E-TRIBALIZED MARKETING: THE INFORMATION PROPAGATION AND EVOLVEMENT AMONG VIRTUAL COMMUNITIES

Meichieh Chen[-], Hirohiko Suwa[-], Ko Fujimura[—], Toshizumi Ohta[-]
Graduate School of Information Systems, University of Electro-Communications
1-5-1 Choufushi, Choufugaoka, Tokyo 182-8585, Japan
NTT Cyber Solutions Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan

ABSTRACT

There is not the slightest doubt that in the web information space the electronic tribes, organized and differentiated by individual interests, show more active and discerning in consumption behaviors. The study of influences among virtual communities we focus on based on two key insights – (a) Marshall McLuhan's electric media theory of tribal society; (b) various business-to-consumer technologies, namely viral marketing, are widely applied into blogosphere for the purpose of acquiring and retaining a large customer base. Through the application program interface of Blogranger, a blog search engine provided by NTT, (Nippon Telegraph and Telephone Corporation), the opinion leaders among all Japanese blog users are identified. Our research involves developing computational models for tracing the dependency of text information by analyzing the information propagation and evolvement of a specific virtual community composed of opinion leaders and their followers. The dynamics of those consumer-generated media will be collected as a real live information database for the future research of database-driven relationship marketing or the enlargement of market shares.

Keywords: blog, virtual community, information propagation, viral marketing.

1. INTRODUCTION

According to Merriam-Webster Collegiate Dictionary, “Blog” is the top word of 2004 and collected as a new entry in the 2005 version dictionary. The State of the Blogosphere April 2007 from Technorati Weblog describes that Technorati is now tracking over 70 million weblogs, and about 120,000 new weblogs being created worldwide each day, 1.5 million posts per day. In terms of blog posts by language, Japanese is the first blogging language at 37% followed by English at 36%, Chinese at 8%, and Italian at 3% [1]. A growing number of blogs tells that the blog age is coming overwhelmingly with lots of potential business opportunities driven from researching in blog user behaviors. To analyze blog behaviors, more and more web portal sites provide blog related searching services, which directs online readers’ click decision making and categorizes communities’ characteristics in some way. Moreover, cooperated with users, application program interfaces provided by various portal sites accelerate the influences on blog reading behaviors.

Blog, which provides support for the issues bloggers deem interesting and important. Along with the development of internet and increasing prevalence and convenience of web-related activities, social network occurring in virtual communities, which

spontaneously transfers individuals' opinions, interests, and desires, is now the hot materials of emerging marketing research. Through analyzing individuals' blogs, researchers tag or categorize into their communities and build up the database or predictive model for the purpose of lifestyle intruding marketing.

More than three decades ago, Marshall McLuhan declared that the new "software" environment—radio, TV, phones, computers—has replaced the old "hardware" environment of books and rules and roads and railways. In the process, people have become "tribalized" and thrust together into new worldwide intimacy (see, e.g. McLuhan 1970). Similarly, a very much greater speed-up, the previous "new software environment" detribalize. The electric technology retribalizes existing social components, especially in market.

The internet or web blog is changing the dominant logic of marketing. Viral marketing exploits existing social networks by encouraging customers to share product information with their friends [2]. The concept of viral marketing grabs marketers' attentions because consumers want to decide for themselves and show increasing resistance to traditional forms of advertising such as TV or newspaper ads. A typical example on political process is the post-Howard Dean effect. Howard Dean's internet strategies alter recent politicians not only believe in advertising campaign but also exploit blog as a new media tool. Based on this logic digital marketing agencies provide the matching service for intruding advertising. Figure 1 shows the arising community around the primary blogger.

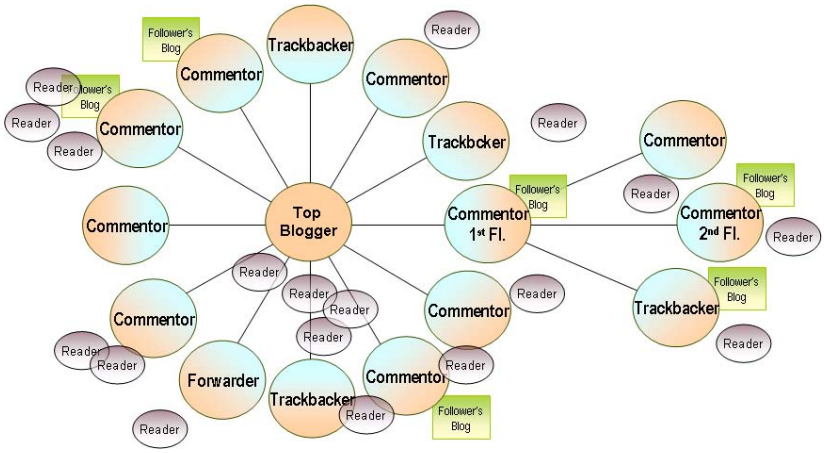

Figure 1. Arising community around top blogger

2. RESEARCH QUESTIONS

However, the influences of virtual individuals and their outputs are not yet verified. While an increasing number social network studies [2] on web blog have shown that the impact of internet on mapping the blogosphere or conduct blogger surveys on their demographic background and motivation on blogging, the influential relationships between bloggers in virtual communities are still in the grey zone. The existing navigation tools are either not able to provide the service solutions regarding tracing the information dependency in virtual communities or too general searched results to

satisfy user's needs.

For the purpose of pursuing the dynamics of bloggers' interaction in the aspect of social economics in the future, we would like to research the characteristics of information propagation in the virtual communities, circle interest tribes from electric diaries based real time database information, discover the linkage of distinguished interest topics.

The first step this paper is to develop a navigation tool which is able to explain what is the key factor by which individuals drive each other in the web space, how does a virtual community to be formed, how to measure these influential factors, and how to make a prediction of information propagation among virtual communities, especially for the purpose of marketing research.

3. DATA AND METHOD

The research purpose of this study firstly is to build up comprehension of the dynamic dependency between virtual individuals; secondly, to give a measurement for scaling the influential relationship; finally, to create a new navigation system, also a visualization mode, for virtual communities' searching more efficiently and the better usage of electric marketing and research.

The first half of this study aims to answer the text mining problems of identifying keywords from blog corpus and grouping them into various conceptual topics. The following figure illustrates the research structure, divided into five different analytical modules.

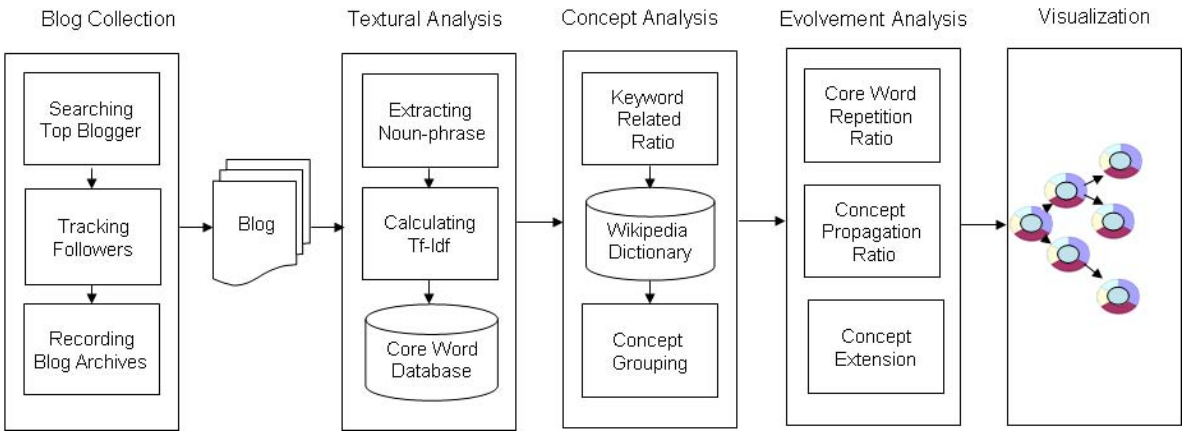

Figure 2. Research structure

3.1 Data Collection

Through the application program interface of Blogranger which is the blog search engine developed by NTT, the opinion leader among all Japanese bloggers is identified. For anyone giving a comment, trackback, or forward action on leader's post is defined as leader's follower. Both leader and follower are the actors in the

blogosphere. Once the positions of leader and follower are fixed, we collect their blog articles from the date they posted over two-week period by using our data collection program. Since most blogs are written as personal diaries, analyzing personal blog content could be an effective way to peek into one's interests, life style, and thoughts. The reason we use two weeks as data collection period is concerning the effective of topic transformation over time.

3.2 Textural Analysis

To process texture analysis, we adopt Mecab, a Japanese morphological analyzer, to extract noun-phrases from each personal blog corpus. Tf-idf (term frequency/inverse document frequency) model is used to weight the importance of each word entry. Given the distributions of importance values (as Figure 4.), word entries with the value of the biggest sub-distribution or over are dismissed. That is, only the most important 15% to 20% noun-phrases will be selected as the core words.

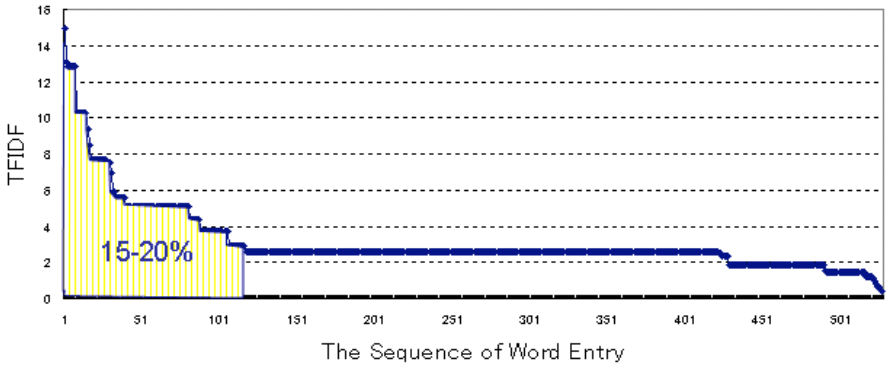

Figure 3. The distribution of important value

3.3 Concept Analysis

The goal of this module is to design a automated system which is able to firstly select any word entry closed to the default keyword, secondly group the filtered-out word entries into distinguishing concepts.

A dictionary based index indicator named Keyword Related Ratio (KRR) is designed for weighting the closeness of default keyword with selected word entries. The number of page returned through web search engine is used as quantitative measurement. In this study, the Goo search engine, which is also developed by NTT, is exploited. So, in this research KRR is a returned ratio that compares the number of page returned by Goo search engine limited in searching data inside Goo domain only where both the word entry and the keyword occur, versus the number of page returned where the word entry occurs. KRR requires two measurements. They are:

- $pr(w_i, k_j)$: The number of page returned only where both the word entry and the keyword occur
- $pr(w_i)$: The number of page returned where the word entry occurs

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

Considering web blog as the subject of this study, we use Wikipedia's list of page titles, given the essence of consumer generated information (CGI) as well. The Wikipedia Dictionary is designed with the same logic. We downloaded Wikipedia's list of page titles up to May 12th, 2007, which includes about 500,000 words and alphabetically selected 5% of all page titles by equal sampling interval method. Any word with the number of page returned by Goo search engine less than 50 is deemed as noise, which takes a big part in the list. Eventually, 4,341 words were filtered and given their KRR values. By observing the keyword spectrum of Wikipedia and the normalized KRR values, we decided the allowance boundary, by which a word with higher KRR value is considered as a keyword related word and vice versa.

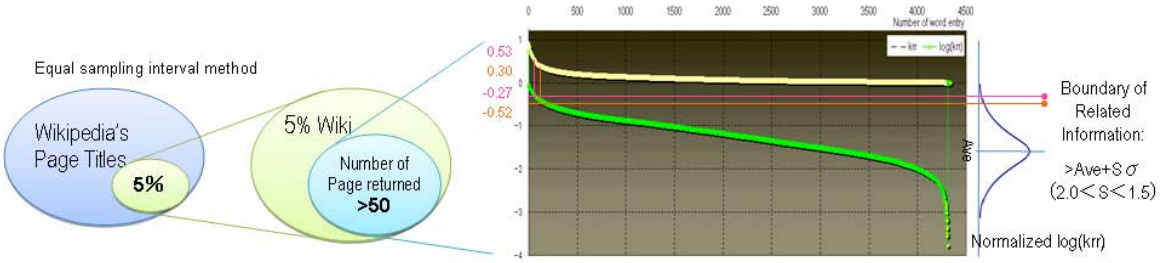

Figure 4. The representative of the Wikipedia dictionary and the KRR boundary decided

After filtering individual's core words related to the concept of the inputted keyword, we started grouping the rest into different concepts. To process concept grouping, we created a concept matrix, composed by KRR values of the rest core words, show the closeness of each other. The concept matrix is a symmetric matrix with the diagonal value 1.

$$A = A^T = \begin{pmatrix} krr(w_1, w_1) & krr(w_1, w_2) & \cdots & krr(w_1, w_n) \\ krr(w_2, w_1) & krr(w_2, w_2) & \cdots & krr(w_2, w_n) \\ \vdots & \vdots & \ddots & \vdots \\ krr(w_n, w_1) & krr(w_n, w_2) & \vdots & krr(w_n, w_n) \end{pmatrix}$$

n : The number of word entries

The concept matrix is also to be seen as a proximity matrix, given the fact that KRR values represent the distances between each word entry. Given a partition of a proximity matrix of similarities into clusters, then the average similarity values within each gives a measure of the extent to which the groups form clusters [3]. In order to maximize the similarity value or minimize the distance measure, matrix permutation is adopted to optimize this proximity matrix. We limited the clustering into several groups. Comparing the density and sum up correlation values or each group, generously one or two group stranded up; the other groups are seen as the gathering of non-influential pieces.

3.4 Evolvement Analysis

By the stage of evolvement analysis, the results of three ratios- core word repetition ratio (CWRR), concept propagation ratio (CPR), and concept extension (CE) explain the difference information dependency among bloggers. CWRR shows the direct information propagation. Opposite to CWRR, CPR said about the experience of indirect propagation. That is, for ones whose interests strongly relate to some specific keyword, are not sharing the same information with others. The final, CER, shows how topics relate to topics.

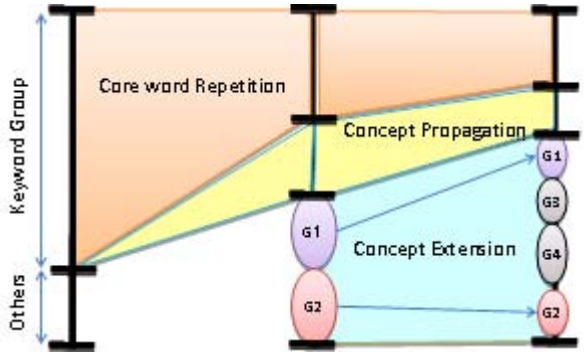

Figure 5. The three dimensions of evolvement analysis

3.5 Visualization

Through network visualization, we have designed a new navigation system for researching information propagation and evolvement in order to make searching more efficient and accurate.

4. RESULT

In this section we use “cooking” as a keyword to demonstrate the result of three case studies. They are the top blogger of “cooking” topic and his/her two followers. Table 1 shows the analysis result. After noises removed, in Top blogger’s core word list, all words are related to “cooking” topic. In Follower_c6’s core word list, 22% of words are related to “cooking”, and 17% of the “cooking” related subgroup is the same as top blogger’s. However, there is none in Follower_c19’s list.

Through optimizing Follower_c6 and Follower_c19’s concept matrixes, distinguished grouping results are found. For Follower_c6, except “cooking” topic, “movie”, “game”, and “arena” topics also play on. Follower_c19 is not interested in”cooking” at all, but enthusiastic about “baseball” and “politics”.

Table 1. The propagation result of three case studies

	TopBlogger	Follower_c6	Follower_c19
CWRR	100%	17%	0
CPR	100%	22%	0
CE		G1(_____……) G2(_____……)	G1_____… …_ G2_____ ……_

5. CONCLUSION AND FUTURE WORK

The area of data mining, or text mining, focus on the study of all the articles expressed by a representative phrase of noun extracted from corpus. Recently, text mining technology combined with network analysis is expecting to provide more comprehensible and explainable results by clustering those fragmental noun phrases. While an increasing number social network studies on web blog have shown that the impact of internet on mapping the blogosphere or conduct blogger surveys on their demographic background and motivation on blogging, here we draw a connection between the measurements of information dependencies of bloggers' online social network.

This research is to give a micro view on the dynamics of information propagation in the virtual communities. By a relatively small amount of data inputted, the analyzed result successfully excludes noises and show more satisfying fit with original blog corpus. The established criteria, using existing web searching results, such as KRR and usage of Wikipedia dictionary, to analyze web archive show more accurate in the information filtering stage.

This paper shows a preliminary study on information propagation and evolvement among virtual communities and aims for designing of an automated system to trace information propagation and evolvement, either for the usage of blog searching or market research on consumer behaviors. The succeeding research will focus on system evaluation and data validation. Also, we would like to have more statistics combing network analysis in the both micro and macro aspect.

6. REFERENCES

[1] Technorati Weblog: http://technorati.com/weblog/
[2] Hurst, M., blog site- data mining: text mining, visualization and social media http://datamining.typepad.com/gallery/blog-map-gallery.html
[3] Borgatti, S.P., Everett, M.G. and Freeman, L.C.2002.Ucinet 6. Harvard: Analytic Technologies.
[2] Adar, E., Zang L, Adamic, L.A., Lukose, R.M., Implicit Structure and the Dynamics of Blogspace, Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, May, 2004.