Estimating Bloggers' Prediction Ability on Buzzwords and Categories

Jianwei Zhang, Yoichi Inagaki, Reyn Nakamoto, and Shinsuke Nakajima

Abstract—**It is a challenging task to find important users from social media. In this paper, we propose an approach to identify prophetic bloggers by estimating bloggers' prediction ability on buzzwords and categories. We conduct a time-series analysis on large-scale blog data, which includes categorizing a blogger into knowledgeable categories, identifying past buzzwords, analyzing a buzzword's peak time content and growth period, and estimating a blogger's prediction ability on a buzzword and on a category. Bloggers' prediction ability on a buzzword is evaluated considering three factors: post earliness, content similarity and entry frequency. Bloggers' prediction ability on a category is evaluated considering the buzzword coverage in that category. For calculating bloggers' prediction ability on a category, we propose multiple formulas and compare the accuracy through experiments. Experimental results show that the proposed approach can find prophetic bloggers on real-world blog data.**

Index Terms—**social media, time-series analysis, buzzword detection, expert finding, prophetic blogger.**

I. Introduction

It is a challenging task to find important users from social media. There are two main directions in past research: finding knowledgeable users by measuring their expertise levels or finding influential users by estimating their influence degrees. The former is usually based on textual content analysis while the latter also makes use of link structure in social networks. We focus on users' prediction ability on future popularity, which has not been investigated in previous works.

The blogosphere is a conductive platform for bloggers to issue posts, share ideas and exchange opinions. The data in the blogosphere is dynamic reflecting information change over time. Potential knowledgeable bloggers with prior awareness of future popular trends may exist in the blogosphere. Identifying these bloggers can bring great values. For example, analysis on their blog entries may help find future trends or communication with them may even help foresee things that will become popular.

We propose an approach to identify important bloggers based on their prediction ability on buzzwords and categories. Buzzwords are the 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 "prophetic bloggers". Bloggers' prediction ability on a buzzword is evaluated considering three factors: post earliness, content similarity and entry frequency. The general idea is based on: (a) The earlier a blogger posted blog entries containing a buzzword, the better prediction ability on the buzzword he may have. (b) The more similar the contents of their past entries to the peak time content of a buzzword at its popularity peak, the more accurate their prediction ability on the buzzword. (c) The larger the quantity of early and similar blog entries containing the buzzword are, the better prophetic blogger he may be. In our previous work [1], bloggers' prediction ability on a category was not fully discussed. In this paper, bloggers' prediction ability on a category is evaluated making use of prediction scores on buzzwords in that category and considering the buzzword coverage. The general idea is that the more buzzwords relative to a category he can well predict, the better prophetic blogger on the category he may be.

For identifying prophetic bloggers, we conduct a time-series analysis on real-world blog data consisting of 150 million entries from 11 million bloggers. Our contributions are summarized as follows:

- We introduce a method for categorizing a blogger into his appropriate potential communities called knowledgeable categories (Section II).
- We develop a method for automatically identifying past buzzwords from historical blog data based on their persistence (Section III).
- We analyze a buzzword's properties by identifying its peak time content and calculating its growth period (Section IV).
- We integrate the necessary factors for evaluating a blogger's prediction ability on a buzzword (Section V).
- We propose multiple formulas for estimating a blogger's prediction ability on a category (Section VI).

II. Categorizing a blogger into knowledgeable categories

We extract potential communities of bloggers called knowledgeable categories (kc), and automatically categorize bloggers into their appropriate kcs. 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 interest in a category explicitly, if he has posted many blog entries related to the category, our method can categorize him into the appropriate kcs automatically.

A. Extracting knowledgeable categories and constructing co-occurrence dictionaries

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 by performing a regular Web search with the search keywords such as "expert in *" and "fan of *". We manually remove inappropriate ones and

J. Zhang is with Iwate University. E-mail: zhang@iwate-u.ac.jp
Y. Inagaki and R. Nakamoto are with kizasi Company,Inc.
S. Nakajima is with Kyoto Sangyo University.

categorize the keywords into 122 categories, ending up with a list of 122 kc names (e.g., "$politics$", "$economy$", "IT").

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. Specifically, n is 400 in our current implementation. The co-occurrence words and their co-occurrence degrees are stored in each co-occurrence dictionary for each kc.

B. Calculating a blogger's knowledge score

A blogger's knowledge score for a kc is calculated by analyzing how often as well as how in-depth he 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.

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

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

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 co-occurrence degree 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.

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

$$Knowledge_{kc}(blg) = \frac{l}{n} \cdot \frac{log(m)}{m} \cdot \sum_{i=1}^{m} Relevance_{kc}(e_i) \qquad (2)$$

where e_i is an entry that blogger blg posted, m is the number of entries that blg posted during the analysis period, n is the number of the co-occurrence words and 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. $log(m)/m$ 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 knowledge score is larger than a given threshold. Moreover, a blogger may be categorized into two or more kcs and thus may have two or more knowledge scores for different categories. For example, if a blogger belongs to both "$politics$" and "$economy$", he has a knowledge score representing his expertise degree in "$politics$" and another one representing his expertise degree in "$economy$". Through the above process, we have a list of knowledgeable bloggers for each kc.

III. Identifying past buzzwords

Before evaluating a blogger's buzzword prediction ability, buzzwords need to be first detected. We identify past buzzwords by analyzing real-world blog data.

A. Determining buzzword candidates

We start with the top-ranked keywords in the daily topic ranking list provided by kizasi Company. 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. We take the top-k ($k = 100$) keywords from each day and then exclude repeated words and periodical words. The remaining keywords become buzzword candidates.

In our approach, we evaluate a blogger's prediction ability for a kc based on their prediction scores on the buzzwords that belong to the kc. In order to associate buzzword candidates (bwc) with kcs, we calculate the similarity between a bwc and each kc. A bwc is associated with a kc if they share many co-occurrence words. For example, bwc "$Abenomics$" [1] and kc "$politics$" have many common co-occurrence words such as "Abe", "$premier$" and "$party$", and thus, "$Abenomics$" can be categorized into "$politics$".

Each bwc is categorized into the top-k ($k = 5$) kcs with the highest similarities. Consequently, given a kc, the set of similar bwcs can also be identified. This categorization result will be used for the subsequent process in Section III.B and Section VI.

B. Determining 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 since it is forgotten by the public soon after the peak.

We extract influential words as buzzwords from buzzword candidates based on their persistence (Figure 1). 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.

IV. Analyzing 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.

A. Extracting peak time content words

The co-occurrence words of the buzzword around the the peak time are the candidates of its peak time content words. However, not all of them are appropriate as the peak time content words. Figure 2 shows the idea of extracting the peak time content words. We select the co-occurrence words whose time-series variation is the most similar to the buzzword's for representing its peak time content. In Figure 2, co-occurrence word xx is more appropriate as the peak time content word than aa, since xx has much more similar variation curve with buzzword bw.

[1] Abenomics refers to the economic policies advocated by Shinzo Abe, the Prime Minister of Japan.

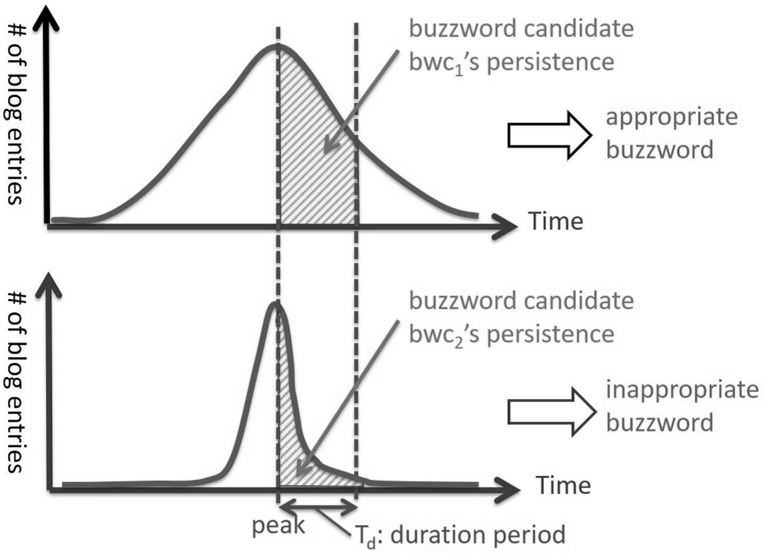

Fig. 1. Buzzword candidates' persistence

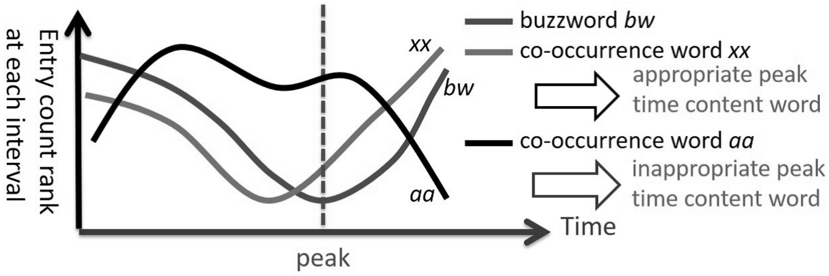

Fig. 2. Determining peak time content words

B. Calculating 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.", the growth period has not begun. Since it only contains ordinary words, this period is inappropriate for analyzing bloggers' prediction ability on the buzzword. If some blog entries such as "$iPhone$ 6 may $adopt$ new $chip$ and $larger$ $display$." begin to appear and some content words from the popularity peak such as "$chip$" and "$display$" are mentioned, the growth period may begin.

Figure 3 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:

$$Similarity(t_i, peak) = \frac{|COW_{t_i} \cap CTW_{peak}|}{min(|COW_{t_i}|, |CTW_{peak}|)} \quad (3)$$

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 3, 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.

V. Calculating a blogger's prediction score on a buzzword

A blogger's prediction score on a buzzword is calculated based on post earliness, content similarity and the quantity

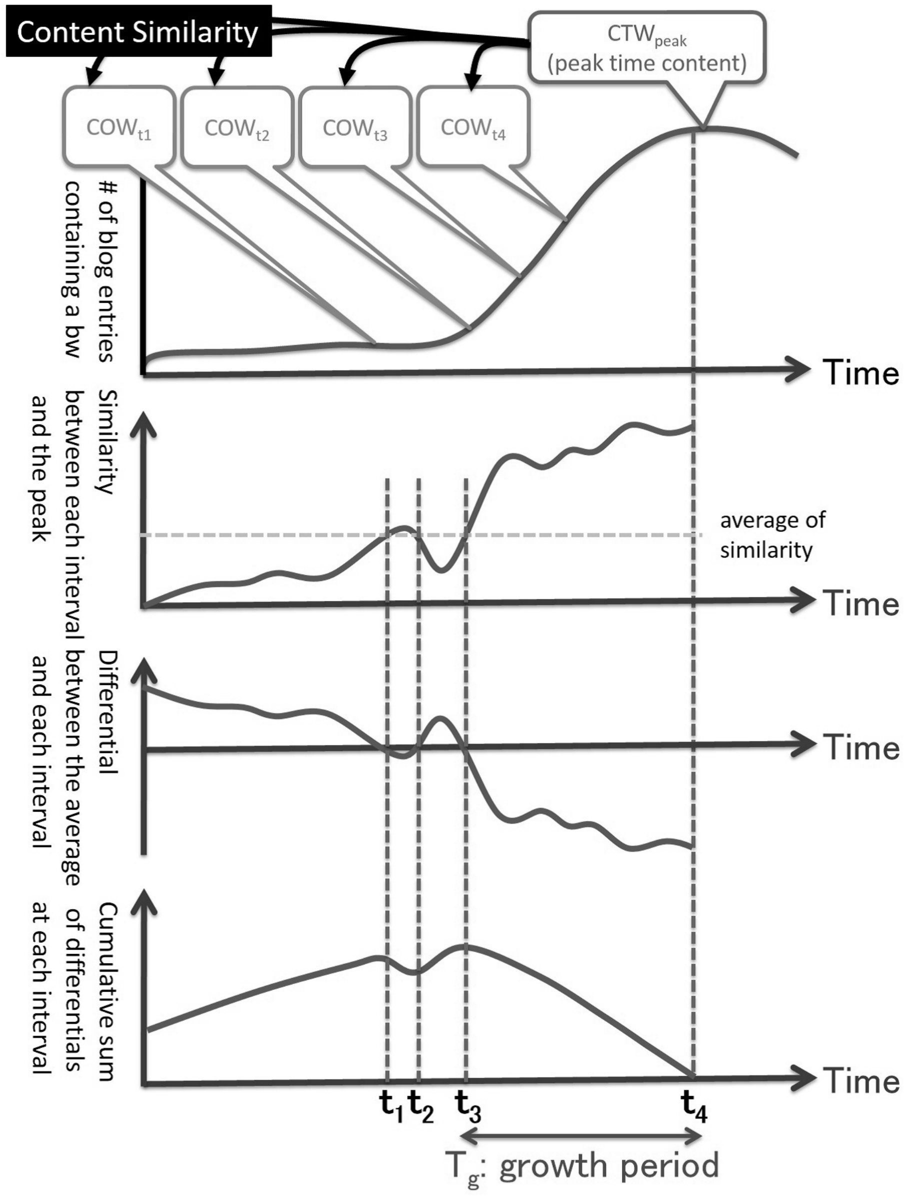
Content Similarity
CTW$_{peak}$
(peak time content)
COW$_{t1}$
COW$_{t2}$
COW$_{t3}$
COW$_{t4}$
# of blog entries containing a bw
Time
Similarity between each interval and the peak
average of similarity
Time
Differential between the average and each interval
Time
Cumulative sum of differentials at each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

Fig. 3. Determining a growth period

of his blog entries containing the buzzword during its growth period.

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 the formula for post earliness of entry e_i for buzzword bw as follows:

$$Earliness_{bw}(e_i) = -log\frac{order(e_i)}{|E_{T_g}|} \tag{4}$$

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. For example, if there are 100 entries containing a buzzword during its growth period, the earliness scores are 2, 1.698, 1.522, ..., 0.008, 0.004, 0, respectively.

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 can be regarded as a good predictor on this buzzword. Thus, we devise the formula for the prediction score of blogger blg for buzzword bw as follows:

$$Pdt_{bw}(blg) = \sum_{i=1}^{m} Earliness_{bw}(e_i) \cdot Similarity(e_i, ptc) \tag{5}$$

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:

$$Similarity(e_i, ptc) = \frac{|D(e_i) \cap CTW_{peak}|}{min(|D(e_i)|, |CTW_{peak}|)} \qquad (6)$$

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.

VI. Calculating a blogger's prediction score on a category

A blogger's prediction ability on a category is evaluated considering his 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.

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 II. Then, we find the buzzwords in this category shown in Section III. Each knowledgeable blogger's prediction score on each buzzword can be calculated by the method described in Section V.

Using the prediction scores on buzzwords, we propose five methods for estimating a blogger's prediction ability on a category. The first method counts the numbers of buzzwords that a blogger can predict. Concretely, 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 multiple top blogger lists as prophetic bloggers on that category. In Figure 4, blg_2 is the best prophetic blogger in that category since he has successfully predicted three buzzwords in that category. blg_6, blg_7 and blg_8 are the next best prophetic bloggers since they predicted the next highest number of buzzwords after blg_2. By this method blg_6, blg_7 and blg_8 have the same rankings since the numbers of buzzwords that they can predict are identical.

In order to meticulously distinguish bloggers' prediction ability on categories, we further propose four calculation formulas. Formula 7 sums up a blogger's prediction scores ($Pdt_{bw}(blg)$) on all buzzwords (bw) that belong to a category (C). Formula 8 introduces a factor $\frac{\log(m+1)}{m+1}$ (m is the number of blog entries containing buzzword bw), which intends to reduce the effect that a blogger posts a large number of entries only related to a specific buzzword. Formula 9 considers the buzzword coverage by introducing l/n where n is the number of buzzwords in a category and l is the number of buzzwords that the blogger can predict. Formula 10 integrates all the factors of the above formulas.

$$Pdt_C(blg) = \sum_{bw \in C} Pdt_{bw}(blg) \qquad (7)$$

$$Pdt_C(blg) = \sum_{bw \in C} \frac{\log(m+1)}{m+1} Pdt_{bw}(blg) \qquad (8)$$

$$Pdt_C(blg) = \frac{l}{n} \sum_{bw \in C} Pdt_{bw}(blg) \qquad (9)$$

$$Pdt_C(blg) = \frac{l}{n} \sum_{bw \in C} \frac{\log(m+1)}{m+1} Pdt_{bw}(blg) \qquad (10)$$

VII. Experimental evaluation

In the 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 IV.B, we calculate the growth period for each buzzword. The growth period of different buzzwords are different, varying from about six months to more than one year.

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 Section V. The ranking list of the top five bloggers with the highest prediction scores on each buzzword is generated. We investigate whether there exist bloggers who appear in more than two buzzwords' top blogger lists for each category. We find that the proposed approach detects eight, seven and six bloggers who appear in more than two top blogger lists for the three categories respectively.

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 has 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.

We compare the accuracy of top-k bloggers ranked by the proposed approach with the method based on the numbers of entries containing any of the ten buzzwords in each category and the method based on bloggers' knowledge scores. Table I shows the comparison results. Since the two methods based on entry numbers and knowledge scores do not take temporal features and prediction abilities into account, the accuracies for identifying prophetic bloggers averaged over the three categories are low (25.6% and 15.1%). Our proposed five methods achieve the average accuracies from 42.9% to 52.6%, outperforming the two comparison methods.

Among the proposed five methods, the proposal using Formula 7 that sums up the prediction scores on all buzzwords in a category does not work better than the other proposals. The proposal using Formula 8 that reduces the effect of the numbers of blog entries performs better than the other proposals for one category $TV\ program$. The proposal using Formula 9 that considers the buzzword coverage performs better than Formula 7 and Formula 8. The highest accuracies are observed for the basic proposal considering the numbers of buzzwords that a blogger can predict and the proposal using Formula 10 that considers both the effect of numbers of blog entries and the buzzword coverage.

Although the basic proposal and the proposal using Formula 10 provide the same accuracies in this experiment, the proposal using Formula 10 can distinguish bloggers more meticulously than the basic proposal. The basic proposal only considers the numbers of top blogger lists a blogger appears in and the numbers are usually small integers which may give many bloggers the same rankings. However, the proposal using Formula 10 calculates bloggers' prediction scores on a category in real numbers, which can better rank bloggers by avoiding many same rankings.

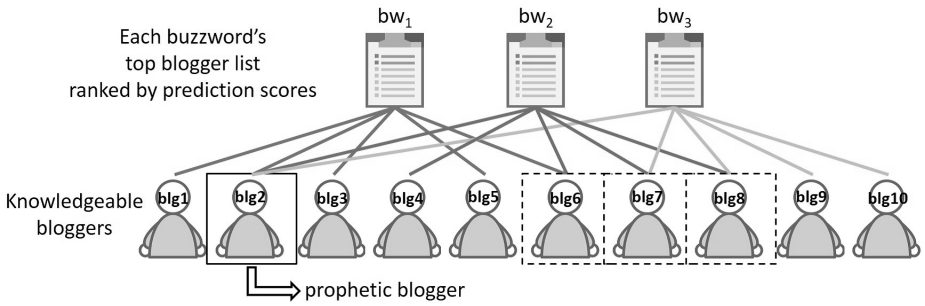

Fig. 4. Evaluating a blogger's prediction ability on a category

TABLE I
ACCURACY COMPARISON

Category	# of	# of	Knowledge	Proposal	Proposal	Proposal	Proposal	Proposal
	bloggers	entries	score	Basic	Formula 7	Formula 8	Formula 9	Formula 10
Movie	8	12.5% (1/8)	0% (0/8)	**62.5%** (5/8)	50.0% (4/8)	37.5% (3/8)	**62.5%** (5/8)	**62.5%** (5/8)
TV program	7	14.3% (1/7)	28.6% (2/7)	28.6% (2/7)	28.6% (2/7)	**42.9%** (3/7)	14.3% (1/7)	28.6% (2/7)
Smartphone	6	50.0% (3/6)	16.7% (1/6)	**66.7%** (4/6)	50.0% (3/6)	50.0% (3/6)	**66.7%** (4/6)	**66.7%** (4/6)
AVG	7	25.6%	15.1%	**52.6%**	42.9%	43.5%	47.8%	**52.6%**

VIII. RELATED WORK

Identification of important users has been widely studied. [2] provided a survey on expert finding within an organization. [3] 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 [4], [5] and the other that identifies influential users [6], [7]. 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 [8], [9] is closely related to our work. 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 [10], videos [11], news [12], search [13], [14], tweets [15], [16], and unrestricted use generated contents [17]. 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.

IX. CONCLUSIONS

In this paper, we proposed the approach to find prophetic bloggers. We focused on temporal and content features of blog data and analyzed bloggers' prediction ability on buzzwords and categories. Bloggers were evaluated on how early, how related, how often and how in-depth they posted blog entries containing the buzzwords in a category. Multiple formulas for estimating bloggers' prediction ability were compared. The experimental results showed our approach could extract prophetic bloggers.

In the future, we will try to develop the methods for identifying future buzzwords from the blog entries posted by prophetic bloggers and implement a practical system that can extract future buzzwords.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number #26330351.

REFERENCES

[1] J. Zhang, S. Tomonaga, S. Nakajima, Y. Inagaki, R. Nakamoto: Prophetic blogger identification based on buzzword prediction ability. *IJWIS* 12(3) (2016): 267-291.

[2] K. Balog, Y. Fang, M. Rijke, P. Serdyukov, L. Si: Expertise Retrieval. *Foundations and Trends in Information Retrieval*, 6(2-3) (2012): 127-256.

[3] S. H. Hashemi, M. Neshati, H. Beigy: Expertise retrieval in bibliographic network: a topic dominance learning approach. In *CIKM 2013*: 1117-1126.

[4] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, G. Vesci: Choosing the right crowd: expert finding in social networks. In *EDBT 2013*: 637-648.

[5] I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, I. Ronen: Mining expertise and interests from social media. In *WWW 2013*: 515-526.

[6] E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts: Everyone's an influencer: quantifying influence on twitter. In *WSDM 2011*: 65-74.

[7] Y. Singer: How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *WSDM 2012*: 733-742.

[8] H. Yin, B. Cui, H. Lu, Y. Huang, J. Yao: A unified model for stable and temporal topic detection from social media data. In *ICDE 2013*: 661-672.

[9] D. Spina, J. Gonzalo, E. Amigo: Learning similarity functions for topic detection in online reputation monitoring. In *SIGIR 2014*: 527-536.

[10] X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, J. Xia: Event detection and popularity prediction in microblogging. *Neurocomputing* 149(2015): 1469-1480.

[11] H. Li, X. Ma, F. Wang, J. Liu, K. Xu: On popularity prediction of videos shared in online social networks. In *CIKM 2013*: 169-178.

[12] R. Bandari, S. Asur, B. A. Huberman: The Pulse of News in Social Media: Forecasting Popularity. In *ICWSM 2012*.

[13] S. R. Kairam, M. R. Morris, J. Teevan, D. J. Liebling, S. T. Dumais: Towards Supporting Search over Trending Events with Social Media. In *ICWSM 2013*.

[14] N. Golbandi, L. Katzir, Y. Koren, R. Lempel: Expediting search trend detection via prediction of query counts. In *WSDM 2013*: 295-304.

[15] L. Hong, O. Dan, B. D. Davison: Predicting popular messages in Twitter. In *WWW (Companion Volume) 2011*: 57-58.

[16] J. Bian, Y. Yang, T. Chua: Predicting trending messages and diffusion participants in microblogging network. In *SIGIR 2014*: 537-546.

[17] M. Ahmed, S. Spagna, F. Huici, S. Niccolini: A peek into the future: predicting the evolution of popularity in user generated content. In *WSDM 2013*: 607-616.