

## VI. CONCLUSION

In this paper, we propose a novel retrieval approach for document similarity search. We have formulated a method to get similar documents based on the concept of the input document by extracting the relevant keywords and key phrases from the document. The experimental results have shown favorable performance of the proposed approach. We have also built a personalized document recommendation system based on the user profile created implicitly with the inputs received from the user.

In future, we will aim to extend the similarity search and recommendations to web pages, video content and to cross-lingual similarity search where information in one language could be used to find similar information in other languages

## REFERENCES

[1] Z. Zhang, H. Cheng, Keyword extracting as text chance discovery, IEEE Fuzzy systems and knowledge discovery conference (FSKD), 2007.

[2] Xin Jiang, Yunhua Hu, Hang Li, A Ranking Approach to Key phrase Extraction, Proc. SIGIR 2009.

[3] M. Grineva, M. Grinev, and D. Lizorkin. 2009. Extracting key terms from noisy and multi-theme documents, Proc. WWW 2009, pages 661–670.

[4] Lee, J.W. and Baik, D.K., A model for extracting keywords of document using term frequency and distribution, Lecture notes in computer science, Springer, Pg. 437—440, 2004.

[5] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning, Kea: Practical automatic key phrase extraction, Proceedings of the 4th ACM conference on Digital Libraries.

[6] Yang Y., Bansal, N., Wisam, D., Panagiotis, I., Koudas, N., Dimitris, P., Query by Document, WSDM '09.

[7] Xiaojun Wan, Jianwu Yang, Jianguo Xiao, Document Similarity Search based on Manifold Ranking of TextTiles, AIRS 2006, LNCS 4182, pp. 14 – 25, 2006.

[8] Jimmy Lin, Michael DiCuccio, Vahan Grigoryan, W.John Wilbur, Navigating information spaces: A case study of related article search in PubMed, Information Processing and Management, 2008, Elsevier

[9] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yonker. Query by image and video content: The qbic system. Computer, 28:23–32, 1995.

[10] Rakesh, A., Sreenivas, G., Alan, H., Samuel, I., Diversifying search results, WSDM '09.

[11] Z. Liu, P. Li, Y. Zheng and M. Sun, Clustering to Find Exemplar Terms for Keyphrase Extraction, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.

[12] Dell, Z., Yisheng, D., Semantic, Hierarchical and Online Clustering of Web Search Results, AP Web 2004, LNCS 3007, pp 69-78, 2004.

[13] Rudi L. Cilibrasi and Paul M. B. Vitanyi, The google similarity distance, IEEE Transactions on Knowledge and Data Engineering, 19(3):370–383, 2007.

[14] D Zhou, S Zhu, K Yu, X Song, BL Tseng, H Zha, Learning Multiple Graphs for Document Recommendation, Proc. WWW 2008.

[15] Kannan, C., Susan, G., Praveen, L., HP Luong, Concept-Based Document Recommendations for Citeseer Authors, Lecture Notes in Computer Science, 2008 – Springer

[16] B. Yang, G. Jeh, Retroactive Answering of Search Queries, Proc. WWW 2006.

[17] S Xu, H Jiang, FCM Lau, Personalized online document, image and video recommendation via commodity eye-tracking, Proc. RecSys 2008

[18] Fei Wang , Sheng Ma , Liuzhong Yang , Tao Li, Recommendation on Item Graphs, Proceedings of the Sixth International Conference on Data Mining, p.1119-1123

# Monitoring Propagations in the Blogosphere for Viral Marketing

Meichieh Chen, Neil Rubens, Fumihiko Anma, Toshio Okamoto
Knowledge Systems Laboratory, University of Electro-Communications, Tokyo, Japan
Email: meichieh@hotmail.com rubens@ActiveIntelligence.org {anma, okamoto}@ai.is.uec.ac.jp

*Abstract*—**Even though blog contents vary a lot in quality, the disclosure of personal opinions and the huge blogging population always attracts marketing's attention on blog information. In this paper, we investigate how marketers can identify the information propagation in degree among blog communities. In this way, topic similarity, relatedness, and word repetition between leader and followers' writing products are considered as the propagated information.**
**The contribution of this paper is twofold. The work presented here is to show how blog content can be economically and feasibly analyzed by existing internet sources such as Wikipedia database and the usage of page return from a Japanese search engine. To this extent, this system, which combined in-link algorithms and text mining analyzes, tracing propagation channels and propagateable information allows analyzing the power of influences in viral marketing. We demonstrated the effectiveness of the system by applying blogger identification, topic identification, and the topic propagations.**

*Index Terms*—**blog, text mining, viral marketing, content based propagation, Wikipedia, thesaurus, page return of search engine**

## I. INTRODUCTION

Blog, which provides support for the issues bloggers deem interesting and important. Along with the development of the internet and increasing prevalence and convenience of web-related activities, social network occurring in virtual communities that spontaneously transfers individuals' opinions, interests, and desires, is now the hot materials of emerging marketing research. Researchers are able to tag or categorize into bloggers communities to build up database or predictive models for the purpose of lifestyle intruding marketing.

As more and more people participate in the blogging behaviors, the blogosphere has become a trend and an important space for people to exchange information. In there, people write blogs to produce information and read blogs to consume information from others. The potential of operating marketing strategy in the blogospheres has been noticed by marketers. In there, people share opinions and experience with others through word of mouth, but the disseminated results is like a virus that continuously spreads and infects more and more people without any further marketing effort [11].

Viral Marketing is a marketing strategy for social network through a persuasive message designed to spread from the opinion leader to followers [10]. As information that flows in the blog spaces is extremely sensitive to trendy topics one needs to continuously monitor which marketing strategies take place. However, most research about monitoring propagation in the blog spaces fell into underestimating the influences of blogging behaviors on viral marketing. Modeling propagation among bloggers based on recommendation, invitation, or any in-link actions, that are only able to analyze the directly accessed channels of viral marketing, may underestimate the influence of blogging behaviors. On the contrary, modeling propagation depending on topic similarity would meet the problem of overestimating the influence of blogging behaviors. To design a monitoring system for indentifying the information propagation with solid evidences of marketing effects is a difficult and subjective process. It requires the understandings of thoughtful marketing strategies to bloggers and the technological capabilities of data processing on this social medium.

In this paper, we investigate how marketers can identify the information propagation in degree among blog communities. The only requirement that we have for the user is that (s)he should decide a interest keyword as to look for the opinion leader in that interest domain. Based on the opinion leader's history of comment receiving, the comment givers, who show the evidences that have obtained knowledge from leader, are traced as the followers. Topic similarity, relatedness, and word repetition between leader and followers' writing products are considered as the propagated information. Such an approach should allow the users to explore their interest domains, trace the evidences of influences from blogging behaviors, and presents the complex social relationship of people to people, topic to topic, and people to topic.

During the design of our approach, special attention, which is given to the data processing of topic identification from blog contents, should allow core words, the reprehensive words of documents, to be extracted from the informal writings in blogs, such as new words, mixed languages, and loose grammar [14]. In order to present topics of interest domains from the special word usage, a simple specification of choosing the thesaurus source could be very helpful. For this purpose we have decided to exploit the database of Wikipedia titles as the thesaurus due to its appropriability and


doi:10.4304/jetwi.4.1.94-105

accessibility. The Wikipedia thesaurus can be used for defining semantic relatedness that resembles all inclusive topics, including trendy topics, interested by internet users. In addition, the Wikipedia thesaurus updates frequently and is easy to be obtained.

In order to experiment with the proposed approach, we have implemented a sample observation that presents blogger identification, topic identification, and the topic propagations among bloggers. The results showed successful topic identification from blog contents. In this way, the propagations among opinion leader to follower bloggers, such as topic propagation, word repetition, and topic evolvement, are able to be measured and promising to be applied in a large size of data.

The contribution of this paper is twofold. The work presented here is to show how blog content can be economically and feasibly analyzed by existing internet sources such as Wikipedia database and the usage of page return from Goo search engine[1]. To this extent, this system, which combined link and topic analyses, tracing propagation channels and propagateable information is supportive to the observation of viral marketing

Section 2 continues with presenting related work on identifying propagation for viral marketing. Section 3 explains the details of our approach, whereas Section 4 introduces the propagation model we have implemented. The results are presented in Section 5. Finally, we draw conclusions in Section 6.

## II. RELATED WORKS

A lot of research has already been done in areas related to recognizing propagation among blog communities for viral marketing. For instance, several analyses and frameworks have been introduced that are designed for monitoring propagations. This section continues with discussing some related work that is relevant for our research.

As the volume of blogs and other public online forums such as message boards increased in the early 2000's, commercial enterprises which base their business model on mining business intelligence from these sources emerged. The methodology consists of a platform combining crawling, information extraction, sentiment analysis and social network analysis [15].

Wu, Huberman, Adamic, and Tyler [17] are the first to use the concept of virus epidemic model to simulate information propagation through email forwarded URLs and attachments within a group of people. The system defines the distance of interest similarity by the node attributes shown on the personal homepages. This research presents a way to analyze information flow that takes into account the observation that an item relevant to one person is more likely to be of interest to individuals in the same social circle than those outside of it. Wu's research points out that the similarity between people is a key factor for information propagation but the method is limited in the extremely high quality information and is hard to be applied in the general social media.

Being an extension of information propagation in a large group of people, Leskovec, Admic, and Huberman [12] present an analysis of a person-to-person recommendation network, consisting of 4 million people who made 16 million recommendations on half a million products to establish how the recommendation network grows over time and how effective it is from the viewpoint of the sender and receiver of the recommendations. They present a model that successfully identifies communities, product and pricing categories for which viral marketing seems to be very effective, while on average recommendations are not very effective at inducing purchases and do not spread very far. Leskovec's research also point out only the relationship of recommendation link between people is not sufficient to explain the reasons for making a purchase decision. On the contrary, Individuals are often impervious to the recommendations of their friends, and resist buying items that they do not want. Those findings are very important to interpret how behaviors of virtual communities could contribute the influences on other individuals in the real world.

More online behaviors especially for blogging behaviors are discussed in Ali-Hasan and Adamic's work [1]. They examined three blog communities in different geographical locations, both by analyzing the network structure of their blogrolls, citations, and comments, and by surveying the bloggers directly. In all three communities, there is strong evidence that blogs do enable relationship formation, with some of those new relationships later extending to other communication media and offline meetings. Compared with previous blog studies that have typically placed more emphasis on blogrolls and citations, Ali-Hasan and Adamic's find that much of the community interaction occurs in comments and is not always reflected in blogrolls and citations.

Viral Marketing is a marketing strategy for social network through a persuasive message designed to spread from the opinion leader to followers [10]. In order to establish deeper understanding of blogging behaviors and its potential applications in marketing, studying the social relations of linkage built up by forward, recommendation, blogrolls, citation, and comment is not sufficient in the aspects of the directional analysis of the opinion leader to followers and the propagated message from leader to followers. More and more research have tent to explore the potential relationship between bloggers through analyzing blog contents. However, applying Natural Language processing to extract useful information could be very complicated and difficult, especially working on blog writing, which includes lots of informal language.

Instead of dealing with bloggers' writing contents, many research try to use tags, which are defined by blog users to be the attributes of bloggers, to explore the similarity of blogs [5]. However, there are several drawbacks of using tag system on blog analysis. Muller

[1] Goo search engine is an internet search engine and web portal based in Japan, which crawls and indexes primarily Japanese language websites.

points the problem that similar tags do not describe similar things [16], because of the compatibility of contextual information [8]. Given the same observation, Hayes [6] suggests that tagging system may work well for social bookmark site, like Del.icio.us[2] where the multiple users tag a unique resource, but not suitable for analyzing blog.

Fujimura, Fujimura, and Okuda [4] present a model to extract community from the enormous amount of web contents based on the clusters made by co-occurring words in the query results. The conducted experiments show the feasibility of their measuring system, which also shows the possibility of the application on analyzing blog contents as a subset of web contents. However, query on blog and query on web are with different purposes. For example, the web queries contain many large web sites (Yahoo! eBay, Hotmail and so on) and higher percentage of political and technology-related queries [14][2]. Mishne & de Rijke did an extensive query log analysis of blog user behavior in terms of queries and page views. Their research determined that most of the named-entity queries for blogs were requests to learn what is being said currently about that entity, while the more general queries were often attempts to find blogs or posts on a topic of interest. Based on their suggestions, the model to extract blog communities should be considered as composing communities of interests.

Some discussions about the importance of interest communities in the blogosphere are based on the potential contributions to marketing. Kale, Karandikar, Kolari, Java, Finin, and Joshi [9] suppose interest similarity could conduct trust building and influence giving among bloggers. They present a model to find "like minded" blogs based on blog-to-blog link sentiment for a particular domain, politics. They identify the polarity (positive, negative or neutral) of the text surrounding links that point from one blog post to another. Rather than passively mining the blogspace for business intelligence, Java et al. propose application of formal influence models to information propagation patterns in the blogspace, to generate CGM. This work attempts to locate a set of influential blogs which, if discussing a certain topic, are likely to maximize the “buzz” around this topic in the rest of the blogspace. From a marketer’s point of view, these sets of blogs constitute an important marketing channel [15].

Based on previous studies, we have known that researching on the relationship among interest communities in the blogosphere has strong potentials in marketing. However, to measure the information propagation among bloggers, involved extracting interest topics from blog contents, is complicated and hard to be generalized. The usage of word co-occurrence is validated in searching similarity in web contents but has to be customized on analyzing blog data. In the section 3, the details of our approaches and the specifications of the model design are introduced.

## III. SPECIFICATIONS OF SETTINGS

The model of propagation is used to monitoring the information flow among bloggers for viral marketing strategies. It searches for the relations between bloggers, including the existence of social relationship such as the actions of giving comments or invitations; and the potential relationship, like sharing similar interests. Usually a blogger’s interests are considered as the topics written in his or her blogs. Then, the patterns of topic propagations indicate the different purposes of marketing strategies.

We make use of easily accessible internet resources such as Wikipedia titles and the page returns of queried words from Goo search engine to formulated word relatedness measures. These usages are based on two considerations: accessibility, which can lessen the complexity of language processing, and appropriability, which can make the interest exaction from blog contents more topic oriented. In terms of the whole procedure the identification of the existing directional social relationship is antecedent, and the potential relationship of content relatedness is consequent. We now continue to describe the framework of our propagation model with its settings.

### *3.1 Propagations for Marketing Strategies*

The information propagation measured by in-link algorithms is used to mine bloggers with relationship of action giving and receiving for identifying the possible marketing channels of viral marketing. Such link propagation consists of an opinion leader blogger, a link relation, and the followers. The leader and follower are the actors who receive and give the link actions such as comment, read, click, forward, and trackback. In our implementation, the opinion leader is the most popular blogger in a topic domain, which can be defined and searched by the existing blog search systems. Then, the followers can be detected by the records of having link relations with the leader.

Content relations between the leader blogger (L) and followers (F1, F2) describing how much knowledge followers accedes from the leader can divide into several layers – blog, content, core word, topic, and propagation. These layers are used in matching propagations, which are done as consecutive steps. First of all, the propagation that is associated with a directional link relation is retrieved. Then, the propagation that is associated with blog contents is based on accumulated blog archives of each blogger. The third step tokenizes contents to identify the representative core words of each blogger’s interests and knowledge. Finally, the participants (L, F1, and F2) and the relations of similarity or relatedness for all core words composed topics are denoted. We illustrate this process with an example, which is depicted in Fig.2.

[2] http://www.delicious.com/

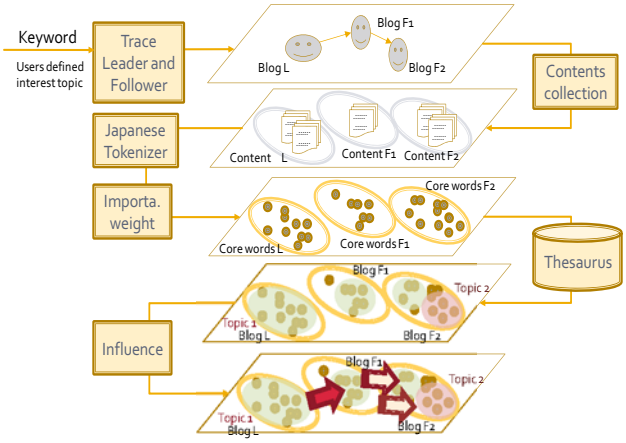



Figure 2. Data processes

Information propagation based on content can be interpreted into three patterns: (1) topic propagation, (2) topic evolvement, and (3) word repetition. Similarity in contents from the leader blogger to follower blogger describing how much knowledge the follower accedes from the leader has been used to sense the degree of direct influence from leader to follower [7][4][5][9].
(1)Topic propagation: For the application of viral marketing, topic propagation is created to trace the introduced message from leader to follower in terms of finding the efficient paths to access the most followers.
(2) Topic evolvement: Contrary to similarity, differences in bloggers' contents mean the different interests had by individuals. Bloggers sharing similar interests easily compose a community, but each blogger is supposed to have one's own distinguishing so that bloggers can exchange their knowledge and interests to influence each other in the viral space. Since leader and follower interact and influence each other, topic evolvement is created to trace the different interests from follower to leader in terms of fining the efficient way to penetrate different interest groups.
(3) Word repetition: Extending to topic propagation, among these followers word repetition is created to sense the imitation of the leader's word usage.
Based on identified relations of links and topics, the proposed propagations can be shown to the user for validation. It is up to the user to validate a certain identified propagation based on the blog contents in which the topic was found. The patterns of propagations based on content information are illustrated in Fig. 3.

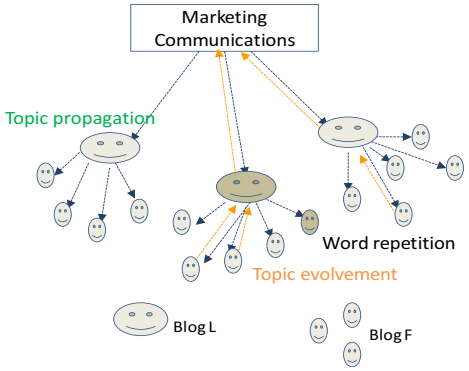



Figure 3. Patterns of propagations based on content information

We use an example to illustrate how topic propagation and topic evolvement interpret influences in viral marketing in Fig. 4.

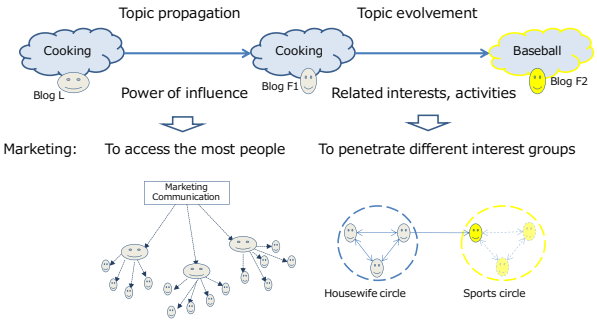



Figure 4. Topic propagation and evolvement for marketing purpose

Having traceable propagations in blog spaces will improves monitoring and decision making of marketing strategy. For instance, a follower of an opinion leader blogger in the topic domain of "cooking" talks about "life aboard", which reflects the blogger's personal experience or opinion. The second layer follower's blog is written about "online auctions". One can speculate that people who live aboard may interest in searching recipe or cooking tactics, moreover, checking for the opportunities of online auction. These records of topic propagation and evolvement describing the trending of bloggers' interests can be applied in the study of social phenomenon and human behavior at scales that before were never possible.

### *3.2 Topic Identification from Blog Content*

As stated earlier, the topic-based information extraction framework which we propose uses resources from internet to facilitate the blog-triggered thesaurus databases. One or more database resources are associated with technologies of semantic processing, which can weigh the importance of words in the contents. The goal of these processing is to enable knowledge engineers and marketing experts to express their knowledge in a simple yet expressive way by extracted topics with propagation patterns. To be able to make use of these topics, their semantics must be defined.

We can distinguish between two kinds of semantic processing: core word extraction and topic identification. Both of these processing are applied to an individual blogger' blog contents. Core word extraction deals with simple word tokenization and importance evaluation. Then, topic identification aims to group semantic related words into one topic, which is the challenging task in our work.

The characteristics of blog are depicted in Table 1, summarized that a blog post represents that blogger's personal opinion and experience. Because of the in Section 2 mentioned drawbacks of informal language used in blog contents such as new words, mixed languages, and loose grammar [13], a proper thesaurus database is necessary for expressing the characteristics of blog contents and the usage of word relatedness. For example, extracted blog topics represent bloggers' personal interests. The measured relatedness should focus

on the "relatedness" of activities or interests, not traditional relatedness of meaningfulness and explanation.

Table 1. Comparison between blog and other social media

| | Web pages | Forums | Blogs | SNS |
|---|---|---|---|---|
| Content | Anything, really, low on sentiment content | Specific topics, low on sentiment content | Personal, diary-like, commentary, observations, sentiments, moods, richer and more complete content | More personal, sentiments, moods. dialogue -like. |
| Links | Static | Close circle, Membership | Links change frequently; different types | Close circle, friendship |
| Timeline | Usually static | Daily based update | Daily based update | Hourly based update |
| Represents | Information | A group's knowledge | A person's life | A person's network |

In order to create more expressive word expressions a comparison of kinds of thesaurus database is exemplified in the table 2. Conventionally, dictionary and news databases are well used as semantic thesaurus. However, thesaurus sourced from dictionary like *ruigo.jp does not cover enough new topics, and the extracted topics are based on the explanatory relatedness and hierarchy of semantics. Even though thesaurus sourced from news like **database of Asahi News includes selected trending topics, the extracted topics are event oriented and based on time serial. Since 2007 Google n-gram has been launched as the hugest thesaurus source, which is based on all words from web pages which is about 20 billion documents, may cover all kinds of topics. Its calculation of word relatedness is based on frequency of co-occurring words in a sentence at most in the distance of 6 grams including noun, aux. verb, adj., verb, particle…. In this way, Google n-gram database cannot really reflect the relatedness of interests or activities.

Table 2. Comparison of thesaurus sources

| | Wikipedia | Web page | Dictionary | Google n-gram | News |
|---|---|---|---|---|---|
| Title subject | word | sentence | word | word in sentence | sentence |
| Double count | low | high | no | high | no |
| Dust | less | huge | no | huge | no |
| Update | frequent | frequent | Out of date | frequent | frequent |
| Size | huge (732,873 D) | too huge (3 billions D) | Middle (470,000 W)* | too huge (20 billion D) | huge (12million D)** |
| User edit | yes | yes | no | - | no |
| Gathering data | open download | take long time | have to buy | have to buy | have to buy |
| Limit for use | open | up to content (difficult to judge) | high (copy right) | - | high (copy right) |
| Quality Writing | middle | low | high | no | high |
| Topic oriented | high (related topics) | no | low (explanation) | no | middle (event oriented) |

In the aspect of expressing characteristics of blog contents, Wikipedia articles as consumer generated media (CGM) share the similar characteristics as blogs that both editors and audiences are internet users. Therefore, thesaurus sourced from Wikipedia could cover the most inclusive topics related to bloggers' interests and activities than other sources. For the concerns of data processing Wikipedia is the most feasible source in terms of the size of documents, dusts in the writings, and the quality of writings. Besides, it is also the most economical approach without limitation for usage. Incorporating thesaurus sourced from Wikipedia makes topic identification more proper and accessible.

## IV. METHODS

This section presents our methods able to identify topic propagation between leader and follower. Methods consist of several processes: detecting leader and follower bloggers, identifying topics of each blogger, and measuring the propagations among leader and follower bloggers. Each process consists of multiple components, i.e. leader and follower identifications, which are described in Section 4.1, core word extraction and topic identification, which are described in Section 4.2, and also topic propagation, topic evolvement, and word repetition, which are described in Section 4.3.

Using leader and follower identifications, the user can detect the opinion leader in a topic domain and the followers based on their relations with that leader. Core word extraction can filter out core words from content bodies of blogs. Topic identification makes users are able to group core words into topic of keyword related, keyword non-related, and noise by using defined relatedness measures. Propagation measures include topic propagation, topic evolvement, and word repetition, which present the close topics follower accedes from leader, the different topics follower derives from leader, and the same ideas follower copies from leader.

### *4.1 Leader and Follower Identifications*

The leader and follower identifications allow users to construct the structure of directional propagation which will be used to collect the propagated information. The proposed leader identification allows users to look for the opinion leader in their interest domain, which is created by modifying Blogranger API[3], a well-known Japanese blog recommendation system, to find the top 10 leader bloggers in a self-defined period. Based on the opinion leader's history of comment receiving, the commentators, who show the evidences that have obtained knowledge from leader, are traced as the followers. Fig. 5 shows the user interface which is used when the users want to find the opinion leader with followers. Such an approach allows the users to explore their interest domains and trace the evidences of influence paths from leader to follower.

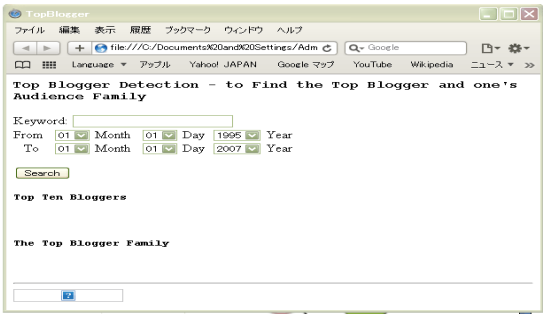



Figure 5. Interface for identifying leader and follower

Given the targeted Leader and follower bloggers' URLs, we collect their blog content bodies within two

[3] http://ranger.labs.goo.ne.jp/TG/

week long from the date when followers access the leader blogger's information. That is, the collected documents allow users to observe the changing interests and activities of that blogger in a period across two weekly life cycles.

### 4.2 *Identifying Topics of Bloggers*

To automatically conceptualize these fragments of contextual information is a challengeable task. Conventionally, the Tf-idf weight is well used in text mining field. However, in the case of mining blog content, the limited information amount of each blog post makes the calculation of the Tf (term frequency) loss in effectiveness, and the speedily increasing blog posts, which are not consistent in content, makes the idf (inverse document frequency) calculation impossible. In order to make good use of mining blog content, we have modified several methods based on the purpose of topic identification.

Identifying topics of each blogger involves two parts of data processing: core word extraction and topic identification. The basic text filtering for content analysis is done by using Mecab[4], which is an environment supporting the research and development of language processing software. Mecab provides its users with text-parsing engine that splits Japanese text into its separate morphemes. It also allows its users to develop their own word database or to extend the existing ones. For example a sentence "Like this case might be extreme" will be split into these morphemes: pre-noun adj. noun aux. verb noun particle noun particle verb aux. verb. In our framework, each content set will be pushed through this software. In order to identify topic, only nouns are extracted from the text bodies. We also extend the exiting word database of noun by adding all the Wikipedia titles to ensure that the new topics are included in the software and can be extracted from the contents.

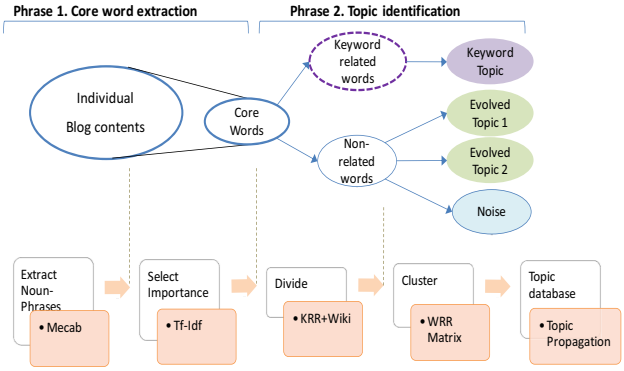



Figure 6. Framework of identify topics of each blogger

The framework of data processing, as depicted in Fig. 6 is comprised of several components, some of which are distributed with Mecab by default, whereas others are designed by us. After extracting nouns from blog contents with Mecab, the importance of each noun in a content body to all blog contents is weighted with Tf-idf (term frequency–inverse document frequency), which is a statistical measure used to evaluate how important a word is to a document in a collection. Through the calculation, only the important nouns of individual content set are selected as the core words to represent one's main topics of interests and activities.

In the process of topic identification, there are two steps: to identify keyword related words as in the same topic belonging of keyword, and for words without topic belonging, to cluster words into distinguished topics. Both steps are associated with the measure of word relatedness, which plays an important role in grouping semantically related two words in the same topic belonging.

The concept of relatedness measure is derived from the cosine similarity, which is often used to compare documents in text mining. Given two vectors, $W_i$ and $W_j$, with attributes of $w_i$ and $w_j$ word occurring, the cosine similarity, is represented using a dot product and magnitude as

$$\text{word relatedness} = \cos(W_i, W_j) = \frac{W_i \cdot W_j}{\|W_i\|\|W_j\|} = \frac{\sum_{k=1}^{n} W_{i_k} \times W_{j_k}}{\sqrt{\sum_{k=1}^{n}(W_{i_k})^2} \times \sqrt{\sum_{k=1}^{n}(W_{j_k})^2}}$$

where n is the size of documents. Each vector has a value for every word appearing in the document, with the value at that position containing the frequency of the word in the current section of the document. Thus, small segments of the document are compared based on their word frequency.

Exploiting the number of page returned from existing search engines provides users a very simple approach to measure word relatedness. The number of page returned is based on the largest document base to measure the general important of term with the concept of inverse document frequency (idf). The size of documents n equals the size of searched web documents by the search engine. The dot product of $W_i$ and $W_j$ can be the frequency of co-occurrence, and the magnitude of $\|W_i\|$and $\|W_j\|$ is the product of the square roots of the frequencies of word $w_i$ and word $w_j$ occurrences. Thus, the relatedness measure, called word relatedness ratio (*wrr*), is defined as the number of page returned of two co-occurring words divided by the numbers of page returned of each single word. We can denote *wrr* as:

$$wrr(w_i, w_j) = \frac{pr(w_i, w_j)}{\sqrt{pr(w_i)pr(w_j)}}$$

$$i, j = 1, 2, \ldots, n; n = \text{set of words}$$

where $pr(w_i)$ is the number of page returned where the word (wi) occurs, $pr(w_i, w_i)$ is the number of page returned where the words (wi, wj) occurs. However, judging similarity is conditionally based on the word itself and the size of the Web [3]. Besides, the quality of web documents does matter in the measuring

[4]http://code.google.com/p/furigana-injector/wiki/MeCab

performance. Strube and Ponzetto [17] pointed that in term of measuring similarity in English Wikipedia performs better than Google Counts than WordNet. In this study, Focusing on Japanese, a well known Japanese search engine, Goo, is adopted. In order to ensure the quality of searched results, the searched range is limited inside the contents of that search engine only.

In the first step of identifying keyword related words as in the same topic belonging of keyword, the keyword is decided by the user as one's interest domain. To measure the relatedness between word and keyword, *wrr* can be modified to keyword related ratio (*krr*), which is denoted as:

$$krr(w_i,k)=\frac{pr(w_i,k)}{\sqrt{pr(w_i)pr(k)}}$$

$$i,j=1,2,\ldots,\mathrm{n};\ \mathrm{n}=\text{set of words}$$

where k as a keyword, $pr(w_i,k)$ equals the number of page returned where the word ($w_i$) and keyword occur.

Subsequently, to decide if the word is in the same topic belonging of keyword, a benchmark of *krr*, which tells how close the related word and keyword should be circulated into a topic group, is necessary. The benchmark is a relative value of whole words' relatedness with keyword. Wikipedia titles as all inclusive words under web users' usage are appropriate to be applied to simulate the word relatedness with the keyword. From the distribution of *krr* with all kinds of words we can divide words into the closest related words and others. The division with a value of *krr* can be the benchmark to divide core words into keyword related group and non-related group.

The second step is for core words in the non-related group, to cluster them into distinguished topics. The values of *wrr* represent the relatedness of all words in pairs, which can be interpreted as the correlation values of core words of a blog content set. The correlation values of core words in pairs can be presented in a matrix, which is denoted as:

$$wrr(W_i,W_j)=\begin{pmatrix} wrr(w_1,w_1) & wrr(w_1,w_2) & \cdots & wrr(w_1,w_n) \\ wrr(w_2,w_1) & wrr(w_2,w_2) & \cdots & wrr(w_2,w_n) \\ \vdots & \vdots & \ddots & \vdots \\ wrr(w_n,w_1) & wrr(w_n,w_2) & \vdots & wrr(w_n,w_n) \end{pmatrix}$$

n : The number of words in a blog data set

Words in the same topic belonging are supposed to have the closest relatedness with each other. In order to group related core words into a topic, we run the optimization model to maximize the sum of *wrr* in each partition with a number of partitions given by the user. Through the calculation of matrix permutation, the optimized composition of topic groups is done. When the average *wrr* score of a partition is less than the total average *wrr* score of the whole contents, words in that partition are consider as noises, without any topic belonging. That is, even though the number of partitions is arbitrarily decided by the user, the result of word belongings to topics is constant.

### 4.3 Propagations Among Leader and Follower

Based on the results of topic identification in each blog data set, the propagations among leader and follower bloggers can be detected. There are three measures of propagations which have to present the degree of information propagation from the opinion leader to followers. Topic propagation, which presents the close topics follower accede from the leader, is calculated as the percentage of the core words in the same topic belonging with keyword to the whole core words. Topic evolvement presents the different topics derived from the leader. Word repetition, which presents the same ideas the follower accedes from the leader, is calculated as the percentage of identical core words used by both leader and follower to the whole core words. These simple propagation measures allow users observe the patters of propagations in the large scale of blog data.

## V. ANALYSIS AND RESULTS

We now continue with analyzing real data to confirm the effectiveness of our propagation model. First, with an inputted keyword, which indicates our interested topic domain, we analyze contents of the leader and follower bloggers in that interest domain and show the propagations between them. In Section 5.1, we introduce the identification of the leader and follower bloggers. After this, we perform the topic identification of individual blog contents in Section 5.2. Finally, Section 5.3 performs measures of propagations among bloggers.

### 5.1 Leader and Follower Identifications

The research purpose of this study is to build up comprehension of the information propagations between virtual individuals. We model propagations in a finest scale to introduce the relationships between bloggers in blog layer and topic layer. Through the application program interface (API) of Blogranger, the opinion leader of all Japanese blogs in that topic domain is identified. We used cooking as the keyword to identify opinion leader (L). The reason to choose this keyword is because cooking is one of the popular topics, which has formed many small and strong virtual communities in the blog spaces.


Figure 7. Identify the opinion leader in cooking area.

Given the searched period is from 2006/4/20 to 2007/4/20, the opinion leader (L) in cooking domain is found as shown in Fig.7. The most popular blog posted on the date 2007/4/9. On that blog there are 125 comments recorded, among these 65 are self-commented by the author; 25 comments are traceable with

commentators' URLs. For a sample demonstration, we selected one with the most comments as the follower in the first layer (F1) from these 25 commentators; then with the same logic the follower in the second layer (F2) is identified.

Once the positions of leader (L) and followers (F1, F2) are fixed, we collect their blog articles from the date they posted over two-week period by using our data collection program. The reason we use two weeks as the data collection period is concerning the effective of topic transformation over weekly life cycle. Since most blogs are written as personal diaries, analyzing accumulated blog content could be an effective way to peek into one's interests, life style, and thoughts.

### 5.2 *Identifying Topics of Blogger*

This section performs the basic content filtering of individual blog content set and present samples of topic identification in each step. Given the collected contents in previous section, the noun phrases are extracted by the Mecab software with our extension of the existing noun phrase database. The customization by us is adding all the Wikipedia titles around 500,000 words (until 2007/5/12) to ensure that the new topics are included in the software and can be extracted from the contents.

Referred to Fig. 5 the framework of topic identification, we input three set of individual blog contents into Mecab and get return of 420 words, 529 words, and 437 words of nouns extracted from L, F1, and F2 respectively. In order to weight the importance of each extracted noun to its origin contents we use Tf-idf algorithm given the occurrence of the noun in content and the size of the origin data set. During our development period, we have seen the right-skewed distribution shape in all the cases that the relatively important words, words with higher Tf-idf values, occupy 15-20% of a data set. For convenience we decide the 15% most important words in the individual contents as the core words of that blogger. In this way, L blogger has 63 core words, F1 blogger has 79 core words, and F2 has 62 core words, respective to their contents.

Subsequently, referred to Fig.5 the challenging task of topic identification is conducted in the later of this section. We introduce the performance of the processing with real data. For the convenience of reading some of the samples are only partially shown due to the huge size of data. In the first step: to identify keyword related words as in the same topic belonging of keyword, we use whole the Wikipedia titles to simulate the word relatedness with keyword, then decide the benchmark of relatedness, which is the reference to judge if a core word is in the keyword related group or non-related group.

The database of Wikipedia titles include 500,000 titles in Japanese, which are considered as nouns. In order to effectively examine titles' relatedness with the default keyword – cooking, we randomly select 5% of 500,000 titles by the sequence of alphabet. Among the selected 25,000 titles, a big portion of these are not meaningful such as repeated Japanese alphabets, punctuation marks, and so on. We eliminate the meaningless titles by examining their frequency of usage in general writings. Therefore, only titles with more than 50 pages returned by the Goo search engine are considered as into our Wikipedia thesaurus database, totally 4,341 titles.

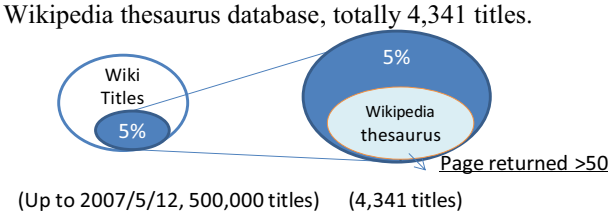



Figure 9. The Wikipedia thesaurus

Finally, we input titles from the Wikipedia thesaurus with the keyword and return their keyword related ratio (*krr*) values, which is introduced in Section 4.2. The Wikipedia thesaurus is selected to be representative of the usage of nouns in the real world. From the distribution of the *krr* and log (*krr*) index of the Wikipedia thesaurus, shown in Fig. 10, we can decide that the relatedness benchmark is given by the top 5% related titles, given that the 5% is the portion of the upper two- standard deviation. That is, the *krr* benchmark is 0.27 by which we can say if any word's *krr* value is bigger than 0.27, this word is strongly related with the keyword –cooking, and vice versa.

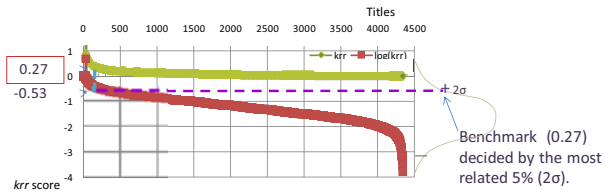



Figure 10. Distribution of the *krr* and log (*krr*) index

The purpose of the first step of topic identification is to find if the blogger has the similar topic belonging with the default keyword. When there is none of core words related to the keyword, or the keyword-related words are excluded, we use matrix permutation to cluster these words into distinguished topics. Section 4.2 explains the *wrr* value and how can the *wrr* values be applied to a correlation matrix of core words in pairs. We utilize an analytic tool called UCINET 6 to run the optimization. Given that the default number of partitions is 4, the function of Optimization in Clustering analysis finds the word composition of maximized sum of *wrr* in each partition. A sample result is shown in Table 4, where core words: Giant (baseball team), fly, game, Chunichi (baseball team), Yakuru (baseball team), professional, team, base on balls, Yankees (baseball team), Hanshin (baseball team), Matsusaka (baseball player), Utsumi (baseball player), baseball, grounder, Matsui (baseball player), standings, leading hitter, Igawa (baseball player), Red Sox (baseball team), batter's box, Okajima (baseball player), and major league, are clustered in a topic; mail magazine, important person, tendency, Hiroshima, Clinton, Yokohama, Russia, Cabinet, center, Diplomacy, Bush, tease, countries, and Abe are clustered in another topic. The leading words of topic 1 are Matsusaka and baseball because of the largest sum of *wrr* in group.

Table 4. Topic identification of blogger F2

| Topic 1 | | Topic 2 | | Noise | | Noise | |
|---|---|---|---|---|---|---|---|
| core word | sum of *wrr* | core word | sum of *wrr* | core word | sum of *wrr* | core word | sum of *wrr* |
| 巨人 (Giant ) | 3.829 | メルマガ (mail magazine) | 1.789 | 飛行機 (airplane) | 1.376 | プラス (plus) | 0.987 |
| フライ (fly) | 3.4 | 要人 (importabt person) | 0.859 | 数字 (number) | 1.54 | 葬儀 (funeral) | 0.98 |
| ゲーム (game) | 4.722 | 傾向 (tendency) | 1.942 | 事務所 (office) | 2.064 | ローマ法王 (Pope) | 0 |
| 中日(Chunichi) | 3.635 | 広島 (Hiroshima) | 2.041 | キャスター (caster) | 1.496 | 倖田來未 (Kumi koudaku) | 1.951 |
| ヤクルト (Yakuruto) | 3.995 | クリントン (Clinton) | 1.653 | ユニット (unit) | 1.778 | 小谷真生子 (Kotami maoko) | 1.413 |
| プロ (professional) | 4.28 | 横浜 (Yokohama) | 1.798 | メッセージ (message) | 2.192 | 最下位 (least significant bit) | 0.98 |
| チーム (team) | 4.318 | ロシア(Russia) | 2.274 | 聡子 (Satoko) | 1.469 | 吉沢 (Yoshizawa) | 0.988 |
| 四球 (base on balls) | 4.572 | 内閣 (Cabinet) | 2.257 | 未来 (future) | 1.926 | 得失点 (goals) | 0 |
| ヤンキース (Yankees) | 3.655 | センター (center) | 2.363 | ひとみ (Hitomi) | 1.746 | | |
| 阪神 (Hanshin) | 3.336 | 外交 (Diplomacy) | 2.41 | マガジン (magazine) | 1.414 | | |
| 松坂 (Matsusaka) | 6.738 | ブッシュ (Bush) | 2.324 | 藤井 (Fujii) | 1.544 | | |
| 内海 (Utsumi) | 2.54 | いじめ (tease) | 2.09 | 希美 (Nozomi) | 1.507 | | |
| 野球 (baseball) | 5.563 | 各国 (countries) | 2.154 | | | | |
| ゴロ (grounder) | 3.268 | 安倍 (Abe) | 2.164 | | | | |
| 松井 (Matsui) | 4.585 | 大統領 (President) | 2.966 | | | | |
| 順位 (standings) | 2.545 | 首相 (Prime minister) | 3.48 | | | | |
| 首位 (leading hitter) | 4.448 | | | | | | |
| 井川 (Igawa) | 4.416 | | | | | | |
| レッドソックス (Red Sox) | 2.875 | | | | | | |
| 打席 (batter's box) | 3.921 | | | | | | |
| 岡島 (Okajima) | 2.28 | | | | | | |
| 大リーグ (major league) | 2.605 | | | | | | |

The whole clustering results of blogger F2's core words are presented in Table 4, including 63 core words, in 4 partitions. Numbers in the right column are the sums of *wrr* scores for each core word. When the average *wrr* score of a partition is smaller than the average of whole the *wrr* score, about 2.5, words in that partition are considered as noises. As the results show, mainly two topics – baseball and politics are interested by the blogger F2. Considering the core words respect to the topic, the topic identification has successfully distinguished words into topics.

### *5.3 Propagations Among Leader and Follower*

As shown in Fig. 11, compared to the clustered topics and the included core words of the leader and two follower bloggers, the leader blogger only concentrates on "cooking" topic; the follower F1 blogger shows interests in not only "cooking" but also other leisure activities such as movie, game, shopping, traveling…; the follower F2 blogger does not show interests in "cooking" but shows strong interest in "baseball" and "politics". Differences between the results of topic domains can be explained by the characteristics of the blogging-behavior base and patterns. The blogger who is identified as an opinion leader in a specific field usually introduces one's professional knowledge in the writings. For example the leader blogger in our case is special in "cooking" area whose writings introduce recipes, eating ideas by using a lot of proper nouns, which is similar to the writings of news reports in terms of full of substantial evidences. It shows a typical blogging behavior of knowledge introducing.

On the other hand, compared with the opinion leader, the follower bloggers show loose topic concentration in terms of showing multi-interest focus and looser performance in identified boundary between topic belongings. For example, both followers F1and F2 have more than one topic concentration shown in their writings. The follower F1 shows looser performance in topic clustering except the propagated topic -"cooking" from the leader. By observing the other two topics of core words, starting with "movie" and "diamond", one could notice that both topics are indicating some related leisure activities such as traveling, shopping and so on, but the differences are hard to be defined. On the contrary, topic identification in the case of follower F2 perfectly defines core words' topic belongings into "baseball" and "politics". These results confirm with our knowledge base that blog writing and blogger's interactions differ by topic. However, the differences are hard to be told by machine. From the results of topic identification of these three cases, one can conclude that the ways bloggers elaborate interests are able to be distinguished: some are knowledge introducing or sharing like L and F2 bloggers'; some are life recording like F1 blogger's.

Finally, the propagation patterns can be measured based on the leader and followers' results of topic identification. Similar topics from the opinion leader to followers contribute the measure of topic propagation. Different topics from followers to the leader contribute the measure of topic evolvement. Identical core words which are used by followers contribute the measure of word repetition. In the sample results of topic identification, topic propagation of "cooking" can be detected from L to F1 blogger. In F2 blogger's blog contents, 22% are related with the leader (L) blogger's. Among this, 17% word repetition indicates that F2 blogger is influenced from reading L's blogs. In terms of topic evolvement of "cooking", " movie" ,"travel" and "shopping" topics are directly evolved from "cooking", while "baseball" and "politics" topics are farther related with "cooking" topic. The summary is denoted in Table 5.

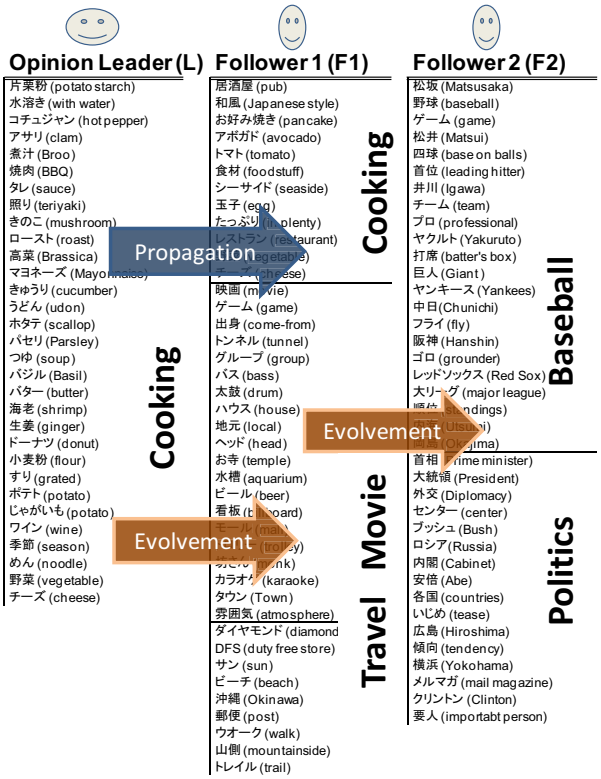



Figure 11. Propagations from the leader to follower bloggers

Table 5. Summary of propagation patterns between bloggers

| | L | F1 | F2 |
|---|---|---|---|
| Topic propagation | "cooking" | "cooking" (22%) | 0 |
| Word repetition | - | 17% | 0 |
| Topic evolvement | "cooking" | "movie","travel", "shopping" | "baseball", "politics" |

While our sample of analysis and results within three-layer interacted bloggers shows that by observing bloggers' word usage and topic concentration different types of blogging behaviors - knowledge introducing and life recording can be distinguished. The measures of topic propagation and word repetition show the degree of influence that followers receive from the opinion leader. Topic evolvement tells the relative distances of bloggers' interest topics. In terms of applications in the viral marketing research, topic propagation is contributive to search for efficient ways to access the most people, and topic evolvement is contributive for the ways to penetrate different interest groups once enough data have been accumulated.

## VI. CONCLUSIONS

In this paper we have proposed the use of exiting internet resources for interest topics extraction from blog content feeds. This approach is implemented with a model combined link and topic analyses with which one and results are limited within three-layer interacted bloggers, it will not harm for the generalization of the method because the interest domain, the default keyword, is designed as replaceable based on the Blogranger's can trace interests based influences using propagation patterns. These propagation patterns make use of bloggers' interaction and topics relatedness, which leverage the existing information propagation models to a higher concretion level in the applications of viral marketing. Topic propagation and word repetition are contributive to efficiently access to most people. Topic evolvement is contributive to penetrate different interest groups, which is also important for providing customized marketing offers by cross promotion of topics from different product fields. In order to approach these topic based propagations, we have developed and presented a system of topic identification specified for blog content feeds, as well as blogger identification focused on directional interactions.

In our work, we make use of combining the online thesaurus database - Wikipedia titles, the search engine - Goo search engine, and the blog recommendation system - Blogranger which reflect the most up-to-date topics in the blogosphere, allow an easy construction and understanding of propagation patterns by users. The contribution of this paper is twofold. Technically, the contribution of our approaches focuses on processing the informal contextual information of blog contents, which general text mining methods could not work well in topic identification. To this extent, this system, which combined in-link algorithms and text mining analyzes, tracing propagation channels and propagateable information allows analyzing the power of influences in viral marketing.

Our system consists of many processes. For each process, we compared relevant methods and chose the best performed one. For example, in the case of measuring word relatedness, we have compared the effectiveness of the number of page returned from search engines such as Google, Yahoo, and Goo. Eventually, Goo returns the best results due to its language focus and relatively small size of data, also less spam information. Other cases like adopting Cosine similarity and matrix optimization are all based on the comparison of existing methods and the applicability of data processing. This paper does not focus on detailed mathematical analysis nor fine algorithm design, rather on the conceptual and methodological system used to approach the analysis of viral marketing. Therefore, we demonstrated the effectiveness of the system by applying blogger identification, topic identification, and the topic propagations.

The effectiveness of the proposed approach mainly depends on the results of topic identification. We conclude that generally, the tool is effective because of its high accuracy, i.e., the topic belonging of each core word is correct by human judge. While our sample of analysis

definition of top blogger in a topic domain. Followed our designs of locating followers and the *krr* benchmark against topic identification will not ruin the capability of the methods as well. Some setting of the methods has

been decided by authors such as the number of partitions during the process of *wrr* matrix optimization, which depends on how meticulous users prefer and will not change results in differentiating topics from noises. Our work aims to provide a method to discover propagations in general cases. In spite of the stability of the system has not yet been verified. It shows promising results to enhance performance of automated generalization processing.

For future research, we suggest a couple of directions. First of all, the topic-triggered actions which are now used for identifying bloggers' interests can also be used for other purposes. For example, we could combine event based thesaurus by including news database, thereby automatically notifying blogger interests with related real-world events in a real-time manner. Secondly, currently we have been focusing on processing representative bloggers' blog contents for two week long instead of entire archives. The reason for this is that the collected documents allow users to observe bloggers' interests and activities in a general life cycle. Moreover, data can be processed in a limited amount of time. However, the drawback of this approach is that except opinion leaders, common bloggers are not eager to update their blogs that is needed for monitoring the change, which is likely to be solved by processing the entire blog archives. Therefore, future research into the possibilities of processing entire blog archives is suggested.

Furthermore, it would be interesting to conduct research on interest chains, as usually, interests are not isolated but they are part of a chain of interest. For instance, the readers who are interested in cuisine may prefer those blogs that share recipes, recommend kitchen utensil, or introduce lifestyle. Among those readers, some may have interests in both cuisine and kids education kinds of blogs [2]. It would be interesting to formulate such chains of interests in order to monitor the developments of specific domains over time. By identifying these patterns of interests, forecasting of what people will be interested can be done. If certain events intrigue people with specific interests, it is likely that people with other interests are also intrigued.

Related future work might include recognizing the difference in the motivation aspects of interest sharing (e.g., knowledge introducing, life recording, etc.) and using this accordingly when updating the knowledge base. Also, one could consider adding the attitude (e.g., negation and approval etc.) support to the topics shared. At the current moment, it is the user who decides if a topic has been correctly found and it needs to trigger bloggers' motivations for blogging behaviors. Hence, a framework of motivations to share interests and attitudes to shared ideas would be desirable, as these enhance performance of automatic processing. Furthermore, it would be worthwhile to investigate blog ranking based on evidence. If a blog or topic is frequently identified, there is more evidence, and thus the blog or topic is more likely to be credible than is the case with less evidence.

Additionally, topic identification needs to be improved. Related future research could include automatic topic adjustment, as current leading words in topics are chosen by relative frequency of word occurrence. However, the popularity of topics in the blog spaces or internet spaces is extremely unbalanced. Automating the adjustment process would improve the stability of our solutions. A possible solution might be to perform generalizations based on hierarchical summarization. By analyzing words with hierarchical structures (e.g., semantic hierarchy, information directory, etc.) choosing a representative word to a topic can be formulated. Users can then validate these topics and associate other applications to them.

Finally, the usage of propagation patterns can be subject to further research. The propagation patterns introduced in this paper are addressed more on text and ontologies. Perhaps, in the future work we could use the output of propagations as input for expressive large-scale graphs and ontologies.

## REFERENCES

[1] Ali-Hasana, N. and Adamic, L., (2007), Expressing Social Relationships On The Blog Through Links And Comments, In ICWSM,, Boulder, Colorado, 2007.

[2] Chen, M. and Ohta, T. (2010), Using Blog Content Depth And Breadth To Access and Classify Blogs. International Journal of Business and Information Volume 5, number 1, June 2010.

[3] Choudhari, R., R. D., and A. (2008), Increasing Search Engine Efficiency Using Cooperative Web. CSSE '08: Proceedings of the 2008 International Conference on Computer Science and Software Engineering - Volume 04, IEEE Computer Society

[4] Fujimura, K., Toda, H., Inoue, T., Hiroshima, N., Kataoka, R., Sugizaki, M. (2006). BLOGRANGER – A Multi-faceted Blog Search Engine, Proc. WWW'06.

[5] Fujimura, S., Fujimura, K., and Okuda, H. (2007). Blogosonomy: Autotagging Any Text Using Bloggers' Knowledge. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI '07.

[6] Hayes, C., Avesani, P., and Bojars, U. (2007), An Analysis Of Bloggers, Topics And Tags For A Blog Recommender System, From Web to Social Web: Discovering and Deploying User and Content Profiles, Lecture Notes in Computer Science, 2007, Volume 4737/2007

[7] Java, A., Kolari, P., Finin, T., and Oates, T. (2006). Modeling the Spread of Influence on the Blogosphere. In WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, at WWW '06: the 15th international conference on World Wide Web, 2006.

[8] Jung, J.J. (2009), Knowledge Distribution Via Shared Context Between Blog-based Knowledge Management Systems: A case study of collaborative tagging. Expert Systems with Applications, Volume 36, Issue 7, September 2009, Pages 10627-10633.

[9] Kale, A., Karandikar, A., Kolari, P., Java, A., Joshi, A., and Finin. T. (2007), Modeling Trust And Influence In The Blogosphere Using Link Polarity. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007), March 2007. Short Paper.
[10] Kirby, J., & Marsden, P. (Eds.). (2006), Connected Marketing: The Viral, Buzz and Word of Mouth Revolution. p107-118, Oxford, Butterworth-Heinemann (Elsevier), 2006
[11] Lake, L. (2003), Word Of Mouth vs. Viral Marketing: What's The Difference?
[12] Leskovec, J., Adamic, L., and Huberman, B.(2007), The Dynamics Of Viral Marketing, ACM Transactions on the Web (TWEB), 1(1), 2007.
[13] Mishne, G. (2006), Information Access Challenges In The Blogspace. In the International Workshop on Intelligent Information Access (IIIA 2006).
[14] Mishne, G., & de Rijke, M. (2006), A Study Of Blog Search, in ECIR 2006.
[15] Mishne, G. (2007), Using Blog Properties To Improve Retrieval. In ICWSM 2007.
[16] Muller, M.J. (2007), Comparing Tagging Vocabularies Among Four Enterprise Tag-based Services, Proceedings of the 2007 international ACM conference on Supporting group work.
[17] Ponzetto, S.P. and Strube, M. (2007), Knowledge Derived From Wikipedia For Computing Semantic Relatedness, Journal of Artificial Intelligence Research 30 (2007) 181-212
[18] Wu, F., Huberman, B. A., Adamic, L.A., and Tyler, J.R. (2004), Information Flow In Social Groups', Physica A, 337:327-335, 2004

**Meichieh Chen** is a post-graduate student in the Department of Social Intelligence and Informatics in the Graduate School of Information Systems at the University of Electro-Communications, Japan. She is pursuing her doctoral thesis focusing on marketing prediction with analysis of blog content. Other research interests include relationship marketing and virtual communities, especially which based on the technology of dynamics network analysis.

**Neil Rubens** is an assistant professor at the Knowledge Systems Laboratory, University of Electro-Communications, Japan. He holds a M.Sc. degree from the University of Massachusetts and a Ph.D. degree from the Tokyo Institute of Technology - both in Computer Science. His research focuses on developing Active Intelligence systems which are systems that are self-adaptable utilizing unsupervised and semi-supervised learning, and active communication and data acquisition. He has applied developed methods to diverse fields such as e-Learning, information retrieval, recommender systems, bioinformatics, and policy analysis. Dr. Rubens authored chapters on the topics of machine learning, active learning and recommender systems published by MIT Press and Springer. His research has received funding from corporations and the governments of Japan, United States and Sweden.

**Fumihiko Anma** is an assistant professor with Graduate school of Information Systems, The University of Electro-Communications. He received his B.E. from Tokyo Institute of Technology in 2000 and his Master Degree and PhD Degree from Shizuoka University, Japan in 2002 and 2005 respectively. His research interests include knowledge computing, e-learning, artificial intelligence, and semantic web. He is a member of Japanese Society for Information and Systems in Education, Japanese Society for Artificial Intelligence and Japan Society for Educational Technology.

**Toshio Okamoto** is a professor in the Graduate School of Information Systems, the University of Electro-Communications. He is also the director of the Center for e-Learning Research and Promotion in UEC and the convener of WG2 (Collaborative Technology) of ISO/JTC1 SC36 (Learning Technologies Standards Committee). He obtained his PhD from Tokyo institute of Technology in 1988. His research areas place on the system development of intelligent web-based educational systems utilizing the artificial intelligence technologies such as e-Learning, web based collaborative learning including recommending and knowledge mining functions.

# Internationally Distributed Living Labs and Digital Ecosystems for Fostering Local Innovations in Everyday Life

Tingan Tang
Aalto University School of Science, Espoo, Finland
Email: tingan.tang@aalto.fi
Zhenyu Wu
Beijing University of Posts and Telecommunications, Beijing, China
Email: shower0512@bupt.cn
Kimmo Karhu
Aalto University School of Science, Espoo, Finland
Email: kimmo.karhu@aalto.fi
Matti Hämäläinen
Aalto University School of Science, Espoo, Finland
Email: matti.hamalainen@aalto.fi
Yang Ji
Beijing University of Posts and Telecommunications, Beijing, China
Email: jiyang@bupt.edu.cn

*Abstract*— **There are an increasing number of information sources and services around us enabling new ways of interacting with our everyday environment. Examples include intelligent devices, sensors embedded in the environment and the emerging Internet-of-Things. Simultaneously users are becoming increasingly involved as information providers and consumers by means of Web 2.0 and social media. While these areas have gained a lot of attention recently and while the research on Digital Ecosystems has also dealt with these phenomena separately there seems to be need for research on the rich and complex ecosystem combining the sensor-based information sources with Web 2.0 and mobile services. In this paper, we propose a Digital Ecosystem architecture, which combines the social media and Internet-of-Things. The architecture is the fruit from the international collaboration between two long-term university Living Lab projects in Finland and in China. It aims at fostering student innovations in their everyday campus lives. We discuss the experiences learnt in the context of this international collaboration and the implications to Digital Ecosystem research.**

*Index Terms*— **Living Labs, Digital Ecosystems, Internet of Things, Ubiquitous Computing, Web of Things, Social Media**

## I. Introduction

With the rapid development of Information and Communication Technology (ICT), digitization has entered into nearly every aspect of people's lives such as entertainment, e-business, e-learning, e-government and e-health. There is an increasing number of information sources and services around us enabling new ways of interacting with our everyday environment in work, at home and in leisure activities. Examples include intelligent devices, sensors embedded in the environment and the emerging Internet-of-Things (IoT). Simultaneously users are becoming increasingly involved as information providers and consumers by means of Web 2.0 and social media. Around these phenomena, some emerging research concepts and research fields have become popular such as Digital Ecosystem (DE), Living Lab and Experiential Computing, which will be explained in more detail in the related research subsection.

In this paper, we describe the development of an internationally distributed DE research environment aiming at providing a Living Lab type of innovation platform for studying the combination of the "intelligent environment" and the users and their activities. Specifically, we describe the underlying architecture of the environment for enabling research in service creation and use over long periods of time in real-life settings.

### A. Related Research

*a) Digital Ecosystem:* The origin of DE concept is related to the concept of Digital Business Ecosystem (DBE) , which was first proposed in Europe as a response to how the European Union could assist the SMEs (Small and Medium Enterprises) to adopt ICT technologies more effectively to improve productivity [1]. Nachira defined DBE as " a 'digital environment' populated by 'digital species' which could be software components, applications, services, knowledge, business models, training

This paper is based on "An internationally distributed ubiquitous living lab innovation platform for digital ecosystem research," by Tang, T. and Wu, Z. and Karhu, K. and Hämäläinen, M. and Ji, Y., which appeared in the Proceedings of the International Conference on Management of Emergent Digital EcoSystems, Bangkok, Thailand, October 2010. © 2010 ACM.

This work was supported in Sino-Finnish science and technology collaboration project No. 2010DFB10570.


doi:10.4304/jetwi.4.1.106-115

modules, contractual frameworks, laws, ...” [1]. The DBE is the combination of technical or digital part (Digital Ecosystem) and business part (Business Ecosystem) [2]. The DBE definition emphasizes the perspective of business.

There are many different emerging definitions for DE. For example, Briscoe and Wilde define DE as “the digital counterparts of biological ecosystems, which are self-organizing and scalable architectures that can automatically solve complex, dynamic problems” [3]. This definition views DE from architecture perspective. Chang and West define DE as “an open, loosely coupled, domain clustered, demand-driven, self-organizing agents’ environment, where each specie is proactive and responsive for its own benefit or profit” [4]. This definition views DE from environment perspective. For the purpose of this paper, we view DE as a technical architecture.

Many emerging DEs have been studied recently. For example, Karhu et al. study a DE where users use Web 2.0 tools to develop new web services [5]. Lawson et al. study a virtual museum DE implemented as Web 2.0 applications [6]. Briscoe and Marinos study the DE from Cloud Computing’s perspective [7]. Innocenti et al. study the DE of a digital preservation system [8]. These studies approach DE from the perspectives of Web 2.0 and Web services.

On the other hand, with the development of sensors, RFID (Radio-Frequency Identification), wireless networks and other enabling technologies, more and more devices and artifacts in daily life such as washing machines and coffee machines have computing and communication capabilities and become new digital species in ICT networks. There is also a lot of research on sensor-based systems in DE area. For example, Mostefaoui and Piranda study the architecture of a multimedia sensor network [9]. Zatout et al. study a hybrid wireless sensor network architecture for monitoring people at home [10]. Liu and Roantree study a precomputing query method for personal health sensor environments to overcome the inefficiency of XML query languages [11].

However, the research to combine the sensor-based systems with Web 2.0 and social media and studying the combination from the ecosystem perspective appears to be scarce.

*b) Living Lab:* The Living Lab concept was initially developed by Prof. William Mitchell, of the MIT MediaLab and School of Architecture [12]. According to the statistics of ENoLL (European Network of Living Labs) website, Living Labs are getting increasing momentum all over the world [13]. During its rapid growth, it has been defined as an environment [14], a methodology [12], [15] and a system [16] for innovation. Although different definitions view Living Lab from different perspectives, two common emphasis points are the central roles of users in innovation and the importance of real-life contexts or living environment of users for innovation. In this paper, we see Living Lab both as an environment and a methodology depending on the context of discussion. We define Living Lab as a user-centric and multi-party collaborative R&D methodology or environment where innovations such as new services are created and validated in multi-contextual real-life environment within individual regions [12], [14].

*c) Experiential Computing:* Traditionally, computing is separated from other forms of human activities and focuses on organizations and business [17]–[19]. With the ubiquity and pervasion of ICT and digitization by sensors, embedded computing, mobile computing and social computing, a new computing paradigm called “experiential computing” has emerged [17]. Experiential computing is defined as “digitally mediated embodied experiences in everyday activities through everyday artifacts with embedded computing capabilities” [19]. There is a lot of call for more research on experiential computing [17], [19].

### B. Research Motivation

First, we illustrate the relationships between the aforementioned three emerging research concepts: DE, Living Lab and Experiential Computing in Fig. 1.

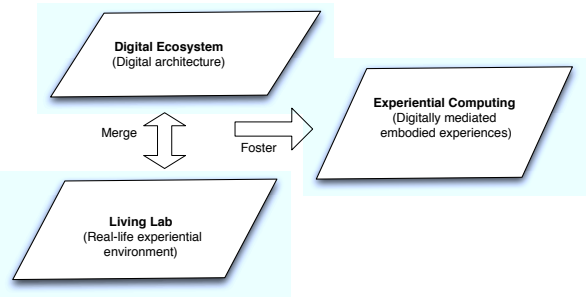



Figure 1. The relationships between Digital Ecosystem, Living Lab and Experiential Computing

DE provides a scalable and self-organizing ICT technical architecture [3], while Living Lab provides a rich experiential environment for data, information and innovation sources [12]. We believe that the combination of DE and Living Lab will foster experiential computing and innovations in everyday life experiences.

As we mentioned in the previous subsection, though there are a lot of DE research on Web 2.0 and sensor-based applications, the study to combine these two parts seems to be scarce. The paper aims at filling the gap in DE research and also responding to the call for more research on experiential computing. We describe a joint research effort carried out by researchers in Europe and China on campus-based Living Labs for studying the combination of ubiquitous and mobile social media services. We propose a DE architecture enabling easy use of the information sources in the environment for locally and situationally relevant services and for collection of quantitative and qualitative data on the use of the services in people’s daily lives.

The key contribution of this paper is by mapping the concepts of Living Lab, social media (Web 2.0 services)

and ubiquitous services (sensor-based services) in the context of internationally distributed DE research.

### C. Structure of the paper

The rest of the paper is organized as follows: First, we introduce the background and current status of the two Living Lab research projects in Finland and in China in Section II and Section III respectively. Section IV presents the DE architecture design and implementation of the ubiquitous campus Living Lab innovation platform. Section V discusses what we have learned in the process of the international project collaboration and the implications of our work to DE research and future work. Finally, Section VI concludes the paper.

## II. Living Lab and Social Media—OtaSizzle Project

### A. OtaSizzle Project Background

Social media such as Wikipedia, Facebook, Twitter and YouTube have become more and more popular in people's digital lives. According to Hitwise, an online trends and analysis website, Facebook surpassed Google Search as the most visited site in US on the week ending March 13, 2010. The popularity of social media has attracted a lot of academic and industrial researchers all over the world to study this phenomenon. It's worthwhile to note that similar developments in social media are taking place also among the extensive Internet and mobile user population in China. For example, the mentioned social media services have Chinese equivalents that are rapidly expanding: Twitter—Sina Weibo, YouTube—Tudou and Youku, as well as the several local Facebook type of services.

With the development of 3G networks and smart phones, social media are increasingly being accessed by mobile clients or browsers. Mobile social media are new focal research areas for researchers. However, although there are already a great many of social media services such as Facebook and Twitter in the market, the data in existing social media services are mostly proprietary and controlled by companies. Therefore, access to relevant data is a major challenge for researchers when studying mobile social media [20]. The aim of OtaSizzle project is not to compete with the existing social media services, but to provide an environment that is designed and well instrumented for supporting research. According to the project Wiki, "OtaSizzle will develop an open experimentation environment for testing mobile social media services. It will be a 'living lab' for thousands of users in Otaniemi, with extensions in greater Helsinki. The project will create prototype mobile social media service platforms and study them with extensive field tests, coupled with quantitative measurements and qualitative analysis. The outcome will be a "packaged" experimentation environment, "SizzleLab" concept" [21].

### B. Current Status of OtaSizzle

The OtaSizzle platform includes core services and end user services. The core services provide some common services such as user profiles, user groups, session management, location information and social networks that are shared by all end user services. End users can keep the same accounts and their social relationships among all end user services. Some core services are provided by the project. For example, the ASI (Aalto Social Interface, http://cos.sizl.org) service is social networking web service built with Ruby on Rails. Some core services are provided by third-party services providers such as the geolocation and localization services provided by OpenNetMap (www.opennetmap.org). On top of the core and enabling services, end users can create many kinds of mobile and Web-based social media services [22]. The end user services can be created by different programming languages such as Ruby, Java, PHP and JavaScript. The communication between the core services and the end user services is based on RESTful (REpresentational State Transfer) HTTP request and response [23]. The overview of OtaSizzle project architecture is shown in Fig. 2.

Currently there are five end user services listed in the Sizl portal (www.sizl.org). Later, all the end user services and applications will be listed in a dedicated marketing place—Aalto Apps. Among the five end user services, Ossi (https://ossi.sizl.org) is a group-centered mobile social media service oriented to high-end mobile phones such as Nokia N97 and iPhone. In Ossi, users can connect with friends, create new groups, join existing groups and initiate discussions [22]. The user interface of Ossi is shown in Fig. 3.


Figure 3. Ossi mobile social media service interface

Kassi (http://aalto.kassi.eu/) is an online resource exchanging social media service. In Kassi, users can post what goods and services they can give and what they need [24]. Currently, Kassi can only be accessed by browsers in computers. Mobile version is under active development. The home page of Kassi is shown in Fig. 4.

NordSecMob (http://nsm.sizl.org/) is practical information sharing social net-working service for NordSecMob (Master Program in Security and Mobile Computing) student community [24]. It can be accessed by computers and mobile phones.

Unlike the Ossi, Kassi and NordSecMob which are developed by the OtaSizzle project, AaltoLunch

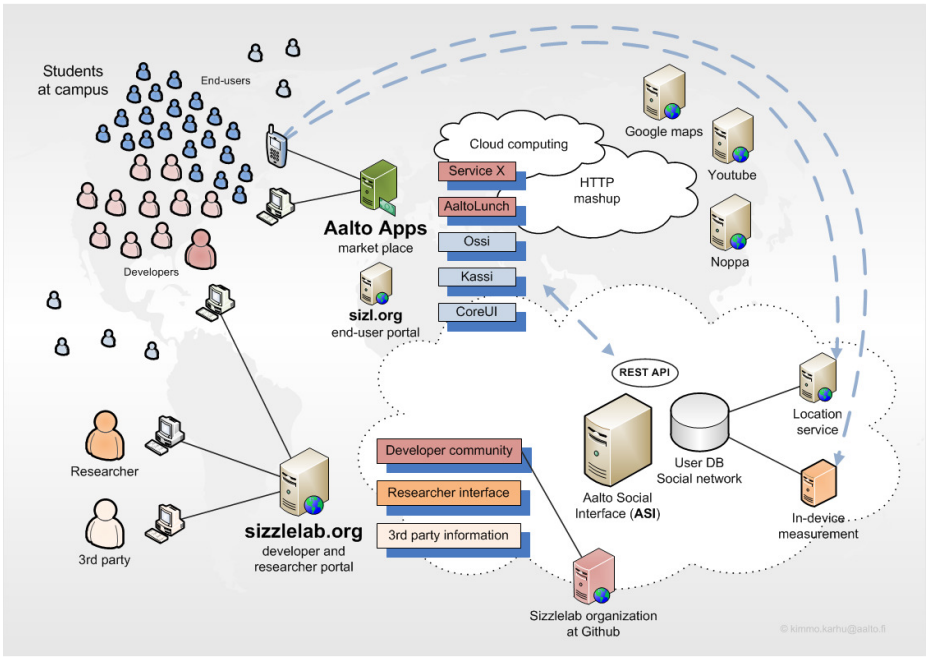



Figure 2. OtaSizzle architecture (adapted from [22])

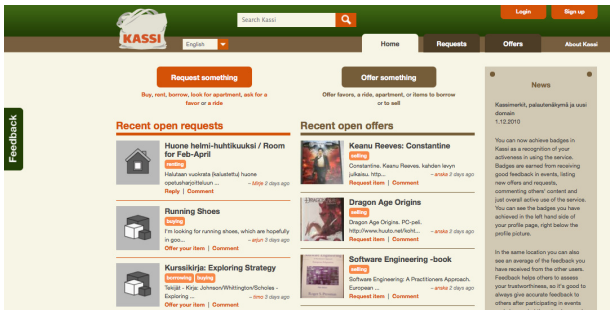



Figure 4. Kassi social resource exchange web service

(http://www.aaltolunch.fi) is the first end user service innovated by students in Aalto University. It's a mobile service providing daily menus of student restaurants on all three Aalto University campuses, in which users can also share their lunch plans with their friends.

The aforementioned four end-user services have been all developed by OtaSizzle project and student teams in Aalto University. An example of external service that has been linked to OtaSizzle environment is the Mobile Hubi (http://m.hubi.fi/mobi/). It is the first external service (developed by VTT, the Technical Research Center of Finland) to join OtaSizzle via the ASI. It provides context-aware services such as public transport route guide and Helsinki area event recommender.

All core services and end user services developed by the project or student teams in Aalto University have been open-sourced under the MIT open source license. The source codes are hosted in the Github (http://github.com/sizzlelab).

It should be noted that the above services themselves are just seeds for further development. The core services and related research tools are among the key elements of OtaSizzle project. The environment and experiments are being partially replicated in China (BUPT, Beijing), in US (UCBerkeley) and in Africa (University of Nairobi) for carrying out comparative studies on the development and use of target services.

The Smart BUPT and joint UBISERVE work, described next, provides complementary services for including sensors, intelligent environment and IoT approach, forming a joint basis for the DE research environment described in this paper.

## III. Living Lab and Internet of Things—Smart BUPT Project

### A. Smart BUPT Project Background

According to ITU (International Telecommunication Union) Reports in 2005, Internet of Things (IoT) is conceptual vision of future Internet where anything can be connected anywhere at anytime by using enabling technologies such as GPS (Global Positioning System), RFID and sensors. Since the proposal of IoT concept, Chinese government has considered it of great importance for research and development and several initiatives have been launched in that area. One of the active Chinese universities in mobile and IoT research is BUPT (Beijing University of Posts and Telecommunications). Smart BUPT project, focusing on IoT research, aims at creating an open campus based innovation platform by combining IoT and Living Lab approaches to facilitate user-driven creation of useful and intelligent services related to their daily activities. In order to lower the technical threshold for users to create mobile ubiquitous services, Smart BUPT project architecture is more based on the concept

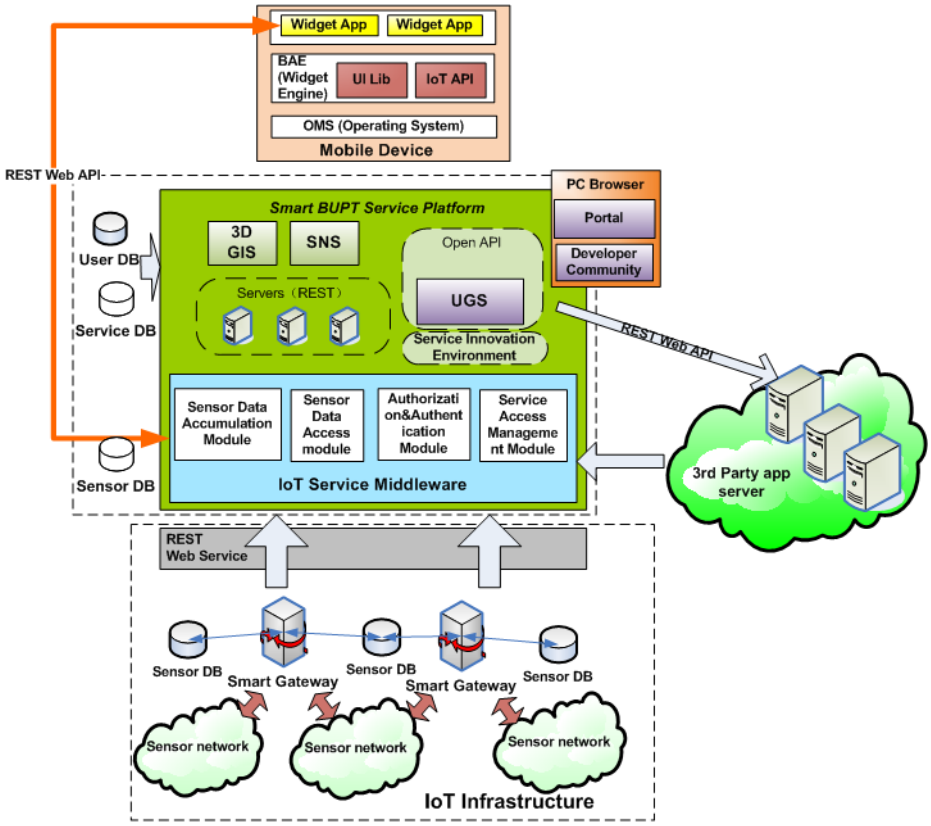



Figure 5. Smart BUPT architecture

of Web of Things (WoT). Similar to the concept of IoT, WoT is based on the vision that everyday devices and objects are connected and fully integrated to the Web by using existing well-accepted Web standards such as HTTP and REST [25]. End user services use Mobile Widget technology which is also based on popular Web technologies such as HTML, CSS and Javascript. The Smart BUPT project architecture is shown in Fig. 5.

### *B. Current Status of Smart BUPT*

Currently, there are three end-user mobile widget applications based on the sensors (temperature sensors and infrared sensors).

One service example is the temperature warning system in which students can check the current temperature and temperature history in the monitoring points such as dormitories. Automatic warning messages will be sent by the system if the temperature in the monitoring points is higher than the given thresholds. Fig. 6 shows the mobile widget end user interface of the temperature warning system.

Another service is the dressing index where students can input their demo-graphic information such as gender, age, weight and height. The system will suggest students what kind of clothes to dress at particular monitoring points such as classrooms. Fig. 7 shows the mobile widget user interface of dressing index service.

The last end user service of the current three sensor-based mobile widget services is the seat occupation rate


Figure 6. Temperature warning system

service, in which students can check the current situation of seat occupation rate (for example, whether the classroom is fully occupied or still seats available) based on the infrared sensors near the doors of monitoring points such as classrooms and library.

There is a RFID-based sub project called Smart Library under the Smart BUPT project, in which students and faculty can use their mobile phones to trace the location of a book by its RFID tag in the library. Another service example that combines location information and 3D maps is called 3D campus navigation. It is based on GPS and GIS technology developed by Terra-IT (http://www.terra-it.cn) company.

Again, the aforementioned services are meant as seeds


Figure 7. Dressing index service user interface

and examples, and the key area of the activity is to develop the infrastructures and environment for long-term research in real-life settings with support for situationally and locally relevant services as in the case of the OtaSizzle counterpart but with special emphasis on enabling use of sensor-based information sources and IoT approach.

## IV. UBISERVE: A Future Ubiquitous Innovation Platform

### *A. Motivations for the combination of OtaSizzle and Smart BUPT*

According to Yoo, people's everyday experiences can be conceptualized as the interactions with four dimensions: time, space, actors and artifacts as shown in Fig. 8 [19].

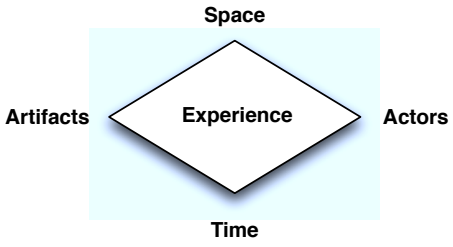



Figure 8. Schematic Framework of Experiential Computing (from [19])

From the perspective of experiential computing, experiential computing is enabled by the mediation of all or part of the dimensions of the aforementioned four dimensions of human experiences through digital technology [19]. For example, according to Yoo, the digitization of physical artifacts can be realized by RFID, sensors and IoT. The digitization of actors has been accomplished partly by the proliferation of social networking sites and social media [19].

From the perspective of DE, Nachira et al. believe that DE is made possible by the convergence of three networks: ICT networks, social networks and knowledge networks [2].

Based on these perspectives, we believe that the combination of the OtaSizzle (focusing on social media) and Smart BUPT (focusing on IoT) is important for both experiential computing research and DE research. The complementary relationship between the projects is shown in Fig. 9. From Fig. 9, we can see that current OtaSizzle project focuses on the combination of Living Lab and social media (the network of people), while Smart BUPT project focuses on the combination of Living Lab and Internet of things (the network of things). The combination of these three parts is the ubiquitous Living Lab service platform—UBISERVE.

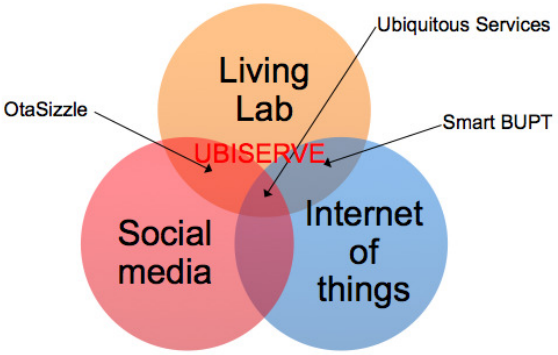



Figure 9. Complementary relationship between the projects

The UBISERVE project (Research on Future Ubiquitous Services and Applications) is "a joint research effort funded by Finland Tekes (the Finnish Funding Agency for Technology and Innovation) which is dedicated to advance research in the field of Future Ubiquitous Services (FUS). The project will strength the collaboration between Finland and China in ICT Alliance through constructing service enabling environments, developing test environments for FUS in real-life settings. The activities include living-labs based research on ubiquitous innovation and constructive research on transmission algorithms and service overlay architectures" [26].

The similarities between OtaSizzle project and Smart BUPT project are as follows:

- Living Lab approach
  - Both are based on Living Lab idea and mobile services platform.
  - Both are first deployed in campus environment.
- Technical similarities
  - Their architectures are similar (layered and modular). There are core services libraries and different end user services.
  - The core services libraries such as OtaSizzle ASI and Smart BUPT sensor libraries are both written by Ruby on Rails.
  - The calls between end user services and core services is by RESTful APIs.
  - Both have location-based services.
- Ecosystem thinking
  - Both are open platforms which provide open APIs to third-party developers.
  - Both are supported by partnering with third-party companies.

Based on the similarities and complementary relationship between OtaSizzle and Smart BUPT, we propose a new campus-based ubiquitous Living Lab innovation platform, which will be described in the next subsection. The main purposes and motivations of the platform are as follows: a) creating of a campus-based environment for creating and studying locally and situationally relevant mobile social media services; b) making use of information sources related to the local environment (sensors, e.g. temperature), context (e.g. location), users own input (e.g. observations/comments) and local information services (e.g. various campus based services) ; c) supporting local service innovation by combining social media and sensor-based services; d) providing a platform for better understanding social networks, user behaviors, and system interactions across different cultures and with likely very different underlying infrastructure and social contexts.

### *B. Architecture of the Ubiquitous Living Lab Innovation Platform*

The architecture of the ubiquitous Living Lab innovation platform is shown in Fig. 10. It is the combination of Living Lab environment and DE architecture. Therefore, we divide the architecture into two parts: Living Lab environment part and DE architecture part. In the Living Lab environment part, we focus on the actors such as end users and developers and their roles in the ecosystem. In the DE architecture part, we focus on the digital species and technical architecture that combines the social media and sensor-based services from original OtaSizzle and Smart BUPT architectures. The architecture is described in three blocks. The leftmost block is developers block. The rightmost block is the researchers block. The middle block is the DE architecture. Above these three blocks, The topmost part is the end users.

*1) Living Lab Environment:* There are different types of actors or players in the Living Lab environment. They are self-organizing and related to each other and maintain the ecosystem collaboratively. Specifically, the actors in the ecosystem are as follows.

*a) End users:* The main end users are students in campus. But this group of actors also include faculty and staff in university as well as the local citizens. End users are not only services consumers and testers but can also act as services co-creators. Their needs and requirements in their daily activities and their experiences are the sources for new services and innovations.

*b) Developers:* End user services developers — They develop all kinds of end user services based on the core and non-core services. They can be project developers, students and third parties.

Core services developers — Core service developers can be project developers such as ASI developers and third parties such as OpenNetMap. They provide DE foundation services for end user services developers.

Third-party services providers — Third-party services providers have different roles in the architecture. Some third-party companies are project sponsors. For example, Nokia company sponsors a batch of N97 mobile phones for OtaSizzle. Elisa company sponsors free bandwidth for OtaSizzle. External entities and companies can be seen as providers of both core services, such as OpenNetMap, and non-core and end user services, such as Facebook and Mobile Hubi. Some companies provide services for researchers, such as Zokem company (http://www.zokem.com/) who provides handset-based data collection and measurement for OtaSizzle user research.

*c) Researchers:* Researchers carry out constructive and empirical research such as service design and implementation, and studies on service usage, adoption and diffusion, carrying out measurements and user behavior analysis in different cultural contexts. Developers support researchers by the development of research tools and facilities such as Ressi. Ressi is a Web-based research tool for researchers to view and download research data in the databases and visualize user activities. On the other hand, such research helps the development of better services for the DE based on research findings and insights.

*2) Digital Ecosystem Architecture:* The middle block is the DE architecture block. In order to illustrate the combination of OtaSizzle (social media part) and Smart BUPT (IoT part), we use different colors to represent components from OtaSizzle and Smart BUPT respectively. Specifically, in the architecture, the components related to OtaSizzle social media are filled in white, while the components related to Smart BUPT IoT are filled in dark gray. If a component is filled in white with dark gray shadow, this means that both OtaSizzle and Smart BUPT architecture contain this component. Examples include the third-party services or this components based on the combination of components from both projects such as the new end user services built on top of both OtaSizzle and Smart BUPT core services.

The DE architecture block contains three sub-blocks or layers. The bottom layer is third-party services networks and sensor networks. Third-party services networks include campus services such as the online course registration service and library service and other third-party company services such as Facebook and Google Maps. The sensor networks include smart objects (sensors, RFID) and wired/wireless networks deployed in the campus areas. Different services and data sources such as sensor data can be combined to create a new service by Web mashup—a Web application that integrates services and data from multiple sources to provide a unique service [27].

The middle layer is the core services layer. The core services include the social network service such as ASI, the sensor-based services and third-party core services such as OpenNetMap service.

The top layer is the end user services layer. The end user services can be built on top of core services and third-party services. Some end user services per se also provide RESTful APIs such as the Kassi service and can

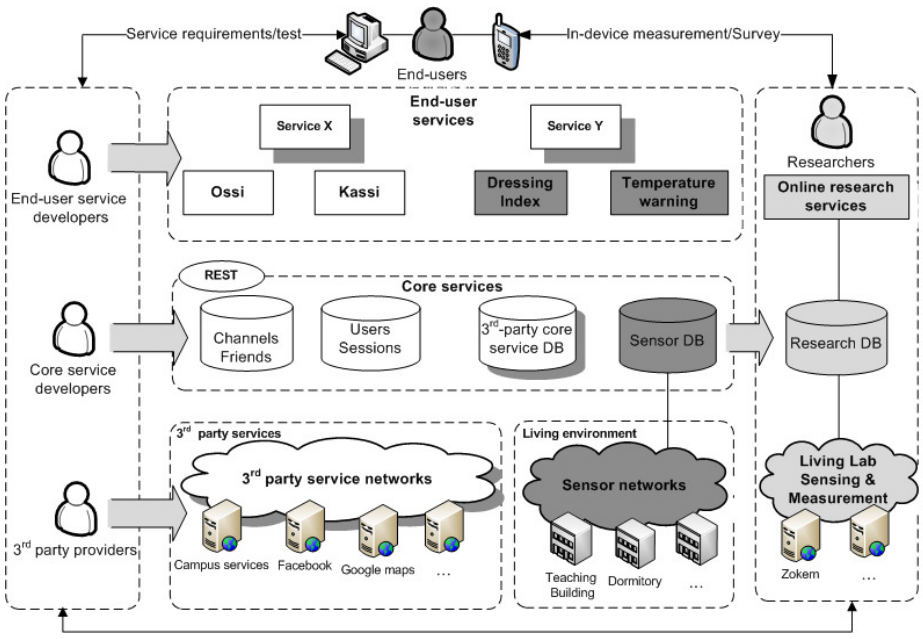



Figure 10. Architecture of the ubiquitous campus Living Lab innovation platform

be the sources for further mashups. In the future, we will also have end user services built on top of OtaSizzle and Smart BUPT core services.

## V. Discussion

### A. What we have learned in the international project collaboration?

Internationally distributed project collaboration is usually challenging, especially in the different cultural contexts like Finland and China. The collaboration between OtaSizzle and Smart BUPT has progressed quite well. For example, the OtaSizzle ASI and Kassi services have been deployed and adapted at BUPT. Next, we summarize the factors we believe to have contributed to this. Also, we recognize the need of future work in development joint methods and tools for comparative studies, and in gaining better understanding on how to take the different regulatory and cultural contexts into account when carrying out long term user studies enabled by linking the environments and research activities.

*a) Open source:* Nachira et al. believe that open source approach is the only possible choice for the DE infrastructure [2]. To facilitate academic research and international collaboration, the simple and flexible license type—MIT license was chosen for OtaSizzle project. MIT license is less restrictive than GPL license and thus more business friendly.

*b) Common core services technologies:* Both OtaSizzle social media core services and Smart BUPT sensor-based core services are built on the common software technologies (e.g. Ruby on Rails). Therefore, it is much easier for the integration and improvement of core services of both projects.

*c) Simple and lightweight service development and mashup technologies:* The RESTful APIs provided by OtaSizzle services, Smart BUPT services and mobile widget development technologies can lower the technical barriers for service development and mashups.

*d) Effective collaboration and communication tools:* Github has been chosen for project source code management and coordinating distributed development. For communication, Email, Skype and Flowdock are used. Regular Skype conference calls between the developers in Finland and China every two weeks are carried to check the current status and the next steps. While common in distributed software development, these settings are also beneficial for distributed research collaboration.

*e) Focus on local needs and situations:* Unlike Facebook which provides the same service globally, the purpose of OtaSizzle and Smart BUPT services is to satisfy the local and situational needs. For example, the OtaSizzle services in Nairobi University have been adapted to SMS-based services to gear to the poor mobile infrastructure there.

*f) The power of social media in marketing services:* Social media provides important channels for services marketing. For example, based on the Google Analytics that was set up for OtaSizzle services such as AaltoLunch, Facebook has quickly surpassed Google search and the aforementioned Sizl portal to be the most important channel for students to find the service.

### B. Implications to Digital Ecosystem Research

The main implications of our work to DE research are as follows: a) we combine the three emerging research concepts: DE, Living Lab and Experiential Computing by applying the DE architecture in the internationally

distributed Living Lab environment to foster local innovations in everyday life experiences; b) we integrate the two complementary aspects, i.e. mobile social media and ubiquitous (IoT) services for DE architecture in the real-life settings. The integration is particularly well suited for having campus-based environment as a research basis; c) we focus on the locality aspect, namely the locally relevant aspects that tie the services to the area and activities (e.g. campus area and activities) and to the physical environment. There is potential service innovation and even development community that can be actively involved both in the use and in the creation of new services in the "miniecosystem of campus" and the interlinking campuses; d) we are developing the basis for multi-contextual/multi-cultural DE studies by linking the Living Labs in Europe/Finland and in China/Beijing with option of others like Africa/Nairobi and US/California, which provides rich academic research opportunities for DE.

### C. Future Work

For the future work, we are integrating OtaSizzle social media core services with Smart BUPT sensor libraries. We will have joint code camps to develop some new end user services based on these core services and other services such as campus services. We are currently conducting similar and comparative user surveys in UCBerkeley, Nairobi University, Aalto University and BUPT. Later, we will have joint research experiments and comparative studies of mobile use and development. We will collect some comparable datasets for user behavior study (e.g. handset based measurement and analysis using Zokem's mobile clients, using server side logging, situational surveys, etc.).

## VI. Conclusions

In this paper, we combine three emerging research concepts: DE, Living Lab and Experiential Computing. We propose a DE architecture for ubiquitous campus Living Lab innovation platform based on the international exchange and collaboration between two Living Lab research projects in Finland and in China. The DE architecture is designed by combining two complementary aspects: social media and ubiquitous sensor-based services in real-life settings. The implications of our work to DE research have been discussed. We hope our work will inspire other researchers to join the effort and build a platform together that enables truly collaborative DE research.

## Acknowledgment

This work has been partly supported by the OtaSizzle research project that is funded by Aalto University's MIDE program. It has also been partly supported by the UBISERVE project funded by Tekes, the Finnish Funding Agency for Technology and Innovation, and the Smart BUPT project. The authors acknowledge the colleagues who have contributed to the development of OtaSizzle and Smart BUPT environments described in this work.

## References

[1] F. Nachira, "Towards a network of digital business ecosystems fostering the local development," 2002.

[2] F. Nachira, P. Dini, and A. Nicolai, "A network of digital business ecosystems for europe: roots, processes and perspectives," *European Commission, Bruxelles, Introductory Paper*, 2007.

[3] G. Briscoe and P. De Wilde, "Digital ecosystems: Evolving service-orientated architectures," in *Bio-Inspired Models of Network, Information and Computing Systems, 2006. 1st*. IEEE, 2007, pp. 1–6.

[4] E. Chang and M. West, "Digital ecosystems a next generation of the collaborative environment," in *the Eight International Conference on Information Integration and Web-Based Applications & Services*, vol. 214, 2006, pp. 3–23.

[5] K. Karhu, A. Botero, S. Vihavainen, T. Tang, and M. Hämäläinen, "A digital ecosystem for boosting user-driven service business," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. ACM, 2009, pp. 246–253.

[6] A. Lawson, P. Eklund, P. Goodall, T. Wray, V. Daniel, and M. Van Olffen, "Designing a digital ecosystem for the new museum environment: the virtual museum of the pacific." Social Innovation Network (SInet), University of Wollongong, 2009, p. 227.

[7] G. Briscoe and A. Marinos, "Digital ecosystems in the clouds: towards community cloud computing," in *Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on*. IEEE, 2009, pp. 103–108.

[8] P. Innocenti, S. Ross, E. Maceciuvite, T. Wilson, J. Ludwig, and W. Pempe, "Assessing digital preservation frameworks: the approach of the shaman project," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. ACM, 2009, pp. 412–416.

[9] A. Mostefaoui and B. Piranda, "Multimedia sensor networks: an approach based on 3d real-time reconstruction," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. ACM, 2009, pp. 188–195.

[10] Y. Zatout, E. Campo, and J. Llibre, "Toward hybrid wsn architectures for monitoring people at home," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. ACM, 2009, pp. 308–314.

[11] J. Liu and M. Roantree, "Precomputing queries for personal health sensor environments," in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. ACM, 2009, pp. 49–56.

[12] M. Eriksson, V. Niitamo, and S. Kulkki, "State-of-the-art in utilizing living labs approach to user-centric ict innovation–a european approach," *development*, 2005.

[13] OpenLivingLabs, "Living labs," Cited Feb 12 2011. [Online]. Available: http://www.openlivinglabs.eu/livinglabs

[14] H. Schaffers, M. Cordoba, P. Hongisto, T. Kallai, C. Merz, and J. Van Rensburg, "Exploring business models for open innovation in rural living labs," 2007.

[15] M. de Leon, M. Eriksson, S. Balasubramaniam, and W. Donnelly, "Creating a distributed mobile networking testbed environment-through the living labs approach," in *Testbeds and Research Infrastructures for the Development of Networks and Communities, 2006. TRIDENTCOM 2006. 2nd International Conference on*. IEEE, 2006, pp. 5–139.

[16] V. Niitamo, S. Kulkki, M. Eriksson, and K. Hribernik, "State-of-the-art and good practice in the field of living labs," in *Proceedings of the 12th International Conference on Concurrent Enterprising: Innovative Products and Services through Collaborative Networks. Italy: Milan*, 2006, pp. 26–28.

[17] R. Jain, "Experiential computing," *Communications of the ACM*, vol. 46, no. 7, pp. 48–55, 2003.
[18] R. Singh and R. Jain, "From information-centric to experiential environments," *Interactive Computation*, pp. 323–351, 2006.
[19] Y. Yoo, "Computing in everyday life: A call for research on experiential computing," *MIS Quarterly*, vol. 34, no. 2, pp. 213–231, 2010.
[20] J. Rana, J. Kristiansson, J. Hallberg, and K. Synnes, "Challenges for mobile social networking applications," *Communications Infrastructure. Systems and Applications in Europe*, pp. 275–285, 2009.
[21] OtaSizzle, "Ubiquitous social media for urban communities," Cited 10 Feb 2011. [Online]. Available: http://mide.tkk.fi/en/OtaSizzle
[22] M. Mäntylä, M. Hämäläinen, K. Karhu, L. A., L. K., N. E., O. A. P. O., S. R., S. E., T. J., T. S., and V. A., "Sizzlelab: Building an experimentation platform for mobile social interaction," ACM, Ed. Bonn, Germany: MobileHCI'09, September15-18 2009.
[23] R. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, Citeseer, 2000.
[24] SizzleLab, "What is otasizzle?" Cited 10 Feb 2011. [Online]. Available: http://www.sizzlelab.org/content/what-otasizzle
[25] V. Stirbu, "Towards a restful plug and play experience in the web of things," in *The IEEE International Conference on Semantic Computing*. IEEE, 2008, pp. 512–517.
[26] UBISERVE, "Introduction," Cited 10 Feb 2011. [Online]. Available: http://jimup.cs.hut.fi/ubiserve/
[27] J. Yu, B. Benatallah, F. Casati, and F. Daniel, "Understanding mashup development," *IEEE Internet Computing*, pp. 44–52, 2008.

**Tingan Tang** holds MSc in engineering from the Guangdong University of Technology, China. He is a researcher and PhD candidate in the Software Business and Engineering Institute at the Aalto University School of Science, Finland.

**Zhenyu Wu** holds MSc in engineering from School of Information and Communication Engineering of Beijing University of Posts and Telecommunications. He is a researcher and PhD candidate in the Mobile Life & New Media Lab, Beijing University of Posts and Telecommunications, China.

**Kimmo Karhu** holds MSc in computer science from the Helsinki University of Technology. He is a researcher and PhD candidate in the Software Business and Engineering Institute at the Aalto University School of Science, Finland.

**Matti Hämäläinen** holds PhD from the Department of Management Science and Information Systems, University of Texas at Austin. He is a Professor in the Software Business and Engineering Institute at the Aalto University School of Science, Finland.

**Yang Ji** holds PhD from the School of Telecommunications Engineering, Beijing University of Posts and Telecommunications. He is a Professor in Mobile Life & New Media Lab at the Beijing University of Posts and Telecommunications, China.

# 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)*

REGULAR PAPERS



RISING SCHOLAR PAPERS