Trend Prediction in Social Bookmark Service Using Time Series of Bookmarks

Takashi Menjo
Graduate School of Information Science
Nagoya University
Nagoya, 464-8601, Japan
menjo@dl.itc.nagoya-u.ac.jp

Masatoshi Yoshikawa
Graduate School of Informatics
Kyoto University
Kyoto, 606-8501, Japan
yoshikawa@i.kyoto-u.ac.jp

ABSTRACT

Social bookmark service is a web-based service which enables its users to manage and share their bookmarks on Web pages. Many bookmarks are aggregated and shared on social bookmarks, so they become useful news sources now. In this paper, we propose a trend prediction method of newly-posted pages, using time sequential data of users and "tags" annotated to bookmarks.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Social Bookmark, Trend Prediction, Ranking Algorithm

1. INTRODUCTION

In the past few years, *social bookmark services* such as del.icio.us [1] and Hatena Bookmark [3] have been becoming popular. Those services enable users to annotate and save their bookmarks and share them with others on the Web. Users can find useful information by looking through aggregated and shared bookmarks which had been hidden in users' local web browsers.

Bookmarks are categorized by *tags* in social bookmark services. Tags are short keywords which users can add to their bookmarks without depending on a controlled vocabulary. In contrast to services like web directories, which the provider classifies its contents based on a predefined taxonomic structure, tag-based categorizing systems have been recently termed *folksonomy*, short for "folk taxonomy." Those systems are now supported not only in social bookmark services but also in weblogs, photo-sharing sites like Flickr [2], video-sharing sites like YouTube [4], and many other user-generated media sites.

Major search engines measure importance of web pages based on hyperlink analysis algorithm such as PageRank [10]. However, link analysis costs expensively and it is not suitable to evaluate new pages because new pages need certain period before they obtain in-links. On the other hand, social bookmark services basically rank shared pages by the number of users who have bookmarked the page. For example, most social bookmarking sites have a *popular list* or a *hot list* on the front page, which display shared pages bookmarked or just being bookmarked by many users recently. Viewing those list, users can easily get a sense of pages which are catching attention of other users

We believe that there are more useful information for ranking in social bookmark services. For example, in the current systems, any user is counted as one user even if s/he usually bookmarks useful pages or s/he is a spammer. However, we will be able to rank the page bookmarked by the former users highly if we take their activities of bookmark into account. As well as tags, we can catch trend topics or interests of users from them.

In this paper, we propose a new trend prediction method in social bookmark services. Our method targets the pages newly shared in the social bookmark, predicts how much attention will be attracted to those pages, and ranks them based on it. We use time series information of bookmarks to evaluate users and tags, and to predict new pages which would collect many markups.

This paper is organized as follows. Section 2 reviews related works on social bookmarks. Section 3 proposes the trend prediction method. Section 4 reports the experimental results. Section 5 draws the conclusion and work to be done.

2. RELATED WORKS

Recently, much research on social bookmark has been done [5, 6, 13, 12, 7, 8], especially on bookmark recommendation [13, 9, 11]. Our goal is to evaluate and rank the pages, rather than to recommend them.

Some research on ranking algorithm related to social bookmarks has been carried out. Hotho [7, 8] proposed the ranking algorithm in social bookmarks, which is inspired by PageRank. Yanbe et al.[14, 15] proposed hybrid ranking algorithm based on PageRank and social bookmarks for improving the precision of web search. These algorithms focus on all the pages shared in a social bookmark or all the pages on the Web; however, we focus on newly bookmarked pages.

3. TREND PREDICTION OF WEB PAGES

3.1 Modeling bookmarks

Figure 1 shows transition of bookmarks on a certain page shared in Hatena Bookmark, aggregated in every six hours

WWW2008, April 21–25, 2008, Beijing, China.
.

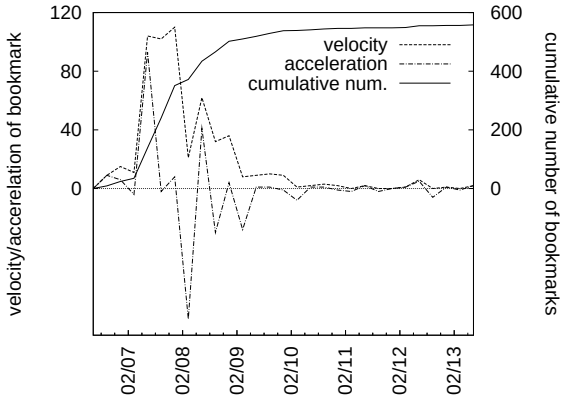

Figure 1: Transition of bookmark.

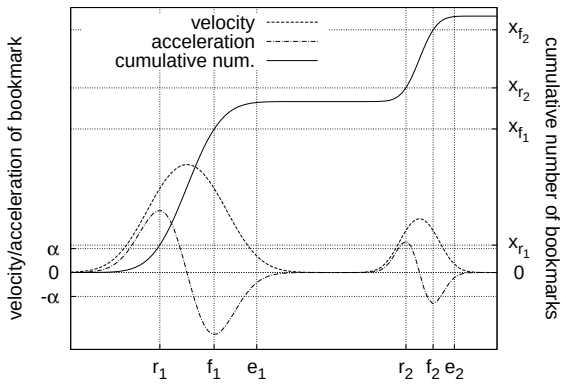

Figure 2: Transition of bookmark (idealized).

during one week from the time the page was shared for the first time. The dashed line indicates the number of users who have bookmarked the page in each time interval, and the chain line indicates the difference of them between two consecutive intervals. Those two lines can be regarded as velocity and acceleration of increase of bookmark, respectively. Also, the solid line shows the cumulative number of bookmarks until each interval. The figure shows that the velocity reached its peak within a day or two from the time the page was shared for the first time, and decreased gradually afterward. Pointed by Biddulph [5] and Golder et al.[6], this pattern is also observed in del.icio.us. In addition, the acceleration took a large local maximum value and minimum value at the time which the velocity reached the beginning of the peak and the ending of it. The cumulative number of bookmarks bursts in this period. Our basic idea is to treat the users who bookmarked the page before the acceleration took a large local maximum value as those who found out the importance of the page.

Now, we will formalize the above mentioned basic idea. First, we model a **bookmark** b as a quadruplet

$$b = (u, p, t, L),$$

where u is a user, p is a page, t is the time when b was shared, L is the set of tags given to b. A quadruplet indicates "when, who bookmarked which page with what tags." A user u can bookmark a page p at most once, so no two bookmarks can have the same pair of u and p. Also, u may not use any tag for a bookmark b, in which case $L = \emptyset$.

Then we denote a sequence of bookmarks on a page p in a time interval T starting from the first time p was bookmarked as

$$B_p^T = [(u_1, t_1, L_1), (u_2, t_2, L_2), \cdots, (u_n, t_n, L_n)].$$

Also, let τ be a unit time period such that $T = l\tau$ for a positive integer l. We denote a sequence of integers obtained by counting the bookmarks in B_p^T in every time period τ as

$$V_p^{(T,\tau)} = [v_1, v_2, \cdots, v_l],$$

that is,

$$v_i = \left|\left\{(u_k, t_k, L_k) \;\middle|\; (u_k, t_k, L_k) \in B_p^T \wedge\ t_1 + (i-1)\tau \le t_k < t_1 + i\tau\right\}\right|. \quad (1)$$

Each v_i is defined as **velocity** of bookmark at $t_1 + i\tau$. Likewise,

$$a'_i = v_{i+1} - v_i \quad (i = 1, 2, \cdots, l-1)$$

can be regarded as acceleration of bookmarks at $t_1 + i\tau$, and we denote their sequence as

$$A_p^{'(T,\tau)} = [a'_1, a'_2, \cdots, a'_{l-1}].$$

However, this sequence, in general, is oversensitive to local deviation of velocities. Hence we apply the following smoothing function to $A_p^{'(T,\tau)}$

$$\begin{array}{lll} a_1 & = (\quad 3a'_1 & +a'_2 \quad) / 5 \\ a_i & = (a'_{i-1}+3a'_i & +a'_{i+1}) / 5 \quad (i = 2, 3, \cdots, l-2) \\ a_{l-1} & = (a'_{l-2}+3a'_{l-1} & \quad) / 5 \end{array}$$

We define the sequence obtained after the smoothing

$$A_p^{(T,\tau)} = [a_1, a_2, \cdots, a_{l-1}]$$

as **acceleration**. Similarly, we denote the **cumulative number of bookmarks** between t_1 and $t_1 + i\tau$ as x_i, i.e.

$$x_i = \sum_{j=1}^{i} v_j \quad (i = 1, 2, \cdots, l).$$

We also denote the sequence of x_i's between t_1 and $t_1 + T$ as

$$X_p^{(T,\tau)} = [x_1, x_2, \cdots, x_l].$$

3.2 Evaluation of bookmark growth

Consider Figure 2 as idealized transition of bookmarks on a certain page. The horizontal axis shows subscript i, and vertical axis shows velocity v_i, acceleration a_i and cumulative number of bookmarks x_i. Both axises are discrete in fact, but shown continuously in the figure for simplicity. In the figure, cumulative number of bookmarks has grown twice. Pointed by Biddulph [5], little ripples of attention recur after the first burst. Each growth can be roughly observed between subscripts $i = r_j, f_j$ ($j = 1, 2$ in the figure), which correspond to the time acceleration reaches a local maximum value and a local minimum value. We call such an interval as j-th **growing interval**.

Given threshold values α, β (> 0) and a positive integer parameter γ, we obtain all the growing intervals of a certain page as described below:

1. Set $j = 1, s = 1$.
2. (Find the starting point of the next growing interval.)
   On $i \geq s$, find a minimal i to be assigned to r_j that satisfies $a_i > a_{i-1}$, $a_i > a_{i+1}$ and $a_i > \alpha$; or finish if such i was not found.
3. (Find the time point when the number of new bookmarks becomes very few.)
   On $i > r_j$, find a minimal i to be assigned to e_j that satisfies $|a_k| < \alpha$ $(i \leq k < i+\gamma)$ and $\sum_{k=i}^{i+\gamma-1} |a_k| < \beta$; or $e_j = l$ if such i was not found.
4. (Find the corresponding ending point of the growing interval.)
   On $r_j < i \leq e_j$, find a minimal i to be assigned to f_j that satisfies (a) and (b) below; or only (a) if such i was not found.
   (a) $a_i < a_{i-1}$ and $a_i < a_{i+1}$
   (b) $a_i < -\alpha$
   Still not found, $f_j = r_j + 1$.
5. $s = e_j + 1$, increment j, and go to step 2.

A threshold α is used to ignore little extreme values that do not seem to be the beginning of some bookmark growths at step 2 (and also the ending of them at step 4.)

Then we value each bookmark growth based on the following idea: it is better for the page to be bookmarked by users more broadly and quickly. Back to Figure 2, that is to say, the less the length of growing interval $f_j - r_j$ and the more the growth of bookmark $x_{f_j} - x_{r_j}$, the growth is more highly valued. So we value j-th growth of p with r_j and f_j as below:

$$g\text{-}score(p, j) = \frac{x_{f_j} - x_{r_j}}{f_j - r_j} \tag{2}$$

3.3 Trend prediction using values of users

In this section we value users using the value of the growth $g\text{-}score(p, j)$.

We consider that the users who noticed the importance of a certain page bookmarked it just before it reached some growing intervals. We denote a set of such users who bookmarked p within time $\gamma\tau$ before j-th growth as $U_{p,j}$. Namely,

$$U_{p,j} = \Big\{ u_k \;\Big|\; \big((u_k, t_k, L_k) \in B_p^{r_j \tau}\big) \wedge \big((u_k, t_k, L_k) \notin B_p^{(r_j-\gamma)\tau} \text{ (if } r_j > \gamma)\big) \Big\} \tag{3}$$

With equations (2) and (3), we define the value of the user u on j-th growth of p, $u\text{-}score_{p,j}(u)$, as below:

$$u\text{-}score_{p,j}(u) = \begin{cases} \dfrac{g\text{-}score(p,j)}{|U_{p,j}|} & (u \in U_{p,j}) \\ 0 & (u \notin U_{p,j}) \end{cases}$$

Then we value users as a whole by aggregating the values of those on every page in certain time interval. We would now also introduce *bookmark precision* of each user, that is the ratio of the grown pages to all the pages the user bookmarked. The larger the ratio is, the higher we want to value the user. Furthermore, we reject the users who bookmarked less than a certain number of pages because of insufficiency of information.

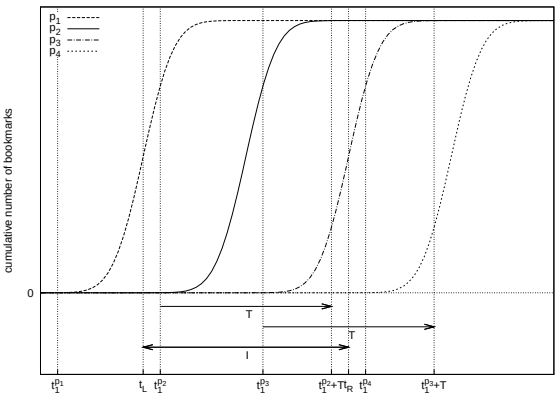

Figure 3: Relationship between T and I.

We denote a set of pages which a user u bookmarked in a time interval $I = [t_L, t_R)$ as

$$P_u^I = \{p \mid (u, p, t, L) \wedge t \in I\}.$$

We also denote a set of pages bookmarked in I for the first time and by u just before any growths of them as

$$P_u^{I,r} = \left\{ p \;\middle|\; u \in \bigcup_j U_{p,j} \wedge t_1^p \in I \right\}. \tag{4}$$

With these equations, we define the value of a user u as

$$u\text{-}score(u) = \begin{cases} \dfrac{|P_u^{I,r}|}{|P_u^I|} \sum_p \sum_j u\text{-}score_{p,j}(u) & (|P_u^I| \geq N) \\ 0 & (|P_u^I| < N) \end{cases}$$

where N is a positive integer parameter for rejection.

Now we complement T, I and equation (4). Figure 3 shows relationship between T and I on four pages p_1, p_2, p_3 and p_4. We only use the pages bookmarked in I for the first time, namely p_2 and p_3. Then we only use bookmarks from the time the page was bookmarked for the first time until T has passed, indicated by two right-pointing arrows in the figure. But if the arrow of T is over the current time, we only use the bookmarks until now.

Finally, we predict how much attention will be attracted to the page p as below:

$$hotness(p) = \frac{1}{t_{\text{now}} - t_1^p} \sum_{u \in U_p} u\text{-}score(u)$$

where U_p is a set of users who bookmarked p by the current time t_{now}.

3.4 Trend prediction using values of users and tags

In this section we introduce tags to extend our method described in section 3.3. It is to catch trend topics from tags and value pages about them highly. However, it is not always true that those tags which are used frequently express trend topics, because they tend to be common words. So we use *prominency* of tags, instead simply using frequency of tags.

Table 1: Parameters.

parameter	value
T	7 days
τ	1 hour
α	2
β	5
γ	6
N	28
t_{now}	03/01/07 00:00, 03/02/07 00:00, $\cdots$, 03/10/07 00:00
I	$[t_{\text{now}} - 28 \text{ days}, t_{\text{now}})$
I_1	$[t_{\text{now}} - 7 \text{ days}, t_{\text{now}})$
I_2	$[t_{\text{now}} - 1 \text{ days}, t_{\text{now}})$
K	100
T'	7 days

Given two time intervals with the same right endpoints $I_1 = [t_{L_1}, t_R), I_2 = [t_{L_2}, t_R)$ (that $t_{L_1} < t_{L_2}$), we define **prominency** of a tag l as below:

$$prominency(l) = \begin{cases} \dfrac{|I_1||B_l^{I_2}|}{|I_2||B_l^{I_1}|} & (|B_l^{I_1}| \neq \emptyset) \\ 0 & (|B_l^{I_1}| = \emptyset) \end{cases}$$

where B_l^I is a set of bookmarks which was shared in a time interval I and was given l to, namely

$$B_l^I = \{(u, p, t, L) \mid t \in I \wedge l \in L\}.$$

A prominency of l is the relative frequency of it in I_1 to that in I_2. It will be the maximum value when l was only used in I_2.

Finally, we denote the value of a set of tags L as

$$prominency(L) = \begin{cases} \max_{l \in L} prominency(l) & (L \neq \emptyset) \\ 1 & (L = \emptyset) \end{cases}$$

and predict attention of a page p as below:

$$hotness(p) = \frac{1}{t_{\text{now}} - t_1^p} \sum_{u \in U_p} u\text{-}score(u) \cdot \max\{prominency(L_{u,p}), 1\}$$

The minimum value of the term $\max\{\cdot\}$ is 1, so the users who bookmarked p is valued the degree of themselves at least.

4. EVALUATION EXPERIMENT

We carried out experiments to examine the effectiveness of our method. In the experiment, we created a couple of ranked lists of trend pages by our methods and compare them with a list created by the traditional method to examine the effectiveness.

First we set parameters as shown in Table 1 empirically.

Our experiment was carried on a set of Hatena Bookmark [3] data. From Hatena Bookmark, we retrieved 243084 URLs of pages which had been bookmarked for the first time from February 1, 2007 to March 9, 2007. Then we got all the bookmarks until seven days (= T) after the first time each page was bookmarked. Using this dataset, we valued users and tags by the method described in Section 3.

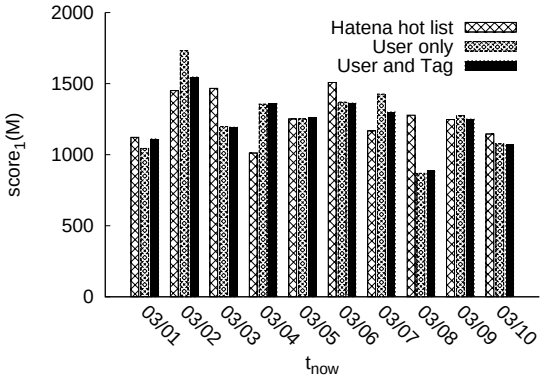

Figure 4: Experiment on $score_1(M)$.

We compared our lists with Hatena Bookmark hot entry list[1]. In this list, the pages bookmarked by a specified threshold number of users were ranked and displayed in reverse chronological order, with the most recent pages on top. The threshold was three as a default, and we used this value to create the hot entry list.

Assumed some t_{now} shown in Table 1 as current times, we created two lists by our methods described in Section 3.3 and Section 3.4, and the hot entry list. Then we removed the pages grown at least once from all the lists. We also removed the pages bookmarked by less than three users from our two lists. Finally we picked out the top 100 (= K) pages from each list to be a set M and evaluate it by equations below:

$$\begin{aligned} score_1(M) &= \sum_{p \in M} bookmark(p) \\ score_2(M) &= \sum_{p \in M} \frac{bookmark(p)}{rank_M(p)} \\ score_3(M) &= \sum_{p \in M} \frac{bookmark(p)}{\log_{10} rank_M(p) + 1} \\ score_4(M) &= \sum_{p \in M} \frac{\log_{10}[bookmark(p) + 1]}{\log_{10} rank_M(p) + 1} \end{aligned}$$

where $rank_M(p)$ is the rank of p in M, $bookmark(p)$ is the number of users who bookmarked p in seven days (= T') from t_{now}. Each equation evaluates a list by the number of users who bookmarked the pages in it. The larger the value, the higher the list is evaluated. The difference of each equation is to what degree it considers the rank of the page.

The results are shown in from Figure 4 to Figure 7. "Hatena hot list" indicated the score of the hot entry list of Hatena Bookmark. "User only" and "User and Tag" show the score of the lists created by our methods described in section 3.3 and 3.4, respectively. Comparing "User only" with "Hatena hot list", we can find that our list is better than that of Hatena (26 of 40). Similarly, "User and Tag" is better than Hatena (24 of 40).

Now, we emphasize that it is disadvantageous for us to evaluate our methods using Hatena Bookmark hot entry list. It can be natural for the pages ranked high in Hatena hot list to be bookmarked by more users because the list is displayed

[1]http://b.hatena.ne.jp/entrylist?sort=hot

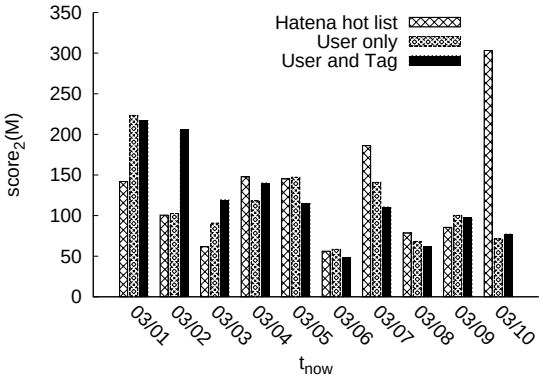

Figure 5: Experiment on $score_2(M)$.

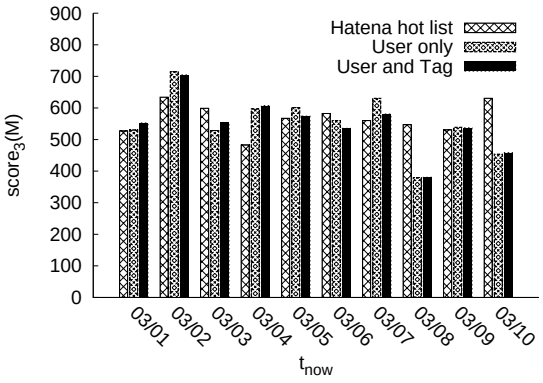

Figure 6: Experiment on $score_3(M)$.

on the front page of Hatena Bookmark and viewed by many users. Despite that our lists are not to be so, they are not inferior to the list of Hatena. The fact indicates that the results can be better if our lists are exposed to the users.

5. CONCLUSIONS

Social bookmark services have become popular not only as tools of annotating and sharing users' bookmarks, but also as new information sources. The services display the pages which is bookmarked by many users recently, ranking them by the number of users bookmarking each page. There are, however, more useful information for ranking. Current social bookmark services are not fully utilizing such information.

In this paper we proposed a trend prediction method using time series of bookmarks. We first define the growth of the page, treat the users who bookmarked it just before some growth as those who found out the importance of the page and value them. Besides, we introduced the prominency of tags to catch trend topics from them. We carried out experiments and found that our methods were superior to Hatena Bookmark hot entry list.

As a future work, we would like to introduce interests of the users to our method. A user may be interested in a certain topic but not be familiar with another. Corresponding to this, the activity of bookmark may be different. We think that the difference may appear in the difference of tag usage.

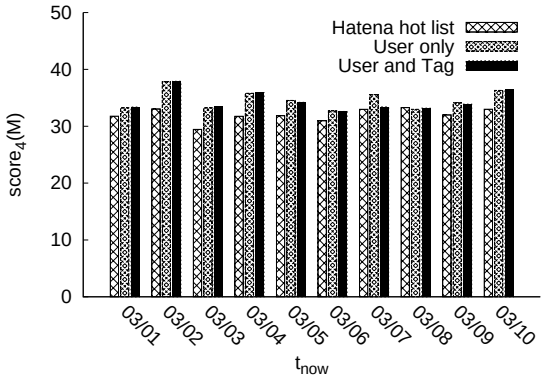

Figure 7: Experiment on $score_4(M)$.

6. REFERENCES

[1] del.icio.us. http://del.icio.us/.

[2] Flickr. http://www.flickr.com/.

[3] Hatena Bookmark. http://b.hatena.ne.jp/.

[4] YouTube. http://www.youtube.com/.

[5] M. Biddulph. Introducing del.icio.us. XML.com, Nov 2004. http://www.xml.com/pub/a/2004/11/10/delicious.html.

[6] S. A. Golder and B. A. Huberman. The Structure of Collaborative Tagging Systems. Information Dynamics Lab, HP Labs, Aug 2005.

[7] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information Retrieval in Folksonomies: Search and Ranking. In *The Semantic Web: Research and Applications; 3rd European Semantic Web Conference, Eswc 2006 Budva, Montenegro, June 11-14, 2006, Proceedings*, pages 411–426, 2006.

[8] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Trend Detection in Folksonomies. In *Semantic Multimedia: First International Conference on Semantic and Digital Media Technologies, SAMT 2006 Athens, Greece, December 6-8, 2006 Proceedings*, pages 56–70, 2006.

[9] S. Niwa, T. Doi, and S. Honiden. Web Page Recommender System based on Folksonomy Mining for ITNG'06 Submissions. In *Third International Conference on Information Technology: New Generations (ITNG'06)*, pages 388–393, 2006.

[10] L. Page and S. Brin. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, Apr 1998.

[11] A. Sasaki, T. Miyata, Y. Inazumi, A. Kobayashi, and Y. Sakai. Web Content Recommendation System based on Similarities among Contents Cluster of Social Bookmark. In *DBWeb 2006*, pages 59–66, 2006.

[12] K. Shiratsuchi, S. Yoshii, and M. Furukawa. Finding Unknown Interests Utilizing the Wisdom of Crowds in a Social Bookmark Service. In *2006 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 421–424, 2006.

[13] Y. Xu, L. Zhang, and W. Li. Cubic Analysis of Social Bookmarking for Personalized Recommendation. In *Frontiers of WWW Research and Development –*

APWeb 2006: 8th Asia-Pacific Web Conference, Harbin, China, January 16-18, 2006, Proceedings, pages 733–738, 2006.

[14] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can Social Bookmarking Enhance Search in the Web? In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2007)*, pages 107–116, 2007.

[15] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Towards Improving Web Search by Utilizing Social Bookmarks. In *Proceedings of the 7th International Conference on Web Engineering (ICWE 2007)*, 2007.