International Journal of Web Information Systems

Twitter user tagging method based on burst time series

Shuhei Yamamoto Kei Wakabayashi Noriko Kando Tetsuji Satoh

Received 22 March 2016
Revised 19 May 2016
Accepted 19 May 2016

Twitter user tagging method based on burst time series

Shuhei Yamamoto
Graduate School of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Japan
Kei Wakabayashi
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan
Noriko Kando
Information and Society Research Division, National Institute of Informatics, Tokyo, Japan, and
Tetsuji Satoh
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan

Abstract

Purpose – Many Twitter users post tweets that are related to their particular interests. Users can also collect information by following other users. One approach clarifies user interests by tagging labels based on the users. A user tagging method is important to discover candidate users with similar interests. This paper aims to propose a new user tagging method using the posting time series data of the number of tweets.

Design/methodology/approach – Our hypothesis focuses on the relationship between a user's interests and the posting times of tweets: as users have interests, they will post more tweets at the time when events occur compared with general times. The authors assume that hashtags are labeled tags to users and observe their occurrence counts in each timestamp. The authors extract burst timestamps using Kleinberg's burst enumeration algorithm and estimate the burst levels. The authors manage the burst levels as term frequency in documents and calculate the score using typical methods such as cosine similarity, Naïve Bayes and term frequency (TF) in a document and inversed document frequency (IDF; TF-IDF).

Findings – From the sophisticated experimental evaluations, the authors demonstrate the high efficiency of the tagging method. Naïve Bayes and cosine similarity are particular suitable for the user tagging and tag score calculation tasks, respectively. Some users, whose hashtags were appropriately estimated by our methods, experienced higher the maximum value of the number of tweets than other users.

Originality/value – Many approaches estimate user interest based on the terms in tweets and apply such graph theory as following networks. The authors propose a new estimation method that uses the time series data of the number of tweets. The merits to estimating user interest using the time series data do not depend on language and can decrease the calculation costs compared with the above-mentioned approaches because the number of features is fewer.

Keywords Twitter, Burst, Hashtag, Time series data, User tagging

Paper type Research paper

Emerald

International Journal of Web Information Systems
Vol. 12 No. 3, 2016
pp. 292-311

1744-0084
DOI 10.1108/IJWIS-03-2016-0012

This work was supported by Grant-in-Aids for scientific Research No. 25280110 and No. 15J05599 and NII's strategic open-type collaborative research.

1. Introduction

Twitter, which is one of the internet's most popular social media services, had 280 million active users per month at the end of September 2014 (Twitter, 2014). On it, users post very short articles called tweets and share them with others. Many tweets provide instantaneous information during concerts, baseball games and television and reflect user interests in and reactions to new products. Twitter users can easily obtain beneficial information by following other accounts that post interesting tweets. Therefore, the research topic estimating user interests from tweets has been attracting much research attention (Yamaguchi *et al.*, 2012; Ma *et al.*, 2014; Wu *et al.*, 2010).

Many approaches estimate user interest based on the terms in tweets and apply such graph theory as following networks. We propose a new estimation method that uses the time series data of the number of tweets. The merits to estimating user interest using the time series data do not depend on language and can decrease the calculation costs compared with the above-mentioned approaches because the number of features is fewer.

Our hypothesis focuses on the relationship between a user's interests and the posting times of tweets: as users have interests, they will post more tweets at the time when events occur compared with general times. Our hypothesis is reflected in Figure 1, which shows the number of tweets on May 21, 2012, as orange lines and on the weekdays as blue dotted lines. An annular solar eclipse occurred in Japan on the morning of May 21. Many users were interested in it and concurrently tweeted. The number of tweets on May 21, 2012, exploded compared with the average number of weekday tweets. Such a state is called a burst in previous research, and we observe the burst timestamp of each user's tweets.

In our previous research (Mizunuma *et al.*, 2014), we detected many bursts and burst events by checking tweet texts and times. Some bursts have relevance to television programs, for example, *Lupin III: The Castle of Cagliostro*, *Smile PreCure!* and *Tetsuko's Room*, and bursts were caused by televised sports events as well. Justin Bieber's appearance on a Japanese television program caused a burst. In addition, a burst occurred 3 min after television news announced the arrest of Takahashi, the last Aum fugitive from the sarin gas attack on the Tokyo subway in 1995. All these examples suggest a strong association between bursts and television broadcasting. Furthermore, bursts have relevance to uncommon natural phenomenon, for example, earthquakes, bomb cyclones, tornadoes, heavy snow and heavy rain. People experiencing a disaster post their situations on Twitter and others use Twitter to disseminate information about the disaster. From these observations, we assume that burst occurrence mechanism consists of four steps as shown in Figure 2. First, any events occur. Second, users

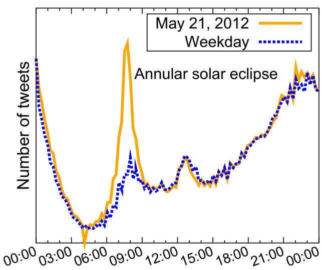

Figure 1. Number of tweets on May 21, 2012

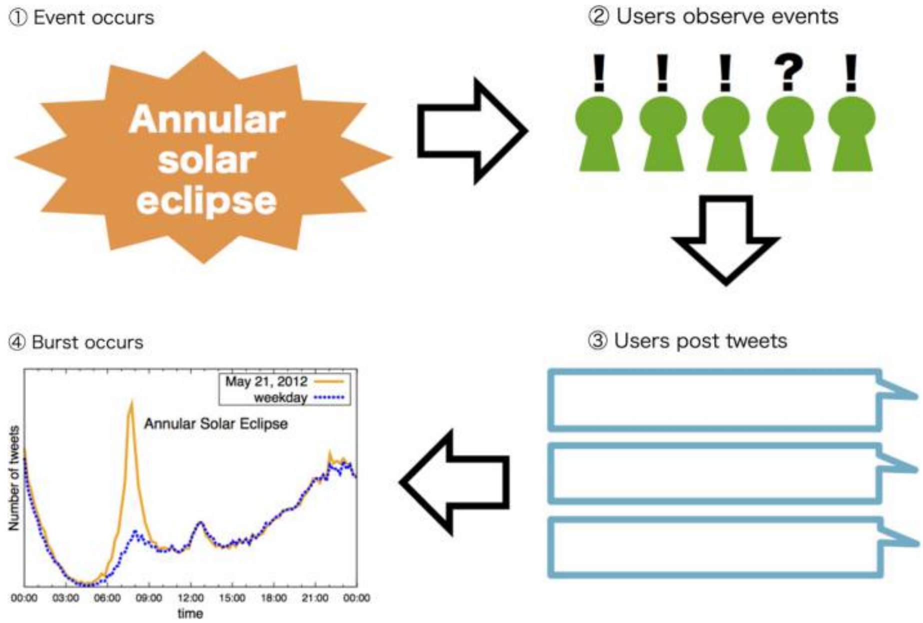

Figure 2.
Burst occurrence mechanism

observe the events. Third, users having interest to the events post many tweets. Finally, we can observe this events as bursts on Twitter.

In this paper, we propose a new user tagging method based on burst time series to tag them by interests. We assume that hashtags are labeled tags to users and observe their occurrence counts in each timestamp. Hashtags are tagged to tweets by users who select suitable hashtags for a tweet's content from among many hashtags. We modify Kleinberg's burst enumeration algorithm (Kleinberg, 2002) to accommodate Twitter. In our method, the maximum burst level is decided using the maximum and average values in the numbers of tweets, and the time series of the burst level is estimated to become the minimum cost value in each user and each hashtag. We manage the burst levels as term frequency in documents and calculate the tag scores in each user by such typical score calculation methods as cosine similarity, Naïve Bayes and term frequency (TF) in a document and inversed document frequency (IDF; TF-IDF).

The remainder of our paper is organized as follows. In Section 2, related works are discussed. In Section 3, we explain the details of the user tagging method based on burst time series. In Section 4, the experimental evaluations of the estimated tags are described using expected reciprocal rank (ERR) and Q-measure. In Section 5, we discuss our proposed method. In Section 6, we conclude our research and describe future works.

2. Related works

2.1 Bursty topic detection on twitter

To immediately detect bursty topics on Twitter, Li *et al.* (2012) detected the burst intervals whose combinations of words rapidly increased. They similarly detected the events for newly obtained document streams, calculated the similarity between these and old events and tracked them. Xie *et al.* (2013) also proposed a topic tracking method called TopicSketch to achieve the same purpose with low calculation costs. Their method detects the bursty topics by concurrently observing all Twitter streams and the documents of each term and each term's pair. Diao *et al.* (2012) detected bursty topics using Time-User-LDA, which is an extension of Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003). They evaluated the accuracy of topic detection among three LDA models and clarified that Time-User-LDA detects with the highest accuracy. Mathioudakis and Koudas (2010) extracted burst keywords from automatically collected tweets and identified trends that fluctuated in real time by creating groups using the co-occurrence of keywords. Wang *et al.* (2007) extracted bursty topics with high correlation by comparing burst patterns among different news streams for various viewpoints. Koike *et al.* (2013) extracted the bursty topics with a correlation between news streams and Twitter by applying a dynamic topic model (Blei *et al.*, 2006) and Kleinberg's burst enumerating algorithm. In our work, we estimate bursty topics using the occurrence counts of hashtags.

2.2 Tagging method for Twitter

Yamaguchi *et al.* (2012) proposed a user tagging method using Twitter lists to discover user topics. Based on their observations, they assumed that the users included identical lists probably posted on the same topic. From experimental evaluations with two data sets, their method effectively acted as a user tagging method. Ma *et al.* (2014) automatically annotated hashtags to tweets. Their probabilistic latent semantic analysis (PLSA)-style models include user, time, and tweet content factors and achieve higher precision than other methods. Wu *et al.* (2010) automatically generated personalized tags to label Twitter's user interests. They extracted keywords from Twitter messages and calculated TF-IDF and TextRank (Mihalcea and Tarau, 2004) scores for them. Huang *et al.* (2010) analyzed Twitter hashtags and reported that they are different from other social media's tags because they disappear after being used for a few days. Influenced by these stimulating previous studies, our work tags hashtags to each user just with burst time series.

2.3 User's authority estimation

Many current studies are related to the detection of popular users in Twitter. Weng *et al.* (2010) estimated user topics using LDA and detected the users who exert great influence on Twitter. They built a network for each topic based on follows and followers and calculated each user's score in each network using TwitterRank, which extended PageRank. They proposed ranking methods of users for every topic. Cha *et al.* (2010) analyzed user features with influence by comparing the number of followers, followees and replies. Those users with maximum influence wield critical power on various topics. They also clarified that influence cannot be obtained by only posting on a single topic. Using input keywords, Pal and Counts (2011) proposed a user ranking method to detect users who possess authority. They created a vector based on the features of a tweet's content, the number of retweets and retweeted and built clusters of both authority and non-authority users. Users in the authority cluster are ranked by the summation of the feature values. Yamaguchi *et al.* (2012) calculate the authority scores of Twitter users based on link analysis. They proposed TURank, which is a Twitter user ranking method, and represented them in a user-tweet graph that models information flow. Users' authority scores are evaluated using ObjectRank (Balmin *et al.*, 2004). In this paper, we estimate hashtags as user interests. Our method can estimate a user's topic authority per hashtag.

2.4 User behavior analysis on twitter

Study on behavior analysis on Twitter is flourishing. Java *et al.* (2007) clarified that the diameter of the network graph based on the follow relationship was 6 node. They also reported that 20 per cent of all tweets were conversational with @, and 13 per cent contained a uniform resource locator (URL) sent to share information. Myers and Leskovec (2014) clarified the catalyst that increases a user's followers based on bursts of retweet diffusion. They analyzed follow networks with timestamps and proposed a model for inferring new followers for each user. To effectively diffuse tweets, Wang *et al.* (2013) estimated not only user interests but also diffusion capability. They recommend the optimal address to diffuse one's own tweets. Yamaguchi *et al.* (2015) assumed that a list name plays the role of a folksonomy tag for users included in each list, and they analyzed tagging networks by using lists on Twitter. Their analysis clarified that the number of bilaterally tagging user pairs is major in friend relationships despite the number of them being minor in Twitter. Yang and Counts (2010) compared blogs and Twitter from the viewpoint of their information diffusion structures. They concluded that users who tweeted less than 30 times a month have shorter tweet intervals than blog post-intervals, and a larger number of tweets denote a smaller difference between the two intervals.

Several studies on Twitter have focused on the communication functions of replying and retweeting. Chalmers *et al.* (2011) analyzed inter-tweet intervals and tweet frequencies for all non-replies and replies. They clarified that posting intervals are different between replies and non-replies. Kwak *et al.* (2010) created retweet trees, which are composed of the connections among retweets, and analyzed the relationship between users and the distance from the retweet (RT) tree's seed. Ghosh *et al.* (2011) analyzed retweeting activity using two features, time-intervals and user entropy, and identified five retweeting categories:

(1) automatic/robotic activity;
(2) newsworthy information dissemination;
(3) advertising and promotion;
(4) campaigns; and
(5) parasitic advertisements.

Yamaguchi *et al.* (2014) analyzed transitions in posting activity on the basis of feature values such as the number of tweets, replies and retweets in each timestamp. They split users into several clusters by K-means clustering using these feature values and calculate the transition probability between clusters on the basis of sequences of cluster numbers.

3. Tag estimation method

3.1 Overview

As tags attached to users, we use hashtags that are labeled to a tweet by users based on its contents. By observing both number of user tweets and the hashtags, we can tag them without terms in user tweets. Our method is compared of two phases. First, we extract the timestamps of the burst levels from each user and each hashtag. Second, we calculate the score between the users and the hashtags using the burst timestamps. Hashtags with high score are more suitable as user tags.

Our hypothesis is that users frequently post when their interests are aroused. We can clarify the kinds of interests by observing the occurrence counts of hashtags in each timestamp because they frequently appear when events occur that are related to it. Therefore, we can tag the interests by counting the overlaps of the burst times between users and hashtags. As methods for counting overlaps, we use the three typical scoring functions: cosine similarity, Naïve Bayes and TF-IDF.

This section of our paper consists of the following parts. Section 3.2 extracts the burst times from both the tweet streams of the users and hashtags. Section 3.3 calculates the scores among users and hashtags based on burst time series.

3.2 Burst extraction

Our goal here is to extract the time series of burst level $\mathbf{b} = \{b_1, b_2, \cdots, b_T\}$ from the number of tweets series $\boldsymbol{n} = \{n_1, n_2, \cdots, n_T\}$, where T denotes the number of timestamps.

We extract the burst times based on Kleinberg's burst enumerating algorithm, which enumerates the burst times of documents with specific keywords that increase in the streams. This algorithm enumerates the burst time series by minimizing cost function $c(\mathbf{b}|\mathbf{n})$ as follows:

$$c(\mathbf{b}|\mathbf{n}) = \left(\sum_{t=0}^{T-1} \tau(b_t, b_{t+1})\right) + \left(\sum_{t=1}^{T} \sigma(n_t, \lambda_{b_t})\right),$$

where $\tau(b_t, b_{t+1})$ denote function to block the state transition from b_t to b_{t+1} because it is unnatural to discontinuously change the burst state in continuous times. $\tau(b_t, b_{t+1})$ is defined as follows:

$$\tau(i,j) = \begin{cases} (j-i)\gamma & j > i \\ 0 & j \leq i \end{cases},$$

where γ is a cost parameter to control the state transition. We chose $\gamma = 1$ based on an original paper.

Function σ gives the cost for staying at burst level b_t. In the original paper, this function was defined with a binomial distribution, because Kleinberg detected the burst times of documents in which some keywords increase. In our case, we calculated $\sigma(n_t, \lambda_{b_t})$ with a Poisson distribution, where n_t is generated based on parameter λ at burst level b_t. $\sigma(n_t, \lambda_{b_t})$ is defined as follows:

$$\sigma(n_t, \lambda_{b_t}) = -\frac{\lambda_{b_t}^{n_t} e^{-n_t}}{n_t!}.$$

We take the logarithm of both sides as follows:

$$\log \sigma(n_t, \lambda_{b_t}) = -n_t \log \lambda_{b_t} + n_t + \log n_t!,$$

where λ_0 denotes average value, which is calculated by the maximum likelihood estimation for the Poisson distribution as follows:

$$\lambda_0 = \frac{1}{T1}\sum_{t=1}^{T} n_t, \; T1 = |\{1 \leq t < T{:}n_t \geq 1\}|,$$

where T1 denotes the number of timestamps and $n_t \geq 1$.

λ_x is defined as follows: $\lambda_x = \lambda_0 \cdot e^x$, $(0 \leq x < L)$. When n_t is large, burst level b_t also takes a large value to minimize the cost function $c(\mathbf{b}|\mathbf{n})$.

Finally, we decide the maximum burst level L using λ_0 and the maximum number of tweets:

$$L = \left\lceil \log \frac{M}{\lambda_0} \right\rceil, \; M = \max_{1 \leq t \leq T} n_t,$$

where $\lceil A \rceil$ denotes a ceiling function to return an integer under A. If M is lower than $2 \cdot \lambda_0$, L reaches to 1 and burst level b_t is 0 every minute.

3.3 Hashtag score calculation

In this section, we calculate the hashtag score for each user based on the burst time series extracted in the previous section.

3.3.1 Cosine similarity. The simplest idea of a score calculation method between hashtags and users is to calculate their cosine similarity. Cosine similaritycos(u,h) between user u and hashtag h is calculated as follows:

$$\cos(u,h) = \frac{\Sigma_{t=1}^{T} b_{u,t} \cdot b_{h,t}}{\sqrt{\Sigma_{t=1}^{T} b_{u,t}^2} \cdot \sqrt{\Sigma_{t=1}^{T} b_{h,t}^2}},$$

where $b_{u,t}$ and $b_{h,t}$ denote the burst level of user u and hashtag h at time t, respectively.

3.3.2 Naïve Bayes. Naïve Bayes is a one of the most effective and typical classification method. The fundamental Naïve Bayes in a case of document classification calculates the term likelihoods and class prior probability using training data as an assumption of the term independence in documents.

Here, we believe that a time series assumes the terms and the burst level assumes the term frequency in the documents. Posterior probability $p(h|u)$ of hashtag h in given user u is defined as follows:

$$p(h|u) = p(h)p(u|h) = p_h \prod_{t=1}^{T} p_{t,h}^{b_{u,t}}.$$

We take the logarithm of both sides as follows:

$$\log p(h|u) = \log p_h + \sum_{t=1}^{T} b_{u,t} \log p_{t,h},$$

where p_h denotes the probability of hashtag h and is calculated as follows:

$$p_h = \frac{\Sigma_{t=1}^{T} b_{h,t} + 1}{\Sigma_{t=1}^{T} b_{h',t} + H},$$

where H is the number of hashtags in the datasets, $b_{u,t}$ denotes the burst level of user u at time t and $p_{t,h}$ denotes the likelihood that hashtag h occurs at time t and is calculated as follows:

$$p_{t,h} = \frac{b_{h,t} + 1}{\Sigma_{t=1}^{T} b_{h,t'} + T}.$$

3.3.3 TF-IDF. The TF-IDF method, which calculates the term's weight, is often used in information retrieval systems. TF works as the term importance in a document, and IDF evaluates the generality of a term.

Here, we incorporate TF-IDF into our user tagging model based on burst time series. We can directly replace TF by burst level $b_{h,t}$ of hashtag h at time t ($tf_{h,t} = b_{h,t}$). IDF idf_t of time t is calculated as follows:

$$idf_t = \log \frac{H}{\Sigma_{t=1}^{T} \delta(b_{h,t})}, \quad \delta(i) = \begin{cases} 1 & i \geq 1 \\ 0 & i = 0 \end{cases}.$$

By considering the user's burst time series as a query, $score(u,h)$ between user u and hashtag h is calculated as follows:

$$score(u,h) = \sum_{t=1}^{T} \mathrm{tf}_{h,t} \cdot \mathrm{idf}_t \cdot b_{u,t}.$$

4. Experimental evaluations

4.1 Dataset

4.1.1 Hashtag. In this section, we explain the data set overview that we used in our experimental evaluations. We exhaustively collected tweets from April 1, 2012, to June 4, 2013, (430 days) using the Twitter search application programming interface (API) in Japanese and set the observation time to 1-hr periods. Number of times T is 10,320 (= 430 · 24).

From these, we extracted the hashtags written only by alphanumeric characters and underscores ("_"). The hashtag distribution in the number occurrence is shown in Figure 3. The horizontal axis is the number of hashtag occurrences, and the vertical axis is the number of hashtags in each number of occurrences. Even though hashtags are used only once, over 10^7 exist because users freely create them.

We extracted the hashtags where the number of occurrences exceeds 5,000 because tags with fewer occurrences are not suitable for users. We show the hashtag data set abstract in Table I. The number of total and unique hashtags decreased to 80 and 0.05 per cent by filtering the number of occurrences.

In our collected hashtags, the top ten with their total number of occurrences is shown in Table II. The highest value is #sougollow at 44,225,513. Hashtags frequently appear including "follows". Users demand that new follow relationships be built. Figure 4 shows #sougofollow's time series of the amount of tweets and burst levels. This hashtag's burst level remained at zero because there were no times when the number of tweets greatly increased. Therefore, in our methods based on burst time series, such hashtags are not tagged to any user.

The top ten hashtags with their total number of burst levels are shown in Table III. The highest value is #rbooks at 6,139. #RakutenIchiba appeared in both Tables II and III. Figure 5 shows its time series of tweet and burst levels. Maximum burst level L of this

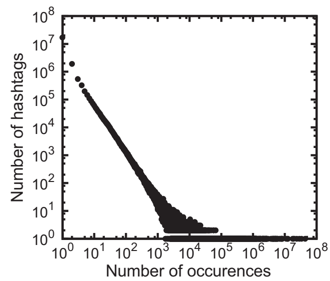

Figure 3.
Number of hashtags in each number of occurrence

Table I.
Collected results of hashtags

Threshold	Total number		Unique number	
All	1,015,053,250	100%	21,208,151	100%
Over 5,000	876,935,832	86.4%	10,367	0.05%

Table II.
Top ten hashtags with total number of occurrences

Rank	Hashtag	Total
1	#sougofollow	44,225,513
2	#followmeJP	33,830,393
3	#followme	28,894,908
4	#countkun	23,794,617
5	#RakutenIchiba	18,782,057
6	#followback	17,740,886
7	#nowplaying	12,501,181
8	#teamfollowback	12,359,201
9	#follow	11,125,344
10	#autofollow	8,324,476

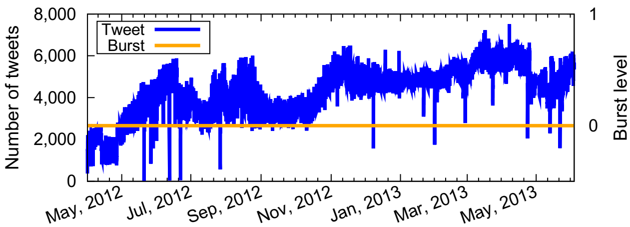

Figure 4.
Time series of tweets and bursts of #sougofollow

hashtag was two. The number of #RakutenIchiba's tweets greatly increased from February 2013. Therefore, the burst levels continued to remain at two for a long time, and the total number of burst levels with #RakutenIchiba became higher than other hashtags.

4.1.2 Users. To evaluate the effectiveness of our tagging methods, we collected 20 users who have high popularity in some topics in Twitter. They are shown in Table IV with their feature values and descriptions. The # columns is the total number of tweets from our data collecting period. The maximum value in each column is shown in bold. @nhk_news posted the most tweets at 52,240 among the evaluated users. The λ_0 and M columns are the average value of the tweets and their maximum value in one hour, respectively. @mt3776fujisan showed the highest values for these features among the evaluated users.

Figures 6 and 7 show @Japan_Olympic's and @Nintendo's time series of the amount of tweets and burst levels, respectively. Although the number of tweets with two users is vary fewer than Figures 4 and 5, we can obtain the burst level comparable to Figures 4 and 5. We calculate the score with same scale between users and hashtags using burst levels.

4.2 Evaluation procedure

4.2.1 Collecting accurate hashtags for each user. We evaluated the effectiveness of three methods using a typical evaluation method for information retrieval systems. First, every proposed method calculated the hashtags scores for each user. Second, the first author looked at the top 100 hashtags of each user by every method, combined with the user's tweets, profile and followees and determined the relevance of each hashtag by the following three levels of relevance criteria:

Rank	Hashtag	Total
1	#rbooks	6,139
2	#japanese	5,242
3	#agpr	4,878
4	#njslyr	4,412
5	#yauc	4,390
6	#cho_ag	4,378
7	#logsoku	4,343
8	#RakutenIchiba	4,339
9	#animeJP	4,332
10	#tokyomx	4,330

Table III. Top ten hashtags with total number of burst levels

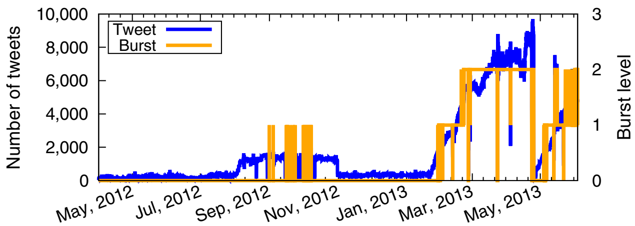

Figure 5. Time series of tweets and bursts of #RakutenIchiba

Screen name	#	λ_0	*M*
@nhk_news	*52,240*	5.794	66
@Reuters_co_jp	17,055	3.028	16
@oricon	12,585	2.181	19
@mt3776fujisan	8,950	*12.277*	*93*
@gizmodejapan	8,903	1.614	48
@jleague	6,782	8.607	32
@nico_nico_info	5,259	1.610	12
@asahi_shogi	4,374	2.243	19
@hochi_baseball	3,278	1.745	13
@tsukubais	2,971	1.869	17
@tenkixjp_jishin	2,564	1.334	11
@shop_TSUTAYA	2,215	1.725	14
@toeikotsu	1,318	1.373	9
@Yahoo_weather	1,204	2.098	6
@Nintendo	1,043	2.376	31
@Pokemon_cojp	1,010	2.686	22
@JAXA_jp	993	1.806	10
@Japan_Olympic	631	1.399	7
@DQ_PR	594	1.632	8
@AbeShinzo	301	1.427	13

Note: The maximum value in each column is shown in italic

Table IV.
Feature values for evaluation users

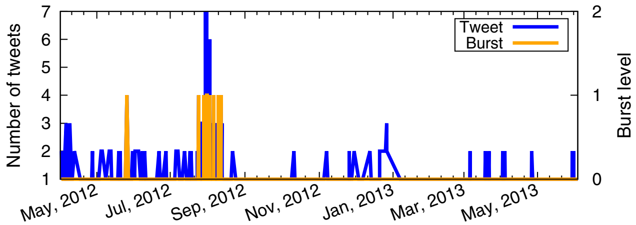

Figure 6.
Time series of tweets and bursts of @Japan_Olympic

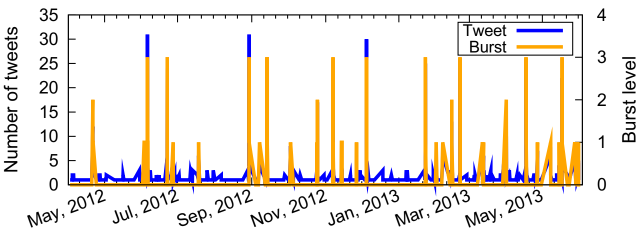

Figure 7.
Time series of tweets and bursts of @Nintendo

(1) *Highly relevant (L2)*: This hashtag has relevance and is suitable as a user's tag because it easily identifies user interests.

(2) *Relevant (L1)*: Even though the hashtag has relevance, only identifying the user interests from it is difficult.

(3) *Irrelevant (L0)*: This hashtag does not have any relevance.

4.2.2 Metrics. We evaluate the relevance of top ten hashtags that were calculated by our methods. Our evaluation uses the ERR (Chapelle *et al.*, 2009) and the Q-measure (Sakai, 2004).

ERR, which is used in navigational information retrieval system evaluations, determines high value when highly relevant documents exist at the top of the rankings. This metric is suitable to evaluate of tagging tasks because we tag the top hashtags to each user. Up to ranked K, which is set to 100 in our evaluation, the ERR is caluclated as follows:

$$\mathrm{ERR}@K = \sum_{r=1}^{K}\left(\frac{1}{r}\cdot p(r)\prod_{i=1}^{r-1}(1-p(i))\right),$$

where $p(r)$ denotes the user stopping probability at rank r and is defined as follows:

$$p(r) = \frac{2^{g(r)}-1}{2^{mg}},$$

where $g(r)$ denotes the relevance grade at rank r's hashtag; our case is selected from $\{0,1,2\}$; and mg denotes a maximum relevance grade, and our case is $mg = 2$.

On the other hand, ERR does not evaluate the hashtag scoring adequacy because it evaluates by weighting the top ranking. Therefore, we calculate Q-measure to evaluate multiple relevance values. This metric can evaluate the overall correctness in a ranking. Up to the top ranked K, which is set to 100 in our evaluation, the Q-measure is calculated as follows:

$$Q@K = \frac{1}{K}\sum_{r=1}^{K} I(r)BR(r),$$

$$I(r) = \begin{cases} 1 & \text{rank r's hashtag has relevance} \\ 0 & \text{otherwise} \end{cases},$$

$$BR(r) = \frac{C(r)+\beta cg(r)}{r+\beta cg^{*}(r)},$$

where $C(r)$ denotes the number of relevant hashtags up to rank r. $cg(r)$, which denotes the cumulative gain until rank r, is calculated as follows: $cg(r) = \Sigma_{k=1}^{r} g(k)$, $cg^{*}(r)$ is also cumulative gain for an ideal ranking list obtained by arranging the relevance grade of all the hashtags, β denotes a parameter that controls the user's endurance during information retrieval behavior; we set $\beta = 1.0$, which is generally used.

4.3 Experimental results

We show the ERR@10 and Q-measure@10 values of each user in Table V. The cosine, Naïve Bayes and TF-IDF columns are the evaluation values of our methods explained in Section 3.3.1, 3.3.2, and 3.3.3, respectively. The maximum values of each user in each metric are shown in bold. The L0, L1 and L2 columns are the numbers of irrelevant, relevant and highly relevant hashtags decided by the first author. The far right column shows examples of the highly relevant hashtags.

Table V. ERR@10 and Q-measure@10 values of each user by each method

Screen name	ERR@10 (%) Cosine similarity	ERR@10 (%) Naïve Bayes	ERR@10 (%) TF-IDF	Q-measure@10 (%) Cosine similarity	Q-measure@10 (%) Naïve Bayes	Q-measure@10 (%) TF-IDF
@nhk_news	79.5	78.6	7.5	23.5	14.2	1.2
@Reuters_co_jp	79.5	77.2	20.0	22.4	14.9	5.8
@oricon	84.4	84.4	0.0	20.0	20.0	0.0
@mt3776fujisan	13.1	84.4	18.1	7.1	20.0	8.5
@gizmodojapan	81.9	78.8	19.5	36.5	31.0	5.4
@jleague	46.2	79.6	11.8	77.3	32.1	6.7
@nico_nico_info	25.0	10.7	39.8	3.3	1.6	7.9
@asahi_shogi	54.7	75.0	37.5	26.7	10.0	6.0
@hochi_baseball	78.8	7.5	17.2	22.7	1.4	7.3
@tsukubais	79.9	75.0	75.0	27.2	10.0	10.0
@tenkijp_jishin	86.3	86.3	86.2	94.4	85.4	65.9
@shop_TSUTAYA	29.7	11.4	0.0	11.1	5.3	0.0
@toeikotsu	79.2	77.0	28.7	23.2	17.2	15.6
@Yahoo_weather	0.0	25.8	37.5	0.0	6.7	5.0
@Nintendo	86.3	86.3	86.2	73.8	50.0	40.0
@Pokemon_cojp	58.5	85.9	85.2	40.0	30.0	32.5
@JAXA_jp	2.8	37.5	0.0	1.5	6.0	0.0
@Japan_Olympic	30.8	85.8	86.2	32.9	42.5	45.8
@DQ_PR	25.0	12.5	0.0	6.7	2.5	0.0
@AbeShinzo	0.0	0.0	12.5	0.0	0.0	2.5
Mean	51.1	58.0	33.5	27.5	20.0	13.3
SD	31.5	32.4	31.5	25.9	20.4	17.7

@Nintendo's and @tenkijp_jishin's ERR@10 values by every method showed the maximum value among all users. In Q-measure@10, @tenkijp_jishin also showed the highest value among all users for every method. In cosine similarity and Naïve Bayes, @AbeShinzo showed the lowest value (at 0.0) in both ERR@10 and Q-measure@10. Naïve Bayes achieved the highest average value in ERR@10. On the other hand, in Q-measure@10, cosine similarity achieved the highest average value. The number of highly relevant hashtags of @AbeShinzo, @JAXA_jp, @DQ_PR, @asahi_shogi and @shop_TSUTAYA was one.

5. Discussions

5.1 Cosine similarity vs Naïve Bayes

Table V shows that cosine similarity and Naïve Bayes, respectively, achieved maximum average Q-measure and ERR values. TF-IDF had the lowest values in both metrics. As explained in Section 4.2, ERR shows high evaluation values when highly relevant hashtags exist at the top of ranking list. Q-measure shows high evaluation values when hashtags are arranged by relevance grades. Therefore, Naïve Bayes can estimate highly relevant hashtags at the top of the rankings. Cosine similarity can estimate relevant hashtags among all of the rankings. So, Naïve Bayes and cosine similarity are suitable methods for hashtag tagging and hashtag scoring, respectively.

As an example that demonstrates the differences of these two methods, we show the top ten hashtags tagged by these two methods to @jleague in Table VI. The Rel. column is the relevance to each hashtag. From Table V, the @jleague's maximum ERR@10 and Q-measure@10 values were respectively achieved by Naïve Bayes and cosine similarity. The hashtag at Rank 1 by Naïve Bayes is #jleague, which is highly relevant: however, irrelevant hashtags appear five times in this ranking list. For cosine similarity, irrelevant hashtags appear only one time in the top ten: however, highly relevant hashtags appear at Rank 10. As seen above, Naïve Bayes can estimate the highly relevant hashtags at the top of the ranking, and cosine similarity can gather relevant hashtags from all of the rankings.

5.2 Detailed analysis

To analyze weather our methods appropriately identify types of users, Tables VII-X show the top five hashtags with high scores to four extracted users that were combined with relevance judgment results.

The hashtags of @Nintendo were effectively estimated as highly relevant with the top rank by cosine similarity and Naïve Bayes. We extracted #NintendoDirectJP at Rank 1 and #wii_u at Rank 4 by cosine similarity, and these burst time series are shown

	Cosine similarity		Naïve Bayes	
Rank	Hashtags	Rel.	Hashtags	Rel.
1	#avispa	L1	#jleague	L2
2	#kataller	L1	#j_toku	L1
3	#gainare	L1	#imacoconow	L0
4	#fagiano	L1	#REFELCBEAT_AC	L0
5	#thespa	L1	#hanshin	L0
6	#rbooks	L0	#avispa	L1
7	#sanga	L1	#tigers	L0
8	#yamaga	L1	#miil	L0
9	#tochigisc	L1	#fagiano	L1
10	#jleague	L2	#gainare	L1

Table VI. @jleague's top ten hashtags by cosine similarity and Naïve Bayes

Rank	Cosine similarity		Naïve Bayes	
1	L2	#NintendoDirectJP	L2	#NintendoDirectJP
2	L2	#NintendoDirect	L2	#NintendoDirect
3	L2	#Nintendo	L2	#Nintendo
4	L2	#wii_u	L2	#3DS
5	L1	#E3_nico	L2	#wiiu

Table VII. Top five hashtags with high scores for @Nintendo

Rank	Cosine similarity		Naïve Bayes	
1	L2	#npb	L0	#rbooks
2	L1	#seibulions	L0	#autefollow
3	L0	#autefollow	L0	#REFLECBEAT_AC
4	L0	#LTE	L0	#iuranainet
5	L0	#WiMAX	L0	#watch

Table VIII. Top five hashtags with high scores for @hochi_baseball in Figure 8. The vertical and horizontal axes are the burst levels and the time series, respectively. The bottom figure shows @Nintendo's burst time series. The three parts of the figure share a time series with the bottom figure. @Nintendo and #NintendoDirectJP frequently experienced bursts at the same time. Therefore, this hashtag was ranked at the top in both of our methods. A predominant burst occurred in #wii_u in June 2012. Cosine similarity can appropriately estimate such hashtags because it normalizes the total burst levels.

Next, in Table VIII, we confirm that @hochi_baseball by Naïve Bayes could not estimate the relevant hashtags in the top five. Cosine similarity estimated the highly relevant hashtags of #npb at Rank 1, but Naïve Bayes incorrectly labeled #rbooks's hashtags as irrelevant. Figure 9 shows the burst time series of these two hashtags

Rank	Cosine similarity		Naïve Bayes	
1	L0	#MOCOS_kitchen	L2	#fujisan
2	L0	#moco	L2	#mtfuji
3	L0	#shakkin	L2	#ohayo
4	L0	#olive_gohan	L2	#bt_tenki
5	L1	#zip	L2	#simpleweight_jp

Note: The maximum value in each column is shown in italic

Table IX. Top five hashtags with high scores for @mt3776fujisan

Rank	Cosine similarity		Naïve Bayes	
1	L0	#sm18879213	L2	#airjam
2	L0	#pyconjp	L2	#pyconjp
3	L0	#sprk2012	L2	#kinatsu
4	L0	#KOZOS1	L2	#needfollowers
5	L1	#air_jam	L2	#sm18879213

Table X. Top five hashtags with high scores for @AbeShinzo

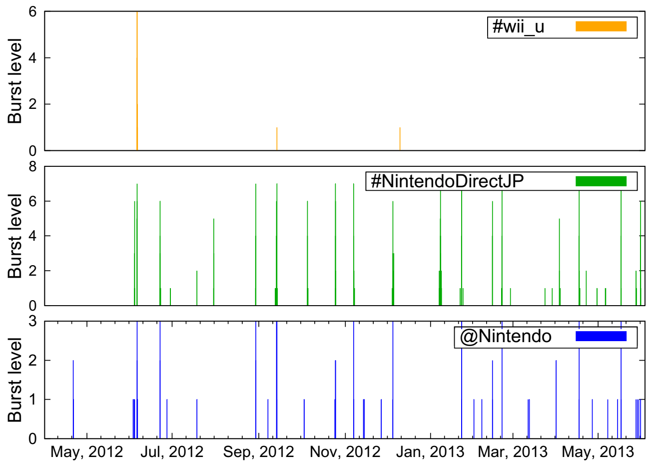

Figure 8. Burst time series of @Nintendo, #wii_u, and #NintendoDirectJP

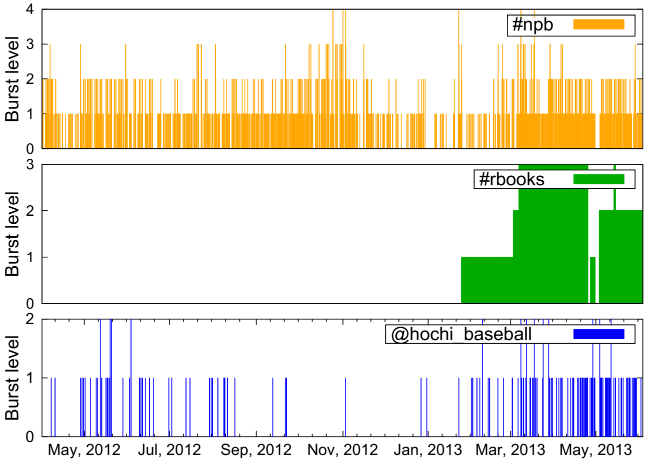

Figure 9.
Burst time series of @hochi_baseball, #npb, and #rbooks

and @hochi_baseball, which often experience bursts at Level 2 from February to May 2013. #rbooks continuously experienced bursts at Level 3 in this period. Moreover, from Table III, #rbooks is the maximum value of the total number of burst levels among all the hashtags. Naïve Bayes calculated high probability p_h for #rbooks because p_h was decided based on the total number of burst levels. Therefore, Naïve Bayes wrongly calculated a high score for #rbooks.

On the other hand, in hashtags estimated to @mt3776hujisan in Table IX, Naïve Bayes could identify #fujisan's highly relevant hashtags, but cosine similarity incorrectly labeled #MOCOS_kichen as irrelevant hashtag. In Figure 10, #MOCOS_kitchen periodically experienced many Level 3 bursts. Naïve Bayes calculates low probabilities for each timestamp because it normalizes each burst by the

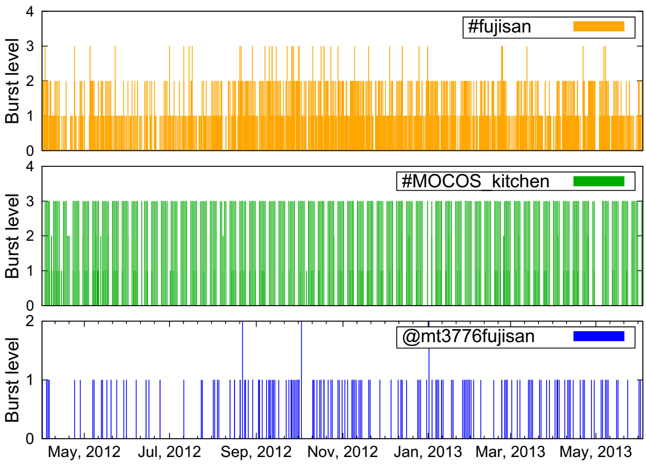

Figure 10.
Burst time series of @mt3776fujisan, #fujisan, and #MOCOS_kitchen

total number of burst levels. However, as cosine similarity directly uses burst levels, such periodical bursts are calculated as high scores by this method.

As an example where our methods failed to effectively estimate, we confirmed hashtags estimated to @AbeShinzo shown in Table X. In both cosine similarity and Naïve Bayes, ERR@10 and Q-measure@10 are zero for this user. In Figure 11, @AbeShinzo experienced few bursts compared with the above users. #pyconjp, which was incorrectly estimated by cosine similarity and Naïve Bayes at Rank 2 to be @AbeShinzo, simultaneously experienced Level 4 bursts with @AbeShinzo's bursts. Our methods cannot identify hashtags for such users who had such few bursts.

Finally, to clarify that our methods work for the feature value with users, we evaluated Spearman's (1904) rank correlation coefficients between user rankings in descending order of each evaluation value and each feature value (Tables XI). The maximum value in each column is shown in bold. The maximum correlation coefficient of the cosine similarity in ERR@10 is the number of tweets at 0.418. In other evaluation values, the maximum correlation coefficients are shown by the maximum value of the number of tweets. In these results, we clarified that our methods based on burst time series have high correlation for the maximum value of the number of tweets of each user.

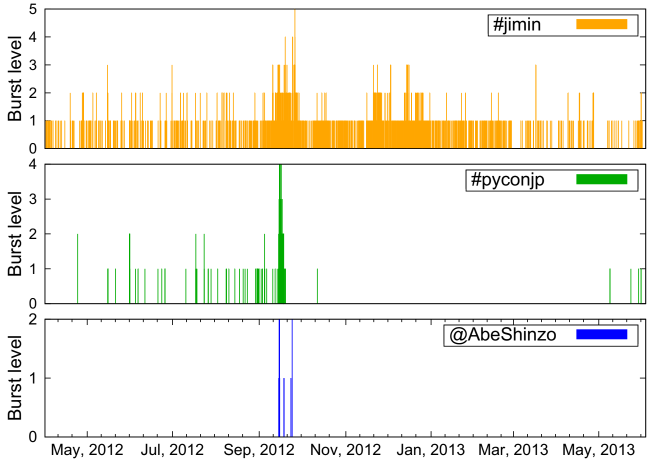

Figure 11. Burst time series of @AbeShinzo, #jimin, and #pyconjp

Table XI. Spearman's rank correlation coefficients among each index

Value	ERR@10 Cosine similarity	ERR@10 Naïve Bayes	ERR@10 TF-IDF	Q-measure@10 Cosine similarity	Q-measure@10 Naïve Bayes	Q-measure@10 TF-IDF
#	*0.418*	0.178	−0.179	0.177	0.197	−0.051
λ_0	0.039	0.265	−0.179	0.108	0.178	−0.080
M	0.377	*0.386*	*−0.126*	*0.401*	*0.344*	*0.051*

6. Conclusion

In this paper, we propose a new user tagging method based on burst time series. Our method consists of two phases. First, we extract the burst time series for each user and each hashtag using Kleinberg's burst enumerating algorithm and estimate the burst levels based on the number of tweets on each time. We calculate the score with same scale between users and hashtags using burst levels. Second, we calculate the score between users and hashtags by such typical score calculation methods as cosine similarity, Naïve Bayes and TF-IDF using the burst level time series.

From our sophisticated experimental evaluations, we demonstrate the high efficiency of our tagging method. Naïve Bayes and cosine similarity are particularly suitable for the user tagging and tag score calculation tasks, respectively. Some users, whose hashtags were appropriately estimated by our methods, experienced higher maximum value of the number of tweets than other users. These results are supported our hypothesis: as users have interests, they will post more tweets at the time when events occur compared with general times.

In future work, we will propose a new tagging model that combines both strengths of Naïve Bayes and cosine similarity.

References

Balmin, A., Hristidis, V. and Papakonstantinou, Y. (2004), "Objectrank: authority-based keyword search in databases", *Proceedings of the VLDB, Toronto*, pp. 564-575.

Blei, D.M. and Lafferty, J.D. (2006), "Dynamic topic models", *Proceedings of the ICML, Pittsburgh, PA*, pp. 113-120.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K.P. (2010), "Measuring user influence in twitter: the million follower fallacy", *Proceedings of the ICWSM, Washington, DC*, pp. 10-17.

Chalmers, D., Fleming, S., Wakeman, I. and Watson, D. (2011), "Rhythms in twitter", *Proceedings of the Social Com, Boston, MA*, pp. 1409-1414.

Chapelle, O., Metlzer, D., Zhang, Y. and Grinspan, P. (2009), "Expected reciprocal rank for graded relevance", *Proceedings of the CIKM, Hong Kong*, pp. 621-630.

Diao, Q., Jiang, J., Zhu, F. and Lim, E.P. (2012), "Finding bursty topics from microblogs", *Proceedings of the ACL*, Stroudsburg, PA, pp. 536-544.

Ghosh, R., Surachawala, T. and Lerman, K. (2011), "Entropy-based classification of 'retweeting' activity on twitter", *CORR*, Vol. 1106, 0346.

Huang, J., Thornton, K.M. and Efthimiadis, E.N. (2010), "Conversational tagging in twitter", *Proceedings of the HT, Seattle, WA*, pp. 173-178.

Java, A., Song, X., Finin, T. and Tseng, B. (2007), "Why we twitter: understanding microblogging usage and communities", *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop, San Jose, CA*, pp. 56-65.

Kleinberg, J. (2002), "Bursty and hierarchical structure in streams", *Proceedings of the KDD, Edmonton, AB*, pp. 91-101.

Koike, D., Takahashi, Y., Utsuro, T., Yoshioka, M. and Kando, N. (2013), "Time series topic modeling and bursty topic detection of correlated news and twitter", *Proceedings of the IJCNLP, Nagoya*, pp. 917-921.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010), "What is twitter, a social network or a news media?", *Proceedings of the WWW, Raleigh, NC*, pp. 591-600.

Li, C., Sun, A. and Datta, A. (2012), "Twevent: segment-based event detection from tweets", *Proceedings of the CIKM, Maui, HI*, pp. 155-164.

Ma, Z., Sun, A., Yuan, Q. and Cong, G. (2014) "Tagging your tweets: a probabilistic modeling of hashtag annotation in twitter", *Proceedings of the CIKM, Shanghai*, pp. 999-1008.

Mathioudakis, M. and Koudas, N. (2010), "Twittermonitor: trend detection over the twitter stream", *Proceedings of the SIGMOD, New York, NY*, pp. 1155-1158.

Mihalcea, R. and Tarau, P. (2004), "Textrank: bringing order into texts", *Proceedings of the EMNLP*, *Barcelona*, pp. 404-411.

Mizunuma, Y., Yamamoto, S., Yamaguchi, Y., Ikeuchi, A., Satoh, T. and Shimada, S. (2014), "Twitter bursts: analysis of their occurrences and classifications", *Proceedings of the ICDS, Barcelona*, pp. 182-187.

Myers, S.A. and Leskovec, J. (2014), "The bursty dynamics of the twitter information network", *Proceedings of the WWW, ACM*, *Seoul*, pp. 278-287.

Pal, A. and Counts, S. (2011), "Identifying topical authorities in microblogs", *Proceedings of the WSDM*, Hong Kong, pp. 45-54.

Sakai, T. (2004), "New performance metrics based on multigrade relevance: their application to question answering", *Proceedings of the NTCIR, Tokyo*.

Spearman, C. (1904), "The proof and measurement of association between two things", *The American Journal of Psychology*, Vol. 15 No. 1, pp. 72-101.

Twitter (2014), "Twitter reports fourth quarter and fiscal year 2013 results", available at: https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=823321 (accessed 1 April 2014).

Wang, B., Wang, C., Bu, J., Chen, C., Zhang, W.Z., Cai, D. and He, X. (2013), "Whom to mention: expand the diffusion of tweets by @ recommendation on micro-blogging systems", *Proceedings of the WWW, Redmond, WA*, pp. 1331-1340.

Wang, X., Zhai, C., Hu, X. and Sproat, R. (2007), "Mining correlated bursty topic patterns from coordinated text streams", *Proceedings of the KDD*, San Jose, CA, pp. 784-793.

Weng, J., Lim, E.-P., Jiang, J. and He, Q. (2010), "Twitterrank: finding topic-sensitive influential twitterers", *Proceedings of the WSDM, New York, NY*, pp. 261-270.

Wu, W., Zhang, B. and Ostendorf, M. (2010), "Automatic generation of personalized annotation tags for twitter users", *Proceedings of the HLT, Stroudsburg, PA*, pp. 689-692.

Xie, W., Zhu, F., Jiang, J., Lim, E.-P. and Wang, K. (2013), "Topicsketch: real-time bursty topic detection from twitter", *Proceedings of the ICDM*, *Dallas, Texas*, pp. 837-846.

Yamaguchi, Y., Amagasa, T. and Kitagawa, H. (2012), "Tagging users based on twitter lists", *International Journal of Web Engineering and Technology*, Vol. 7 No. 3, pp. 273-298.

Yamaguchi, Y., Yamamoto, S. and Satoh, T. (2014), "Behavior analysis methods for twitter users based on transitions in posting activties", *International Journal of Web Information and Systems*, Vol. 10 No. 4, pp. 363-377.

Yamaguchi, Y., Yoshida, M., Faloutsos, C. and Kitagawa, H. (2015), "Patterns in interactive tagging networks", *Proceedings of the ICWSM*, *Oxford*, pp. 513-522.

Yang, J. and Counts, S. (2010), "Comparing information diffusion structure in weblogs and microblogs", *Proceedings of the ICWSM*, *Washington, DC*, pp. 351-354.

Further reading

Domingos, P. and Pazzani, M. (1997), "On the optimality of the simple bayesian classifier under zero-one loss", *The Journal of Machine Learning Research*, Vol. 29 Nos 2/3, pp. 103-130.

Yamamoto, S., Wakabayashi, K., Kando, N. and Satoh, T. (2015), "BUTE: bursty users tagging method estimated by time series data", *Proceedings of the iiWAS, Brussels*, pp. 148-156.

Corresponding author

Shuhei Yamamoto can be contacted at: yamahei@ce.slis.tsukuba.ac.jp