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/

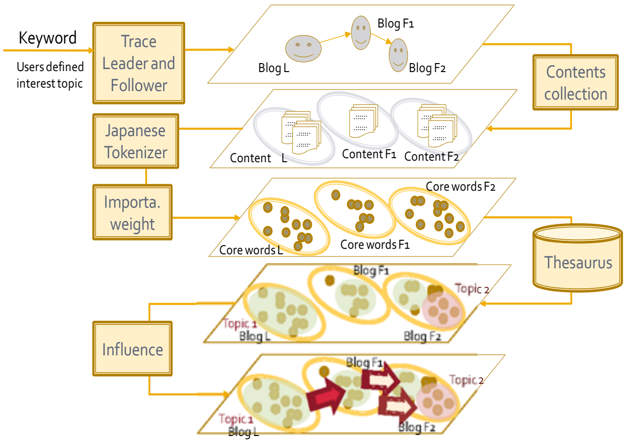


Figure 2. Data processes

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.

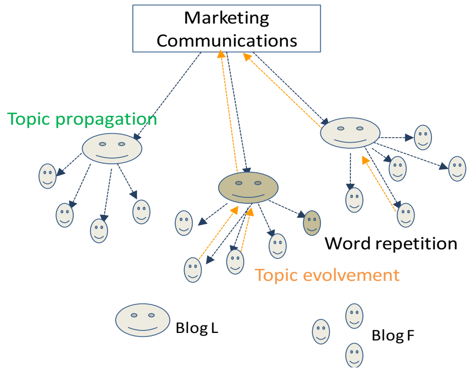


Figure 3. Patterns of propagations based on content information

We use an example to illustrate how topic propagation and topic evolvement interpret influences in viral marketing in Fig. 4.

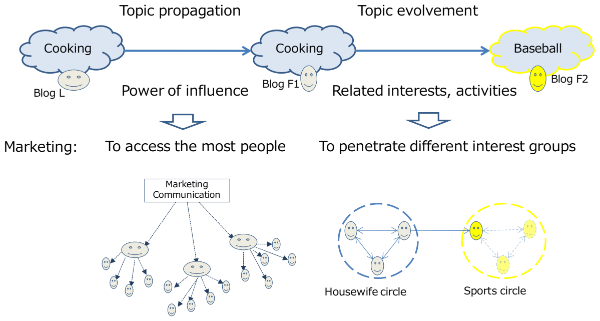


Figure 4. Topic propagation and evolvement for marketing purpose

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.

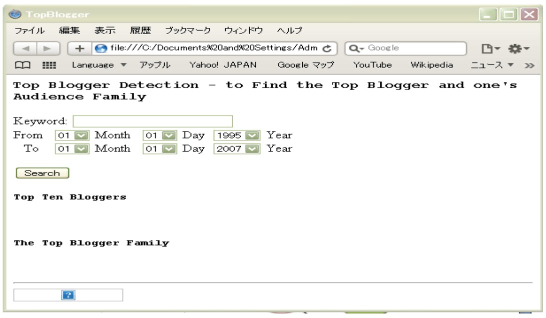
IV. METHODS

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.



Top Blogger Detection – to Find the Top Blogger and one's Audience Family

Keyword:
From 01 Month 01 Day 1995 Year
To 01 Month 01 Day 2007 Year

Search

Top Ten Bloggers

The Top Blogger Family

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.

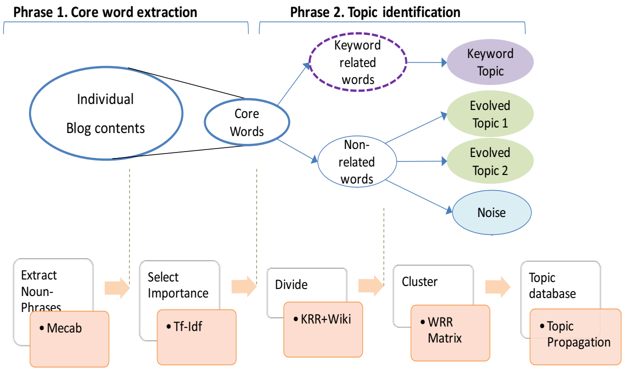


Figure 6. Framework of identify topics of each blogger

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

$$\text{word relatedness} = \cos\left(W_i\,, W_j\right) = \frac{W_i \cdot W_j}{\|W_i\|\|W_j\|}$$
$$= \frac{\sum_{\mathrm{k}=1}^{\mathrm{n}} W_{i_\mathrm{k}} \times W_{j_\mathrm{k}}}{\sqrt{\sum_{\mathrm{k}=1}^{\mathrm{n}}\left(W_{i_\mathrm{k}}\right)^2} \times \sqrt{\sum_{\mathrm{k}=1}^{\mathrm{n}}\left(W_{j_\mathrm{k}}\right)^2}}$$

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 $\mathrm{w_i}$ and word $\mathrm{w_j}$ 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:

$$wrr(w_i, w_j) = \frac{pr(w_i, w_j)}{\sqrt{pr(w_i)\,pr(w_j)}}$$
$$i, j = 1, 2, \dots, \mathrm{n};\ \mathrm{n} = \text{set of words}$$

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:

$$krr(w_i, k) = \frac{pr(w_i, k)}{\sqrt{pr(w_i)pr(k)}}$$

$$i, j = 1, 2, \ldots, \text{n}; \text{n} = \text{set of words}$$

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.

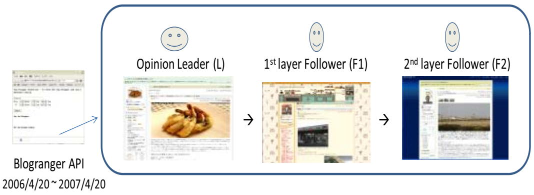


Figure 7. Identify the opinion leader in cooking area.

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.

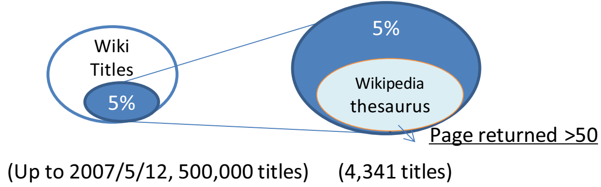


Figure 9. The Wikipedia thesaurus

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.

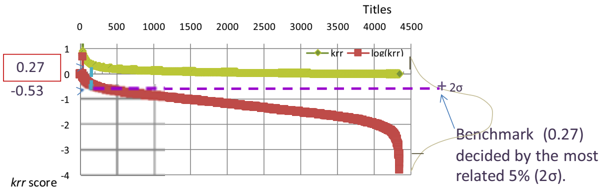


Figure 10. Distribution of the *krr* and log (*krr*) index

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.

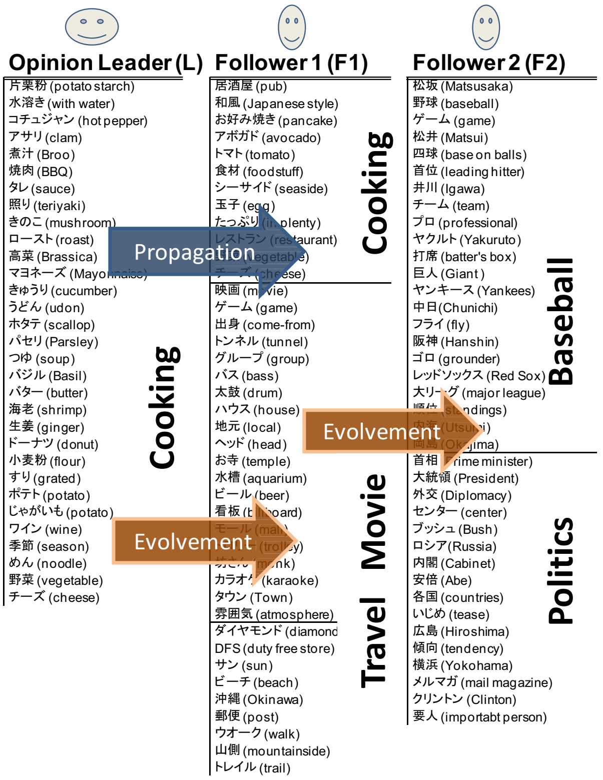


Figure 11. Propagations from the leader to follower bloggers

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.