Early Detection of Buzzwords Based on Large-scale Time-Series Analysis of Blog Entries

Shinsuke Nakajima
Faculty of Computer Science
and Engineering
Kyoto Sangyo University
Kyoto, Japan
nakajima@cse.kyoto-su.ac.jp

Jianwei Zhang
Faculty of Computer Science
and Engineering
Kyoto Sangyo University
Kyoto, Japan
zjw@cc.kyoto-su.ac.jp

Yoichi Inagaki
kizasi Company,Inc.
Tokyo, Japan
inagaki@kizasi.jp

Reyn Nakamoto
kizasi Company,Inc.
Tokyo, Japan
reyn@kizasi.jp

ABSTRACT

In this paper, we discuss a method for early detection of "gradual buzzwords" by analyzing time-series data of blog entries. We observe the process in which certain topics grow to become major buzzwords and determine the key indicators that are necessary for their early detection. From the analysis results based on 81,922,977 blog entries from 3,776,154 blog websites posted in the past two years, we find that as topics grow to become major buzzwords, the percentages of blog entries from the blogger communities closely related to the target buzzword decrease gradually, and the percentages of blog entries from the weakly related blogger communities increase gradually. We then describe a method for early detection of these buzzwords, which is dependent on identifying the blogger communities which are closely related to these buzzwords. Moreover, we verify the effectiveness of the proposed method through experimentation that compares the rankings of several buzzword candidates with a real-life idol group popularity competition.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; H.5.4 [**Hypertext/Hypermedia**]: Navigation

General Terms

Algorithms, Measurement, Experimentation

Keywords

Buzzword detection, Time-series analysis, Blogger community

1. INTRODUCTION

Buzzwords are terms or phrases describing topics that have become popular to general population. They represent the latest trends happening around the world–what people are talking about and what people are currently interested in. This information is very valuable, especially to businesses and marketers; and as such, it would be quite useful if we could detect it at an early stage. However, it is difficult to detect which topics will become the new buzzwords.

Most people become familiar with buzzwords as they are introduced by TV, news or magazines after they have become popular. However, there is often a core set of enthusiasts who talk about these buzzwords before they become popular to the masses. By focusing on this core set of people, we can find potential buzzword candidates.

We categorize buzzwords into two types by their growth patterns: "bursty buzzwords" and "gradual buzzwords." Bursty buzzwords attract people's attention suddenly and become known simultaneously throughout multiple communities, probably because they are featured on TV, news or magazines. Gradual buzzwords begin from a restricted community, spread little by little to other communities, and finally become widely known to most people.

Figure 1 shows the curve variation of blogger numbers for these two types of buzzwords. Some researchers have aimed at extracting bursty buzzwords [3, 16, 13, 14, 11, 2, 1]. However, in this paper we focus on early detection of gradual buzzwords.

In our research, we focus on blogs for the medium of buzzword analysis. Different from the mass media, blogs are an information source in which users can express their ideas, interests, and feelings in real time. Since blogs reflect the people's concern and public opinions, they are a significant source of current trend data. For example, many companies analyze blog entries to determine the opinions of their products. We argue that blog analysis can help detect the forerunners of buzzwords before they become widely known in the world. We use blog data consisting of 81,922,977 blog entries from 3,776,154 blog websites in the past two years provided by the company kizasi.jp [1].

[1] http://kizasi.jp/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA.
Copyright 2012 ACM 978-1-4503-1335-3/12/06 ...$10.00.

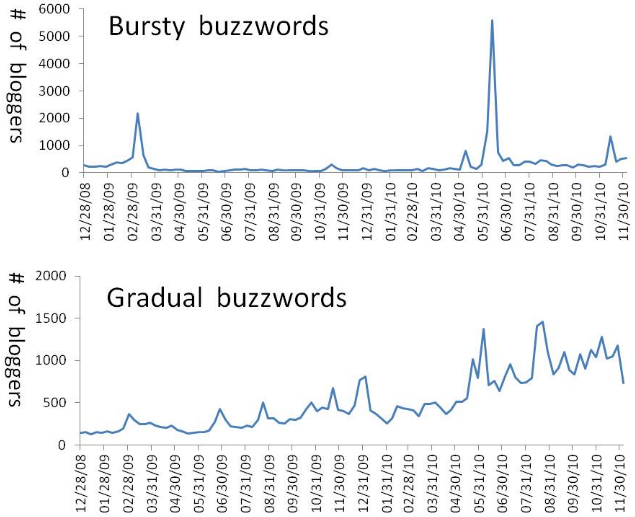

Figure 1: Comparison of curve shapes between bursty buzzwords and gradual buzzwords

We judge whether a topic can grow to become a major buzzword by analyzing the time-series variation of blog entries related to the topic. We focus on two particular features: (1) the increase in the number of bloggers mentioning a topic, and (2) the growth of the topic's popularity between different blogger communities, particularly from restricted blogger communities to general blogger communities. Our previous research [12] extracted 122 blogger communities based on bloggers' interests. These blogger communities are used for the analysis of the growth of topic popularity between communities.

In order to establish the analysis method for early detection of buzzwords, we pick up several major buzzwords from "Words of the Year [2]," observe how they spread throughout the world, and point out what is necessary for a practical application system.

The remainder of this paper is organized as follows. Section 2 presents the feature of the gradual buzzwords. Section 3 describes the formation of blogger communities. Section 4 analyzes the buzzword growth between blogger communities. Section 5 summarizes the necessary functions of a practical system that can detect buzzwords at an early stage. Section 6 reports an experimental result based on the rankings of buzzword candidates. Section 7 reviews related work. Finally, we conclude the paper in Section 8.

2. FEATURE ANALYSIS OF BUZZWORDS

In this paper, we focus on gradual buzzwords–topics that are initially only talked about in a restricted community and gradually spread to other communities. We are not considering bursty buzzwords, which are topics that suddenly become notable in most communities and only last for a short period. For the gradual buzzwords, two features are observed: (1) the increase in the number of bloggers mentioning a topic and (2) the growth of the scope of bloggers. In this section, we describe these two features using two examples of topics that have become buzzwords.

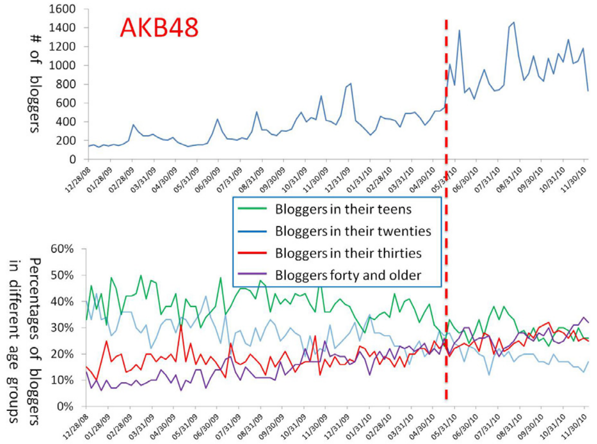

Figure 2: Numbers of bloggers and percentages of bloggers in different age groups (Topic: AKB48)

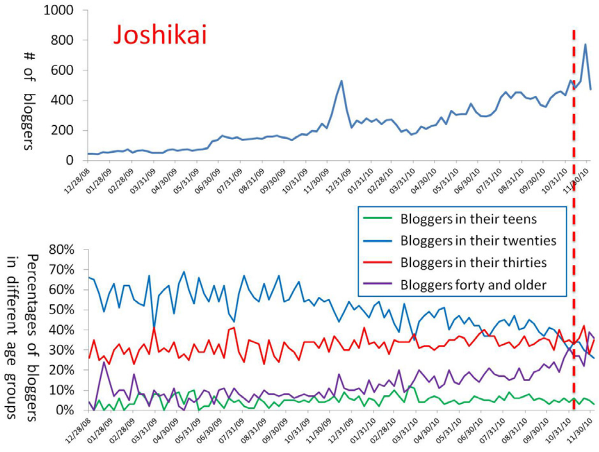

Figure 3: Numbers of bloggers and percentages of bloggers in different age groups (Topic: Joshikai)

2.1 Increase in the number of bloggers

When detecting a new buzzword, the increase in the number of bloggers who mention a topic is the first feature we observe. By definition, a buzzword must be well known, and thus we observe if there is an increase in the number of bloggers who mention the topic. Figure 2 shows the variation of the numbers of bloggers and the percentages of bloggers in different age groups who have talked about the topic "AKB48 [3]" during the period from December 28, 2008 to November 30, 2010. Figure 3 shows the variation for the topic "Joshikai [4]" during the same period. In the

[2] "Words of the Year" is an annual event in Japan that elects some words reflecting the social trends in that year.

[3] "AKB48" is a Japanese female idol group consisting of 48 young members.

[4] "Joshikai" is a Japanese term with the meaning of a girls-only party or event.

upper graphs of two figures, the horizontal axis represents the dates and the vertical axis represents the numbers of bloggers who posted blog entries including the topic keywords. Both “AKB48” and “Joshikai” are the topics that became well known by 2010. From the graphs, it can be observed that these two topics were not talked about by many bloggers in the beginning (December 2008) and after that the numbers of bloggers who talk about them increased gradually.

A potential method for extracting buzzwords is to detect the gradual increase in the number of bloggers. However, although the numbers of bloggers increase on the whole, we also observe that there are repeated cycles of small increases and decreases during this process. Therefore, it is difficult to extract gradual buzzwords at an early stage by only focusing on the increase in blogger numbers. Moreover, even if the numbers of bloggers who talk about a topic (e.g., “AKB48”) increase, it is unknown whether the topic has become a major buzzword which can be recognized by most people, or if it just restricted to a specific community (e.g, idol fans). Therefore, we also analyze the growth of popularity between different blogger communities in the next section.

2.2 Growth of the scope of bloggers

As described in Section 2.1, it is not easy to detect gradual buzzwords only focusing on the increase in the number of bloggers. Therefore, we also observe whether a topic has been recognized by multiple communities. The communities can be blogger groups of different sexes, different ages, different regions, or blogger groups in which bloggers have same interests (e.g., politics, diet, soccer, stock, etc.). The topical feature of a gradual buzzword is that although initially, it is talked about only in a restricted community, it gradually spreads to most communities. Analyzing the growth of a topic to multiple communities is important for detecting the forerunners of gradual buzzwords.

In the lower graphs of Figure 2 and Figure 3, bloggers are divided into four groups by their ages: people between ten and twenty years of age (for simplicity’s sake, we will refer to them as “teens”), people in their twenties, people in their thirties, people forty and older. The horizontal axis represents the dates and the vertical axis represents the percentages of bloggers in each different age group who have talked about the topic. On each date, the sum of the percentages of bloggers in different age groups is 100%.

From the lower graph of Figure 2, we can observe that “AKB48” was mainly talked about by teen bloggers in 2008 and 2009. Actually, “AKB48” is an idol group that is originally popular only within the younger generation. However, by May 2010, the percentages of bloggers in their twenties, thirties and forties have caught up with that of the teen bloggers. The lower graph of Figure 2 reveals that “AKB48” has gradually become known to other age groups. From the lower graph of Figure 3, we can observe that “Joshikai” was familiar only to the bloggers in their twenties in 2008 and 2009. By October 2010, the percentages of bloggers in their thirties and forties age groups have caught up with that of the twenties age group. This indicates that “Joshikai” was initially talked about by the twenties age group and then had gradually become known to other age groups. “Joshikai” often includes drinking alcoholic beverages, and thus, it is less talked about by teen bloggers.

These two figures show that the topics that are originally dominant in a specific age group, then spread to other age groups gradually. It is the typical feature of gradual buzzwords that they are recognized from a restricted community to other general communities. As shown in Figure 2 and Figure 3, May 2010 for “AKB48” and October 2010 for “Joshikai” are the turning points where they expand from a buzzword candidate to a major buzzword. It can be also observed that the numbers of bloggers who talk about these two buzzwords continue increasing even after the turning points.

As mentioned above, it is a key point for early detection of gradual buzzwords to identify the process in which a topic popular only in a specific community grows to become a common topic known to most communities. However, for the growth of buzzwords between communities, it is insufficient to only consider the groups of bloggers of different ages. In the next section, we describe the formation of blogger communities based on bloggers’ potential interests.

3. BLOGGER COMMUNITY FORMATION

Many SNS services require users to register their communities. The users select the communities that they want to take part in according to their own subjective judgment. However, it is possible that a person registers only in the “politics” community, but actually he also has an interest in “artist.”

We extract some popular topics daily discussed in the blog as the names of potential communities, and automatically categorize bloggers into their appropriate potential communities. A potential community in our research is a group of bloggers who take interest in a topic. For example, the “politics” community is the group of bloggers who have an interest in “politics.” Potential communities of bloggers are objectively identified by analyzing bloggers’ entries that they posted. Even if one does not declare his interest in a topic explicitly, if he has posted many blog entries related to the topic, our method can categorize him into the appropriate community automatically.

Next, we briefly describe the formation of bloggers’ potential communities. The details of the method can be found in our previous paper [12].

3.1 Extraction of potential communities’ names and construction of their co-occurrence dictionaries

The names of potential communities are the topics that are often talked about in blogs. The keywords matching the patterns such as “expert in *” and “fan of *” are first extracted from the Web. Then, they are filtered by their occurrence frequency and finally the most appropriate are selected manually. We end up with a list of potential communities (e.g., politics, artist, etc.), and a list of bloggers who have an interest in these topics (e.g., a group of bloggers familiar with politics, or a group of bloggers always paying attention to certain artists, etc.). As of October 2011, 122 potential communities are used as the analysis targets.

For each potential community, a co-occurrence dictionary is automatically constructed. For each keyword representing the community, we extract the top n words that have high co-occurrence frequency with it. Specifically, n is 400 in our current implementation. Figure 4 is an example of the co-occurrence dictionary. For example, the community “politics” has its domain-specific words, such as “pre-

Communities		Co-occurrence words							
i	c_i	j = 1		2		. . .		400	
1	politics	premier	$y_{1,1}$	party	$y_{1,2}$	. . .	. . .	bill	$y_{1,400}$
2	artist	entertainment	$y_{2,1}$	concert	$y_{2,2}$	. . .	. . .	album	$y_{2,400}$
3	computer	windows	$y_{3,1}$	desktop	$y_{3,2}$	. . .	. . .	hardware	$y_{3,400}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
122	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Figure 4: Example of co-occurrence dictionary

mier," "party" and "bill," and the community "artist" has its domain-specific words, such as "entertainment," "concert" and "album." The column c_i shows the names of potential communities, and each row shows their co-occurrence words and corresponding co-occurrence frequency y_{ij}.

3.2 Calculation of bloggers' degrees of interests in community topics

A blogger's score representing the degree of interest in a community topic is calculated by considering how often as well as how in-depth he has posted blog entries related to the community topic. If a blogger has an extensive use of co-occurrence words of a community, a high score is attached to him.

We first calculate $score_{c_i}(e_k)$, the score of a blog entry e_k with regard to the community c_i, as follows:

$$score_{c_i}(e_k) = \sum_{j=1}^{n} x_{ij} \cdot y_{ij} \cdot z_{ij} \qquad (1)$$

where $n = 400$ is the number of the co-occurrence words, $x_{ij} = (n - j + 1)/n$ is the weight of the jth co-occurrence word that decreases as j increases, y_{ij} is the co-occurrence frequency of the jth co-occurrence word, and z_{ij} is a binary value that indicates whether the entry e_k contains the jth co-occurrence word or not.

We next calculate $score_{c_i}(b)$, the score of a blogger b with regard to the community c_i, as follows:

$$score_{c_i}(b) = \frac{l}{n} \cdot \frac{\log(m)}{m} \cdot \sum_{k=1}^{m} score_{c_i}(e_k) \qquad (2)$$

where e_k is an entry that the blogger b posted, m is the number of entries that the blogger b has posted during a given period, $n = 400$ is the number of the co-occurrence words and l is the number of the co-occurrence words that occurred in all the entries posted by the blogger b l/n indicates the coverage ratio of the co-occurrence words that the blogger b has used. $\log(m)/m$ reduces the effect that a blogger frequently posts a large amount of entries, but most of them are the entries unrelated to the target community.

A blogger is categorized into a potential community if his score is larger than a given threshold. In addition, a blogger may be categorized into two or more communities and thus may have two or more scores. For example, if a blogger belongs to both the "politics" community and the "artist" community, he has a score representing his degree of interest in "politics" and another score representing his degree of interest in "artist."

4. ANALYSIS OF BUZZWORD POPULARITY GROWTH BETWEEN COMMUNITIES

4.1 Observation on percentage variation of blog entries from different communities

We pick up six major buzzwords that were elected to "Words of the Year" in Japan, and observe how these topics spread through blogosphere. These six major buzzwords are "AKB48," "Joshikai," "Smartphone," "Android," "Facebook," and "K-POP." Figure 5 shows their variation curves. For each buzzword, the upper graph shows the variation of the total numbers of blog entries in which the buzzword is mentioned. The observation period is two years from August 2009 to July 2011 and the numbers of blog entries are counted on a weekly span. As shown in the upper graphs, for all of the six buzzwords, the total numbers of blog entries continue increasing on the whole, which meets the first feature of gradual buzzwords described in Section 2.1.

The lower graph in Figure 5 shows the percentages of blog entries from four manually selected different communities. These four communities are manually extracted from the 122 potential communities identified in Section 3 based on our visual examination of communities with the remarkable upward tendencies or downward tendencies. It should also be noted that since a blogger may belong to two or more communities, the sum of the percentages of the four communities in every week may exceed 1. Among the four communities, there exist the communities closely related to the target buzzword. They are marked as follows:

AKB48: "Female star" and "Artist"

Joshikai: "Love/Marriage" and "Smiley"

Smartphone: "Internet" and "Cellphone"

Android: "Cellphone" and "Computer"

Facebook: "School life" and "World region"

K-POP: "Artist" and "Korean star"

It can be observed that the percentages of blog entries from the communities closely related to the target buzzword decrease from a high value to a low one, whereas the percentages from the communities weakly related to the target buzzword increase gradually. The observation results reveal the second feature described in Section 2.2: Before a topic becomes a major buzzword, a large fraction of blog entries are posted from restricted blogger communities that are closely related to the target topic. But as the target topic

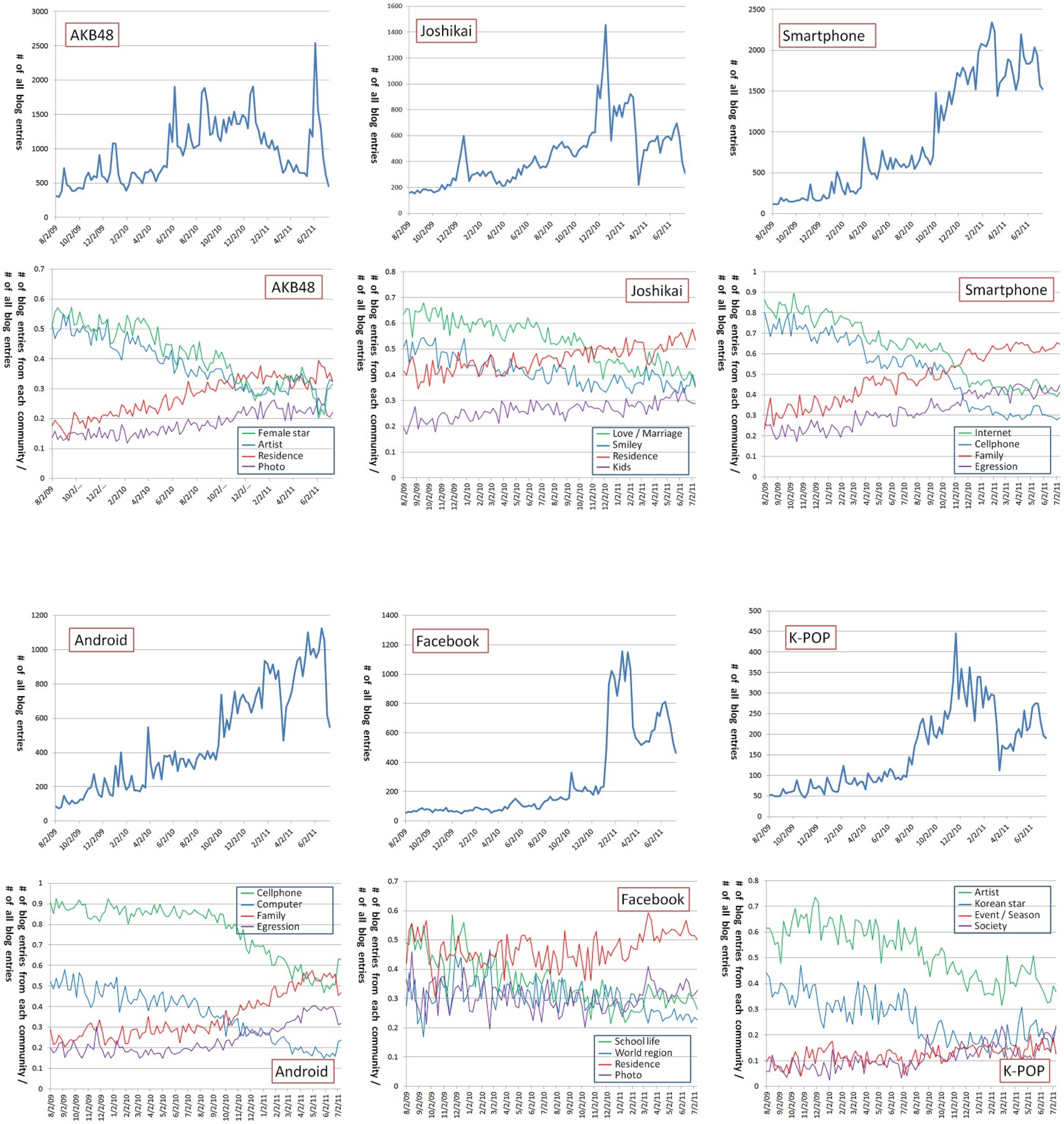

Figure 5: Numbers of blog entries including a buzzword and percentages of blog entries from different communities

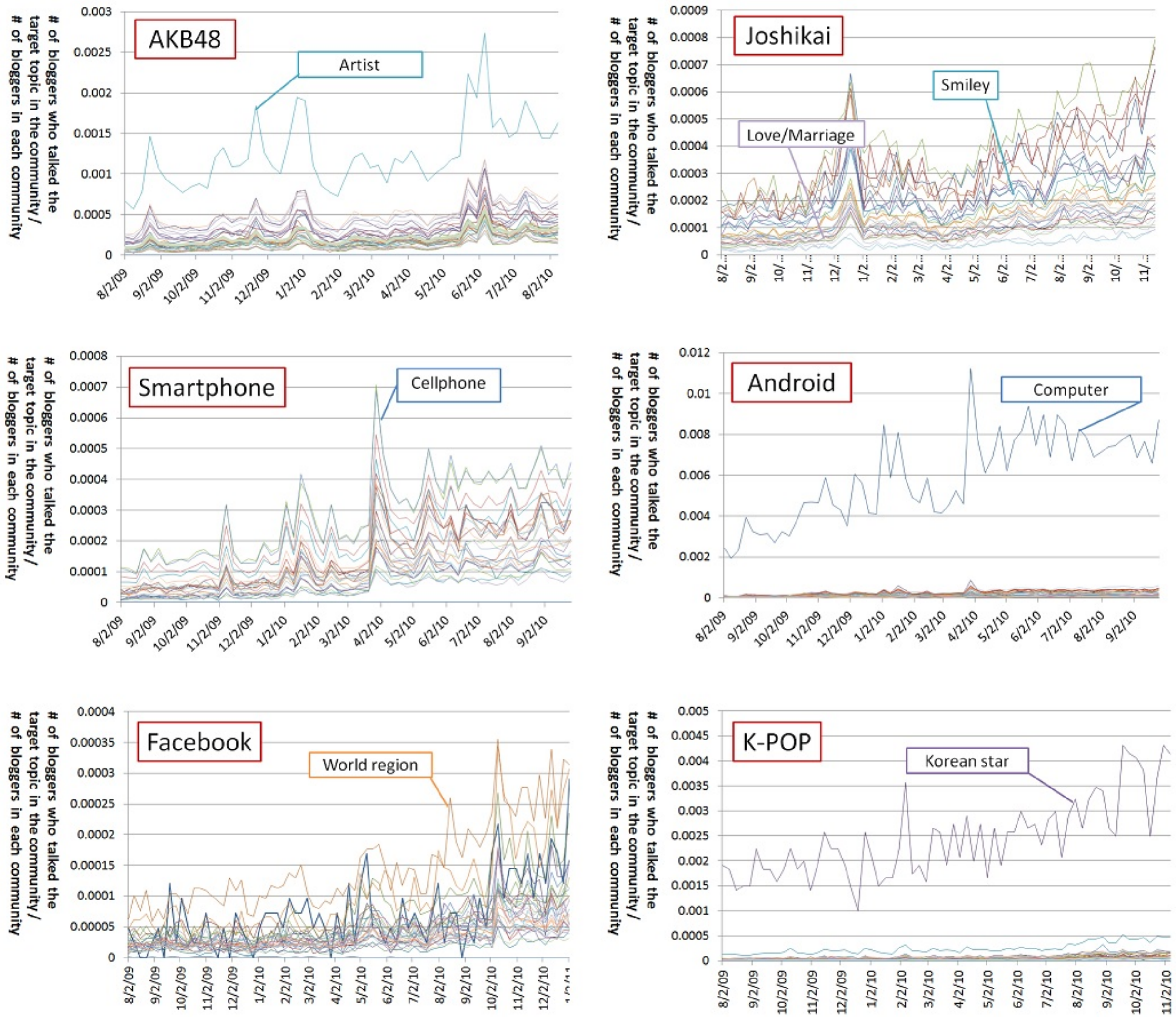

Figure 6: Ratios of bloggers in a community who talked about the target topic to all the bloggers in the community

approaches becoming a major buzzword, it extends pass the original community and is recognized by other communities. This indicates that we can detect a topic's growth process from a buzzword candidate to a major buzzword by analyzing the percentage variation of blog entries from different blog communities.

In the aforementioned examples, "Photo," "Residence" and "Family" appear repeatedly. These are communities which are weakly related to the target buzzwords. They are general groups which are indicative of the general population popularity. They also can be easily identified in advance. On the other hand, we have the communities that are more closely related to a target topic, and are more difficult to identify in advance. For the aforementioned examples, the closely related communities are manually selected. However, automatically extracting them is necessary for implementing a practical system that can early detect buzzwords. We describe a method for automatically identifying the communities closely related to a target topic in the next section.

4.2 Identification of communities closely related to target topics

The percentage of bloggers within a community who have talked about a target topic can be used to identify communities that are closely related. This is the ratio of bloggers who have talked about a target topic to the total number of bloggers within the same community A community can be judged to be related to a target topic if this ratio is large.

We also considered using the percentage of blog entries from each community, as shown in Section 4.1. However, the sizes of various communities vary widely, which means that larger communities with more entries would naturally have a larger percentage without normalization.

Figure 6 shows the ratio of bloggers who have talked about each topic in a community to the total numbers of bloggers in the community. Thirty communities with the largest numbers of blog entries including each topic are compared. In these graphs, the top community is "Artist" for the "AKB48" topic, "Cellphone" for the "Smartphone" topic, "Computer" for the "Android" topic, "World region" for the "Facebook" topic, and "Korean star" for the "K-POP" topic. These are closely related to the corresponding topics as indicated by having a much higher percentages than other communities.

Although this method can not necessarily identify all the communities closely related to target topics (e.g., for the "Joshikai" topic), most of successful examples indicate that it is effective for identifying them. We also plan to further improve this method by integrating with other techniques, for example, analyzing the overlap of co-occurrence words between a target topic and each community name.

In order to detect whether a topic has grown up a major buzzword, it is necessary to identify the communities closely related to the target topic and observe the decrease of the percentages of blog entries from these communities. In the next section, we further analyze how the curves of percentage variation of blog entries from closely related communities change.

4.3 Slope analysis of percentage variation of blog entries

In this section, we analyze the percentage variation of blog entries from different communities during the process in which that a topic grows to become a major buzzword. We then discuss what kind of conditions can indicate the high probability that it will become a major buzzword.

Figure 7 shows the variation for all the six buzzwords. The upper graphs show the increase in the total numbers of blog entries including each buzzword. Especially, the vertical bars mark the time just before the numbers of blog entries increase suddenly. It is before this point that the system must identify the buzzwords to be effective. Therefore, for the six topics, we further analyze the decline in percentage of blog entries from closely related communities up until time marked in the upper graphs. The lower graphs show the percentage decrease of blog entries from the closely related communities. The curves show the actual variation and the straight lines are their approximation based on the least-squares method [6]. The slopes of the straight lines are marked in the lower graphs. As we can see, among the twelve communities (two communities/each topic * six topics), the slopes of ten communities are less than or equal to -0.0003. Only two slopes (-0.00004 for the "Cellphone" community for the "Android" topic, and -0.000005 for the "World region" community for the "Facebook" topic) are larger than -0.0003. Although the further investigation should be done, this value may be considered a rough threshold. Moreover, the period and the scale of percentage decrease should also be further analyzed.

The similar method can be applied to the analysis of percentage increase in the numbers of blog entries from weakly related communities. It will also be further discussed in our future work.

5. MOVING TOWARDS THE IMPLEMENTATION OF THE SYSTEM FOR EARLY DETECTION OF BUZZWORDS

The ultimate goal of our research is to implement a practical system that can detect buzzwords at an early stage. It can be used to the following two applications:

1. When a user specifies a buzzword candidate, the system analyzes whether it will become a major buzzword.
2. When a user specifies a category (e.g., car or digital camera, etc.), the system provides a ranking list of the items included in the category (e.g, various car models or various camera models, etc.) based on the probabilities that they will become major buzzwords.

We summarize the necessary functions for implementing such systems as follows:

1. Automatic identification of communities closely related and weakly related to the target topics.
2. Analysis of percentage variation of blog entries from closely related communities and weakly related communities, including slope coefficients, increase/decrease periods, increase/decrease scales, etc.
3. Calculation of scores for the target topics that indicate the probability that they will become major buzzwords.

For the first function, we have demonstrated the possibility of automatic identification of closely related and weakly related communities in Section 4.2. For the second function, we have analyzed the slopes of approximate straight lines in Section 4.3 and plan to further discuss the increase/decrease periods and scales in the future. The third function is also a task for our future work as well.

6. EXPERIMENTAL EVALUATION WITH RANKING OF BUZZWORD CANDIDATES

As described in Section 5, ranking the items (buzzword candidates) in a category is a useful application, which can sort based on which one is more likely to become a major buzzword. In this section, we report an experimental result of ranking the items to a real-life idol group popularity competition.

In this idol group election, the members were selected based on fans' votes once every year during 2009 - 2011. The members with the most votes from fans were selected. We considered the three elections which were held on July 8, 2009, June 9, 2010, and June 9, 2011. We focused on the top ten people in the final third election. The actual results of the real election were as follows:

Member (1st ranking -> 2nd ranking -> 3rd ranking)

A (1 -> 2 -> 1), B (2 -> 1 -> 2), C (9 -> 8 -> 3),

D (3 -> 3 -> 4), E (4 -> 5 -> 5), F (6 -> 7 -> 6),

G (5 -> 6 -> 7), H (7 -> 4 -> 8), I (27 -> 19 -> 9),

J (out of range -> 11 -> 10)

We see if our method of ranking the members matched the elections results. For each member, we did perform the slope analysis as described in Section 4.3. The graph line for ratio of number of blog entries from a closely related community to total number of blog entries was approximated to a straight line. Then, the members were ranked based on the approximate straight lines' slopes. The member with

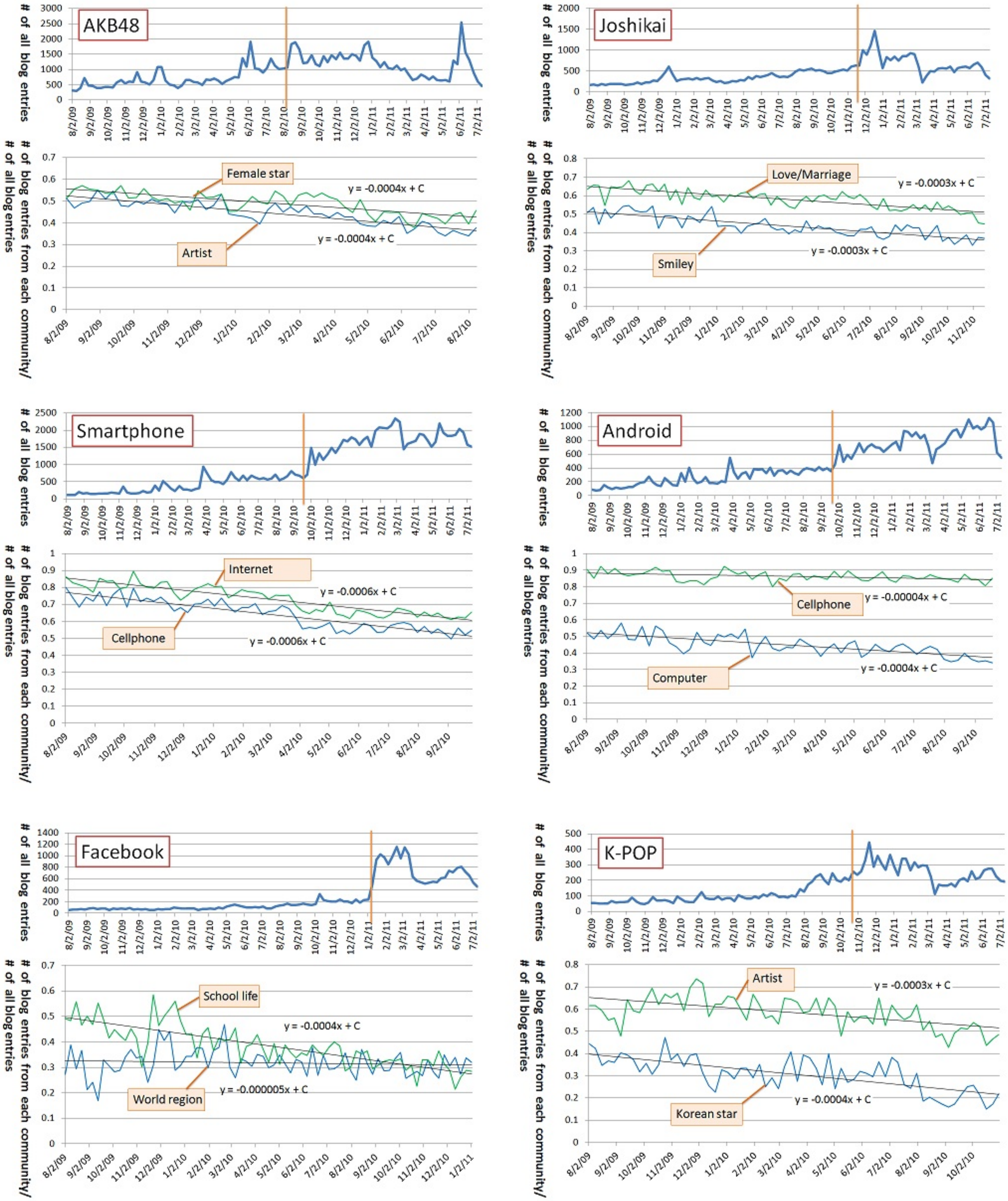

Figure 7: Approximate straight lines of percentage decrease of blog entries from closely related communities

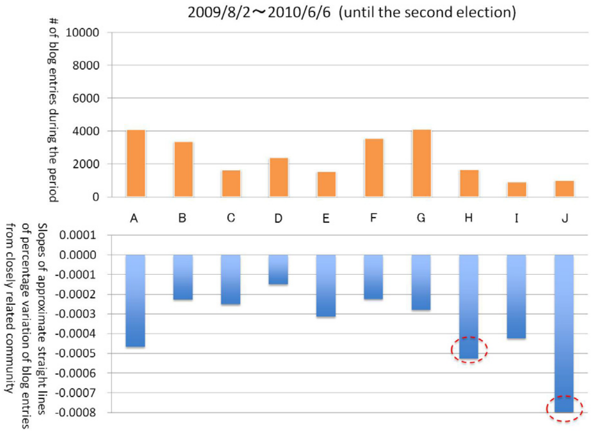

Figure 8: Slope comparison of approximate straight lines of percentage variation of blog entries from a closely related community (until the second election)

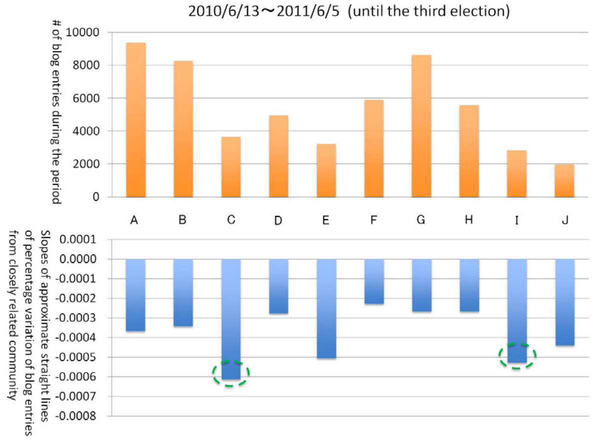

Figure 9: Slope comparison of approximate straight lines of percentage variation of blog entries from a closely related community (until the third election)

a smaller slope is more likely to become a major buzzword. For example, the slope of "-0.0008" means a steeper decrease than "-0.0002," and the member with the slope of "-0.0008" is more likely to become a major buzzword than that with "-0.0002."

This analysis of blog entries was conducted for two periods: one is after the first election and before the second election (August 2, 2009 - June 6, 2010), the other is after the second election and before the third election (June 23, 2010 - June 5, 2011). Figure 8 and Figure 9 show the analysis results for the period until the second election and for the period until the third election respectively. The upper graphs show the total numbers of blog entries in which the name of each "AKB48" member (each buzzword candidate) was mentioned, and the lower graphs show each member's slopes of percentage decrease of blog entries from the closely related community "Artist." By our method, the members with the smaller slopes are considered to be more likely to become a major buzzword. Figure 8 shows J and H were two members with the smallest slopes. In the real second election, their rankings rose sharply: out of range -> 11 and 7 -> 4. Figure 9 shows C and I had the smallest slopes. In the real third election, their rankings also rose sharply: 8 -> 3 and 19 -> 9. The results support our argument in this paper.

The upper graphs in Figure 8 and Figure 9 also show it was difficult to detect the remarkable growth of members only focusing on the increase in the total numbers of blog entries. Only analyzing the total numbers of blog entries can not discriminate whether a topic becomes popular only in a restricted community or has extended to other communities.

7. RELATED WORK

Buzzword detection is closely related to the concept of burstiness. Kleinberg [8] modeled "burst of activity" using an infinite-state automaton in which bursts appeared naturally at state transitions. By assigning costs to state transitions, the model can prevent short bursts and identify the lasting periods of bursts. Yi [16] adopted Kleinberg's model and further proposed an algorithm that detected buzzwords considering their momentum and relative duration. Araujo et al. [2] studied alternative cost functions to the one proposed by Kleinberg and investigated the best distribution of states. Parikh et al. [13] described how to detect bursts in user queries in a large-scale eCommerce system. Lappas et al. [10] built a parameter-free and linear-time approach to identify the time intervals of maximum burstiness for a given term and utilized burstiness information to enhance the search process. There also exist some practical or demonstration systems for detecting buzzwords or trends, such as Yahoo! Buzz Index [1], BlogPulse [3], Buzz of the Day [14], TwitterMonitor [11]. Yahoo! Buzz Index [1] calculates a subject's buzz score based on the percentage of Yahoo! users searching for that subject on a given day, and identifies "leaders" (subjects with the highest buzz scores) and " movers" (subjects with the highest percentage increase in buzz scores from one day to the next). BlogPulse [3] extracted key phrases and key people from blog entries by calculating the ratio of the frequency of occurrence of a phrase or a person name to its average frequency over the past two weeks. TwitterMonitor [11] performed trend detection over the Twitter system by detecting bursty keywords, grouping them as a trend and depicting the evolution of its popularity. These mainly seek to detect bursty buzzwords, the first type categorized in Section 1, by analyzing term frequency, user popularity or time varying patterns, etc. Our research aims to extract gradual buzzwords, the second type categorized in Section 1, by analyzing the growth of topics between communities.

Kumar et al. [9] extracted bursts of activity within blog communities. Gruhl et al. [4] studied the dynamics of information propagation from individual to individual within blog communities. Different from the information propagation between bloggers within a blog community in their research, we focus on the topic growth from restricted communities to other general communities. Also, the communities in their research were identified by links between blogs, whereas the communities extracted in our research are based on bloggers' interests in topics. Based on a document clustering method considering similarity and novelty [7], Ishikawa et al. developed a system to visualize the transition of topics extracted from news articles [5]. Takamura et al. [15] proposed a method for document stream summarization, solving the problems that (1) similar documents can mention different events if they are temporally distant and (2) documents on a single topic can be posted with some temporal delay. Their idea will help us identify the documents related to target topics so as to improve the accuracy of our system.

8. CONCLUSIONS

In this paper, we discussed what is necessary for early detection of buzzwords by analyzing large-scale time-series data of blog entries. More specifically, we analyzed how some topics that had become major buzzwords spread based on 81,922,977 blog entries from 3,776,154 blog websites posted in the last two years. As a result, we found that during the process in which a buzzword candidate becomes a major buzzword, the percentages of blog entries from the communities closely related to the target topic decreased gradually, whereas the percentages of blog entries from the weakly related communities increased gradually. Based on the observation and analysis results, we discussed the necessary functions of the system for early detection of buzzwords. Moreover, the experimental results of ranking comparison between ten buzzword candidates verified the feasibility of our proposition.

In the future, we will further establish the methods for automatically extracting closely related and weakly related communities, and an equation for calculating the probability that a buzzword candidate will become a major buzzword. Also, more evaluation experiments will be conducted for other categories. Finally, we will implement a practical system that can extract buzzwords at an early stage.

9. ACKNOWLEDGMENTS

This work was supported in part by the MEXT Grant-in-Aid for Scientific Research(C) (#23500140).

10. REFERENCES

[1] Yahoo! Buzz Index. http://buzzlog.yahoo.com/overall/

[2] L. Araujo and J. J. M. Guervos, "Automatic Detection of Trends in Time-stamped Sequences: An Evolutionary Approach," *Soft Computing*, Vol. 14, No. 3, pp. 211–227, 2010.

[3] N. S. Glance, M. Hurst and T. Tomokiyo, "BlogPulse: Automated Trend Discovery for Weblogs," In *WWW 2004 workshop*.

[4] D. Gruhl, R. Guha, D. L-Nowell, A. Tomkins, "Information Diffusion Through Blogspace," In *WWW 2004*.

[5] Y. Ishikawa and M. Hasegawa, "T-Scroll: Visualizing Trends in a Time-Series of Documents for Interactive User Exploration," In *ECDL 2007*.

[6] T. Kariya and H. Kurata, "*Generalized Least Squares*," Wiley, 2004.

[7] S. Khy, Y. Ishikawa and H. Kitagawa, "A Novelty-based Clustering Method for On-line Documents," In *World Wide Web Journal*, Vol. 11, No. 1, pp. 1–37, 2008.

[8] J. M. Kleinberg, "Bursty and Hierarchical Structure in Streams," In *SIGKDD 2002*.

[9] R. Kumar, J. Novak, P. Raghavan and A. Tomkins, "On the Bursty Evolution of Blogspace," In *WWW 2003*.

[10] T. Lappas, B. Arai, M. Platakis, D. Kotsakos and D. Gunopulos, "On Burstiness-aware Search for Document Sequences," In *SIGKDD 2009*.

[11] M. Mathioudakis and N. Koudas, "TwitterMonitor: Trend Detection over the Twitter Stream," In *SIGMOD 2010*.

[12] S. Nakajima, J. Zhang, Y. Inagaki, T. Kusano and R. Nakamoto, "Blog Ranking Based on Bloggers' Knowledge Level for Providing Credible Information," In *WISE 2009*.

[13] N. Parikh and N. Sundaresan, "Scalable and Near Real-Time Burst Detection from eCommerce Queries," In *KDD 2008*.

[14] N. Parikh and N. Sundaresan, "Buzz-Based Recommender System," In *WWW 2009*.

[15] H. Takamura, H. Yokono and M. Okumura, "Summarizing a Document Stream," In *ECIR 2011*.

[16] J. Yi, "Detecting Buzz from Time-Sequenced Document Streams," In *EEE 2005*.