PAPER Special Section on Info-Plosion

Social Bookmarking Induced Active Page Ranking

Tsubasa TAKAHASHI[†a)], *Nonmember*, Hiroyuki KITAGAWA[†,††], *Fellow, and* Keita WATANABE[†], *Nonmember*

SUMMARY Social bookmarking services have recently made it possible for us to register and share our own bookmarks on the web and are attracting attention. The services let us get structured data: (URL, Username, Timestamp, Tag Set). And these data represent user interest in web pages. The number of bookmarks is a barometer of web page value. Some web pages have many bookmarks, but most of those bookmarks may have been posted far in the past. Therefore, even if a web page has many bookmarks, their value is not guaranteed. If most of the bookmarks are very old, the page may be obsolete. In this paper, by focusing on the timestamp sequence of social bookmarkings on web pages, we model their activation levels representing current values. Further, we improve our previously proposed ranking method for web search by introducing the activation level concept. Finally, through experiments, we show effectiveness of the proposed ranking method.

key words: *social bookmark, activation analysis, ranking algorithm, web mining*

1. Introduction

Today's information society features the conveyance of varied information in diverse ways. This makes it possible to easily and quickly get useful information. Acquisition is not restricted to passive receiving; we can freely express our experience, reviews and comments, and also communicate with others over the web. CGMs like blogs, wikis and social networks are familiar ways we express ourselves on the web. Unfortunately, not all contents have informative value for each of us. This means we are forced to determine whether or not web contents are reliable and/or informative. Web search engines such as Yahoo!* and Google** are wonderful tools for information retrieval and knowledge discovery. Whenever people want to explore their interests and/or preferences, a web search engine is often their first choice, rather than a friend or the dictionary. Such search engines estimate web page values using hyperlinks among web pages such as PageRank [1] and HITS [2]; other methods include term frequencies, click logs, and a variety of factors. However, there is no way to include reader evaluations and/or preferences. In actual society, we prefer something that is recommended by many people or has been reviewed by experts.

Recently, web services called Social Bookmarks (SBMs) are attracting attention. SBM allows us to register web pages in much the same way the bookmark tool in web browsers works, and assign space to manage our bookmarks. A web page registered on SBM is refereed by user interests, preferences, convenience and a variety of individual review points. It seems, therefore, that the page has some informative value. Additionally, if a page is bookmarked by many users, it seems more informative, and also seems to have reliable content. We are inclined in real life to try things that many others have already tried. So a bookmark from an SBM user to a web page is a vote, and the number of bookmarks to web pages in SBM is a barometer of web page values. Yanbe et al. proposed a web page ranking method based on the number of bookmarks [3]. They showed the informativeness of SBM to improve web page rankings. Heymann et al. analyzed SBM's powers of influence on web searches [4].

We proposed a ranking method based on bookmark relationships between bookmarkers and web pages [5]. Our previous method, called S-BITS, ranks web pages based on a bipartite graph between bookmarked web pages and bookmarking users like the HITS algorithm. It estimates a web page value based on who bookmarks the page, and also estimates a bookmarker value based on what pages that person bookmarks. When a bookmarker is knowledgeable and/or trustworthy, a page gets a high score if that person bookmarks the page. In past experiments, our method yielded better results than an existing web search engine and SBRank, which Yanbe proposed [3].

In SBM, there are many users who do not update or delete bookmark information pointing to obsolete web pages. And there are many obsolete web pages: old news and press releases that are not worth a look. For that reason, we cannot simply say that the number of bookmarks indicates the current informative value of a web page. We need to look not only at the number of bookmarks, but also at the freshness of web pages. An easy way to estimate freshness is based on when the last bookmark appeared [6]. However, the current value of each web page depends on its complete bookmark history. To judge the current value of a web page based on SBM, we should look at how the page has been bookmarked until now, and how frequently the page has been bookmarked. One way to observe the degree of obsolescence is to compare page bookmarking rates

Manuscript received September 11, 2009.
Manuscript revised January 4, 2010.

[†]The authors are with the Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba-shi, 305–8573 Japan.

[††]The author is with the Center for Computational Sciences, University of Tsukuba, Tsukuba-shi, 305–8577 Japan.

a) E-mail: tsubasa@kde.cs.tsukuba.ac.jp

DOI: 10.1587/transinf.E93.D.1403

*http://www.yahoo.com/

**http://www.google.com/

between the past and present. We estimate the informative value of each page on a particular date in SBM. We call it the "Activation Level." The activation level shows how attractive a page is on a particular date. Analyzing time-series variation of social bookmarking frequency, we can estimate the activation level of a web page. We propose an activation level model using the Hidden Markov Model (HMM) [7]. Activation levels suggest the current values of web pages registered in SBM. Further, we improve S-BITS, our previous work, by introducing this activation level concept. Our new method is named S-BITS*. Our experiments show that S-BITS* yields better ranking. These experiments suggest that our estimation method of activation levels is effective and is similar to the human sense of worth.

The remainder of this paper is organized as follows: Section 2 overviews and observes social bookmarkings. Section 3 introduces S-BITS, our previous approach. Section 4 models activation levels of web pages based on social bookmark information. Section 5 proposes S-BITS*, which is a web page ranking method introducing the activation level concept. Section 6 proposes another ranking method, which is used in our experiments for comparison. Section 7 presents and discusses experimental results. Section 8 discusses related work. Finally, Sect. 9 concludes the paper and outlines future work.

2. Social Bookmarkings

Social Bookmark (SBM) services are recently a hot topic. They allow individuals to bookmark and annotate web pages. A user can manage bookmarks by deciding a set of tags. Services like SBM are so called collaborative tagging systems. They have the same character as *Folksonomy*, which means "people's management," and are a manually annotated resource.

del.icio.us[†], at the top of SBM services, started in 2003. It now has as many as 5.3 million users with over 180 million unique URLs registered[††]. SBM has been popular since around 2004, and many services like *Hatena Bookmark*[†††] (Japanese Top Service Provider) have appeared.

Bookmark data, which is a collection of tuples (url, username, timestamp, set of tags), is available in SBM (Fig. 1). The data shows user interests for certain pages. A user, using personal knowledge and intuition, can annotate a page with tags that show keywords and/or categories. Timestamps indicate when users first showed interest in the web page. In this paper, we model bookmark information b as follows:

$$b = (p, u, t, A)$$
$$A = \{a_1, a_2, \ldots, a_n\}$$

where p is a page, u is a user, t is a timestamp, A is a tag set for annotation, and a_i is a tag.

Because SBM is a social network of people, user interests are diverse and knowledge quality varies widely. We show the distribution of bookmark counts obtained from Hatena Bookmark, which we collected. The collected data were made public in June 2009. Pages number around 600 thousand, with bookmarks approaching 10 million (Fig. 2). Looking at the total number of public bookmarks, many web pages have few bookmarks. This suggests that bookmarked pages are diverse and user interests in SBM are also diverse. A web page registered in an SBM is refereed by the user's own review points. If a page is bookmarked by many users, it seems more informative and seems to have reliable contents. Our tendency in life is to want to try something many users have tried. So a bookmark from an SBM user to a web page is a vote, and the number of bookmarks to web pages is a barometer of the web page's value.

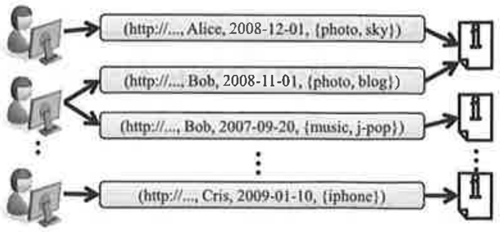
(http://..., Alice, 2008-12-01, {photo, sky})
(http://..., Bob, 2008-11-01, {photo, blog})
(http://..., Bob, 2007-09-20, {music, j-pop})
(http://..., Cris, 2009-01-10, {iphone})

Fig. 1 Social bookmarkings.

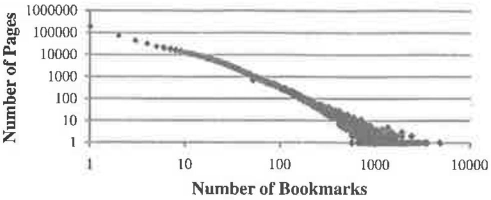

Fig. 2 Numbers of bookmarks.

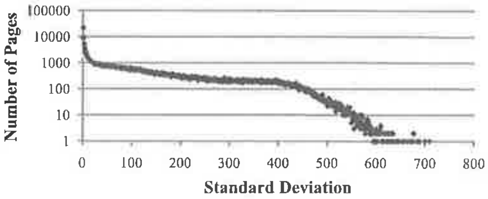

Fig. 3 Standard deviations of the bookmark timestamps.

The deviations in bookmark timestamps for each page are distributed as in Fig. 3. We see that many web pages have small deviations. This means that many web pages in SBM were bookmarked in a specific short duration. A web page that was bookmarked only in a past particular time frame will not be informative now. Moreover, there are pages with a huge number of bookmarks, but they were bookmarked over only a few days; after that, the pages were not bookmarked again (Fig. 4).

By focusing on a time-series variation of social bookmarks, we can assess time-series variation of SBM user attention. We can then discover the value of bookmarked pages on the temporal axis. SBM has many obsolete web

[†] http://delicious.com/

[††] http://blog.delicious.com/blog/2008/11/delicious-is-5.html

[†††] http://b.hatena.ne.jp/

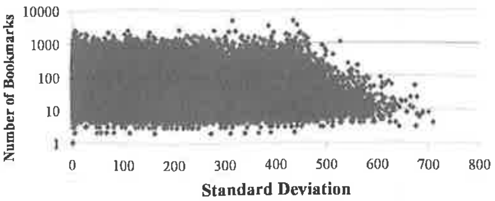

Fig. 4 Standard deviation v.s. number of bookmarks.

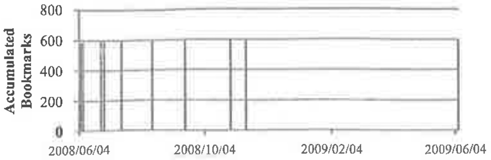

Fig. 5 Bookmark sequence of short-life Web page.

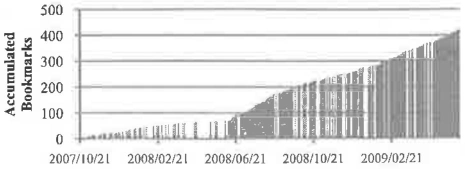

Fig. 6 Bookmark sequence of time-homogeneous Web page.

pages like old news and press releases. They are attractive early, but they soon turn obsolete (Fig. 5). We should not only look at the number of bookmarks when we measure page values; we should look at how the page has been bookmarked up to the present. Looking at the time-series variation of bookmarks for the pages, we can estimate whether the page is still informative and fresh, or obsolete. Temporal analysis, therefore, is important in estimating the current values of web pages in SBM.

However, not only are there web pages with a temporal axis but also web pages with time-homogeneous contents, such as portal sites and reference documents. Page freshness, therefore, depends on not only on the age of web pages but also page content. A time homogeneous page may have been accessed frequently through its SBM and it tends to be bookmarked constantly and continually by various users (Fig. 6). But there is no guarantee that its bookmarkings will continue in the future. On the other hand, a time temporal page is accessed with high frequency at a particular time, but attention then declines over time (Fig. 5). We can say that the history of bookmark arrivals reflects time-homogeneousness and time-temporariness.

We now suggest an easy way to introduce web page freshness and obsolescence. The method is to use exponential aging. The function is as follows:

$$TimeWeight(t) = T_0\, e^{-\frac{log2}{t_{1/2}}(t'-t)} \qquad (1)$$

where t is a timestamp of a bookmark, t' is the timestamp of the evaluation date, $t_{1/2}$ is the half-life period and T_0 is the upper limit of this aging function. Using this function, the value of freshness for each bookmark dissipates over time. Introducing this function, the scoring calculation of S-BITS [5] as $p\_score_i^k = \sum_{b_{ji} \in B} u\_score_j^{k-1}$, $u\_score_i^k = \sum_{b_{ij} \in B} p\_score_j^{k-1}$ transforms as follows:

$$p\_score_i^k = \sum_{b_{ji} \in B} TimeWeight(t_{ji})\; u\_score_j^{k-1} \qquad (2)$$

$$u\_score_i^k = \sum_{b_{ij} \in B} TimeWeight(t_{ij})\; p\_score_j^{k-1} \qquad (3)$$

We call the above method Aging S-BITS; we then use it for a baseline of our proposed method in the experiments.

Freshness and obsolescence, however, depend on web page content and topics, because potential attractiveness differs with content and topic. The potential attractiveness is reflected on the potential bookmark frequency. In the following section, we model the activation concept based on the potential bookmark frequency of a web page, and then estimate page freshness and obsolescence.

3. S-BITS

S-BITS estimates a web page value based on who bookmarks the page, and also estimates a user expertise based on what pages he bookmarks. We apply the HITS concept to the relationship between pages and users in SBM to create bipartite graph G. In graph G, the evaluation similar to the majority rule by bookmarking as a user's vote is available. Further, in weighting estimations from trustworthy users who have a high rate hub score, S-BITS evaluates page scores based on user expertise.

HITS extends the objective page set using link structures of the pages. In S-BITS, taking into account tag sets that frequently co-occur (called *frequent tag sets*), the objective page set is extended to pages which are annotated with tag sets that enclosing *frequent tag sets*. We use an extraction of Maximal Frequent Item Sets [8] to discover set of *frequent tag sets*.

The detailed algorithm of S-BITS is as follows:

1. A query q is issued by the user. Then, giving q to a search engine system, the top M pages (page set P_0) are collected. From the SBM data, the bookmark information $b_{ji}(= (p_i, u_j, t_{ji}, A_{ji}))$ for each page $p_i \in P_0$ is collected (bookmark set B). The set of users (user set U) who bookmark pages in P_0 is collected. Set of tag set V, which consists of all tag sets A_{ji}, is obtained.
2. To collect relevant pages a set V' of frequent tag sets from V is extracted [†]. When a user in U has bookmarked a page (not included in P_0) by A_{km}, which encloses a frequent tag set $F \in V'$, the page p_k is collected. Then, merging the collected pages to P_0 yields extended page set P, the merged set. The resulting graph is G' (Fig. 7).

[†] V is a set of tag sets A_{ji}. Regarding each A_{ji} as a transaction, we extract *Maximal Frequent Item Sets* as *frequent tag sets*.

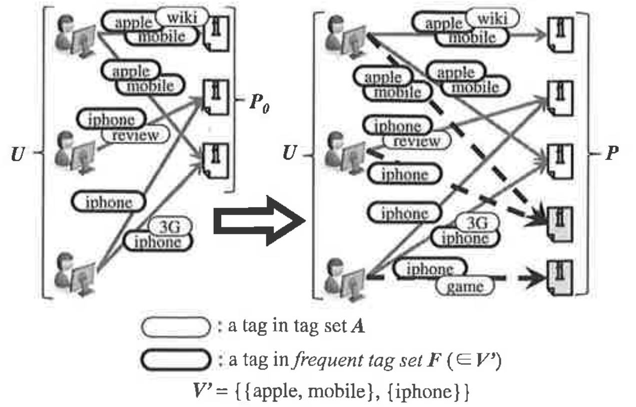
U
apple
wiki
mobile
iphone
review
3G
P0
U
P
game
: a tag in tag set A
: a tag in frequent tag set F (∈ V')
V' = {{apple, mobile}, {iphone}}

Fig. 7 Graph construction of S-BITS.

Algorithm 1 Scoring Algorithm of S-BITS

```
$p\_score^0 := \{1, 1, 1, ..., 1\}$ $u\_score^0 := \{1, 1, 1, ..., 1\}$
$k := 0$
repeat
    $k := k + 1$
    for all $p_i \in P'$ do
        $p\_score_i^k := \sum_{b_{ji} \in B'} u\_score_j^{k-1}$
    end for
    for all $u_i \in U'$ do
        $u\_score_i^k := \sum_{b_{ij} \in B'} p\_score_j^{k-1}$
    end for
    normalize($p\_score^k$) and normalize($u\_score^k$)
until $|p\_score^k - p\_score^{k-1}|_1 < \epsilon_p$ and $|u\_score^k - u\_score^{k-1}|_1 < \epsilon_u$
return $p\_score^k$ and $u\_score^k$
```

3. Finally, authority scores and hub scores are calculated on graph G' (Algorithm 1). This calculation is similar to the HITS algorithm. Then, based on the authority scores, the rank of web pages is created.

S-BITS does not take timestamps of bookmarks into account, and cannot estimate whether or not a page is fresh. Some obsolete web pages, therefore, may be ranked higher if they have a huge number of bookmarks.

4. Activation Level Estimation

In SBM, a web page has been bookmarked in sequence by its users. In other words, a sequence of bookmarks arrives in temporal order. When bookmark arrivals become more frequent than usual, attention to the bookmarked page will be increasing. Conversely, when arrivals become less frequent, attention may have diminished and the page may be obsolete. This paper defines activation level as the comparative difference in bookmark frequency between its baseline and a specific time. When bookmarks are frequent, the page has a high activation level; when they are infrequent, the activation level is low.

Kleinberg's algorithm [9] identifies bursts in the document stream, which is a sequence of documents. Specifically, it detects topic bursts in e-mail document streams. In Kleinberg's work, the word "burst" is a state in which particular topic documents arrive with high frequency as opposed to a normal state in which the arrival rate is baseline. If we want to detect a burst of an event in a document stream, we explore only burst states from the document stream.

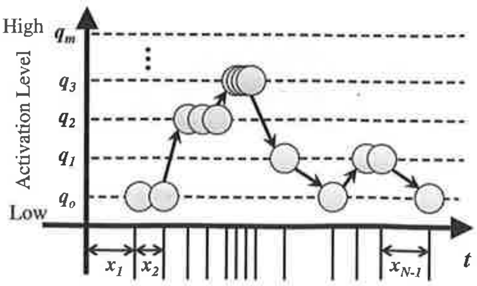

Fig. 8 Activation level based on arrival gaps.

Our proposal to estimate activation level for web pages in SBM is based on Kleinberg's algorithm. In the next subsection, we illustrate Kleinberg's algorithm. We then propose an activation level estimation method for web pages in SBM.

4.1 Kleinberg's Algorithm

Kleinberg assumes that a randomized model for generating a sequence of document arrival times is based on exponential distribution. This assumption means that documents arrive randomly. Supposing that assumption, a document stream that includes N documents within period T follows a uniform distribution. Each document arrives chronologically, and then an inter-arrival gap x_k in time between document $k+1$ and k is distributed according to the exponential density function $f(x) = \alpha e^{-\alpha x}$ ($\alpha > 0$). The expected value of the gap is α^{-1}, and then α represents the rate of document arrivals. The document stream, which is the sequence of N documents, has the inter-arrival gaps $\mathbf{x} = (x_1, x_2, ..., x_{N-1})$. In this stream, the average gap $\hat{g}$ is formulated as T/N.

Kleinberg constructs a probabilistic automaton $\mathcal{A}$ with multiple states that have a normal state and activated states having a frequent arrival rate of documents (Fig. 8). When $\mathcal{A}$ is in normal state q_0, messages are emitted slowly, with gap x between consecutive documents distributed independently according to $f_0(x) = \alpha_0 e^{-\alpha_0 x}$. When $\mathcal{A}$ is in activate state q_i, messages are emitted frequently, with gap distributed independently according to $f_i(x) = \alpha_i e^{-\alpha_i x}$. In Kleinberg's model, q_i is given by a document arrival rate $\alpha_i = \hat{g}^{-1}\beta^i$, where $\hat{g}^{-1}$ is the average arrival rate for this document stream, and β is a parameter representing the resolution of each state. In keeping with the above, a document in interval x_k can be considered to be in a normal state if $f_0(x_k) > f_1(x_k)$. Conversely, a document in interval x_k can be considered to be in an activated state if $f_0(x_k) < f_1(x_k)$. Moreover, if $f_j(x_k) < f_l(x_k)$ ($j < l$), then its document can be considered to be in a more activated state. Documents in activated states are in burst.

Additionally, this algorithm calculates transition costs $\tau(j, l)$ for every state transition from j to l, where $j < l$. This automaton, therefore, requires transition costs from lowerintensity burst to higher-intensity burst. Introducing this cost function, state transitions do not happen easily. Kleinberg then introduces cost function as $\tau(j, l) = (j - l)\gamma \ln N$, where γ is a parameter representing the degree of transition difficulty. When the costs are calculated, the sequence of states that minimizes the total cost is selected. The minimum transition cost until the arrival time of the k-th document is calculated as follows:

$$C_l(k) = -\ln f_l(x_k) + \min_j(C_j(k-1) + \tau(j, l)) \qquad (4)$$

Kleinberg applies the Hidden Markov Model (HMM) [7] for this probabilistic automaton $\mathcal{A}$. State q_i is an inner state in HMM, and sequence of gaps $\mathbf{x}$ is the probabilistic outputs for each state. To discover bursts from a document stream, the Viterbi algorithm [10], a well-known estimation algorithm, is used to extrapolate an optimal sequence of state transitions satisfying minimum cost as Formula (4) from a sequence of outputs.

4.2 Modeling Activation Levels

As discussed above, when bookmark frequency is higher than usual, the activation level is high, suggesting that the attention toward a bookmarked page is on the increase. Conversely, when bookmarks become less frequent, the activation level is lower, suggesting that the attention might have decreased, indicating obsolescence. The activation level is defined as the comparative difference in bookmark frequencies between the baseline and the specific time. Just as with the document stream in Kleinberg's algorithm, we now treat a sequence of bookmarks that occur chronologically. There are gaps x_t between timestamps of the k-th bookmark and $k+1$-th bookmark. In all N bookmarks of a web page, a sequence of gaps $\mathbf{x} = (x_1, x_2, ..., x_{N-1})$ shows time-series variation of people's attention to its page.

We estimate the current value of a web page based on the time-series variation of its bookmarks. Because pages have different content characteristics, the way attention is attracted differs. When we estimate whether or not a page is obsolete, and attracting attention or not, we should consider an individual baseline for each. A factor representing that difference is the average bookmarking rate of a web page. For its baseline, we define the average bookmark gap $\hat{g}$ as follows:

$$\hat{g} = \frac{1}{\frac{1}{2}N} \sum_{k=\frac{1}{4}N}^{\frac{3}{4}N} \chi_k \qquad (5)$$

where $\chi = (\chi_1, \chi_2, ..., \chi_{N-1})$ ($\chi_k \in \mathbf{x}$) is a sequence of bookmark gaps, sorted by length of the gap. Consequently, $\hat{g}^{-1}$ means the average bookmarking rate. We exclude highly frequent spans and infrequent spans from this calculation to be robust, because gaps in burst and obsolescence might occur incidentally. Further, a distribution of bookmark timestamps in SBM is sensitive because the number of bookmarks may not be very large. We use the interquartile range $[\frac{1}{4}N, \frac{3}{4}N]$, which is a robust estimator of the distribution [11], for the range of this calculation.

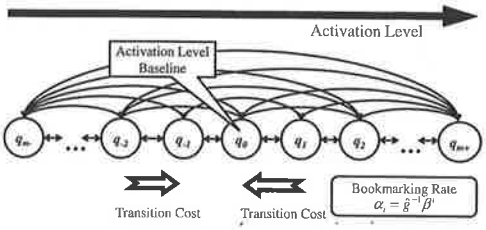

Fig. 9 A model of activation levels in SBM.

State q_0 is the normal state of a web page. Its state has average bookmarking rate $\hat{g}^{-1}$ and the baseline of activation level 0 of the page. Just as above, each state q_i has a bookmarking rate $\alpha_i = \hat{g}^{-1}\beta^i$. In state q_i, a bookmark appears with gap x between consecutive bookmarks distributed independently according to $f_i(x) = \alpha_i e^{-\alpha_i x}$ as with Kleinberg's algorithm. If a web page state is q_i, its activation level is i. If its activation level is more than 0, it seems activated. Conversely, its activation level is less than 0, so it seems deactivated. Probabilistic automaton $\mathcal{A}$, which shows activation level transitions of a web page, has internal states $\mathbf{q} = \{q_{m_-}, q_{m_-+1}, ..., q_{-1}, q_0, q_1, ..., q_{m_+-1}, q_{m_+}\}$ (Fig. 9).

Not only are there web pages having time axis, such as news scripts and some hot issues, but there are also web pages with time-homogeneous contents. A time-homogeneous page tends to be bookmarked constantly and continually. That means it keeps its freshness longer than a temporal page, and it should not turn obsolete easily. When a transition from activation level j to l occurs, the transition cost is as follows:

$$\tau(j, l) = |l - j|\gamma \log N \log SD \quad \gamma > 0 \qquad (6)$$

where SD is the standard deviation of bookmark timestamps of the web page. We set the cost function so as not to transit easily to obsolescence. If a page has been bookmarked for a long time, it should keep the current state. Conversely, if a page is bookmarked only temporally, the activation level should be easy to change. Parameter SD archives these goals.

This paper introduces eleven-level states $\mathbf{q} = \{q_{-5}, ..., q_{-1}, q_0, q_1, ..., q_5\}$ and set $\beta = 4$ and $\gamma = 10$.

4.3 Estimation of Activation Levels

From an SBM service, we can only get a current snap shot of the web page's bookmark history. To estimate the current activation level of a web page, we assume that a hypothetical $N+1$-th bookmark is newly created and its sequence of bookmark gaps turns $\mathbf{x}' = (x_1, x_2, ..., x_{N-1}, x_N)$. The Viterbi algorithm is then applied to $\mathbf{x}'$; it extrapolates an optimal sequence of state transitions $\mathbf{s} = (s_1, s_2, ..., s_{N-1}, s_N)$. Finally, state s_N estimates current activation level $act(p)$ of web page p.

We show examples of estimating activation levels in

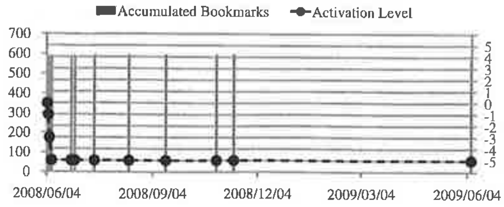

Fig. 10 Examples of activation level estimation 1.

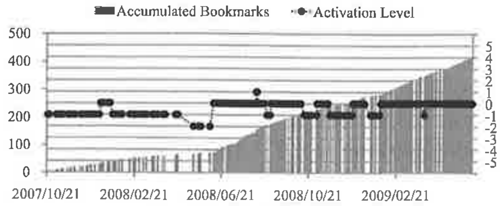

Fig. 11 Examples of activation level estimation 2.

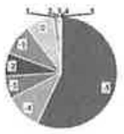

Fig. 12 Distribution of activation levels.

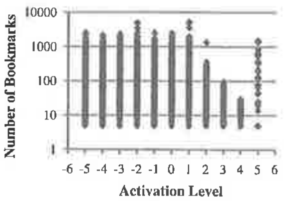

Fig. 13 Correlation between activation levels and num. of bookmarks.

Figs. 10 and 11[†]. Further, we show distribution of activation levels in Figs. 12 and 13. They are based on analysis of Hatena Bookmark data as of June 7, 2009. The page demonstrated in Fig. 10 received over 500 bookmarks in the first two days. After that, however, the frequency of bookmarkings for this page became extremely low. Currently, no one bookmarks this page. In the estimated activation levels for this page, a high activation level is maintained only for the first few days; after that it is at low activation level. That means this page seems obsolete at present. The page shown in Fig. 11 has been bookmarked regularly. The estimated activation levels for this page stay around the average.

Figure 12 shows the distribution of activation levels. It shows that many web pages are currently estimated obsolete. Figure 13 shows the correlation between activation levels and the number of bookmarks. Many pages received a huge number of bookmarks, but are regarded obsolete now.

5. S-BITS*

This section proposes a method to improve our previous S-BITS work by introducing the activation level concept. The improved method is called S-BITS* and takes into account the activation level of each page in the page and user value calculation step.

We now look at web page ranking incorporating the activation level concept. In this work, an activation level is the integer value from −5 to 5. We normalize activation levels into [0, 1]. To normalize them, we use the Sigmoid function (Fig. 14), which is a sampling function used in neural network computation; it is well-known in the area of pattern recognition. The Sigmoid function is defined as follows:

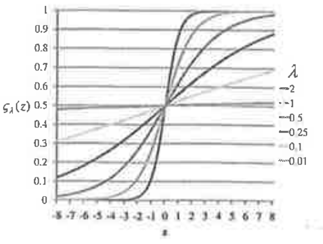

Fig. 14 Sigmoid function.

$$\varsigma_\lambda(z) = \frac{1}{1 + \exp(-\lambda z)} \tag{7}$$

where if λ is close to 0, it approaches asymptotically to $\varsigma_\lambda(z) = 0.5$, and if λ is close to ∞, it approaches asymptotically to a step function. Introducing normalization with this function, we can assign a weight of freshness factor based on the activation level.

To improve S-BITS by introducing the activation level concept, we modify formulas of evaluating page scores and user scores as follows:

$$p\_score_i^k = \sum_{b_{ji} \in B} \varsigma_\lambda(act(p_i) + 1)\ u\_score_j^{k-1} \tag{8}$$

$$u\_score_i^k = \sum_{b_{ij} \in B} \varsigma_\lambda(act(p_j) + 1)\ p\_score_j^{k-1} \tag{9}$$

Applying these formulas to the S-BITS algorithm [5], page scores and user scores are calculated as in S-BITS. Finally, ordering by page scores, a web page ranking is created. We call this improved method S-BITS*.

6. SBRank*

Yanbe [3] proposed SBRank, which indicates how many users bookmarked a page, and estimated effectiveness of SBRank as an indicator of web search. Yanbe re-ranks the top M web pages obtained from a web search engine based on the SBRank values of those web pages. Yanbe usually uses SBRank value with multiple factors. Yanbe further suggests integrating SBRank and PageRank [1] into a ranking method. The activation level concept can be applied to Yanbe's method. We therefore improve SBRank as well, then compare performance with the original SBRank. We call this improved method SBRank*. SBRank* considers the activation level concept as follows:

$$SBRank^*(p) = \varsigma_\lambda(act(p) + 1)\ SBRank(p) \tag{10}$$

[†] Figures 10 and 11 are based on the following URLs:
Fig. 10: http://www.softbankmobile.co.jp/ja/news/press/2008/20080604_01/
Fig. 11: http://ipodtouchlab.com/

Table 1 Top 3 results of S-BITS.

Rank	Page Title	BMs	Act.	SD
1	About "iPhone 3G" (Press Release)	593	-5	10.53
2	APPLE - iPhone -	184	-3	89.38
3	iPod · iPod touch lab	417	0	158.9

Table 2 Top 3 results of S-BITS*.

Rank	Page Title	BMs	Act.	SD
1	iPod · iPod touch lab	417	0	158.9
2	iPhone 3G Wiki*	165	-1	107.5
3	iPhone.Walker	75	-1	87.25

where $act(p)$ is the activation level of page p, $SBRank(p)$ is the value of SBRank of page p. Similar to Yanbe's method, we re-rank web pages based on the SBRank* value. Introducing the activation level concept, we try to raise the ranks of fresh web pages and to lower the number of obsoletes.

7. Experiments

This section presents and discusses experimental results. We measure effectiveness of S-BITS*, the proposed method. Judging the results of its rankings, we discuss effectiveness of the proposed model of activation levels.

For the experiment, we need to decide an appropriate value of the parameter λ in the sigmoid function. To find a better value, we did a preliminary experiment; we found that $\lambda = 1.0$ is close to human intuitiveness. In the experiment below, this value is used.

We compared and analyzed five ranks: the S-BITS* rank, the S-BITS rank, the SBRank* rank, the SBRank rank and the original Yahoo! rank. We used the Yahoo! Web Search API [12] as the Web Search API. We collected SBM data in *Hatena Bookmark*. Details of the collected data are as follows:

- Number of pages: around 600 thousand
- Number of bookmarks: around 10 million

Collected SBM data are based on data of *Hatena Bookmark* as of June 7, 2009. These data are stored in a relational database. For these web pages, we calculated activation levels in advance. We used the HMM Tool Kit (HTK) [13].

The initial page sets for S-BITS*, S-BITS, SBRank* and SBRank are obtained from the Yahoo! Web Search API. We took the top 200 pages from that API. The Yahoo! ranking is also based on the Yahoo! Web Search API. To evaluate each method manually, we recruited examinees.

- Examinees: 20 people
- Number of Queries: 10 queries (5 queries per examine, 100 cases total)
- Query keywords:
  iphone, java, ruby, php, web design,
  How to write thesis[†], English Learning[†],
  Recipe[††], Diet[††], Project Management[††]

We asked examinees to distinguish whether or not the page is fresh and whether or not it is informative. To avoid the psychological effects of any particular method, we prepared a web page list that has a set of URLs that includes the top 20 results of each method. The lists have no duplication and no descriptions, and mixed list order is random.

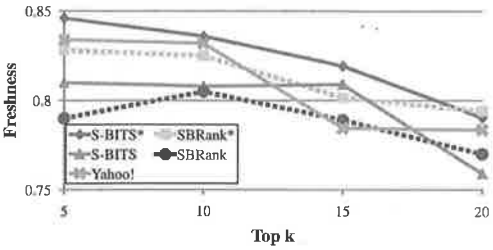

Fig. 15 Freshness.

Specifically, when the query keyword is "iphone," S-BITS* showed good performance. Tables 1[†††] and 2[††††] show the top 3 results of S-BITS and S-BITS*, where *BMs* is the number of bookmarks of its web page, *Act.* is the activation level and *SD* is the standard deviation of its bookmark timestamp distribution. In the ranking of S-BITS, the page describing the past press release falls in the first rank. In SBM, such web pages attracted many bookmarks. But those bookmarks resulted from temporary user attention, so at present, people feel the page is obsolete. In the activation estimation, the activation level of that page is estimated at −5, suggesting that the page is obsolete. The methods based on bookmark count, such as SBRank and S-BITS, cannot detect such web pages. Introducing the activation level concept, however, overcomes this weakness.

Freshness at the top k ranking level is measured as the percentage of web pages that are judged as fresh by an examinee within the top k result web pages. Fig. 15 shows the average freshness values at the top k ranking level for each method. S-BITS* shows higher freshness than the others. Yahoo! shows higher freshness similar to that of S-BITS*. SBRank* is higher and also out-performs SBRank, which is the baseline method of SBRank*. S-BITS and SBRank perform poorly compared with the others. Since SBRank and S-BITS have no way to include the activation level concept, they cannot out-perform the others. Yahoo! is a sophisticated search engine system that evaluates and ranks pages by mixed values. High freshness, therefore, can be achieved. S-BITS* and SBRank* improve baseline methods by intro-

[†] These keywords are written in Chinese characters.

[††] These keywords are written in KATAKANA, Japanese characters.

[†††] Web pages listed in Table 1 are based on the following URLs:
Rank 1: http://www.softbankmobile.co.jp/ja/news/press/2008/20080604_01/
Rank 2: http://www.apple.com/jp/iphone/
Rank 3: http://ipod touch lab.com/

[††††] Web pages listed in Table 2 are based on the following URLs:
Rank 1: http://ipod touch lab.com/
Rank 2: http://iphone.wikiwiki.jp
Rank 3: http://iphonewalker.net/

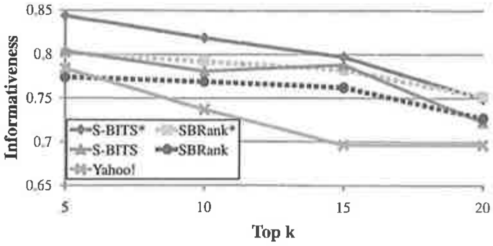

Fig. 16 Informativeness.

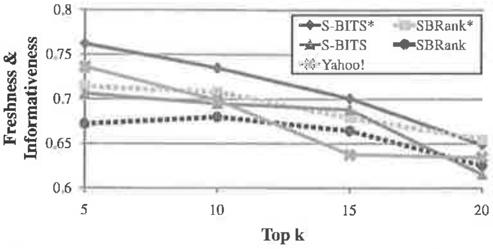

Fig. 17 Freshness and informativeness.

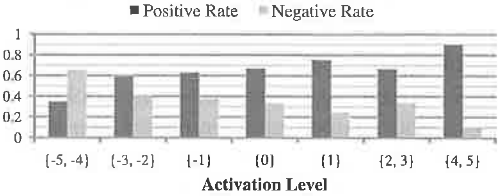

Fig. 18 Relevance between activation levels and examinees' postings.

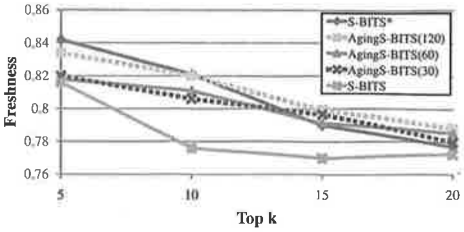

Fig. 19 Freshness comparison among 3 types of S-BITS.

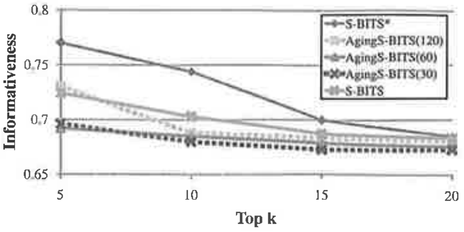

Fig. 20 Informativeness comparison among 3 types of S-BITS.

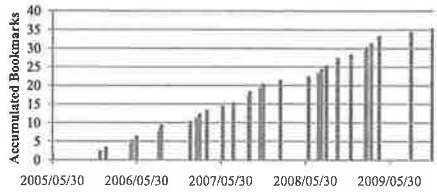

Fig. 21 Bookmarking history example for slow- and long-time bookmarked Web page.

ducing the activation level concept.

Informativeness at the top k ranking level is measured similarly to freshness. Fig. 16 summarizes the result. S-BITS*, our proposed method, shows higher informativeness. SBRank*, S-BITS and SBRank, which rank web pages using SBM data, are better than Yahoo! Focusing on the activation level of the web page, we can improve user satisfaction regarding informativeness.

Then, looking at both informativeness and freshness, S-BITS* performs best of the five methods, and SBRank* improves SBRank (Fig. 17). We can, therefore, say that our new method can improve both informativeness and freshness.

Next, we evaluated whether activation levels of the proposed model are reasonable (Fig. 18). In this experiment, we asked examinees to distinguish freshness. Using the results, we evaluated relevance between activation levels and examinee postings. Results show a positive posting rate and a negative posting rate in all postings for each set of web pages, which have the same activation level. For web pages having activation levels greater than 0, 65% positive postings were achieved. Web pages having activation levels smaller than 0 increased negative postings along with declining activation level. We can, therefore, say that our proposed modeling of activation levels for a web page in SBM is valid.

We also compared our proposed method with a baseline method. The baseline is Aging S-BITS, which introduces the concept of aging for bookmarks. The computational way of Aging S-BITS is described in Sect. 2. We set the parameters of Aging S-BITS as follows: $T_0 = 1.0$ and $t_{1/2} = \{30, 60, 120\}$. We used the same data set and same queries as in the above experiments. In this experiment, we compared S-BITS, S-BITS* and Aging S-BITS as well as the above comparison [†]. Figures 19 and 20 show evaluation results of freshness and informativeness. In the freshness result, baselines show better results than S-BITS, but S-BITS* is best of all. However, baselines had no advantages in informativeness. Baselines yielded high scores to recent highly bookmarked pages based on the half-life point $t_{1/2}$. That resulted because, in that scoring, the potential popularity, such as bookmark frequency, was not taken into account; the web pages that were bookmarked in the slow rate for a long time may be estimated as obsolete pages. For instance, the web page "Java -TECHSCORE-"[††] (Fig. 21), which was bookmarked by 35 users since May 2005, is ranked in 2nd place on S-BITS*, 4th place on S-BITS, 9th place on Aging S-BITS(120) and Aging S-BITS(60), and 12th place on Aging S-BITS(30). This page is a very informative web page because it is given positive votes by all examinees. Nevertheless, this page was bookmarked at less frequent intervals,

[†] This experiment was held in December, 2009.

[††] http://www.techscore.com/index.html

Table 3 Execution time.

Query	Method	API	Init.	Ext.	Rank
ruby	S-BITS*	1.63	1.60	7.83	1.50
ruby	S-BITS	1.61	1.59	7.74	2.36
ruby	SBRank	1.65	1.59	–	0.024
php	S-BITS*	1.65	3.48	8.27	5.99
php	S-BITS	1.62	3.45	8.19	2.12
php	SBRank	1.74	3.48	–	0.011
java	S-BITS*	2.34	1.14	7.31	0.785
java	S-BITS	1.70	1.09	7.30	1.09
java	SBRank	1.73	1.11	–	0.0049
iphone	S-BITS*	2.32	3.72	8.48	2.79
iphone	S-BITS	1.78	3.69	8.39	7.30
iphone	SBRank	1.66	3.70	–	0.016
web design	S-BITS*	1.60	0.419	16.62	1.95
web design	S-BITS	1.83	0.42	16.20	3.42
web design	SBRank	1.69	0.39	–	0.051
Recipe	S-BITS*	1.74	3.57	9.08	2.40
Recipe	S-BITS	2.026	3.54	9.019	2.52
Recipe	SBRank	1.85	3.55	–	0.010
Thesis	S-BITS*	1.85	0.55	7.27	0.64
Thesis	S-BITS	1.78	0.53	7.22	19.45
Thesis	SBRank	1.76	0.53	–	0.0047
English	S-BITS*	1.85	1.66	7.33	3.36
English	S-BITS	1.88	1.63	7.32	2.14
English	SBRank	1.81	1.64	–	0.0027
Diet	S-BITS*	1.87	0.377	7.39	0.19
Diet	S-BITS	1.73	0.34	7.375	0.52
Diet	SBRank	2.00	0.35	–	0.0036
PM	S-BITS*	1.80	0.15	7.38	0.13
PM	S-BITS	1.83	0.13	7.36	1.21
PM	SBRank	1.76	0.12	–	0.0038
Over All	S-BITS*	1.86	1.67	8.70	1.97
Over All	S-BITS	1.78	1.64	8.61	4.21
Over All	SBRank	1.76	1.65	–	0.013

so it could not get a high score in Aging S-BITS. S-BITS*, which introduces the activation concept, estimates web page freshness values based on the potential bookmark frequency, so it can give valid scores to such web pages. For these reasons, baselines could not do better than S-BITS*.

Finally, we discuss computational costs at runtime. Table 3 shows average execution times ([sec.]) of four parts for S-BITS, S-BITS* and SBRank. First, the time for the API is from the time a query is issued to the time results are received from the Yahoo! API. Second, the Init. includes database access to get SBM data and create a bipartite graph. The third time Ext. includes running association rule mining to find frequent tag sets and access the database to get additional data based on these tags. And last, in the Rank time, each method calculates scores and ranks pages.

These profiles were executed on an Ubuntu 9.10 Linux machine with an Intel(R) Core(TM)2 Quad CPU Q6700 (2.66 GHz) and 3.2 GB of main memory. We implemented the web search system with Ruby 1.8.7 and PostgreSQL 8.4. We ran 100 trials consisting of 10-time trials for each 10 queries[†], and used *benchmark.rb*[††] for this profiling.

The execution times for the API and Init. showed similar results among the three methods. Ext., which is an operation specific to our proposed methods, is the most time-consuming part of runtime. But it can be improved by optimizing indexes of the data. In the ranking part, S-BITS consumed a long execution time because it is based on the convergence calculation of the HITS algorithm. However, S-BITS* revealed that execution time could be reduced. S-BITS* calculates scores using the weights of activation levels. As a result of the weighting, the number of convergence trials might be decreased. Our proposed methods have higher computational costs than the other methods, but, as discussed before, we are able to improve user informativeness.

8. Related Work

With the spread of social bookmarking, research into Folksonomy has increased. Golder and Huberman [14] analyzed the structure of social bookmarking in detail and reported regularities for a user's tagging behavior and the nature of tagging. Xian Wu et al. [15] described the exploitation of social bookmarking and suggested a probabilistic generative model for developing a semantic web. Bao et al. [16] demonstrated that keyword associations based on social annotations can improve web searches. Yanbe et al. [3] proposed SBRank, which indicates how many users bookmarked a page, and estimated the effectiveness of SBRank as an indicator of web search. They further suggested integrating SBRank and PageRank [1] into a ranking method. Schmitz et al. [17] applied association rule mining to a tripatite graph of Folksonomy. Heymann et al. [4] provided an empirical analysis of how social bookmarkings can influence Web search. Hotho et al. [18], [19] proposed FolkRank, which is related to PageRank, to discover topic-specific trends within folksonomies. By focusing on the temporal aspect of folksonomies, they discover the popularity change of their elements. Capocci et al. [20] focused on the temporal statistics of tag arrival times; then they found correlation and collaboration patterns. Menjo [21] proposed a hotness predication method for the newly submitted pages in SBM services. Chi and Mytkowicz [22] used information theory to understand the impact of population change and content growth over time. Elizeu Santos-Neto et al. [23] discussed tag reuse characteristics by tracing user tag usages.

Kleinberg's algorithm [9] is well-known for identifying bursty topics from document streams such as blogs and news documents. Research into activation analysis has been approached in various ways. Cui and Kitagawa [24] proposed a novel topic activation analysis scheme that incorporates both document arrival rate and relevance. They then proposed an incremental scheme more appropriate for a document streaming environment. Fujiki et al. [25] proposed an approach to detect word bursts in blogs and BBS. Fung et al. [26] proposed a method to detect bursty events based on feature distributions with no tuning parameters.

We have proposed a method to model activation levels

[†] These queries are same as the above expriments. Some queries are shortened as follows: How to write thesis : Thesis, English Learning : English, Project Mangement : PM.

[††] http://www.ruby-lang.org/ja/man/html/benchmark.html

of web pages in SBM, and to estimate their current activation levels. Existing work on activation analysis focuses on topic bursts in documents such as e-mail, blogs and BBS. We focus on the bookmark timestamps of web pages in SBM, and have proposed a method to detect obsolete pages as well as activated pages. Application of the activation level concept to web page ranking is also the main contribution of this work.

9. Conclusion

Focusing on timestamps, which show bookmarked times in SBM, and bookmark frequency, we modeled the activation level of a web page in SBM. Based on this activation level concept, we proposed a method to estimate the activation level for web pages. Further, we proposed a web page ranking method using this activation level concept. With the introduction of the activation levels, S-BITS*, the newly proposed method, outperforms its S-BITS baseline method. Experiments confirmed the effectiveness of the improved ranking methods and also confirmed our proposed activation level model. Future work is to consider community based activation level estimation and to develop a way to estimate the optimal parameter values.

Acknowledgements

This research has been supported in part by the Grant-in-Aid for Scientific Research from MEXT (# 21013004).

References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Technical Report, Stanford Digital Library Technologies Project, 1998.

[2] J. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM, vol.46, no.5, pp.604–632, 1999.

[3] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, "Towards improving web search by utilizing social bookmarks," Proc. 7th International Conference on Web Engineering (ICWE 2007), pp.343–357, Springer-Verlag Berlin Heidelberg, 2007.

[4] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?," Proc. international Conference on Web Search and Web Data Mining (WSDM 2008), pp.195–206, ACM New York, NY, USA, 2008.

[5] T. Takahashi and H. Kitagawa, "S-BITS: Social-bookmarking induced topic search," Proc. Ninth International Conference on Web-Age Information Management (WAIM 2008), pp.25–30, 2008.

[6] T. Takahashi and H. Kitagawa, "A ranking method for web search using social bookmarks," Proc. 14th International Conference on Database Systems for Advanced Applications, pp.585–589, Springer, 2009.

[7] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–286, 1989.

[8] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," Proc. 1993 ACM SIGMOD International Conference on Management of Data, pp.207–216, ACM New York, NY, USA, 1993.

[9] J. Kleinberg, "Bursty and hierarchical structure in streams," Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.91–101, ACM New York, NY, USA, 2002.

[10] G. Forney, Jr., "The viterbi algorithm," Proc. IEEE, vol.61, no.3, pp.268–278, 1973.

[11] J. Tukey, "Exploratory data analysis," 1977.

[12] "Yahoo! Web Search API," http://developer.yahoo.co.jp/search/web/V1/webSearch.html

[13] "HTK Web-Site." http://htk.eng.cam.ac.uk/

[14] S. Golder and B. Huberman, "The structure of collaborative tagging systems," Arxiv Preprint cs.DL/0508082, 2005.

[15] X. Wu, L. Zhang, and Y. Yu, "Exploring social annotations for the semantic web," Proc. 15th International Conference on World Wide Web (WWW 2006), pp.417–426, ACM New York, NY, USA, 2006.

[16] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," Proc. 16th International Conference on World Wide Web (WWW 2007), pp.501–510, ACM New York, NY, USA, 2007.

[17] C. Schmitz, A. Hotho, R. Jaschke, and G. Stumme, "Mining association rules in folksonomies," Proc. IFCS 2006 Conference, pp.261–270, Springer, 2006.

[18] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information retrieval in folksonomies: Search and ranking," The Semantic Web: Research and Applications, vol.4011, pp.411–426, 2006.

[19] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Trend detection in folksonomies," Proc. First International Conference on Semantics And Digital Media Technology (SAMT), pp.56–70, Citeseer, 2006.

[20] A. Capocci, A. Baldassarri, V. Servedio, and V. Loreto, "Statistical properties of inter-arrival times distribution in social tagging systems," Proc. 20th ACM Conference on Hypertext and Hypermedia, pp.239–244, ACM, 2009.

[21] T. Menjo and M. Yoshikawa, "Trend prediction in social bookmark service using time series of bookmarks," Proc. WWW2008 Workshop on Social Web Search and Mining (SWSM 2008), 2008.

[22] E. Chi and T. Mytkowicz, "Understanding the efficiency of social tagging systems using information theory," Proc. Nineteenth ACM Conference on Hypertext and Hypermedia, pp.81–88, ACM New York, NY, USA, 2008.

[23] E. Santos-Neto, D. Condon, N. Andrade, A. Iamnitchi, and M. Ripeanu, "Individual and social behavior in tagging systems," Proc. 20th ACM Conference on Hypertext and Hypermedia, pp.183–192, ACM, 2009.

[24] C. Cui and H. Kitagawa, "Topic activation analysis for document streams based on document arrival rate and relevance," Proc. 2005 ACM Symposium on Applied Computing (SAC 2005), pp.1089–1095, ACM New York, NY, USA, 2005.

[25] T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura, "Identification of bursts in a document stream," Proc. 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004), 2004.

[26] G. Fung, J. Yu, P. Yu, and H. Lu, "Parameter free bursty events detection in text streams," Proc. 31st International Conference on Very Large Data Bases, pp.181–192, VLDB Endowment, 2005.

Tsubasa Takahashi received a bachelor's degree in information engineering and a master's degree in engineering, all from the University of Tsukuba, in 2008 and 2010, respectively. His research interests include information retrieval and data mining.

Hiroyuki Kitagawa received a B.Sc. degree in physics and M.Sc. and Dr. Sc. degrees in computer science, all from the University of Tokyo, in 1978, 1980, and 1987, respectively. He is currently a full professor at the Graduate School of Systems and Information Engineering and at the Center for Computational Sciences, University of Tsukuba. He is Chairperson of the Department of Computer Science, University of Tsukuba and also Chairperson of the Institute of Information Sciences and Electronics, University of Tsukuba. His research interests include the integration of information sources, data mining, stream-based ubiquitous data management, distributed data processing architecture, Web data management, XML, and scientific databases. He has published more than 170 papers in refereed journals and international conference proceedings. He is a Fellow of IPSJ, Trustee of DBSJ, and a member of ACM, IEEE Computer Society, and JSSST.

Keita Watanabe received a bachelor's degree in information science from the University of Tsukuba in 2009. He is currently a masters degree candidate at the Graduate School of Systems and Information Engineering, University of Tsukuba. His research interests include information retrieval and data mining.