Finding Forerunner Bloggers by Temporal Analysis of Blogger Posts

Shinsuke NAKAJIMA† Adam JATOWT‡ Inagaki YOICHI▪ Reyn NAKAMOTO▪
Jianwei ZHANG† and Katsumi TANAKA‡

†Kyoto Sangyo University Motoyama,Kamigamo,Kita-Ku,Kyoto-City, Japan
‡Department of Social Informatics, Kyoto University Yoshida Honmachi, Sakyo-ku, Kyoto, Japan
▪ Kizasi Company 24-1, Hakozakicho, Nihonbashi, Chuuo-ku, Tokyo, Japan

E-mail: †{nakajima, zjw}@cc.kyoto-su.ac.jp ‡{adam, tanaka}@dl.kuis.kyoto-u.ac.jp
▪ {inagaki, reyn}@kizasi.jp

Abstract Evaluating credibility of online information has recently started to become popular research topic. As it is often difficult for users to manually estimate the level of trustworthiness of encountered content, providing automatic support for assessment of credibility levels of online information should improve user satisfaction. In this paper, we approach the problem of credibility evaluation from the author viewpoint. In particular, we propose analyzing temporal characteristics of posting behavior of bloggers. Our hypothesis is that precursor bloggers who persistently post content ahead of the community are often credible authors. The topics discussed by such bloggers frequently become common topics inside the community of bloggers. We propose categorization of blogging behavior depending on the time delay and content similarity between published posts and demonstrate several methods for calculating blogger scores.

Keyword Information Credibility, Detecting Important Bloggers, Blogger Community

1. Introduction

With the advent of Web 2.0 personal blogs have become popular means for expressing ones opinions and for communicating with other people. Bloggers who play influential roles in their communities constitute trustworthy source of information. The posts of such bloggers should be ranked higher by blog search engines.

Current measures of blogger expertise seem to be based on textual analysis of their contributions to the community or on analysing link structures within blogosphere. However, such measures do not seem to take the time factor into consideration. Therefore, the information about time at which a blogger contributes his posts is generally not used to sufficient degree, even though it may provide valuable evidence regarding blogger's leadership and activity level. For example, a blogger who is consistently ahead of a community in terms of new topics (i.e. he or she writes about the topics that later become popular within the community) should be regarded as a kind of a community leader who sets or predicts the topics of subsequent discussions. By comparing the synchronization level of the blogger's posts with common community topics we attempt to measure the general capability of bloggers to shape community opinions and influence common interests.

In this paper we discuss several methods for characterizing and detecting bloggers within communities according to their temporal posting patters. In general, two blogger categorizations can be proposed. The first one is based only on the posting pattern of individual bloggers, while the second one captures the relation between blogger posts and community topics over time. According to the former one bloggers are categorized as frequent and non-frequent contributors depending on their posting frequency and as regular and irregular depending on the periodicity levels of their posts.

The second categorization, which is the focus of this paper, divides bloggers into:

- forerunners
- repeaters
- late-followers

The forerunners category contains bloggers who write about topics that are later widely discussed in the community. Either they are first in noticing and reporting some events or they have enough expertise and popularity to introduce novel topics into the blogosphere that later become common to the community. In whichever case, this kind of bloggers constitutes important part of the community information chain. Our main objective is to propose means for automatically detecting such bloggers within their communities. The repeaters are bloggers who

repeat the current topics within the community and, in general, actively participate in ongoing discussions. In some sense, they are "digesting" and propagating topics that are at given moment "hot" in the community. In contrast to forerunners they do not have strong opinion formulating power and they do not influence or "foresee" the future direction of the community. Finally, the late-followers are bloggers who post content on topics that were popular within the community in the past. Either they are late in the sense of catching up with popular topics within the community (e.g., as a result of being unable to write posts) or they simply become interested in old, previously popular content.

The above categorization is only a crude one and denotes rather extreme types of posting behavior. Many bloggers may only partially belong to a given category or may belong to more than one category. Moreover, a separate category of bloggers could be proposed embracing bloggers who always publish content on the same or similar topics or whose contributions are quite different from current, future and past community interest even though they are initially classified as community members. Such bloggers can be excluded from the analysis.

Our proposed measures for categorization are independent on the connectivity between bloggers, the number and frequency of their connections and so on. Our method could be actually considered as a complement to other approaches for finding influential bloggers.

The remainder of the paper is structured as follows. In the next section we briefly describe the related work. In Section 3 we discuss the way to calculate bloggers scores and to categorize them into different contributor classes. In section 4 we demonstrate results of preliminary experiments. We conclude the paper in the last section.

2. Related Work

Kumar et al. [4] studied evolution and dynamics of blog communities while Gruhl et al. [2] investigated the diffusion of information through blogosphere. Nakajima et al. [6] proposed detecting influential bloggers in blog-based discussions by analyzing the linking patterns in blogosphere and by considering the increase in the number of posts. Juffinger et al. [3] demonstrated method for calculating the coverage of news in blogs and the level of their synchronization as a measure of blog credibility. Our approach is unique in that we compare content and relative time points of user posts and community topics to categorize bloggers into different categories according to the roles they play in the community.

In information quality theory, accuracy, authority, objectivity, currency and coverage of information are the most frequently used evaluative criteria [4]. For example, the objectivity involves determining whether web content represents facts or opinions, while currency is the measure of how up-to-date the content is. Our work concerns the last two criteria, currency and coverage, in the sense that it provides means for detecting up-to-date sources that cover popular topics within communities.

Despite the presence of various educational guidelines and tools for evaluating credibility of online information, many users are unprepared and do not possess sufficient skills to properly assess the quality of online information [5]. Consequently, few of them perform rigorous evaluation of the quality of online information. Therefore, automatic tools for supporting users in the judgment of web content quality are becoming increasingly necessary. In our hypothesis, bloggers who are forerunners are credible sources of information. While the opinions and contributions of such bloggers may not be always understood or highly evaluated by the community at the beginning, often later they become widely adopted and discussed.

3. Methodology

In this section we describe several methods for calculating the level of synchronization of blogger posts and community topics.

In order to score bloggers according to their synchronization patterns with community topics the following preprocessing steps need to be done:

1. Detection of topical community of bloggers
2. Detection of community topics at equally spaced time units (e.g., every month) and their representation in vector space model
3. Representation of blogger posts as a sequence of vectors in the vector space model
4. Calculation of similarities between the sequence of blogger posts and the series of community topics

The first step, the blogger community detection, can be performed by using any existing algorithms [1,7] and is outside of the scope of this paper.

Detection of community topics over time can be performed in a variety of ways using standard methods. For example, in order to select such topics one could employ burst detection on the feature space derived from

the content of blogger posts agglomerated within each time unit. Keyphrase extraction algorithms could be used here too. In a straightforward solution which we have adopted in our implementation, the community topics are found by calculating top co-occurring words to queries that represent the community. That is, for a given community (e.g. iphone users) the term specific to this community (e.g. "iphone") is issued into a blog search engine with a temporal constraint of time unit t_i. Then the top co-occurring terms to this term are selected as the community topics representative for t_i. Co-occurrence is computed using the standard Jaccard coefficient. Community topics are represented in a vector space with weights depending on their co-occurrence values with query. Thus community topics C_i are given for points $t_{1 \ldots} t_N$ where $t_i \in T$ where T is a fixed time period of analysis (e.g. 1 year).

In order to represent blogger posts we use tf-idf weighting scheme. Suppose that there is a sequence of posts P_j of a given blogger which were created at time points $t_1 \ldots t_M$ where $t_j \in T$. Note that we neglect bloggers whose posting behaviour is infrequent. Threshold of minimum posts by a blogger during T is set to 50 posts per year. Thus only bloggers with the number of posts above this threshold are considered. Note that in contrast to community topics blogger posts are usually not equally distributed in time dimension,

Given the community topics and blogger posts, we calculate similarity value denoted as $sim(P_j,C_i)$ between each post and each community topic using cosine similarity. Figure 1 shows the concept of calculating similarities between pairs of community topics and blogger posts. Although community topics are calculated periodically in this representation, our approach can perform well also in case of unequally distributed samples of community topics over time.

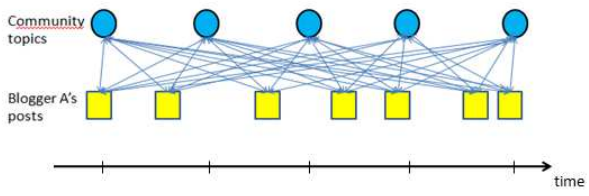

Figure 1 Community topics and blogger posts over time.

Below, we describe different measures for scoring bloggers.

Linear Scoring for Finding Forerunner Bloggers based on Time Analysis

The first method aims at detecting forerunner bloggers. It calculates blogger score as a weighted average of the similarity values between his or her posts and community topics:

$$S = \frac{1}{\sum_{j=1}^{j=M}\sum_{i=1}^{i=N} \varpi(i,j)} \sum_{j=1}^{j=M}\sum_{i=1}^{i=N} \left[\varpi(i,j) * sim\left(P_j, C_i\right)\right] \qquad (1)$$

$\omega(i,j)$ is a temporal weight assigned to each cosine similarity measure $sim(P_j,C_i)$. Here, $\omega(i,j)$ is defined as follows:

$$\varpi(i,j) = \frac{1}{2} * \frac{t_i - t_j}{|T|} + 0.5 \qquad (2)$$

where $|T|$ is the size of time period T.

The temporal weight $\omega(i,j)$ is a linear function with respect to the length of the time interval between a given post and given community topic. Figure 2 shows this function on the graph where the horizontal axis represents the time difference $t_i - t_j$. Note that $\omega(i,j)$ equals 0.5 when $t_i = t_j$, that is when the timestamp of the blogger's post t_j and the time point of the community topic t_i are same. Also, $\omega(i,j)$=1 when $t_i - t_j = |T|$ and $\omega(i,j)$=0 when $t_i - t_j = -|T|$. Thus blogger has the higher score the more his or her posts are similar to the community topics that appear at later time than the timestamps of the posts.

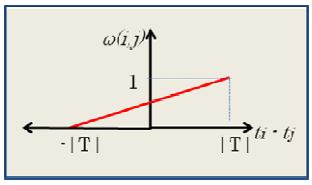

Figure 2 Temporal weight used in Equation 1.

Relative Scoring for Finding Forerunner Bloggers based on "Past-future Comparison"

This scoring method rewards correct future predictions of community topics and penalizes late copying activity. It assigns higher scores to bloggers who are forerunners in the community and lower scores to bloggers who are repeaters. The difference with the previously described method is that this method penalizes past repetitions. That is, bloggers who repeat the topics that were long time ago discussed in the community have diminished scores.

$$S = \frac{1}{M * N} \sum_{j=1}^{j=M} \sum_{i=1}^{i=N} \left[\frac{\varpi(i,j) * sim^{F}(P_j, C_i) + 1}{\varpi(i,j) * sim^{P}(P_j, C_i) + 1} \right] \quad (3)$$

$$\begin{aligned} sim^{F}(P_j, C_i) &= sim(P_j, C_i) \text{ when } t_i > t_j \\ sim^{F}(P_j, C_i) &= 0 \text{ when } t_i \le t_j \\ sim^{P}(P_j, C_i) &= sim(P_j, C_i) \text{ when } t_i < t_j \\ sim^{P}(P_j, C_i) &= 0 \text{ when } t_i \ge t_j \end{aligned} \quad (4)$$

$$\begin{aligned} \varpi(i,j) &= \frac{t_i - t_j}{|T|} \text{ when } t_i > t_j \\ \varpi(i,j) &= -\frac{t_i - t_j}{|T|} \text{ when } t_i \le t_j \end{aligned} \quad (5)$$

The temporal weight, $\omega(i,j)$, increases along with the distance from the "current" time point (see Figure 3).

Note that this scoring will not work well in case when blogger posts have both very low future-related similarity, $sim^{F}(P_j,C_i)$ and very low past-related similarity, $sim^{P}(P_j,C_i)$. To remedy this problem threshold levels for both future-related and past-related similarities should be used.

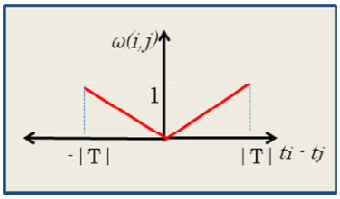

Figure 3 Temporal weight used in Equation 3.

Scoring for Finding Repeaters

Repeating behavior may be actually considered as a measure of activity of bloggers and in certain cases finding such bloggers may be useful. For example, if one wishes to be up-to-date with popular community topics, he or she can pay attention to the posts contributed by such bloggers. The function described below assigns top scores for repeater type bloggers.

$$S = \frac{1}{\sum_{j=1}^{j=M} \sum_{i=1}^{i=N} \varpi(i,j)} \sum_{j=1}^{j=M} \sum_{i=1}^{i=N} \left[\varpi(i,j) * sim(P_j, C_i) \right] \quad (6)$$

$$\begin{aligned} \varpi(i,j) &= 0 \text{ when } t_i \le t_j - x \text{ or } t_i > t_j \\ \varpi(i,j) &= 1 \text{ when } t_i > t_j - x \text{ and } t_i \le t_j \end{aligned} \quad (7)$$

Figure 4 shows the function representing temporal weight $\omega(i,j)$. Bloggers who report recent topics (less than x months with timeline of monthly granularity) that are common within the community will have high scores assigned. Active bloggers that are uptodate with community topics should be then detected. Note that by increasing the value of x relaxes the condition of repeating behaviour by allowing longer delays in topic reporting.

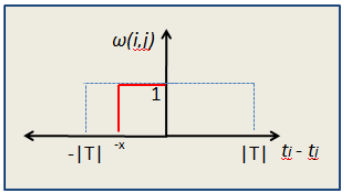

Figure 4 Temporal weight used in Equation 7.

Scoring for Finding Late-followers

Lastly we show the method for finding late-follower type of bloggers. Such bloggers are characterized by contributing posts that are similar to old or obsolete topics which are no longer popular within their communities. The calculation method is similar to the one for finding repeaters. Equation 6 is are also used for this approach in the same form as before, however Equation 7 is modified as follows:

$$\begin{aligned} \varpi(i,j) &= 0 \text{ when } t_i > t_j - x \\ \varpi(i,j) &= 1 \text{ when } t_i \le t_j - x \end{aligned} \quad (8)$$

Figure 5 shows the function assigned to temporal weight $\omega(i,j)$ in this case.

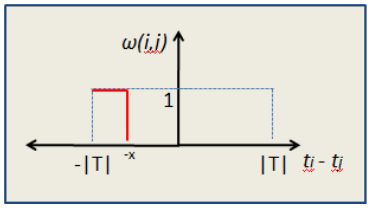

Figure 5 Temporal weight used in Equation 8.

4. Experiments

In this section we show the results of preliminary experiments. The experiments were done in January 2011. We have selected 7 topics and collected top 400 related terms for each month during 2 years period for each topic. The terms were selected according to the method mentioned in Section 3, that is, by calculating co-occurrence values for manually chosen topic-specific search keywords using at least 1000 blog entries. We have used here Kizasi company blog search engine[1].

We then calculated similarities of blogger posts to the

[1] http://www.kizasi.jp

derived community topics. For each topic we show the average similarities of selected 9-10 bloggers in Figures 6-12. The bloggers were selected according to the method described in [7]. This method finds knowledgeable bloggers based on how often and how in-depth they write about certain topics. In short, if a blogger extensively uses topic-specific keywords then his or her score will be computed high.

The vertical axes in the Figures 6-12 represent the average similarities while the horizontal ones represent the time difference between blogger posts and community topics. For example, if a blogger post and community topic has the same time points ($t_i = t_j$) then the corresponding similarity value will be assigned to the point 0. If the community topic was calculated 1 year before the post, then their similarity will be assigned to point -12. The graphs show average similarities on monthly granularity since bloggers could write multiple posts in a month.

Looking at Figure 6 we can see that on average the top bloggers seem to constitute representative examples of forerunners type. The curves mildly decrease from the right to the left hand side. Especially blogger 9 frequently posted about topics that later became popular in the community, while he rarely discussed old topics.

Figure 7 shows another example of technology related topics. Most of knowledgeable bloggers who posted about "MacBook Air" did it before the topic become popular within their communities. These topics are relatively easy to clearly distinguish representative forerunner type bloggers as they represent technology objects that were expected long time before their official releases.

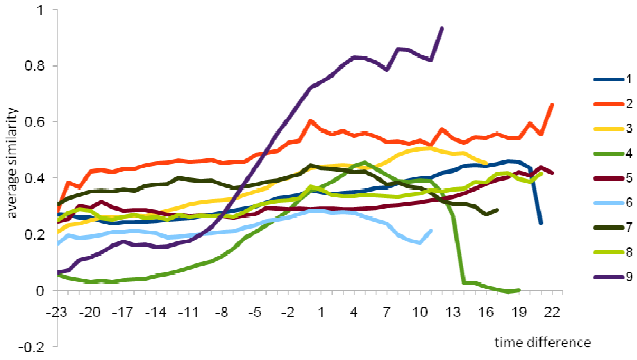

Figure 6 Average similarities for "iphone".

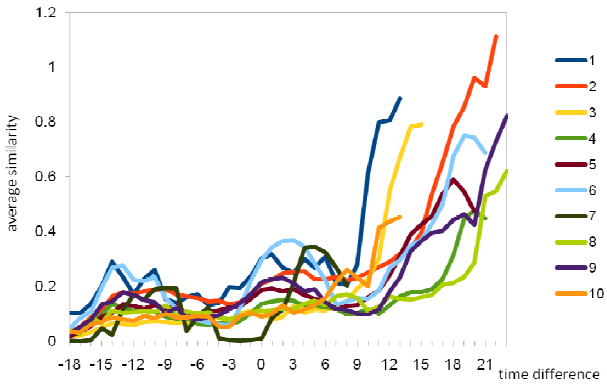

Figure 7 Average similarities for "macbook air".

Figures 8, 9 and 10 demonstrate examples of knowledgeable bloggers for "politics", "futenma" and "obama", respectively. Futenma is a shortcut for Marine Corps Air Station Futenma located in Okinawa, Japan. The military base stirs much controversy among locals and has become a hot political topic lately. We can observe the similarity plots are different than the ones in Figures 6 and 7. There are no clear examples of forerunner type bloggers. Rather almost all of the bloggers match more the repeater type. Especially for the topic "futenma" we can see that some of the top-scored bloggers correctly predicted the subsequent hot topics within the community. It may be because "futenma" is more precise topic than "politics".

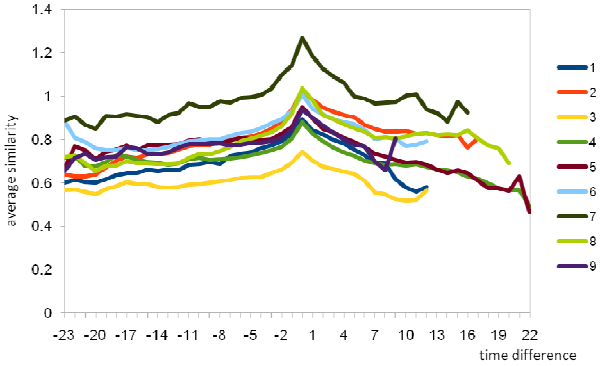

Figure 8 Average similarities for "politics".

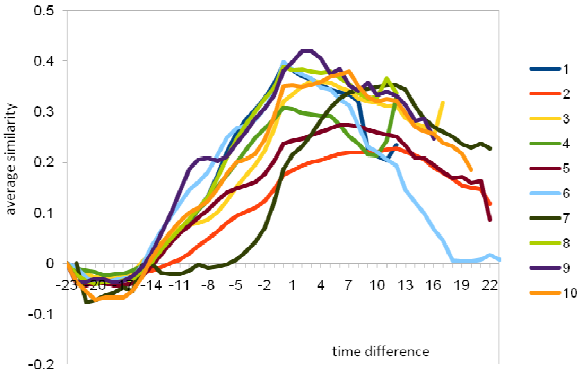

Figure 9 Average similarities for "futenma".

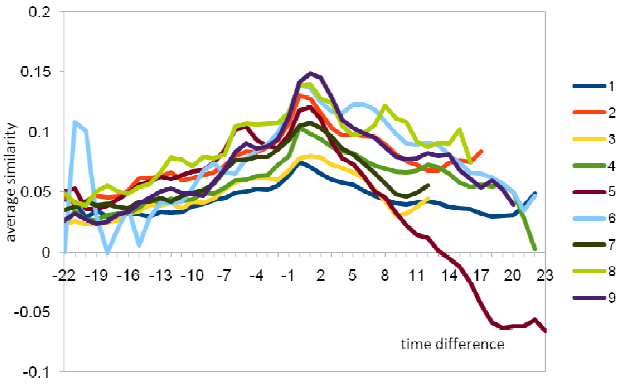

Figure 10 Average similarities for "obama".

We have tried also ambiguous topics such as "mother" (see Figure 11). As it is very difficult to predict future events and hot topics for this kind of community, most of the knowledgeable bloggers are of late-repeaters type.

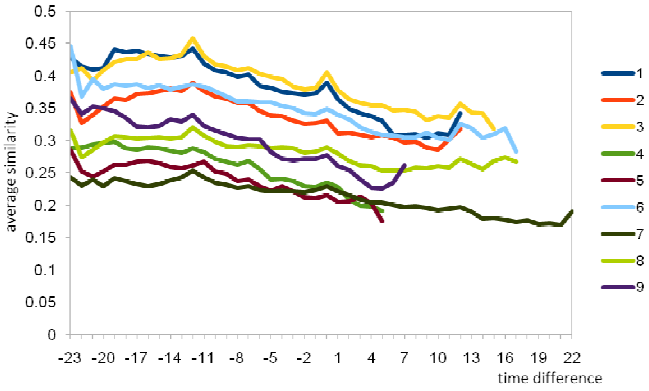

Figure 11 Average similarities for "mother".

Lastly we show an example of periodical topics. Figure 12 demonstrates results obtained for topic "hanshin". Hanshin is associated with several meanings in Japan, although the most common one is the name of popular baseball team, Hanshin Tigers. The selected bloggers for this community topic have periodical pattern type with the period equal to 1 year following the periodical nature of sports events. This particular example demonstrates potential weakness of our method, that is, poor ability to differentiate between new and old topics in case of periodically occurring topics. We are going to approach this problem in the future work.

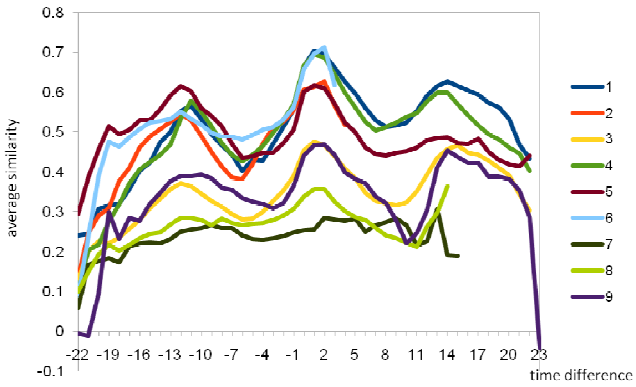

Figure 12 Average similarities for "hanshin".

5. Conclusions

In this paper we have discussed different categories of bloggers from the viewpoint of their temporal posting patters in relation to community topics. We argued that bloggers that consistently post ahead of the community in terms of community topics are the ones that set tone of discussions in the community. We have demonstrated several methods for distinguishing such bloggers by calculating the synchronization levels between their contributions and the community evolving themes. Lastly, we demonstrated examples of characteristic blogger types.

Acknowledgments

This research was supported the National Institute of Information and Communication Technology, Japan, the MEXT Grant-in-Aid for Young Scientists B (#22700096) and by the Microsoft IJARC CORE6 Project, "Mining and Searching the web for Future-related Information".

References

[1] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 150–160, (2000).

[2] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, "Information Diffusion Through Blogspace", Proc. of the 13th World Wide Web Conference (2004).

[3] A. Juffinger, M. Granitzer and E. Lex. "Blog credibility ranking by exploiting verified content", Proc. of the 3rd Workshop on Information Credibility on the Web (WICOW 2009), 51-58, (2009).

[4] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, "On the Bursty Evolution of Blogspace", Proc. of the 12th

World Wide Web Conference (2003).

[5] M.J. Metzger. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *JASIST*, 58(13), 2078-2091, (2007).

[6] S. Nakajima, J. Tatemura, Y. Hino, Y. Hara and K. Tanaka, “Discovering Important Bloggers based on Analyzing Blog Threads”, Proc. of WWW 2005 the 2nd Annual Workshop on the Weblogging Ecosystem (2005).

[7] S. Nakajima, J. Zhang, Y. Inagaki, T. Kusano, R. Y. Nakamoto, Blog Ranking Based on Bloggers' Knowledge Level for Providing Credible Information. Proc. Of WISE 2009, 227-234, (2009).