ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 2, No.14 , March 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Discovery of Potential Topics from Blog Articles by Machine Learning

Yoshiaki YASUMURA[1], Yuhei KOSAKA[2] , Hiroyoshi TAKAHASHI[2] and Kuniaki UEHARA[2]

[1]College of Engineering, Shibaura Institute of Technology, Japan
Saitama City, Japan
yasumura@shibaura-it.ac.jp

[2] Graduate School of System Informatics, Kobe University, Japan
Kobe City, Japan
kosaka, hiroyoshi, uehara@kobe-u.ac.jp

Abstract

This paper presents a method for potential topic discovery from blogsphere. We define a potential topic as an unpopular phrase that has potential to become a hot topic. To discover potential topics, this method builds a classifier to detect potentiality of a topic from topic frequency transitions in blog articles. First, this method extracts candidates of potential topics from categorized blog articles because categorization enables us to extract specialists. To extract potential topics from the candidates, a classifier for detecting potential topics is built from topic frequency transition data. For this learning, we propose two types of learning methods: supervised learning and semi-supervised learning. Though supervised learning provides more precise results, it requires enormous size of labeled data. Creating labeled data is costly and difficult. On the other hands, semi-supervised learning can build classifier from small size of labeled data and a lot of unlabeled data. Experimental results with real blog data show the effectiveness of the proposed method.

Keywords: *Web Mining, Machine Learning, Blog.*

1. Introduction

Blog is a media that individual can easily provide commentary and news on a particular subject. Since the number of bloggers increases rapidly, a lot of blog articles are updated daily. Thus, blog articles reflect the trend of the real world. This fact enables us to analyze market trend by monitoring information on blogs [1-3].

So far, a lot of methods are proposed for monitoring and analyzing information on blogs. One of the most popular methods is detecting a burst of a word in a document stream of blogs [4-9]. Since burst words are viewed as hot topics in the blogsphere, detecting burst provides market analyzer the trend in the blogsphere. However, most burst topics are not valuable information from the view point of marketing because they are already popular in the blogsphere. Valuable topics are described in a few blogs and have a potential to become a hot topic. We call such topics “potential topics”. The system that can discover potential topics as early as possible is required for marketing.

Therefore, in this paper, we develop a system for discovering potential topics from the blogsphere. This system is based on machine learning and blog categorization. A machine learning technique creates a predictor for discovering potential topics. This predictor is built from the topic frequency transition in the blogsphere. Blog categorization extracts specialists who can describe potential topics earlier in their blogs.

This system first extracts candidates of potential topics by filtering general phrases. Next, a predictor for detecting potential topics is built by learning from the data of topic frequency transition in the blogsphere. Finally, the predictor shows users potential topics candidates. In this paper, we propose two types of learning methods: supervised learning and semi-supervised learning.

2. Potential Topic Discovery

In this section, we present key ideas of our method. First, we describe potential topics and their usefulness. Second, we present key ideas for discovering potential topics.

2.1 Potential Topics

Valuable information for marketing can create or capture new demand. The system that can detect such information as early as possible is required for marketing. One of the systems for this purpose is hot topic extraction by detecting a burst of a word in the blogsphere. A burst of a word is defined as sharp increase in frequency of the word. However, most burst topics are not valuable information because they are already popular. Fig.1 shows an example of burst detection. This graph charts the topic frequency transition of the phrase ``subprime loan problem'' in the

blogsphere. From the graph, this phrase is first described in blogs in March, 2007. After that, the phrase increased in frequency gradually, and burst in October, 2007. If burst of the phrase is detected at that time, it is not valuable information. This is because subprime loan problem was already reported on TV and newspaper, and stock price began to fall at that time. However, the phrase ``subprime loan problem'' is valuable information if it was detected before burst. Such a valuable phrase is called a potential topic. We define a potential topic as an unpopular phrase that has potential to become a hot topic. In this paper, we develop a system for discovering potential topics as early as possible.

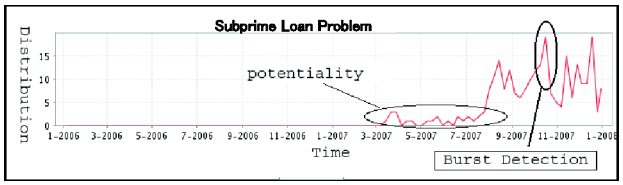

Fig.1 An example of burst detection

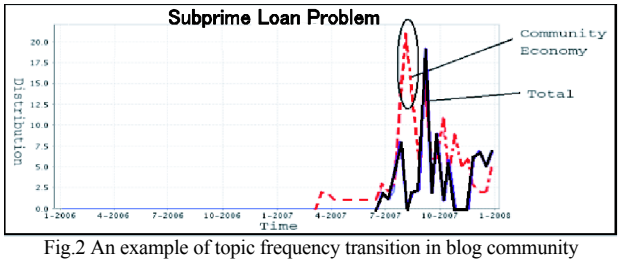

Fig.2 An example of topic frequency transition in blog community

2.2 Key Idea

Our method is based on two key ideas. The first idea is to build a predictor for discovering potential topics by machine learning technique. The second idea is focusing on specialists who can describe potential topics earlier.

In order to discover the hot topics earlier, we try to detect potentiality of a phrase by analyzing the topic frequency transition before burst. To do this, we built a predictor for detecting potential topics. The predictor learns from the topic frequency transition in the old data of the burst topics. Fig. 1 shows an example of potentiality of a topic. The topic frequency before bursting gradually increased. This is potential to become a hot topic. For creating the predictor, this method learns from topic frequency transitions in blog articles.

In order to discover potential topics as early as possible, we extract specialists who can describe potential topics of the category earlier in their blogs. To extract them, we classify blog articles into some category. By analyzing the topic frequency transition in each category, we can detect potential topics earlier. Fig.2 shows an example of topic frequency transition of the phrase ``subprime loan problem'' in the economy category and the total of blogs. From Fig 2, the bloggers in the economy category first described the phrase in the blogs and the burst in the economy category is detected before bursting in the total of the blogs. This example shows that blog categorization is effective for discovering potential topics earlier. This is because the bloggers in a category are the specialists of the category. Thus, we attempt to discover potential topics by using blog categorization.

Based on the above ideas, we develop a system for discovering potential topics. Fig. 3 presents an overview of our system. The system discovers potential topics by the following procedure.

1. The system classifies blog articles into some categories to extract specialists in the category.
2. The system extracts topics as candidates of potential topics from the blogs in each category.
3. The system filters the extracted topics by using Document Frequency value (DF) because some extracted topics are too general such as “Christmas”.
4. The system determines whether the candidate topics are potential topics or not. For this decision, a predictor is built by learning from the topic frequency transition.

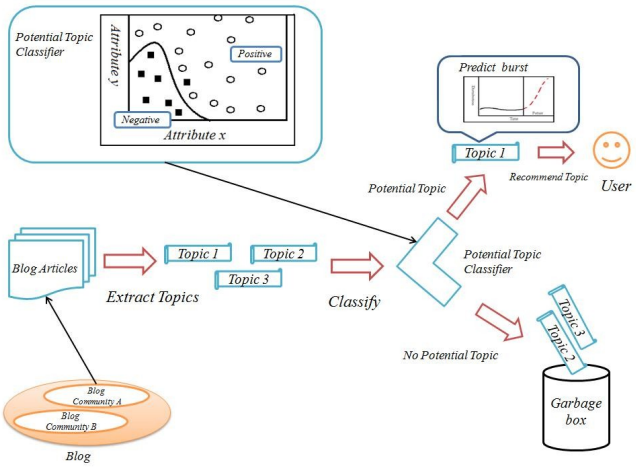

Fig.3 Overview of the proposed method

3. Potential Topic Discovery by Learning

In this section, we present a method for discovering potential topics. First, the system classifies blog articles into some categories for extracting specialists. Second, the system extracts candidates of potential topics from blog articles. Finally, the system detects potential topics from candidates by machine learning.

3.1 Blog categorization

Here, we describe blog categories and a method for categorization. Bloggers describes commentary and news in their blogs. The contents of the blogs depend on the blogger's preferences. For example, the blogger who likes football often describes articles on football, and the articles usually contain details on football. Thus, the bloggers that have common preference is viewed as specialists of the category. For this reason, we classify blogs into some categories.

The classifier for this categorization is built by machine learning. The training data for this learning is manually created as the articles labeled their category. Table 1 shows the categories used in this system. In the system, a blog article can be labeled multi-category. The system uses Naive Bayes as a classifier for blogs.

In this system, we classify bloggers into categories according to their blogs. If the rate of the blogs classified into a particular category is over the threshold, the blogger is classified into the category.

Table 1: Blog Category

Parent Category	Subcategory
Politics Economy	Politics, Economy, News
Sport	Baseball, Football, Golf
Music	Pops, Jazz, Classical music
Food	Restaurant, Recipe, Food
Entertainment	Movie, TV program, Entertainer
Art	Literature, Picture, Fashion
Vehicle	Car, Train, Motorcycle, Bicycle
Gamble	Horse Racing, Casino
Life	School, Family, Love, Business
Pet	Dog, Cat, Gardening
Technology	Computer, Internet, Science
Hobby	Toy, Military, Cartoon, Game

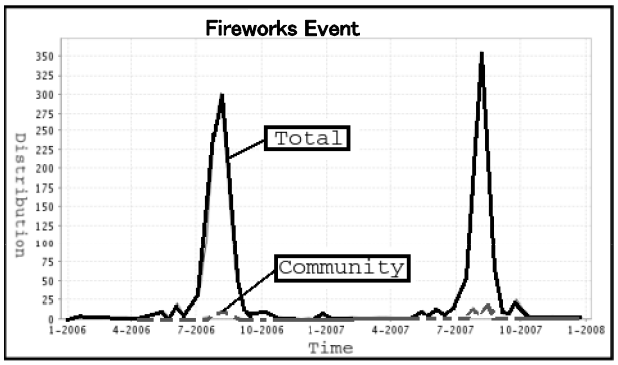

Fig.4 Frequency transition of "fireworks Event"

3.2 Filtering General Term

In this system, candidates of potential topics are nouns and simple noun sequences. So the system collects all nouns and simple noun sequences as candidates of potential topics. However, the candidates are too many for discovering potential topics effectively. To reduce the candidates, the system eliminates general terms from the candidates. For example, Fig.3 shows an example of the general phrase "fireworks event". Since fireworks events are usually held in summer, the phrase bursts every summer. However, detecting the burst of the phrase is not valuable information, because everyone knows the fact.

For eliminating general term, the system uses DF (Document Frequency) value. If a term has high DF value, it appears in many documents. So the system eliminates terms that have high DF value. By eliminating general terms, we can obtain unpopular terms (e.g. subprime loan problem) as candidates of potential topics.

3.3 Data for Learning

From the candidates of potential topics, the system extracts the topics that have potential to burst. The candidates of potential topics extracted by the above procedures consist of potential topics and low-frequency terms. A low-frequency term is a word that described in few blogs or in a small community of blogs such as ``Federal Open Market Committee''. Since low-frequency terms are not valuable for marketing, the system classifies the candidate terms into potential topics and low-frequency terms.

For the classification, we build a classifier for detecting potential topics based on the topic frequency transition. The classifier is built by machine learning technique from manually labeled training data. For building the classifier, we choose attributes of training data by considering the

assumption that a potential topic should have the following features.

- The potential topic is described not only in a particular category but also in the other categories.
- The frequency of the potential topic increases gradually.
- The potential topic continuously appears in the blogs.
- It is not long from the first time the potential topics appear in the blogs.

Considering the above features of potential topics, we choose attributes for classification as below.

- The frequency of the topic in that day, the three days before, the seven days before, and thirty days before.
- The number of continuously appearance of the topic.
- The number of the day from the first appearance of the topic.
- The total number of the bloggers who described the topic.
- The number of the increased bloggers from the three days before, the seven days before, and the thirty days before.

Instances in the training data are labeled by detecting burst. If burst is detected, the instance is positive. If burst is not detected, the instance is negative.

3.4 Classifier Built by Supervised Learning and Semi-Supervised Learning

From the created data from blog articles, the system built a classifier for discovering potential topics. Training data is created from old data of blog articles. In this paper, we propose two types of learning method: supervised learning and semi-supervised learning.

In supervised learning setting, a classifier is built from the training data that all instances are labeled. The classifier built by supervised learning can predict potential topics more precisely. However, manually labeling training data is costly and difficult. Fig.5 shows an example which is difficult to label. This graph shows topic frequency transition. We can label a few day before bursting as positive (potential topic). On the other hand, we can label days that have low frequency as negative (not potential topic). However, it is hard to label the days between them. Hence we adopt semi-supervised learning for building a predictor of potential topics. In this learning, difficult instances are dealt as unlabeled instances. Using semi-supervised learning, labeling cost is reduced.

For semi-supervised learning, we adopt Tri-Training [10] that builds three classifiers and classifies by majority vote. This learning method first creates three datasets by bootstrap sampling of labeled data and builds classifiers from them. The built classifiers classify unlabeled instances one another. Then the classifiers are built from the data that contains unlabeled instances labeled by the other classifiers. All unlabeled instances are labeled by repeating this procedure.

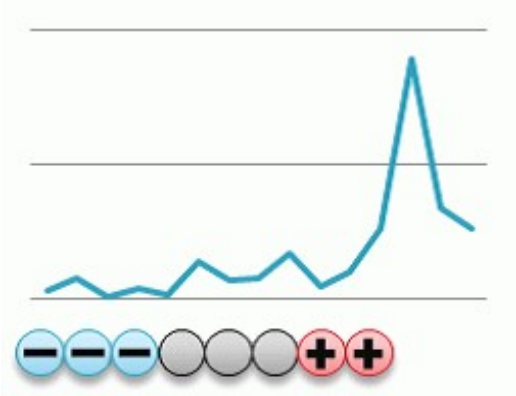

Fig. 5 Semi-supervised learning for potential topic detection

4. Experiment

We conducted evaluation experiments for assessing the effectiveness of our system. In this experiment, we evaluate our system by using actual blog data.

4.1 Experimental Setting

For this experiment, we collect blog data from January 2008 to December 2010. The blog data contains about 200,000 articles written by 2496 bloggers. Training data is the manually labeled topics. The number of the positive topics is 40. The number of the negative topics is about 3500. The rest of the instances are unlabeled. In this experiment, we use the C4.5 decision tree as a base learner for detecting potential topics. We evaluate our method by 10-fold cross-validation. Since the data is divided into subsets based on topics for cross-validation, the instances of the same topic are included in the same subset.

4.2 Experimental Results of Supervised Learning

We evaluate our predictor with precision and recall measures. The precision of prediction is 78.4% and the recall is 83.4%. This result indicates that our method can detect potential topics at higher rate, and that our system is effective for discovering potential topics. Fig.6 and Fig.7 show examples that potential topics are detected by our method before burst. Fig. 6 shows the topic frequency transition of the phrase "Billy's boot camp", which is an exercise program developed by Billy. Fig. 7 represents the topic frequency transition of the phrase "Hatsune Miku", which is singing synthesizer application software. In these examples, the frequency of the topics is gradually increased and then burst. The system can detect the topics like these. Fig.8 shows an example that potential topics are not detected by our method. Fig.8 represents the topic frequency transition of the phrase "Zero Interest Rate", which is the policy of Bank of Japan in 2006. In this case, the frequency of the topic is low and suddenly burst. It is difficult for the system to detect the topics like this. This topic is introduced in TV, and the bloggers watching the TV program describe the topic word in their blogs. The topic that the system cannot detect is shown in Fig. 9. The topic "Chuetsu Earthquake" is burst at the first appearance. The topic is a bursting topic but not a potential topic, because the topic has no potentiality. This is because the topic is reported in the mass media first. However, the topics reported in the mass media are not valuable information for marketing.

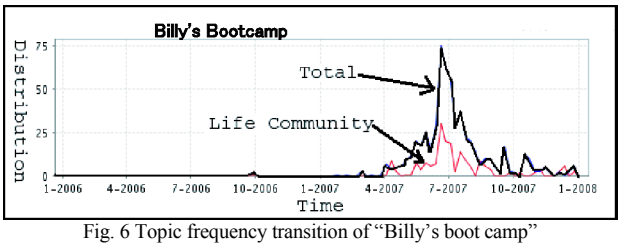

Fig. 6 Topic frequency transition of "Billy's boot camp"

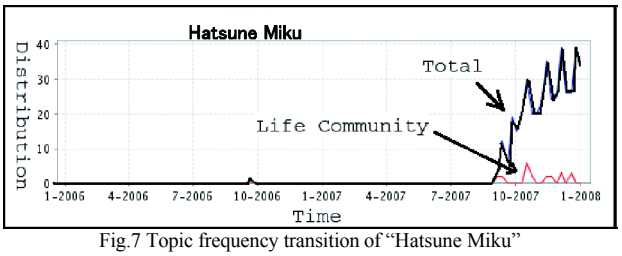

Fig.7 Topic frequency transition of "Hatsune Miku"

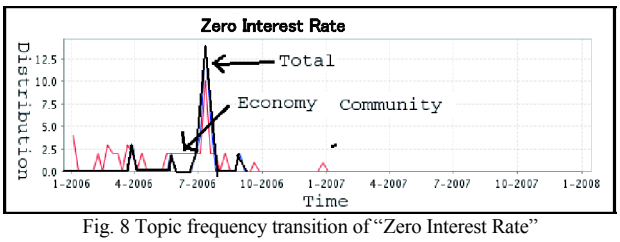

Fig. 8 Topic frequency transition of "Zero Interest Rate"

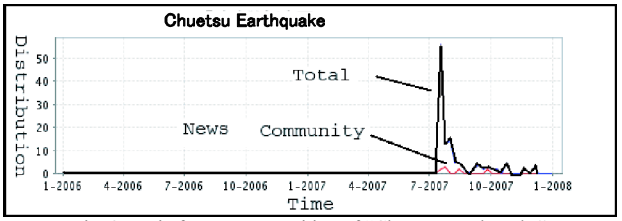

Fig. 9 Topic frequency transition of "Chuetsu Earthquake"

4.3 Experimental Results of Semi-Supervised Learning

We evaluate our predictor with precision, recall and F-measure. Fig. 10 shows the result according to the size of positive instances. In the figure, 1 means that one day before bursting is labeled as positive and 7 means that seven days before bursting is labeled as positive.
From the figure, recall is almost 0.9 in the case that the size of positive instances is over 10. This result show that the system can discover almost potential topics if sufficient size of positive instances are available. However, precision is almost 0.2 in the most cases. From this result, the system extracted some negative instances as potential topics. This mistake is classified into two types. One type is that the system simply mistakes in labeling. The other

type is that the system extracts potential topics that never burst. The latter case is rare, but it has valuable information. For practical use, recall is more important than precision, because the system shows the users the potential topics and the users decide to use the potential topics at last.

Fig. 11, Fig.12 and Fig. 13 show the results of the topic "monster hunter", which is famous game, in the case that the positive label size is one, ten and thirty, respectively. In Fig. 11, the prediction is scattered because the learning is not well. This is caused by insufficiency of positive instances. In Fig. 12, the system predicted well, and the positive instances are just before bursting. In Fig. 13, the system predicted more instances as positive. From these results, the size of positive instances is important for predicting.

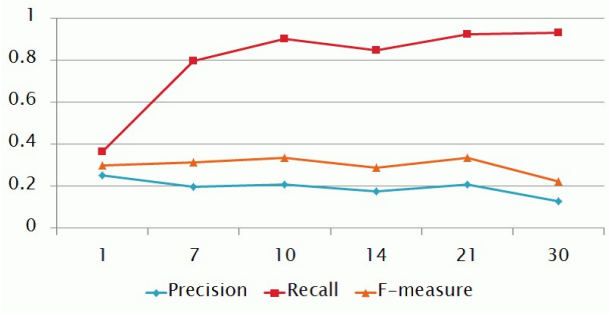

Fig. 10 Precision and recall of semi-supervised learning

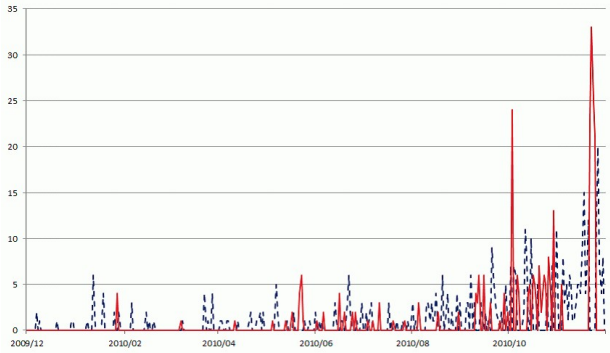

Fig. 11 The result example: positive size is one

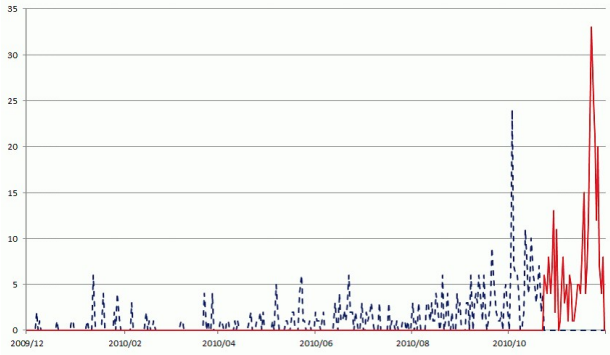

Fig. 12 The result example: positive size is ten

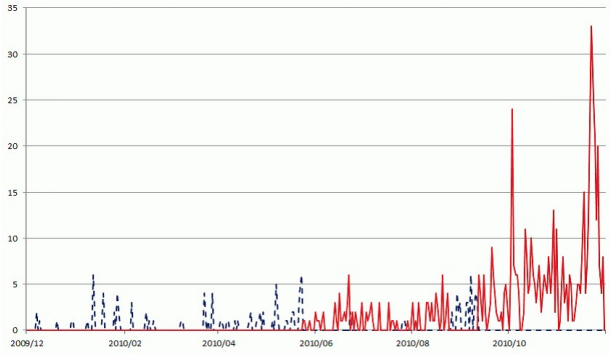

Fig. 13 The result example: positive size is thirty

5. Conclusions

In this paper, we proposed a method for discovering potential topics earlier by learning and blog categorization. Blog categorization extracts specialists who can describe potential topics earlier. From the topic frequency transition in categories, the system discovers the potential topics. We conduct experience using actual blog data. The results of supervised learning setting show that we obtained higher precision and recall for predicting potential topics. This is because labeled data enables us to create a more precise predictor. On the other hands, the classifier built by semi-supervised learning achieves higher recall and lower precision for predicting potential topics. In practical use, recall is more important than precision because users can make decision on marketing analysis from many potential topic candidates. This result indicates that our system is effective for potential topic discovery.

For future work, we try to raise the precision of potential topic discovery.

References

[1] N. Agarwal, H. Liu, L. Tang, P. S. Yu: Identifying the influential bloggers in a community, In Proc. of the International Conference on Web search and Web data mining, pp.207-218 (2008).

[2] Y. Cheng, G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, C. Chen: Model bloggers' interests based on forgetting mechanism, Proc. of the international conference on World Wide Web, pp.1129-1130, (2008).

[3] X. Ding, B. Liu, P. S. Yu: A holistic lexicon-based approach to opinion mining, Proc. of the international conference on Web search and web data mining, pp.231-240 (2008).

[4] G. P. C. Fung, J. X. Yu, P. S. Yu, H. Lu: Parameter free bursty events detection in text streams, Proc. of the international conference on very large data bases, pp.181-192 (2005).

[5] T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura: Identification of bursts in a document stream, Proc. of the

First International Workshop on Knowledge Discovery in Data Streams, pp.54-64 (2004).
[6] N. Bansal, N. Koudas: BlogScope: a system for online analysis of high volume text streams, In Proc. of the 33rd international conference on Very large data bases, pp.1410-1413 (2007).
[7] J. Kleinberg: Bursty and hierarchical structure in streams, In Proc. of the International Conference on Knowledge Discovery and Data Mining, (2002).
[8] R. Kumar, J. Novak, P. Raghavan and A. Tomkins: On the bursty evolution of blogspace, In Proc. of International World Wide Web Conference, (2003).
[9] K. Mane and K. Borner: Mapping topics and topic bursts in PNAS, In Proc. of National Academy of Sciences, (2004).
[10] Zhi-Hua Zhou and Ming Li: Tri-training: Exploiting unlabeled data using three classifiers, IEEE Trans. Knowledge Data Engineering, Vol. 17, pp. 1529-1541, (2005).

Yoshiaki YASUMURA received the PhD degree from Osaka University, Japan, in 1998. From 1998 to 2004, he was a research associate at Tokyo Institute of Technology. From 2004 to 2010, he was an assistant professor at Kobe University. He is currently a professor in the College of Engineering at Shibaura Institute of Technology. His main research interests are in machine learning, Web Intelligence, and computer vision.

Yuhei KOSAKA received the MS degree from Kobe University, Japan, in 2009. His main research interests are in Web Intelligence and machine learning.

Hiroyoshi TAKAHASHI received the MS degree from Kobe University, Japan, in 2011. His main research interests are in Web Intelligence and machine learning.

Kuniaki UEHARA received his B.E., M.E. and D.E. degrees in information and computer sciences from Osaka University, Japan.
He was an assistant professor in the Institute of Scientific and Industrial Research at Osaka University and was a visiting assistant professor at Oregon State University. Currently, he is a professor in the Graduate School of System Informatics at Kobe University, Japan. His conducting research is in the areas of machine learning, data mining, and multimedia processing.