International Journal of Web Information Systems

Prophetic blogger identification based on buzzword prediction ability

Jianwei Zhang Seiya Tomonaga Shinsuke Nakajima Yoichi Inagaki Reyn Nakamoto

Prophetic blogger identification based on buzzword prediction ability

Jianwei Zhang
Faculty of Industrial Technology, Tsukuba University of Technology, Tsukuba, Japan

Seiya Tomonaga
Drecom Co., Ltd, Tokyo, Japan

Shinsuke Nakajima
Faculty of Computer Science and Engineering, Kyoto Sangyo University, Kyoto, Japan, and

Yoichi Inagaki and Reyn Nakamoto
Kizasi Company, Inc, Tokyo, Japan

Received 28 March 2016
Revised 19 May 2016
Accepted 19 May 2016

Abstract

Purpose – Identifying important users from social media has recently attracted much attention in the information and knowledge management community. Although researchers have focused on users' knowledge levels on certain topics or influence degrees on other users in social networks, previous works have not studied users' prediction ability on future popularity. This paper aims to propose a novel approach to find prophetic bloggers based on their buzzword prediction ability.

Design/methodology/approach – The main approach is to conduct a time-series analysis in the blogosphere considering four factors: post earliness, content similarity, entry frequency and buzzword coverage. Our method has four steps: categorizing a blogger into knowledgeable categories, identifying past buzzwords, analyzing a buzzword's peak time content and growth period and, finally, evaluating a blogger's prediction ability on a buzzword and on a category.

Findings – Experimental results on real-world blog data consisting of 150 million entries from 11 million bloggers demonstrate that the proposed approach can find prophetic bloggers and outperforms others that do not take temporal features into account.

Originality/value – To the best of the authors' knowledge, our approach is the first successful attempt to identify prophetic bloggers. Finding prophetic bloggers can bring great values for two reasons. First, as prophetic bloggers tend to post creative and insightful information, analysis on their blog entries may help find future buzzword candidates. Second, communication with prophetic bloggers can help understand future trends, gain insight into early adopters' thoughts on new technology or even foresee things that will become popular.

Keywords Advanced web applications, Social media, Buzzword detection, Expert finding, Prophetic blogger, Time-series analysis

Paper type Research paper

This work was partially supported by JSPS KAKENHI Grant Numbers #26330351, #26870090.

International Journal of Web Information Systems
Vol. 12 No. 3, 2016
pp. 267-291
© Emerald Group Publishing Limited
1744-0084
DOI 10.1108/IJWIS-03-2016-0013

1. Introduction

Identifying important users from social media is a challenging task and could be of great value in many applications. Past research has been conducted mainly in two ways: measuring expertise levels for finding knowledgeable users and determining influence degrees for finding influential users. The former is usually based on textual content analysis, whereas the latter also makes use of link structure in social networks. However, previous works have not studied users' prediction ability on future popularity.

The blogosphere is a conductive platform for bloggers to issue posts, share ideas and exchange opinions. The data in the blogosphere are dynamic reflecting information change over time. Potential knowledgeable bloggers with prior awareness of future popular trends may exist in the blogosphere and, if identified, can provide valuable information.

In this paper, we conduct a time-series analysis on real-world blog data consisting of 150 million entries from 11 million bloggers in the past two years provided by Kizasi Company[1] and propose a novel approach (Zhang *et al.*, 2015a) to find important bloggers based on their prediction ability on buzzwords terms or phrases describing topics or events that have become well known to general population. We call the bloggers who are knowledgeable and have high prediction ability as "prophetic bloggers". We take four factors into account: post earliness, content similarity, entry frequency and buzzword coverage. The general idea is based on the following points:

(1) The earlier a blogger posted blog entries containing a buzzword, the better prediction ability on the buzzword he or she may have.

(2) The more similar the contents of his or her past entries to the peak time content of a buzzword at its popularity peak, the more accurate is his or her prediction ability on the buzzword.

(3) The larger the quantity of early and similar blog entries containing the buzzword are, the better prophetic blogger he or she may be.

(4) The more buzzwords relative to a category he or she can predict, the better prophetic blogger on the category he or she may be.

Figure 1 shows an example that explains our idea. Here, five bloggers (blg_1-blg_5) and their blog entries containing two arbitrary buzzwords (bw_1 and bw_2) are shown in this example. The entries are laid out from the oldest to the most recent. The number of entries containing buzzword bw_1 reaches a peak at time t_8. Bloggers blg_1, blg_2 and blg_3 mention bw_1 before the peak and, thus, rate high in post earliness. Moreover, as blogger blg_1 mentions buzzword bw_1 before blogger blg_2, blg_1 has a better prediction ability in terms of post earliness than blg_2. The same holds for blg_2 vs blg_3. Bloggers blg_4 and blg_5 have low prediction ability based on post earliness, as they mention bw_1 only at or after its peak. Next, words *xx*, *yy* and *zz* are bw_1's peak time content words. For example, the peak time content words of buzzword "iPhone 6" may be "A8 chip", "4.7 inches" and "biggerthanbigger", which reflect its distinctive features when "iPhone 6" becomes popular. Words *aa*, *bb* and *cc* are the unrelated words to bw_1. blg_2's entries are more similar to the bw_1's peak time content than the ones from blg_3 and from blg_1, as blg_2 mentions more peak time content words of bw_1 than blg_3 and blg_1. Thus, blg_2 scores better than blg_3 and blg_1 in terms of content similarity. In addition, as blg_2 posts more

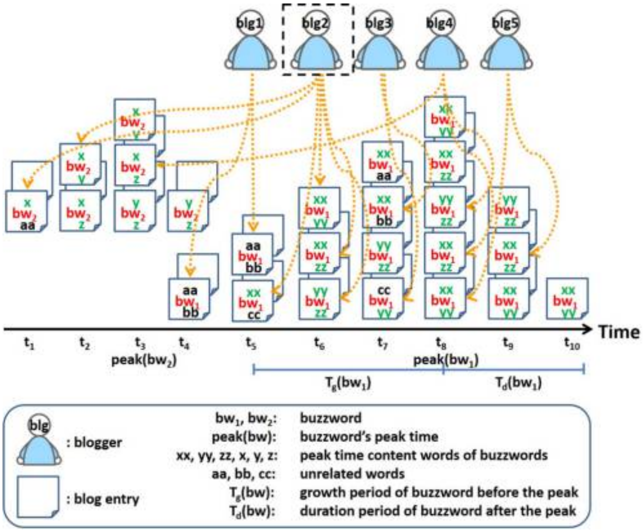
blg1
blg2
blg3
blg4
blg5
Time
t1
t2
t3
t4
t5
t6
t7
t8
t9
t10
peak(bw2)
peak(bw1)
Tg(bw1)
Td(bw1)
blg : blogger
: blog entry
bw1, bw2: buzzword
peak(bw): buzzword's peak time
xx, yy, zz, x, y, z: peak time content words of buzzwords
aa, bb, cc: unrelated words
Tg(bw): growth period of buzzword before the peak
Td(bw): duration period of buzzword after the peak

Figure 1.
Example of differing bloggers' prophetic ability

entries containing bw_1, he or she is also higher in terms of entry frequency. Based on the three measures, we can say that blg_2 has higher prediction ability on bw_1 than others. Furthermore, as blg_2 has mentioned not only buzzword bw_1 (e.g. "iPhone 6") but also buzzword bw_2 (e.g. "Galaxy S6" or "Xperia Z3") before they become popular, he or she is also good in terms of buzzword coverage and, thus, can be regarded as a prophetic blogger on the category that bw_1 and bw_2 belong to (e.g. "smartphone").

To evaluate bloggers' buzzword prediction ability, some preprocessing steps need to be done. Our contributions are summarized as follows:

- We introduce a method for categorizing a blogger into his or her appropriate potential communities called knowledgeable categories (Section 2).

  We assume that a prophetic blogger must first be a knowledgeable blogger. People have different knowledge levels on various categories. For example, a blogger who is an expert in "politics" is not necessarily knowledgeable in "IT". We extract knowledgeable categories such as "politics" and "IT" and categorize a blogger into his or her appropriate categories by calculating his or her knowledge scores related to the extracted categories. Only the bloggers knowledgeable in a category can be considered as prophetic blogger candidates for that category. On the other hand, a knowledgeable blogger is not necessarily a prophetic blogger because he or she may have no prediction ability. As shown in Figure 1, bloggers blg_4 and blg_5 are knowledgeable in the category "smartphone", but they are not prophetic bloggers, as they mention the hot topics only at or after the peak.

- We develop a method for automatically identifying past buzzwords from historical blog data based on their persistence (Section 3).

  As we want to analyze a blogger's prediction ability on buzzwords, how we extract buzzwords is an important problem. First, the top-ranked keywords in the daily topic ranking list provided by Kizasi Company are used as buzzword candidates. Then, we categorize buzzword candidates into their knowledgeable categories. Next, we evaluate whether they are appropriate as buzzwords by considering their persistence. Finally, buzzword candidates that have more entries during a certain duration period (T_d) after its peak in each category are selected as buzzwords in that category.

- We analyze a buzzword's properties by identifying its peak time content and calculating its growth period (T_g) (Section 4).

  For calculating the factor of content similarity to the peak time content, we need to first identify the peak time content words of a buzzword. The peak time content words should not only frequently co-occur with the buzzword but also reflect its distinctive features at the peak. We devise a method to extract peak time content words based on, but not only, co-occurrence analysis. The other necessary analysis is looking at a buzzword's growth period. In contrast to a buzzword's duration period after its peak, the growth period is the time before the peak, at the time when blog entries start to relate to the peak time content. For example, for considering bw_1's growth period, t_4 should be passed over, as at that period, the peak time content words of bw_1 have not yet appeared. The entries which contain "I really want to buy an iPhone 6" do not indicate the start of the growth period of buzzword "iPhone 6", because its peak time content words such as "A8 chip" and "4.7 inches" have not been mentioned.

- We integrate the necessary factors for evaluating a blogger's prediction ability on a buzzword and on a category (Section 5).

  The post-earliness of blog entries containing a buzzword during its growth period is first calculated. Then, a blogger's prediction ability on the buzzword is calculated by integrating post earliness, content similarity and the quantity of his or her blog entries. A blogger's prediction ability on a category is evaluated based on a comprehensive consideration of his or her prediction ability on the buzzwords in the category.

2. Categorization of a blogger into knowledgeable categories

We extract potential communities of bloggers called knowledgeable categories (*kc*) and automatically categorize bloggers into their appropriate *kc*s. A potential community in our research is a group of bloggers who are knowledgeable in a *kc*. For example, the "politics" community is the group of bloggers who are knowledgeable in the "politics" category. Potential communities of bloggers are objectively identified by analyzing bloggers' entries that they posted. Even if one does not declare his or her interest in a category explicitly, if he or she has posted many blog entries related to the category, our method can categorize him or her into the appropriate *kc*s automatically.

2.1 Extraction of knowledgeable categories

Each *kc* is represented by a keyword that is often mentioned in the blogosphere. This keyword becomes the name of the *kc*. They are extracted as follows:

- We perform a regular Web search by using the search keywords such as "expert in *" and "fan of *"[2].
- The retrieved keywords are then filtered by their occurrence frequencies.

- We further remove duplicate and inappropriate ones, resulting in obtaining about 14,000 keywords.
- Finally, we manually categorize the keywords into 122 categories, ending up with a list of 122 *kc* names (e.g. "politics", "economy" and "IT").

2.2 Construction of co-occurrence dictionaries

For each *kc*, a co-occurrence dictionary is automatically constructed. For each keyword representing the *kc*, we extract the top *n* words that have the highest co-occurrence degrees from all blog entries over the past two years. Specifically, *n* is 400 in our current implementation. Many methods for calculating co-occurrence degree have been proposed from the simplest co-occurrence frequency to more complicated mutual information (MI)-score (Church and Hanks, 1990) and LogLog-score (Kilgarriff and Tugwell, 2002). Based on the observation in our preliminary experiments, we adopt the LogLog-score, a compromise between co-occurrence frequency which is apt to extract ordinary words, and MI-score, which is apt to bring barely comprehensible words:

$$LogLogscore = log\frac{N_{xy} \cdot N}{N_x \cdot N_y} \cdot logN_{xy}, \tag{1}$$

where N_x and N_y are the numbers of their occurrence, N_{xy} is the co-occurrence frequency of x and y and N is the number of all words. The co-occurrence words and their co-occurrence degrees are stored in each co-occurrence dictionary for each *kc*.

Figure 2 is an example of the co-occurrence dictionary. The column *kc* shows the names of knowledgeable categories. Each row shows the domain-specific words of a *kc* and their corresponding co-occurrence degrees β. For example, "politics" is a *kc* name and has its co-occurrence words such as "Abe"[3], "premier", "party", etc.

2.3 Calculation of a blogger's knowledge score

A blogger's knowledge score for a *kc* is calculated by analyzing how often and how in-depth he or she has posted blog entries related to the *kc*. If a blogger has an extensive use of co-occurrence words of a *kc*, a high score is attached to him or her.

We first calculate $Relevance_{kc}(e_i)$ – the relevance score of a blog entry e_i for a *kc* – as follows:

$$Relevance_{kc}(e_i) = \sum_{j=1}^{n} \alpha_j \cdot \beta_j \cdot \gamma_j, \tag{2}$$

where n is the number of the co-occurrence words ($n = 400$), $\alpha_j = (n - j + 1)/n$ is the weight of the jth co-occurrence word that decreases as j increases, β_j is the LogLog-score of the jth co-occurrence word and γ_j is a binary value that indicates whether the entry e_i contains the jth co-occurrence word or not.

kc	j = 1		2		. . .		400	
politics economy IT ⋮	Abe market information ⋮	$\beta_{1,1}$ $\beta_{2,1}$ $\beta_{3,1}$ ⋮	premier currency computer ⋮	$\beta_{1,2}$ $\beta_{2,2}$ $\beta_{3,2}$ ⋮	. . . . ⋮	. . . . ⋮	party policy innovation ⋮	$\beta_{1,400}$ $\beta_{2,400}$ $\beta_{3,400}$ ⋮

Figure 2. Example of co-occurrence dictionary

We next calculate $Knowledge_{kc}(blg)$ – the knowledge score of a blogger *blg* for a *kc* – as follows:

$$Knowledge_{kc}(blg) = \frac{l}{n} \cdot \frac{log(m)}{m} \cdot \sum_{i=1}^{m} Relevance_{kc}(e_i), \tag{3}$$

where e_i is an entry that blogger *blg* posted, *m* is the number of entries that *blg* posted during the analysis period (typically the past two years), *n* is the number of the co-occurrence words, *l* is the number of the co-occurrence words that occurred in all entries posted by *blg*, *l*/*n* indicates the coverage ratio of the co-occurrence words that *blg* has used and *log*(*m*)/*m* reduces the effect when a blogger frequently posts a large amount of entries, but most of them are the entries unrelated to the *kc*.

A blogger is categorized into a *kc* if his or her knowledge score is larger than a given threshold. Moreover, a blogger may be categorized into two or more *kc*s and, thus, may have two or more knowledge scores for different categories. For example, if a blogger belongs to both "politics" and "economy", he or she has a knowledge score representing his or her expertise degree in "politics" and another one representing his or her expertise degree in "economy". Through the above process, we have a list of knowledgeable bloggers for each *kc*.

3. Identification of past buzzwords

Before evaluating a blogger's buzzword prediction ability, buzzwords need to be first detected. We identify past buzzwords by analyzing blog data over the past two years.

3.1 Selection of buzzword candidates

We start with the top-ranked keywords in the daily topic ranking list provided by Kizasi Company. Every day, the company publishes a list of top-ranked keywords for the day. These are the keywords that have the highest ratios of the number of bloggers who mentioned them in the past two days to the number of bloggers who mentioned them in the past two years. The number of bloggers rather than the number of entries is adopted to avoid the influence of a single person repeatedly posting the same keyword in many entries.

We take the top-k ($k = 100$) keywords from each day over the past two years and then exclude repeated words and periodical words. Periodical words are the ones that appear at regular intervals, such as "payday" every month, "New Year" every year and "Olympic" every four years. As these kinds of words are known well before, they indicate little of a blogger's prediction ability and should be removed. The remaining keywords become buzzword candidates.

3.2 Categorization of a buzzword candidate into knowledgeable categories

In our approach, we evaluate a blogger's prediction ability for a *kc* based on his or her prediction scores on the buzzwords that belong to the *kc*. To associate buzzword candidates (*bwc*) with *kc*s, we calculate the similarity between a *bwc* and each *kc*. Three measures – Simpson's coefficient, Jaccard's coefficient and cosine similarity – are compared in our experiments (Section 6.2):

$$SimpsonSim(bwc, kc) = \frac{|COW(bwc) \cap COW(kc)|}{min(|COW(bwc)|, |COW(kc)|)}, \quad (4)$$

$$JaccardSim(bwc, kc) = \frac{|COW(bwc) \cap COW(kc)|}{|COW(bwc) \cup COW(kc)|}, \quad (5)$$

$$CosineSim(bwc, kc) = \frac{\sum_i cow_i(bwc) \cdot cow_i(kc)}{\sqrt{\sum_i cow_i(bwc)^2}\sqrt{\sum_i cow_i(kc)^2}}. \quad (6)$$

COW(*bwc*) and *COW*(*kc*) appearing in *SimpsonSim* (Formula 4) and *JaccardSim* (Formula 5) are the co-occurrence word sets of *bwc* and *kc*, respectively. *COW*(*bwc*) consists of the top 400 words that have the highest co-occurrence degrees with the *bwc* extracted from all blog entries over the past two years, whereas *COW*(*kc*) consists of the top 400 words that have the highest co-occurrence degrees with the *kc* extracted from the blog entries posted by the bloggers in the *kc* over the past two years. It means that a *bwc* is associated with a *kc* if they share many co-occurrence words. For example, *bwc* "Abenomics"[4] and *kc* "politics" have many common co-occurrence words such as "Abe", "premier" and "party", and, thus, "Abenomics" can be categorized into "politics".

In *CosineSim* (Formula 6), $cow_i(bwc)$ and $cow_i(kc)$ are the weights of co-occurrence words of *bwc* and *kc*, respectively. Here, the co-occurrence ranks of each co-occurrence word are incorporated. More specifically, the co-occurrence words of a *bwc* are ordered and then the top *n* words are saved. In our implementation, *n* is set to 400. Each word's weight is then calculated as $cow_i(bwc) = n - i + 1$, where *i* is the rank in the top *n* words. Similarly, the co-occurrence words of *kc* are ordered, and each $cow_i(kc)$ is calculated using the co-occurrence word rank for the *kc*. The cosine is then taken by comparing the weights of each common term using cosine similarity.

Each *bwc* is categorized into the top-k ($k = 5$) *kc*s with the highest similarities. Consequently, given a *kc*, the set of similar *bwc*s can also be identified. This categorization result will be used for the subsequent process in Sections 3.3 and 5.3.

3.3 Detection of buzzwords based on their persistence

Among buzzword candidates, there are also some burst words that disappear immediately after the peak. This kind of word is not a buzzword, as it is forgotten by the public soon after the peak.

We extract influential words as buzzwords from buzzword candidates based on their persistence (Figure 3). First, for each buzzword candidate, we extract the number of blog entries over the past two years containing it and break it into a given interval (e.g. one week). Then, we smooth the variation curve by forming its moving average and confirm the popularity peak. We think the peak of a buzzword candidate is the time point of its highest public recognition. Next, a buzzword candidate's persistence is evaluated by counting the total number of blog entries containing it during a specified duration period T_d (e.g. six months) after the peak. If the number of entries containing a buzzword candidate during T_d is small, it is of low persistence. In contrast, if a buzzword candidate has a large number of entries containing it during T_d, it has high persistence. From each *kc*, we select the top-k ($k = 10$) buzzword candidates with the highest persistence as the buzzwords representing each *kc*.

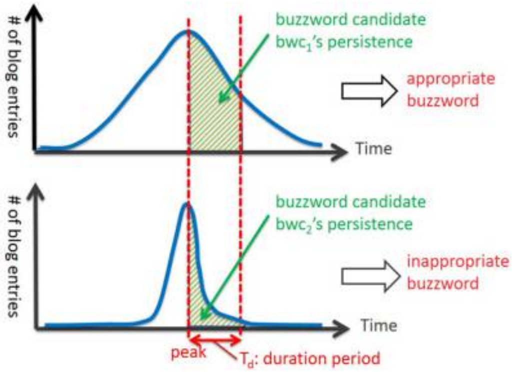

Figure 3.
Buzzword candidates' persistence

4. Analysis of past buzzwords' properties

We identify the peak time content of a buzzword represented by a set of its peak time content words and determine each buzzword's growth period by analyzing the content similarity between the content at each period (e.g. at intervals of one week) before the peak and the peak time content.

4.1 Extraction of peak time content words

Figure 4 shows how we determine the peak time content words. For a buzzword, we identify its peak time content words according to the following steps:

- We get a weekly count of the number of blog entries containing the buzzword and sort them by the highest count (e.g. fifth highest count of 104 weeks).
- We extract $n = 400$ co-occurrence words of the buzzword from the blog entries in the three weeks around the popularity peak.
- For each co-occurrence word extracted at Step 2, we get a weekly count of the number of entries containing the co-occurrence word and sort them by the highest count.
- We calculate the Spearman's rank correlation coefficient (Gregory and Dale, 2014) between the buzzword entry count order from Step 1 and the co-occurrence entry count order from Step 3. The Spearman's rank correlation coefficient is a non-parametric measure between two variables, which assesses how well the relationship between them can be described:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} D^2}{N^3 - N}, \tag{7}$$

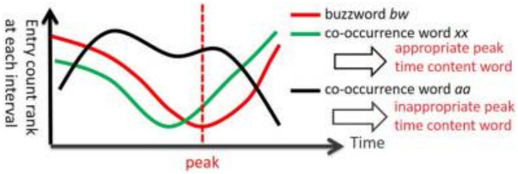

Figure 4.
Determining the peak time content words

where D is the difference of corresponding ranks between the buzzword entry count order and the co-occurrence entry count order and N is the number of pairs (i.e. the number of weeks) in the time series.

- The top-k ($k = 30$) co-occurrence words having the highest correlation coefficients are extracted as the peak time content words for the buzzword.

Rather than the co-occurrence words with the highest co-occurrence degrees, we select the co-occurrence words whose time-series variation curves of entry count orders are the most similar to those of the buzzword for representing its peak time content. In Figure 4, co-occurrence word *xx* is more appropriate as the peak time content word than *aa*, because *xx* has a much more similar variation curve with buzzword *bw*.

4.2 Calculation of growth period based on content similarity

A buzzword's growth period dates from its peak back to the time point when the contents of blog entries start to be similar to the peak time content. For example, if buzzword "iPhone 6" starts to be mentioned in unspecific entries such as "I really want to buy an iPhone 6" or "When will iPhone 6 be released?", the growth period has not begun. As it only contains ordinary words and no content words from the popularity peak are mentioned, this period is inappropriate for analyzing bloggers' prediction ability on the buzzword.

Figure 5 shows the idea of identifying the growth period. For determining the starting point of the growth period, we calculate the content similarity between each period before the peak (at intervals of one week) and the peak time. Specifically, for each period t_i before the peak, we extract the set of co-occurrence words (COW_{t_i}) from the blog entries containing the buzzword posted during each t_i and calculate its similarity with the set of peak time content words (CTW_{peak}) as follows:

$$Similarity(t_i, peak) = \frac{|COW_{t_i} \cap CTW_{peak}|}{min(|COW_{t_i}|, |CTW_{peak}|)}. \tag{8}$$

Then, we calculate the average of $Similarity(t_i, peak)$ before the peak and specify the starting point of the growth period by using the following criterion.

After accumulating the differentials between the average $Similarity(t_i, peak)$ and each interval's $Similarity(t_i, peak)$, the time point when the cumulative sum has the largest value is specified as the starting point of the growth period.

As shown in Figure 5, there are cases where the similarity curve slightly surpasses (t_1) and subsequently falls below the average (t_2). If we were to use the simple intersection of the similarity curve and the average line, the starting point would be set too early (t_1 or t_2). Instead, we adopt the accumulative sum of the differentials between the average and each interval and, thus, avoid this problem. The starting point is the time when the accumulative sum becomes the highest (t_3). Note that different buzzwords have different growth periods, and a growth period of a buzzword is analyzed on the entries posted by all bloggers, independent of any individual blogger.

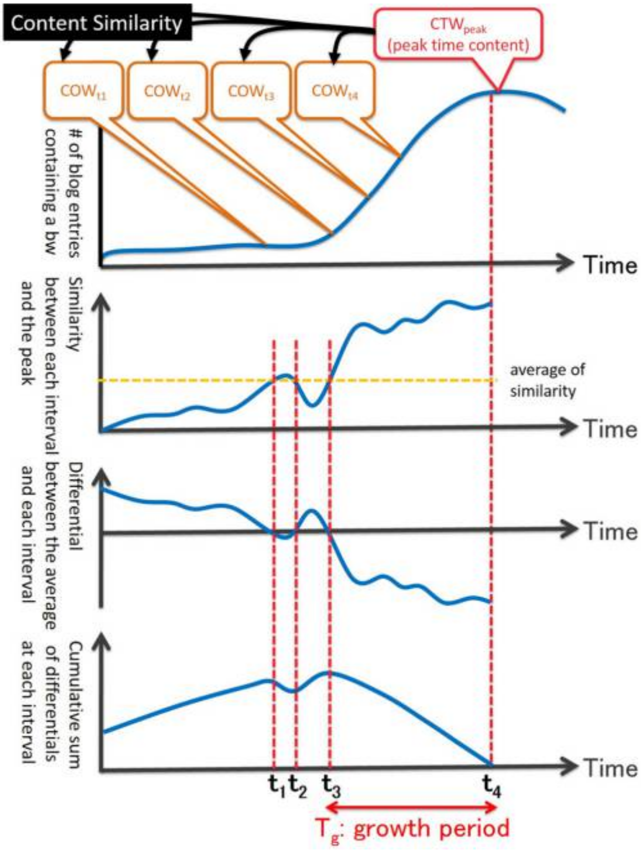
Content Similarity
CTW$_{peak}$
(peak time content)
COW$_{t1}$
COW$_{t2}$
COW$_{t3}$
COW$_{t4}$
# of blog entries containing a bw
Time
Similarity between each interval and the peak
average of similarity
Time
Differential between the average and each interval
Time
Cumulative sum of differentials at each interval
Time
t$_1$ t$_2$ t$_3$ t$_4$
T$_g$: growth period

Figure 5.
Determining a growth period

5. Identification of prophetic bloggers

A blogger's prediction score on a buzzword is calculated based on post-earliness, content similarity and the quantity of his or her blog entries containing the buzzword during its growth period. Moreover, his or her prediction ability on a category is evaluated considering his or her prediction scores on the buzzwords that belong to that category. The knowledgeable bloggers with high prediction ability on a category are identified as prophetic bloggers.

5.1 Calculation of post-earliness of blog entries for a buzzword

We assign a score of post-earliness to each entry containing the buzzword posted during its growth period. All entries containing the buzzword during its growth period are sorted according to their post-dates. An entry posted at the starting point of the growth period should receive the highest earliness score and an entry posted at the end of the growth period (i.e. the popularity peak of the buzzword) should receive the lowest earliness score. Thus, we devise two formulas for post earliness of entry e_i for buzzword *bw* as follows:

$$Earliness_{bw}(e_i) = \frac{\left|E_{T_g}\right| + 1 - order(e_i)}{\left|E_{T_g}\right|}, \tag{9}$$

$$Earliness_{bw}(e_i) = -log\frac{order(e_i)}{\left|E_{T_g}\right|}, \tag{10}$$

where E_{T_g} is the set of all entries containing buzzword *bw* during the growth period T_g and $order(e_i)$ is the appearance order of entry e_i in the set. Formula 9 distributes the earliness scores of entries more evenly during the growth period, whereas Formula 10 assigns much larger earliness scores to the entries posted at and near the starting point of the growth period. For example, if there are 100 entries containing a buzzword during its growth period, the earliness scores of entries for Formula 9 with the orders from 1 to 100 are 1, 0.99, 0.98, …, 0.03, 0.02 and 0.01, whereas the earliness scores for Formula 10 are 2, 1.698, 1.522, …, 0.008, 0.004 and 0, respectively (Figure 6).

5.2 Calculation of a blogger's prediction score on a buzzword

A blogger's prediction score on a buzzword is calculated by integrating his or her three factors: post earliness, content similarity and the number of his or her blog entries (Figure 7):

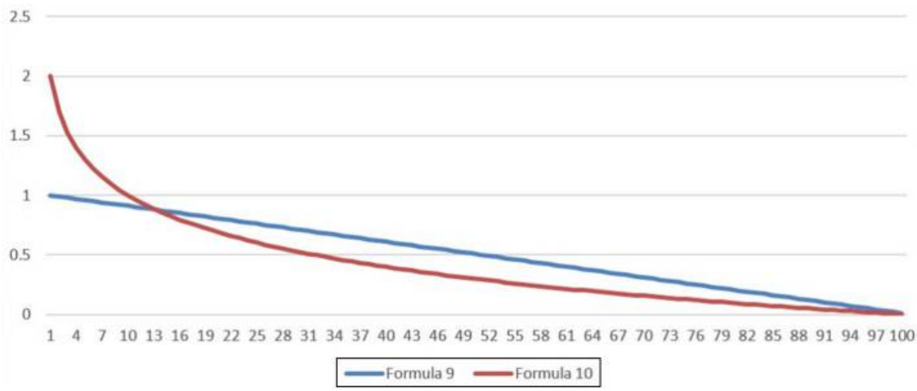

Figure 6.
Earliness scores for two formulas

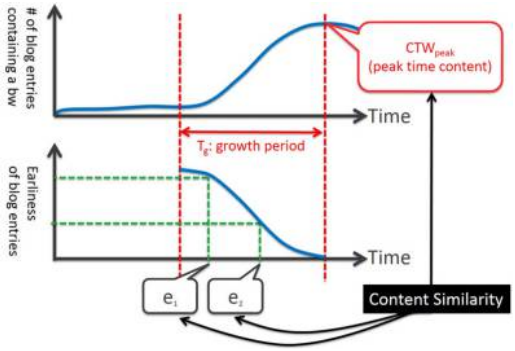
# of blog entries containing a bw
CTW$_{peak}$ (peak time content)
Time
T$_g$: growth period
Earliness of blog entries
Time
e$_1$
e$_2$
Content Similarity

Figure 7.
Calculating a blogger's prediction score on a buzzword

$$Prediction_{bw}(blg) = \sum_{i=1}^{m} Earliness_{bw}(e_i) \cdot Similarity(e_i, ptc), \tag{11}$$

where e_i is one of m entries containing buzzword *bw* that blogger *blg* posted during its growth period, $Earliness_{bw}(e_i)$ is e_i's earliness score and $Similarity(e_i, ptc)$ is its content similarity to the peak time content *ptc* of buzzword *bw*.

The content similarity between entry e_i and the peak time content *ptc* is calculated as follows:

$$Similarity(e_i, ptc) = \frac{|D(e_i) \cap CTW_{peak}|}{min(|D(e_i)|, |CTW_{peak}|)}, \tag{12}$$

where $D(e_i)$ is the set of words appearing in e_i and CTW_{peak} is the set of peak time content words of the buzzword.

If a blogger posted many blog entries containing a buzzword similar to its peak time content at the early stage of its growth period, he or she can be regarded as a good predictor on this buzzword.

5.3 Evaluation of a blogger's prediction ability on a category

As prophetic blogger candidates for a category, we first select the top-k (k = 300) knowledgeable bloggers with the highest knowledge scores in this category calculated in Section 2. Then, we find the buzzwords in this category shown in Section 3.2. Each knowledgeable blogger's prediction score on each buzzword can be calculated by the method described in Section 5.2. Consequently, for each buzzword, we can prepare a top-k (k = 5) blogger list in which the bloggers have the highest prediction scores on it. We regard the bloggers who appear in many of a category's buzzwords' top blogger lists as prophetic bloggers on that category.

We do not simply sum up a blogger's prediction scores on all buzzwords that belong to a category as the criterion for evaluating a blogger's prediction ability on the category, as we aim to distinguish a blogger who posted related entries containing each of several buzzwords before the peak from a blogger who posted the same number of related entries containing only one buzzword. In Figure 8, blg_2 is the best prophetic blogger in that category, as he or she has successfully predicted three buzzwords in that category. blg_6, blg_7 and blg_8 are the next best prophetic bloggers, as they predicted the next highest number of buzzwords after blg_2.

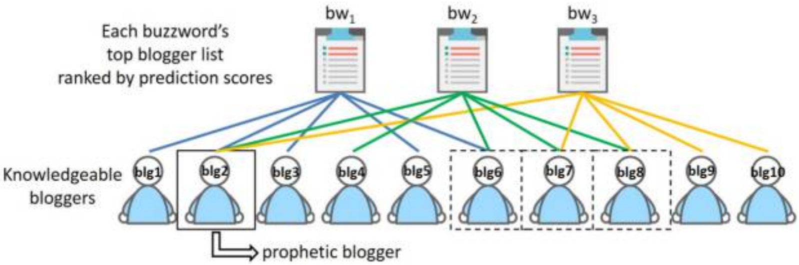

Figure 8.
Evaluating a blogger's prediction ability on a category

6. Experimental evaluation

6.1 Knowledgeable blogger identification

We randomly select 20 *kc*s (Figure 9) from the ones extracted by the method described in Section 2.1. For each *kc*, the top five knowledgeable bloggers extracted by the method in Section 2.3 are evaluated by four human evaluators. Each evaluator browses the bloggers' entries and judges whether they are appropriate as a knowledgeable blogger with respect to the *kc* in question. Accuracy is the ratio of the number of appropriate knowledgeable bloggers to all the top five bloggers. The results are shown in Figure 10. Each bar for kc_1-kc_{20} is the accuracy of appropriate knowledgeable bloggers for that *kc* averaged by four evaluators, and the bar for ave_{20} is the average accuracy of the 20 *kc*s. The average accuracy of 0.91 indicates that most of the bloggers highly ranked by our method are knowledgeable bloggers.

6.2 Buzzword categorization

We select 40 buzzwords for evaluating the categorization quality. For a buzzword, we calculate the similarity between its co-occurrence word set and the co-occurrence word set of each *kc*. Three measures of Simpson's coefficient (Formula 4), Jaccard's coefficient (Formula 5) and cosine similarity (Formula 6) are compared.

For each buzzword, the top-k ($k = 20$) *kc*s extracted by each measure are collected and shown to eight evaluators. Each evaluator selects the five most appropriate *kc*s for the buzzword and sorts them by appropriateness. Consequently, for each buzzword, a ranking list of the top five *kc*s is generated by summarizing eight evaluators' selection.

We compare the ranking list of the top five *kc*s generated by each of the three similarity measures with the human-generated ranking list using the Normalized Discounted Cumulative Gain (*nDCG*) measure (Wang *et al.*, 2013). *DCG* measures the gain of a ranking list, which is accumulated from the top of the list to the bottom with the gain of each item in the ranking list lowering as you go down the list. nDCG evaluates how closely an ideal ranking can be reproduced. *DCG* and nDCG are defined as follows:

kc_1: internal soccer	kc_2: maneuvers	kc_3: cake	kc_4: K-1
kc_5: vehicle	kc_6: idol	kc_7: stock	kc_8: liquor
kc_9: high school soccer	kc_{10}: professional baseball	kc_{11}: horse racing	kc_{12}: art
kc_{13}: Southern All Stars	kc_{14}: Shiki Theatre	kc_{15}: idol	kc_{16}: animation
kc_{17}: Apple	kc_{18}: sommelier	kc_{19}: linux	kc_{20}: diet

Figure 9. 20 *kc*s for evaluation of knowledgeable blogger identification

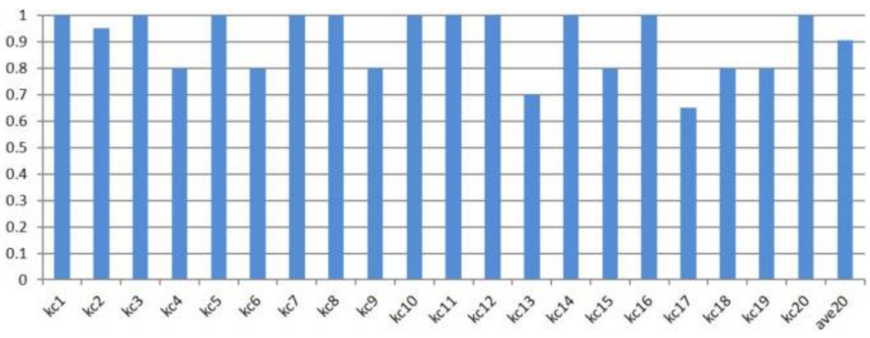

Figure 10. Accuracy of knowledgeable blogger identification

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i)}, \tag{13}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p}, \tag{14}$$

where p is the number of items in the ranking list, rel_i is the relevance of the ith item in the ranking list and $IDCG_p$ is the highest value of DCG_p when the ranking list is ideal. Because we consider the ranking list of the top five *kc*s, p is five in this experiment. $IDCG_p$ is the highest value of DCG_p for the human-generated ranking list, and rel_1-rel_5 of first *kc* to fifth *kc* are set to 5, 4, 3, 2 and 1, respectively.

For example, if the human-generated ranking list of *kc*s is (A, B, C, D, E) and a ranking list generated by a measure is (B, A, E, X, D), $IDCG_p$, DCG_p and $nDCG_p$ are calculated as follows:

$$\begin{aligned} IDCG_5 &= 5 + \frac{4}{log_2 2} + \frac{3}{log_2 3} + \frac{2}{log_2 4} + \frac{1}{log_2 5} = 12.324 \\ DCG_5 &= 4 + \frac{5}{log_2 2} + \frac{1}{log_2 3} + \frac{0}{log_2 4} + \frac{2}{log_2 5} = 10.492 \\ nDCG_5 &= \frac{10.492}{12.324} = 0.850 \end{aligned}$$

Table I shows the nDCG scores of *kc* ranking lists for the 40 buzzwords generated by the three similarity measures. For some buzzwords (e.g. bw_1, bw_2 and bw_3), all three measures generate the *kc* ranking lists close to the human-generated one, resulting in high nDCG scores. The average nDCG scores of 40 buzzwords are 0.570, 0.569 and 0.616, respectively, for Simpson's coefficient, Jaccard's coefficient and cosine similarity. For a buzzword, the highest nDCG score is italicized. Cosine similarity achieves a little higher nDCG scores than the other two similarity measures. We investigate the *kc* ranking lists for the buzzwords (e.g. bw_{39} and bw_{40}) whose nDCG scores are low. These buzzwords are categorized into incorrect *kc*s because the buzzword and *kc* share common co-occurrence words which have different meanings in different domains. For example, although buzzword "Kindle" and *kc* "stationery" share the common co-occurrence words such as "file" and "paper", they are used in different meanings, and "Kindle" should not be associated with "stationery". Solution to this kind of a problem is the subject of our future work.

Table I.
nDCG for *kc* categorization

Buzzword	SimpsonSim	JaccardSim	CosineSim
bw_1	0.923	0.923	*0.971*
bw_2	0.860	0.860	*0.876*
bw_3	0.850	0.850	*0.902*
⋮	⋮	⋮	⋮
bw_{39}	0.181	0.178	*0.194*
bw_{40}	*0.125*	*0.125*	0.112
AVG	0.570	0.569	*0.616*

6.3 Buzzword persistence

To evaluate our buzzword persistence method, we select 80 buzzwords and compare our calculated score with the degree of recognition of buzzwords derived from a human-answered questionnaire.

As shown in Table II, the 80 buzzwords are divided into four groups according to their peak time, and each group of 20 buzzwords are shown to Kizasi Company's users. For each buzzword, the users are asked to answer four types of recognition degrees (Table III). $q1$, $q2$, $q3$ and $q4$ are related to the recognition of semantic contents, impression in the world, impression in the network and impression in the blogosphere, respectively. Four options associated with points 4, 3, 2 and 1 can be selected, representing that they think that the buzzword's recognition on each type is high, relatively high, relatively low or low. Although the different groups have slightly different numbers of respondents, each buzzword receives the answers from about 360 users (Table II). For each buzzword, its recognition degree on each of four types is calculated by averaging the points from all the respondents' answers. Furthermore, for each type of recognition degree, the 80 buzzwords are sorted by their average points, and the consequent ranking list of buzzwords based on human-generated recognition degrees is used as an ideal ranking list. In this experiment, p in $nDCG_p$ is 80 and rel_i is its average point of the ith buzzword.

Buzzword	Peak time	# of respondents
bw_1-bw_{20}	2012/09/01-2012/11/20	359
bw_{21}-bw_{40}	2012/11/21-2013/01/30	362
bw_{41}-bw_{60}	2013/02/01-2013/04/10	384
bw_{61}-bw_{80}	2013/04/11-2013/06/30	367

Table II. Buzzwords for persistence evaluation

No.	Question	Answer option	Point
q_1	Recognition of semantic content	I can provide a detailed description	4
		I know it to a certain extent	3
		I have ever heard of it	2
		I do not know it	1
q_2	Recognition in the world	It is widely popular	4
		It is somewhat popular	3
		It is popular on a small scale	2
		It leaves no impression on me	1
q_3	Recognition in the network	It is widely reported in news or SNS	4
		It is a somewhat hot topic	3
		I have ever seen it in the network	2
		I have never seen it in the network	1
q_4	Recognition in the blogosphere	I posted entries about it frequently	4
		I posted entries about it sometimes	3
		I posted entries about it once or twice	2
		I have never posted entries about it	1

Note: SNS = Social network service

Table III. Questionnaire on recognition investigation

For the comparison targets, we sort the 80 buzzwords according to the total number of entries containing each buzzword during the entire period, including the time before and after their own peak (baseline), during 6, 12 and 24 weeks after the peak. Table IV shows the nDCG scores of the ranking lists based on the four measures with respect to the human-generated ranking list. For each type of recognition degree, the highest nDCG score is italicized. There is no obvious difference among the different duration periods. However, the consideration of a buzzword's persistence – the influence after its peak – works better than the baseline that is not distinguished from the periods before or after the peak. The highest nDCG scores indicate that our method of measuring buzzword persistence aligns with people's recognition degrees.

6.4 Buzzword peak time content

We select 13 buzzwords for analyzing the peak time contents. For each buzzword, we extract all co-occurrence words from the blog entries posted during the three weeks around its peak, that is, all words having more than one co-occurrence with the buzzword during the three weeks. Each co-occurrence word is shown to Kizasi Company's users and is evaluated on whether it can be thought as a related word to the buzzword and an appropriate peak time content word of the buzzword.

Table V shows the numbers of evaluators, co-occurrence words, related words and peak time content words for the buzzwords. A co-occurrence word is labeled as a related word if all its evaluators agree that it is related to the buzzword. A co-occurrence word is labeled as a peak time content word if all its evaluators agree that it reflects the buzzword's peak time content. The numbers of co-occurrence words, related words and peak time content words vary widely for different buzzwords. On average, 680 co-occurrence words, 166 related words and 29 peak time content words are extracted for a buzzword. The great difference between the numbers of related words and peak

Table IV.
nDCG for persistence

No.	Baseline # of all entries	Persistence $T_d = 6$ weeks	Persistence $T_d = 12$ weeks	Persistence $T_d = 24$ weeks
q_1	0.898	0.911	0.912	*0.914*
q_2	0.887	0.904	0.901	*0.905*
q_3	0.892	0.902	0.902	*0.905*
q_4	0.916	0.931	0.935	*0.939*

Table V.
Co-occurrence words, related words and peak time content words

buzzword	# of evaluators	# of co-occurrence words	# of related words	# of peak time content words
bw_1	2	204	163	18
bw_2	3	263	42	23
⋮	⋮	⋮	⋮	⋮
bw_{12}	3	1,252	99	9
bw_{13}	3	1,589	482	69
AVG	3	680	166	29

time content words indicates that only a part of related words is appropriate as the peak time content words.

For each buzzword, we evaluate the quality of the top-k ($k = 10$ and $k = 100$) words extracted by each of the following three measures: co-occurrence frequency, LogLog-score and Spearman's rank correlation coefficient (described in Section 4.1). As the evaluation is done by only about three evaluators and the orders of related words and peak time content words are not given, an ideal ranking list of related words or peak time content words for a buzzword is difficult to generate. Thus, in this experiment, we evaluate the results using three metrics: precision, recall and *F*-measure instead of *nDCG*. The results averaged for the 13 buzzwords are shown in Table VI. For each metric and each measure, the highest score is italicized. Although the words extracted by co-occurrence frequency and LogLog-score as related words have higher precision, recall and *F*-measure values than those extracted by Spearman's rank correlation coefficient, our proposed measure, which calculates the Spearman's rank correlation coefficient between the buzzword entry count order and the co-occurrence entry count order, can extract more appropriate peak time content words than the other two measures, which only take co-occurrence into account.

6.5 Buzzword growth period

In this experiment, we select three categories: movie, TV program and smartphone. For each category, ten buzzwords are manually listed up. Based on the method described in Section 4.2, we calculate the growth period for each buzzword. The result is shown in Table VII. The growth periods of different buzzwords are different, varying from about six months to more than one year. These categories, buzzwords and calculated growth periods are used for the identification of prophetic bloggers in the next section.

6.6 Prophetic blogger identification

For each of the three categories, the top 300 bloggers with the highest knowledge scores are first extracted. For each of the ten buzzwords in each category, the prediction scores of the 300 bloggers on the buzzword are calculated using the method described in

Metric	Related words		Peak time content words	
Measure	Top 10	Top 100	Top 10	Top 100
Precision				
Co-occurrence frequency (%)	*77.7*	43.8	26.2	9.4
LogLog-score (%)	75.4	*47.9*	30.8	12.9
Correlation coefficient (%)	51.5	44.0	*37.7*	*15.3*
Recall				
Co-occurrence frequency (%)	*7.8*	35.2	10.7	37.5
LogLog-score (%)	*7.8*	*40.3*	12.4	52.2
Correlation coefficient (%)	4.9	37.4	*15.3*	*57.8*
F-measure				
Co-occurrence frequency (%)	*0.136*	0.349	0.157	0.144
LogLog-score (%)	*0.136*	*0.396*	0.168	0.199
Correlation coefficient (%)	0.087	0.364	*0.221*	*0.232*

Table VI. Precision, recall and *F*-measure of related words and peak time content words extracted by each of the three measures

Table VII. Buzzwords' growth periods

Category	Buzzword	Start	Peak
Movie	bw_1	2013/11/03	2014/7/13
	bw_2	2012/01/08	2012/08/19
	⋮	⋮	⋮
	bw_{10}	2012/08/05	2013/4/28
TV program	bw_{11}	2013/03/17	2013/09/22
	bw_{12}	2013/05/05	2013/09/22
	⋮	⋮	⋮
	bw_{20}	2013/04/21	2013/12/22
Smartphone	bw_{21}	2013/04/14	2013/09/01
	bw_{22}	2011/08/28	2012/09/16
	⋮	⋮	⋮
	bw_{30}	2012/06/10	2013/09/22

Section 5.2. The ranking list of the top five bloggers with the highest prediction scores on each buzzword is generated. Next, we investigate whether there exist bloggers who appear in multiple buzzwords' top blogger lists for each category, that is, whether there exists any blogger who can predict multiple buzzwords. Both of the two formulas (Formulas 9 and 10) proposed in Section 5.1 for the calculation of post-earliness are used for verifying the effectiveness in our experiments. The column charts in Figures 11 and 12 show the number of bloggers who appear in more than n (n = 1,2,3,4) buzzwords' top blogger lists, that is, the number of bloggers who can predict more than n (n = 1,2,3,4) buzzwords. For all three categories, the proposed approach detects the bloggers who equally have high prediction scores on multiple buzzwords. The number of extracted bloggers for the three categories is on the same level. Also, the number of bloggers extracted by Formulas 9 and 10 has no great difference.

We ask two evaluators to browse the entries posted by these bloggers and judge whether they are prophetic bloggers. The judgment criterion is whether the bloggers have posted some entries which contain buzzwords' peak time content words before the peak. The bloggers who are regarded as prophetic bloggers by both evaluators are used as the true prophetic bloggers for the evaluation of identification accuracy. The column charts in Figures 13 and 14 show the number of true prophetic bloggers extracted, respectively, by

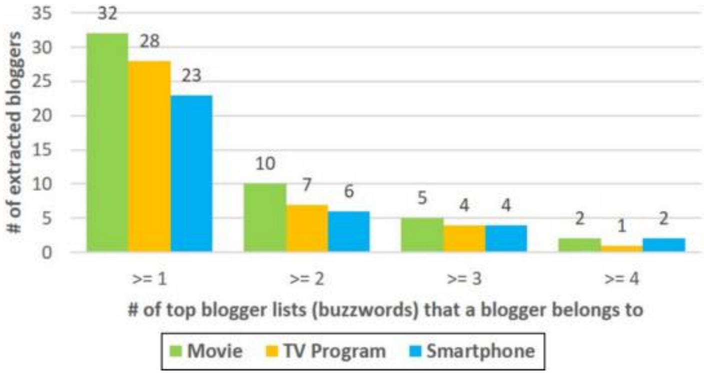

Figure 11. Number of identified prophetic bloggers extracted by Formula 9

Formulas 9 and 10. Except the case of "movie and # *of top blogger lists* ≧ 2", the number of true prophetic bloggers extracted by Formula 10 are larger than or equal to the ones extracted by Formula 9. The line charts in the two figures show the ratios of the true prophetic bloggers to the prophetic bloggers identified by our approach. Except the case of "TV program for Formula 9", the accuracy increases as we increase the required number of top blogger lists that a blogger must belong to. When the minimum number of top blogger lists that a blogger must belong to is set to three for "smartphone" and four for "movie", the

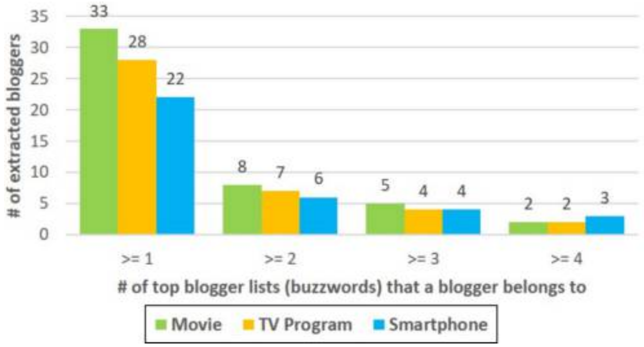

Figure 12. Number of identified prophetic bloggers extracted by Formula 10

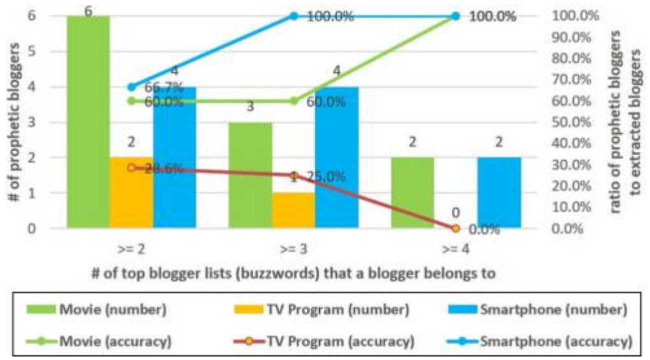

Figure 13. Number and accuracy of true prophetic bloggers extracted by Formula 9

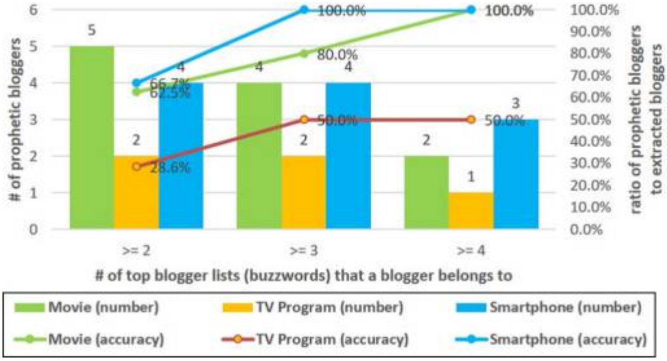

Figure 14. Number and accuracy of true prophetic bloggers extracted by Formula 10

probability that the identified prophetic bloggers are true prophetic bloggers is 100 per cent. Even when the accuracy is relatively low, such as for "TV program", we are still successful in finding true prophetic bloggers by our approach. Overall, the accuracy of Formula 10 is higher than that of Formula 9. As Formula 10 can obtain more true prophetic bloggers and has higher accuracy than that of Formula 9, we can adopt the results (Figures 12 and 14) of Formula 10 as the comparison targets for the next analysis.

We compare the accuracy of top-k bloggers ranked by the proposed approach with the method based on the number of entries containing any of the ten buzzwords in each category and the method based on bloggers' knowledge scores. *k* is set to eight, seven and six for the three categories, which are the number of bloggers who appear in more than two top blogger lists extracted by our approach (Formula 10). Table VIII shows the comparison results. For each category, the highest accuracy is italicized. Our approach works much better than other two methods, as both of them do not take temporal features into account. This indicates that the comprehensive consideration of the four factors – post earliness, content similarity, entry frequency and buzzword coverage – is effective for finding prophetic bloggers.

For all the true prophetic bloggers who appear in more than two top blogger lists identified by our approach (Formula 10), we also investigate their ranks in the ranking list of knowledge scores. Table IX shows the comparison results between the ranks of the true prophetic bloggers identified by the prediction scores and the ones by the knowledge scores. As there is no obvious correlation between them, the prophetic bloggers extracted by our approach cannot be detected only by the knowledge scores.

Table VIII. Identification accuracy compared to other methods

Category	# of bloggers	Our approach	# of entries	Knowledge score
Movie	8	*62.5%* (5/8)	12.5% (1/8)	0% (0/8)
TV program	7	*28.6%* (2/7)	14.3% (1/7)	28.6% (2/7)
Smartphone	6	*66.7%* (4/6)	50.0% (3/6)	16.7% (1/6)
AVG	7	*52.6%*	25.6%	15.1%

Table IX. Rank comparison of bloggers extracted by our approach and by knowledge scores

Category	Blogger	# of top blogger lists	Rank by our approach	Rank by knowledge scores
Movie	m1	4	1	38
	m2	4	1	217
	m3	3	3	6
	m4	3	3	11
	m5	2	5	31
TV program	t1	4	1	12
	t2	3	2	118
Smartphone	s1	4	1	77
	s2	4	1	2
	s3	4	1	11
	s4	3	4	263

7. Related work

Identification of important users has been widely studied. Balog *et al.* (2012) provided a survey on expert finding within an organization. Hashemi *et al.* (2013) addressed the problem of expertise retrieval in a bibliographic network. There is also research aimed at finding important users from social media. We classify them into two types: one that extracts knowledgeable users (Balog *et al.*, 2008; Bozzon *et al.*, 2013; Guy *et al.*, 2013) and the other that identifies influential users (Agarwal *et al.*, 2008; Cai and Chen, 2010; Weng *et al.*, 2010; Cha *et al.*, 2010; Bakshy *et al.*, 2011; Wu *et al.*, 2011; Goyal *et al.*, 2008; Singer, 2012).

Balog *et al.* (2008) adopted two language modeling-based approaches to find expert bloggers and showed that the blogger model considering global aspects of blogs outperformed the posting model which only considered highly relevant posts. Bozzon *et al.* (2013) addressed the problem of selecting knowledgeable users according to the information about social activities and showed that the analysis of social activities, social relationships and socially shared contents helped improve the effectiveness of expert finding. Guy *et al.* (2013) distinguished between expertise in the topic and interest in it and compared these two semantics across the different social media applications.

Agarwal *et al.* (2008) and Cai and Chen (2010) investigated influential users in the blogosphere. Agarwal *et al.* (2008) illustrated that active bloggers were not necessarily influential, and influential bloggers could impact fellow bloggers in various ways. Cai and Chen (2010) proposed a model to mine influential bloggers according to their interest domains, the comments to their posts and their authority in the network. Weng *et al.* (2010), Cha *et al.* (2010), Bakshy *et al.* (2011) and Wu *et al.* (2011) analyzed Twitter to find influential twitterers. Weng *et al.* (2010) proposed a TwitterRank algorithm to measure the influence of users in Twitter taking both the topical similarity between users and the link structure into account. Cha *et al.* (2010) compared three measures of influence: indegree, retweets and mentions and showed that the top users based on the three measures had little overlap. Bakshy *et al.* (2011) emphasized that although the most influential users were often the most cost-effective, sometimes, ordinary influencers – individuals who had average or even less-than-average influence – could realize the most cost-effective performance. Wu *et al.* (2011) exploited the feature of Twitter lists to distinguish between elite users and ordinary users and investigated the information flow among them. Goyal *et al.* (2008) and Singer (2012) proposed the theoretical models to identify influential users. Goyal *et al.* (2008) focused on social action propagation and proposed a frequent pattern-based approach for identifying the leaders when it came to setting the trend for performing various actions. Singer (2012) designed a model to provide appropriate incentives for social network users to become early adopters – the subset of individuals who received incentives and would subsequently influence others. Different from the previous works which focus on the expertise degree and influence degree of users, we attempt to find important users by analyzing users' buzzword prediction ability.

Topic or event detection (Sakaki *et al.*, 2010; Jin *et al.*, 2010; Asur *et al.*, 2011; Becker *et al.*, 2011; Yin *et al.*, 2013; Spina *et al.*, 2014) is closely related to our work. Especially, Jin *et al.* (2010) proposed an algorithm to rank Web documents by their possibility to be the topic initiator – the document that initiated the topic or was the first to discuss about the topic. Asur *et al.* (2011) investigated what factors caused the formation and persistence of trends and found that the resonance of the content contributed to trend creation and its propagation more than user activity and the number of followers. Yin *et al.* (2013) proposed a unified model to distinguish burst topics from stable topics. These works motivate us to analyze the lifespan of buzzwords: the starting point of buzzwords, the peak of buzzwords and the duration period after its peak.

Another related line of research is popularity prediction. Future popularity is predicted for different types of data such as events (Zhang *et al.*, 2015b), videos (Figueiredo *et al.*, 2011; Pinto *et al.*, 2013; Li *et al.*, 2013), news (Lerman and Hogg, 2010; Bandari *et al.*, 2012), search (Kairam *et al.*, 2013; Golbandi *et al.*, 2013; Radinsky *et al.*, 2013), tweets (Mathioudakis and Koudas, 2010; Hong *et al.*, 2011; Bian *et al.*, 2014) and unrestricted use of generated contents Ahmed *et al.* (2013). Although future popularity has been noticed in these researches, it is not used for finding important users. We link buzzword popularity analysis results to finding prophetic bloggers.

8. Conclusions

In this paper, we studied the problem of finding prophetic bloggers who are sensitive to future popular trends. We focused on temporal and content features of blog data and proposed an approach to analyze bloggers' buzzword prediction ability. Bloggers are evaluated on how early, how related, how often and how in-depth they have posted blog entries containing the buzzwords in a category. In the experimental evaluation, our approach accurately extracted prophetic bloggers and also outperformed those methods which did not take temporal features into account.

Although we conducted the experiments on identifying the growth periods for 30 buzzwords from three categories and subsequently used them for the identification of prophetic bloggers, the effectiveness of growth period calculation has not been verified independently. We plan to compare the identification quality based on the blog entries within and outside the growth period. Another interesting direction is to identify future buzzwords from blog entries posted by prophetic bloggers. It is also future work to implement a practical system that can extract future buzzwords.

Notes

1. Kizasi Company is a company specializing in mining and analyzing of social media. Kizasi develops technology allowing for better understanding these data – including search, trend analysis and categorization of such social media.
2. The equivalent Japanese search keyword and grammar are used.
3. Abe is the current Prime Minister of Japan.
4. Abenomics refers to the economic policies advocated by Shinzo Abe, the Prime Minister of Japan.

References

Agarwal, N., Liu, H., Tang, L. and Yu, P.S. (2008), "Identifying the influential bloggers in a community", *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, CA*, pp. 207-218.

Ahmed, M., Spagna, S., Huici, F. and Niccolini, S. (2013), "A peek into the future: predicting the evolution of popularity in user generated content", *WSDM '13 Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, Rome*, pp. 607-616.

Asur, S., Huberman, B.A., Szabo, G. and Wang, C. (2011), "Trends in social media: persistence and decay", *ICWSM'11 Proceedings of the Fifth International Conference on Weblogs and Social Media*, *Barcelona, Catalonia*.

Bakshy, E., Hofman, J.M., Mason, W.A. and Watts, D.J. (2011), "Everyone's an influencer: quantifying influence on twitter", *WSDM '11 Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, *Hong Kong*, pp. 65-74.

Balog, K., Fang, Y., Rijke, M., Serdyukov, P. and Si, L. (2012), "Expertise retrieval", *Foundations and Trends in Information Retrieval*, Vol. 6 Nos 2/3, pp. 127-256.

Balog, K., Rijke, M. and Weerkamp, W. (2008), "Bloggers as experts: feed distillation using expert retrieval models", *SIGIR'08 Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *Singapore*, pp. 753-754.

Bandari, R., Asur, S. and Huberman, B.A. (2012), "The pulse of news in social media: forecasting popularity", *ICWSM'12 Proceedings of the Sixth International Conference on Weblogs and Social Media*, *Dublin*.

Becker, H., Naaman, M. and Gravano, L. (2011), "Beyond trending topics: real-world event identification on twitter", *ICWSM'11 Proceedings of the Fifth International Conference on Weblogs and Social Media*, *Barcelona, Catalonia*.

Bian, J., Yang, Y. and Chua, T. (2014), "Predicting trending messages and diffusion participants in microblogging network", *SIGIR'14 Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, *Gold Coast, QLD*, pp. 537-546.

Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M. and Vesci, G. (2013), "Choosing the right crowd: expert finding in social networks", *EDBT'13 Proceedings of 16th International Conference on Extending Database Technology*, *Genoa*, pp. 637-648.

Cai, Y. and Chen, Y. (2010), "Mass: a multi-facet domain-specific influential blogger mining system", *ICDE'10 Proceedings of the 26th International Conference on Data Engineering*, *Long Beach, CA*, pp. 1109-1112.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P.K. (2010), "Measuring user influence in twitter: the million follower fallacy", *ICWSM'10 Proceedings of the Fourth International Conference on Weblogs and Social Media*, *Washington, DC*.

Church, K.W. and Hanks, P. (1990), "Word association norms, mutual information, and lexicography", *Computational Linguistics*, Vol. 16 No. 1, pp. 22-29.

Figueiredo, F., Benevenuto, F. and Almeida, J.M. (2011), "The tube over time: characterizing popularity growth of youtube videos", *WSDM '11 Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, *Hong Kong*, pp. 745-754.

Golbandi, N., Katzir, L., Koren, Y. and Lempel, R. (2013), "Expediting search trend detection via prediction of query counts", *WSDM '13 Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, *Rome*, pp. 295-304.

Goyal, A., Bonchi, F. and Lakshmanan, L.V.S. (2008), "Discovering leaders from community actions", *CIKM '08 Proceedings of the 17th ACM Conference on Information and Knowledge Management*, *Napa Valley, CA*, pp. 499-508.

Gregory, C.W. and Dale, F.I. (2014), *Nonparametric Statistics: A Step-by-Step Approach*, 2nd ed., Wiley, Hoboken, NJ.

Guy, I., Avraham, U., Carmel, D.S., Ur, M.J. and Ronen, I. (2013), "Mining expertise and interests from social media", *WWW'13 Proceedings of the 22nd International World Wide Web Conference*, *Rio de Janeiro*, pp. 515-526.

Hashemi, S.H., Neshati, M. and Beigy, H. (2013), "Expertise retrieval in bibliographic network: a topic dominance learning approach", *CIKM '13 Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, *San Francisco, CA*, pp. 1117-1126.

Hong, L., Dan, O. and Davison, B.D. (2011), "Predicting popular messages in twitter", *WWW '11 Proceedings of the 20th International Conference Companion on World Wide Web*, *Hyderabad*, pp. 57-58.

Jin, X., Spangler, W.S., Ma, R. and Han, J. (2010), "Topic initiator detection on the world wide web", *WWW '10 Proceedings of the 19th International Conference on World Wide Web*, *Raleigh, NC*, pp. 481-490.

Kairam, S.R., Morris, M.R., Teevan, J., Liebling, D.J. and Dumais, S.T. (2013), "Towards supporting search over trending events with social media", *ICWSM'13 Proceedings of the Seventh International Conference on Weblogs and Social Media*, *Cambridge, MA*.

Kilgarriff, A. and Tugwell, D. (2002), "Sketching words", in Correard, M.H. (Ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, EURALEX, Innsbruck, pp. 125-137.

Lerman, K. and Hogg, T. (2010), "Using a model of social dynamics to predict popularity of news", *WWW '10 Proceedings of the 19th International Conference on World Wide Web*, *Raleigh, NC*, pp. 621-630.

Li, H., Ma, X., Wang, F., Liu, J. and Xu, K. (2013), "On popularity prediction of videos shared in online social networks", *CIKM '13 Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, *San Francisco, CA*, pp. 169-178.

Mathioudakis, M. and Koudas, N. (2010), "Twittermonitor: trend detection over the twitter stream", *SIGMOD '10 Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, *Indianapolis, Indiana*, pp. 1155-1158.

Pinto, H., Almeida, J.M. and Goncalves, M.A. (2013), "Using early view patterns to predict the popularity of youtube videos", *WSDM'13 Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, *Rome*, pp. 365-374.

Radinsky, K., Svore, K.M., Dumais, S.T., Shokouhi, M., Teevan, J., Bocharov, A. and Horvitz, E. (2013), "Behavioral dynamics on the web: learning, modeling, and prediction", *ACM Transactions on Information Systems*, Vol. 31 No. 3, 16.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), "Earthquake shakes twitter users: real-time event detection by social sensors", *WWW '10 Proceedings of the 19th International Conference on World Wide Web*, *Raleigh, NC*, pp. 851-860.

Singer, Y. (2012), "How to win friends and influence people, truthfully: influence maximization mechanisms for social networks", *WSDM '12 Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, *Seattle, WA*, pp. 733-742.

Spina, D., Gonzalo, J. and Amigo, E. (2014), "Learning similarity functions for topic detection in online reputation monitoring", *SIGIR '14 Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, *Gold Coast, QLD*, pp. 527-536.

Wang, Y., Wang, L., Li, Y., He, D. and Liu, T. (2013), "A theoretical analysis of ndcg type ranking measures", *COLT'13 Proceedings of the 26th Annual Conference on Learning Theory*, *Princeton University, NJ*, pp. 25-54.

Weng, J., Lim, E., Jiang, J. and He, Q. (2010), "Twitterrank: finding topic-sensitive influential twitterers", *WSDM '10 Proceedings of the Third ACM International Conference on Web Search and Data Mining*, *New York, NY*, pp. 261-270.

Wu, S., Hofman, J.M., Mason, W.A. and Watts, D.J. (2011), "Who says what to whom on twitter", *WWW '11 Proceedings of the 20th International Conference on World Wide Web, Hyderabad*, pp. 705-714.

Yin, H., Cui, B., Lu, H., Huang, Y. and Yao, J. (2013), "A unified model for stable and temporal topic detection from social media data", *ICDE'13 Proceedings of the 29th IEEE International Conference on Data Engineering, Brisbane*, pp. 661-672.

Zhang, J., Tomonaga, S., Nakajima, S., Inagaki, Y. and Nakamoto, R.Y. (2015a), "Finding prophets in the blogosphere: bloggers who predicted buzzwords before they become popular", *iiWAS'15 Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, Brussels*, pp. 100-109.

Zhang, X., Chen, X., Chen, Y., Wang, S., Li, Z. and Xia, J. (2015b), "Event detection and popularity prediction in microblogging", *Neurocomputing* Vol. 149 (Part C), pp. 1469-1480.

About the authors

Jianwei Zhang is currently a Research Associate in the Faculty of Industrial Technology, Tsukuba University of Technology, Japan. He received his BS degree from North China Electric Power University, China, in 2000, and MS and PhD degrees from University of Tsukuba, Japan, in 2005 and 2008, respectively. He is a member of Association for Computing Machinery, Information Processing Society of Japan and The Database Society of Japan. His research interests include Web mining, Web information system, information retrieval, information recommendation and deaf support system. Jianwei Zhang is the corresponding author and can be contacted at: zhangjw@a.tsukuba-tech.ac.jp

Seiya Tomonaga graduated from Kyoto Sangyo University in 2015. He then joined the Drecom Co. Ltd and is working in the technology division as of the end of March, 2016.

Shinsuke Nakajima is currently a Professor in the Faculty of Computer Science and Engineering, Kyoto Sangyo University, Japan. He was a Visiting Scholar at the Department of Computer Science at the University of California, Santa Barbara, from September 2015 to September 2016. He received his BS and MS degrees from Kobe University, Japan, in 1995 and 1997, respectively, and PhD degree from Kyoto University, Japan, in 2004. He is a member of Association for Computing Machinery, Information Processing Society of Japan and The Database Society of Japan. His research interests include Web mining and recommender system.

Yoichi Inagaki graduated from The University of Tokyo in 1990. He then joined the CAC Corporation, working in the technology research division. From 1996 to 1998, he was a Visiting Researcher at Stanford University's Computer Science Department. In 2005, he began development of a new social media search engine technology, which formed the foundation of the startup company, Kizasi Company. From early 2007, he became the company's chief technology officer.

Reyn Nakamoto is currently a Research and Development Engineer at Kizasi Company, Inc. He graduated from Oregon State University with a bachelor of Science in computer science in 2003. He then received his master of Science in information science from Nara Institute of Science and Technology in 2008. He then joined Kizasi Company in 2008.