Prediction of Social Bookmarking based on a Behavior Transition Model

Tadanobu Furukawa,
Seishi Okamoto
Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka,
Nakahara-ku, Kawasaki-shi,
Kanagawa 211-8588, Japan
{tfuru,
seishi}@labs.fujitsu.com

Yutaka Matsuo
Faculty of Engineering, The
University of Tokyo
2-11-16 Yayoi, Bunkyo-ku,
Tokyo 135-8656, Japan
matsuo@biz-model.t.u-tokyo.ac.jp

Mitsuru Ishizuka
Graduate School of
Information Science and
Technology, The University of
Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 135-8656, Japan
ishizuka@i.u-tokyo.ac.jp

ABSTRACT

We propose an algorithm to predict users' future bookmarking using social bookmarking data. It is a problem that primitive collaborative filtering cannot exactly catch users' preferences in social bookmarkings containing enormous items (URLs) because in many cases user's adoption data is sparse. There can be various influences on bookmarking such as effects from the environment and changes in user preference. We use temporal sequence among the bookmarking-users to represent word-of-mouth and among the bookmarked-URLs to represent user's interest, and model each sequential order as a continuous-time Markov chain. This idea comes from diffusion of innovation theory. A transition probability from a state (user/URL) to another state is defined by the transition rate calculated from the time taken for the transition. We predicted user's preferences through a combination of estimating the most likely transition between users using URLs as input and between URLs using users as input. We conducted evaluation experiments with a social bookmarking service in Japan called Hatena bookmark. The proposed algorithm predicts users' preferences with higher accuracy than collaborative filtering or simple transition models based on either user or URL.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Algorithm, Experimantation

Keywords

social tagging, information flow, collaborative filtering, recommender system, Markov chain

SAC'10 March 22-26, 2010, Sierre, Switzerland.
Copyright 2010 ACM 978-1-60558-638-0/10/03 ...$10.00.

1. INTRODUCTION

As it becomes easier for people to post information on the web through blogs, online bulletin boards, and content-sharing websites, the amount of available content continually increases. In addition, there are various online services such as the online store Amazon.com [1] and the movie sharing-site YouTube [2], and recently new services for cloud computing are appearing. The web is approaching a state of infoglut. On one hand, the increase of content that is useful or attractive to users is good; on the other hand, much of the content is useless, including spam blogs (splog), and gets in the way of accessing the useful content.

Recommender systems are one type of solution to infoglut; they extract useful content for users and suggest them in a form adjusted to the interests of the users. They have been used to recommend many types of content: not only web pages, but also products and movies. Recommender systems are classified into two main types: content-based filtering and collaborative filtering [1]. The former recommends items which have similar properties to a user's tastes. The latter calculates a prediction for the active user (the user who receives the prediction) from ratings by like-minded users (users who share the same rating patterns as the active user). The user's tastes must be learned to structure a recommender system.

Recently social bookmarking sites such as del.icio.us [3] and Digg [4] are receiving attention as systems to organize web pages. Social bookmarking is different from bookmarking in browsers; users publish their bookmarked pages and can share them with the general public. Most social bookmark services encourage users to organize their bookmarks with informal tags. Thus web pages are categorized based on tags or on the users who bookmarked them. A user can find interesting web pages by referring to the bookmarks of users who have interests similar to the user, or by referring to web pages with tags that are similar to the tags attached to web pages the user has bookmarked.

Bookmarking is a behavior to save web pages which a user wants to reread later, and thus it reflects the user's tastes. In this paper, we analyze a user's tastes by social book-

[1] http : //www.amazon.com/
[2] http : //www.youtube.com/
[3] http : //del.icio.us/
[4] http : //digg.com/

marking data and propose a method to predict web pages which the user will save on the social bookmarking site. If our predictions are accurate, we will be able to recommend interesting web pages to users.

Because new items arrive on the web frequently, it is important to recommend just the new and worthwhile items out of the large number available. Many research studies have proposed time-line based methods and have achieved some success. For example, a method may be based on transitions of a user's preference, or on the order of adopters [20, 23]. However a user's bookmarking behavior can be influenced by numerous factors such as word-of-mouth from others and characteristics of the items. A single transition model cannot consider all these factors.

Thus we propose a method to predict a user's preferences through a combination of estimating the most likely transitions between users using URLs as input and between URLs using users as input. We conducted experiments to predict user's bookmarking behavior with a dataset from a Japanese social bookmarking service called Hatena bookmark. Our method provides better accuracy than existing methods with one transition model or collaborative filtering.

This paper is organized as follows: the next section explains the characteristics of social bookmarks and the definition of transition on social bookmarking. In Section 3, we describe a model to predict users' preferences with the transition model. We conduct evaluation experiments and show the effect of our model in Section 4. After a describing related studies about recommender systems in Section 5, we conclude the paper.

2. RELATED WORK

Recommender systems, predicting which items a user will like and suggesting them, have been studied since the early 1990s [7, 21]. Many recommendation methods have been proposed based on different kinds of information or scenarios [1]. For example, Fab [3] builds a preference profile for a user based on contents which he has adopted, and recommends web pages with a collaborative filtering method based on the similarity of his preferences to other persons. A study to predict which blogs a user reads willingly was performed by examining comments and the trackback network structure [2]. In addition, various online services such as Amazon.com or YouTube have their own techniques for recommending books or movies from the user's browsing history on their sites [15].

Recommender systems are classified into two main types: collaborative filtering and content-base filtering. In particular, collaborating filtering can use various kinds of data, so numerous collaborative filtering algorithms have been designed to identify users' preferences [19]. Some research has studied how to measure the similarity between users [13]. Other studies have used complex machine-learning methods [6]. Collaborative filtering suffers from the cold-start problem: new users have to build up profiles before the filtering is effective. To solve this problem, [4, 18] proposed using bots or hybrid systems with content-based filtering.

However it is difficult to extract a user's tastes from web page recommendations. It is necessary to use the original data of the services or to browse web pages through a proxy server to obtain the user's preferences from these recommendation methods. In contrast, social bookmarking services allow users to save their favorite web pages in their bookmarks according to their tastes. Thus these services have attracted attention in recent years as a way to learn the interests of the user or to follow trends in the web. Golder performed statistical analysis of Folksonomy including the social bookmark services [8]. Hotho et al. extracted the tastes of users abstractly using tag information and web page ranks [10]. Markines et al. proposed a web page recommendation method using social bookmarking data [16]. They scored web pages by similarity between users, popularity, and novelty. Noll et al. proposed a technique to personalize search which used social bookmarking data [17]. Studies to improve the precision of a search or a recommendation by resolving ambiguities in tags freely attached by users have also been done [5, 12, 14].

In this paper, we propose a method to predict a user's preferences through the combination of estimating the most likely transitions between users using URLs as input and between URLs using users as input. In recent years, the number of services including chronological order information, such as blogs, folksonomies, and online stores, has increased. Consequently recommendation techniques based on changes in the time of adoption by users have been proposed. Pavlov et al. predicted a user's tastes by applying a maximum entropy model to chronological data [20]. Shani et al. and Iwata et al. made a prediction based on a Markov model [23, 11, 9]. Song et al.[24] also captured the sequential order among users who adopted an item in a Markov process and calculated the transition rate based on diffusion of innovation theory [22]. We model the process of a user deciding to save a web page by applying Markov process analysis techniques to both transitions between items and between users in the social bookmark data, and predict the user's preferences.

3. TWO TRANSITION PROCESSES IN SOCIAL BOOKMARKING

In this paper, we propose an algorithm based on the two transition processes between users and between items in social bookmarking. This section explains these transition processes.

3.1 Transition process between users

We assume that there are innovators who bookmark pages earlier than others and followers who bookmark the same pages later than innovators in social bookmarking. We define this sequence, user c_j saves the page after user c_i, as a transition process between users $c_i \rightarrow c_j$.

We observed this transition process in our inspection of the data of a Japanese social bookmarking service called Hatena bookmark [5]. We define the user who first saves each page as an innovator. The number of innovators was 11,058 for all 484,741 web pages. Figure 1 shows the changing rate of increase in cumulative number of URLs (eq. 1) caused by including more innovators, in descending order of $|S_{\dot{c}}|$. $|S_{\dot{c}}|$ denotes the number of web pages where the innovator is $\dot{c}$. In the summation of eq. 1, $\dot{c}$ denotes an innovator and $\dot{C}$ denotes the set of innovators.

$$\frac{1}{484741}\sum_{\dot{c}\in\dot{C}}|S_{\dot{c}}| \qquad (1)$$

[5] http : //b.hatena.ne.jp/

About 80% of all pages are first saved by 1,888 innovators who comprise only about 12% of all users. This means there can be some transition patterns of users.

In addition, there is a tendency for the innovator to be different for each topic. Table 1 shows the relation between innovator and the pages which the innovator saved earlier than anyone else. User #579 tends to be an innovator for pages about software technology; user #2558 tends to be an innovator for pages about IT related news. Thus the transitions between users can be different for different web page topics.

3.2 Transition process between web pages

We assume some kind of a sequence between bookmarked web pages, and we treat that sequence as a transition process between web pages. We define the transition process between pages $s_i \rightarrow s_j$ as a tendency for page s_j to be bookmarked after page s_i by many users. We calculate the average time of bookmarking for page s as follows, where $|s|$ is the number of bookmarks for s and $t_{c_i \cdot s}$ is the time of bookmarking of s by user c_i. Fig. 2 shows the time-lag of all bookmarkings of every page from $\bar{t}_s$.

$$\bar{t}_s = \frac{1}{|s|} \sum t_{c_i \cdot s}$$

Note that we use only 268,870 bookmarking data consisting of 7,727 users who saved over 30 pages and 7,918 pages saved by over 30 users.

About half of all bookmarkings (133,080) were saved within 100 hours before or after the average bookmark time, and 207,756 bookmarkings accounting for about 80% of the total were saved within 600 hours before or after the average bookmark time. The probability for a page to be bookmarked suddenly falls after the average bookmark time, so that we can see that many pages are not bookmarked in the long term. Therefore we can expect that there are sequences between the pages; many users begin to save a new and popular page when it appears.

We define the URLs which each user saved first as initial pages. Figure 3 shows the changing rate of increase in the cumulative number of initial pages (eq. 2) caused by increasing the number of innovators in descending order of $N_{\dot{s}}$. ($\dot{s}$ denotes an initial page, $\dot{S}$ denotes the set of initial pages, $|C_{\dot{s}}|$ denotes the number of users whose initial page is $\dot{s}$.)

$$\frac{1}{268870} \sum_{\dot{s} \in \dot{S}} |C_{\dot{s}}| \qquad (2)$$

3,876 users accounting for about half of all users have saved at first 254 pages, which is only about 3% of all 7,918 pages. This means there can also be some transition patterns of bookmarked web pages.

4. BEHAVIOR TRANSITION MODEL

Our prediction method is based on these two transition processes. We predict the probable URLs that the users will bookmark by modeling these processes. For the modeling of the transition process, we use the information diffusion theory based on the diffusion rate suggested in [24]. In this section, we explain the technique of [24] and subsequently propose our method.

Table 1: Titles of web pages which the innovator saved before anyone else.

Innovator ID	Title of web page
579	(software technology) • [HOWTO] To use ADO on Visual Basic or VBA in Excel data • VB related technology - DirectX, ADO.NET, Excel, SAPI • Accessing to local data and remote data on ClickOnce applications
2558	(IT related news) • Internet shopping 2006 (digest version) • Survey of Video Research Ltd. - 40% of uses of Wikipedia is searching a personal name • Survey in U.S. - Advertising market in the Web continues to grow. Banner ads are on a decline.
3666	(blog posts of content arranged from writings in an internet forum) • Thread to collect funny jokes • Why is the boom of creating Flash movies by amateurs over? • Thread to collect pictures of magnificent scenery

4.1 Information flow modeling based on diffusion rate

4.1.1 Model of transition process based on CTMC

The method of [24] (rate-based information flow model, RIF) captures a transition process between users where a different user adopts (the bookmark in this paper) an item (the web page in this paper) with a continuous time Markov chain (CTMC). The transition probability between user i and j means how likely it is that j will adopt an item after i adopts the item. This is based on diffusion of innovation theory [22].

By the diffusion of innovation theory, adopters are classified into 5 categories based on the time that he/she adopts it in the process of an innovation occurring and diffusing in a community. They are (1) innovators, (2) early adopters, (3) early majority, (4) late majority, and (5) laggards. At first, innovators, the novelty hunters, adopt an innovation and then early adopters with high social status in the community judge its value and follow. Then the innovation diffuses successively into the early majority, late majority, and conservative laggards. In terms of CTMC models, innovators and early adopters have a high transition probability to others, and late majority and laggards have a lower transition probability to others (Fig. 4).

In a CTMC, $X(t+\delta)$, the state at time $t+\delta$, does not depend on history $x(h)$ $(0 \leq h \leq t)$ but only on the state at time t.

$$\begin{aligned} P\{X(t+\delta) = j | X(t) = i, X(h) = x(h), 0 \leq h \leq t\} \\ = P\{X(t+\delta) = j | X(t) = i\} \end{aligned}$$

Therefore the transition probability matrix $\mathbf{Q}$ is calculated as follows where $P_{ij}(\delta)$ denotes the transition probability

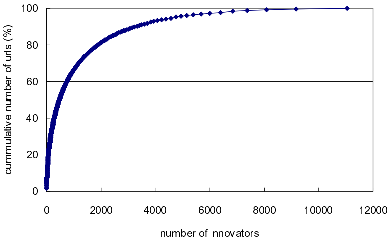

Figure 1: Cumulative number of users who saved initial URLs.

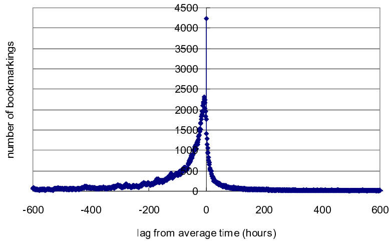

Figure 2: Time-lag from average time of bookmarking.

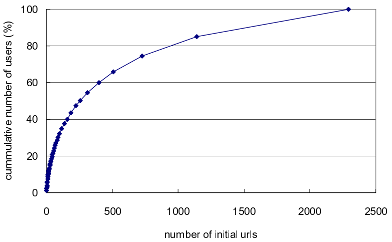

Figure 3: Cumulative number of URLs saved by innovators.

from state i to j in time δ [6].

$$\mathbf{Q} = \begin{pmatrix} q_{00} & q_{01} & \cdots \\ q_{10} & q_{11} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

$$\begin{aligned} q_{ij} &= \lim_{\Delta t \to 0} \frac{P\{X_{t+\Delta t} = j | X_t = i\}}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t)}{\Delta t} \ (i \neq j) \end{aligned}$$

When the chain leaves state i with rate q_i, it must enter some other states. The rate q_i is called the out-state rate.

$$q_{i,i} = -q_i = -\sum_{j \neq i} q_{i,j} \quad (3)$$

The relation between transition rate q and transition probability P is

$$P_{ij} = \frac{q_{i,j}}{q_i} \ (i \neq j), \quad 0 (i = j) \quad (4)$$

We calculate a transition probability from data to model the transition process with a CTMC, and then calculate the transition rate matrix. We calculate the rate of not transiting to any other state using the time T_i that the item remains in the state (user) i (meaning nobody adopts the item after i).

$$\frac{1}{q_i} = T_i$$

We calculate a transition probability from state i to j as follows.

$$P_{ij} = q_i \exp(-q_i t_{ij})$$

t_{ij} means time taken to transit from i to j; this transition is approximated by a Poisson process. We can model the transition process between users with transition rate matrix $\mathbf{Q}$.

4.1.2 Prediction with transition model

The prediction of adoption is based on the score of a utility function [1]. The utility function measures the usefulness (the predictive rating, or whether to adopt or not in the future) of an item to a user. RIF integrates the transition probability for a calculation of the utility based on the transition model. At first the transition probability $P(t)$ when only t passes is calculated as follows.

$$P(t) = e^{t\mathbf{Q}}$$

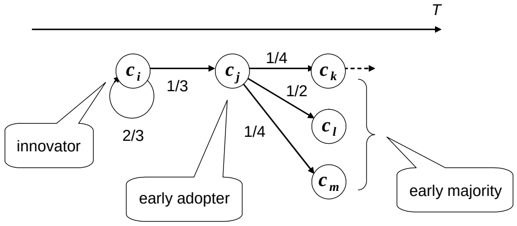

Figure 4: Transitions between users.

Table 2: Prediction with RIF.

```
algorithm Prediction with RIF
input
  $\mathcal{S} = (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n)$
       ($\mathcal{A}_i$: user $c_i$'s adoption data with timestamp)
output
  $L(\tau)$: utility score
begin
1) Estimate the out-state rate by Eq.3
2) Estimate transition probability and $q_{ij}$ by Eq.4
3) Generate transition rate matrix $\mathbf{Q}$
4) Calculate the utility by Eq.5
end
```

The state transition that can occur before time τ passing from an initial state is calculated by integrating the transition probability. RIF uses this value of this integral (following $L(t)$) for the utility and recommends the items which have high utility and have not been adopted yet. $\mathbf{I}$ is the identity matrix.

$$P'(t) = P(t)\mathbf{Q}, \ P(0) = \mathbf{I}$$

$$L(\tau) = \int_0^\tau P(t)dt \quad (5)$$

4.2 Two extensions

Because RIF classifies the users who easily lead a transition and those who are unlikely to lead a transition by considering the transition rate for every user, we think it is effective to apply RIF to social bookmarking data in which there are the relations of the innovators and followers. We propose to extend the prediction model by adding two features.

- Prediction by transitions between web pages.
- Clustering of users and web pages.

[6] We assume the chain is time-homogeneous and the transition probability does not depend on the initial state.

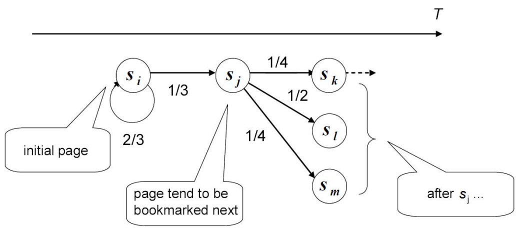

Figure 5: Transition between web pages.

4.2.1 Prediction based on transition between web pages

The first extension is to use the transition process between web pages which we expect to exist. Transitions between users are straightforward as can be seen in Fig. 3, but there are surely some initial pages that are frequently saved at first. Therefore, we apply RIF to transitions between pages. We can do a prediction based on the probability of sequential adoption: users tend to adopt one item after adopting another item (Fig. 5).

4.2.2 Clustering of users and web pages

The second extension is to cluster users and web pages. The purpose of this clustering is to use the supposed tendency of transitions between some items to be common among some users.

This is based on a tendency for the topics in which a user is likely to become the innovator to be different for every user. In addition, in [24] they use the data of MovieLens [7] (256 ratings par movie and 166 ratings per user on average) and the log data of a knowledgebase system to support sales staff with registered documents. Compared to our social bookmarking data (6 bookmarkings per page and 205 bookmarks per user on average), those data are abundant. When we apply the model to sparse data, there is a possibility that we cannot model transitions adequately.

Therefore we build a transition model after clustering users and items as follows.

- User clustering: We build transition models between pages for every user cluster. Each transition model includes pages which users in the cluster have bookmarked as states.

- Web page clustering: We build transition models between users for every web page cluster. Each transition model includes users who bookmarked the web pages in the cluster as states.

We cluster pages based on the similarity among users who bookmark the page, and cluster users based on the similarity among pages which the user bookmarked.

4.3 Combined prediction model of transitions between users and web pages

We show the algorithm of our method below. We estimate the utility by combining estimates from each transition model between users and pages through applying RIF.

1. Cluster users to build a transition model between pages, and cluster pages to build a transition model between users.

Table 3: Prediction with proposed model.

```
algorithm Prediction with proposed model
input
  $\mathcal{S} = (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n)$
      ($\mathcal{A}_i$: user $c_i$'s adoption data with timestamp)
output
  $\forall c, \forall s, u(c, s)$: utility score
begin
1) Cluster users and web pages.
2) Estimate utility scores for each clusters in both transition models by Eq.5
3) Estimate the combined utility score by Eq.6
end
```

2. Calculate the transition probability and transition rate matrix in every cluster for both transition models.

3. Estimate $u_c(c, s)$, the utility of item s to user c from the transition rate matrix of the transition between users model. Estimate $u_s(c, s)$, the utility of item s to user c from the transition rate matrix of the transition between pages model.

4. Assume that the transitions between users and the transitions between pages are independent events, and estimate $u(c, s)$, the utility of item s to user c, by the product of the utilities from both transition models.

$$u(c, s) = (1 + u_c(c, s)) \cdot (1 + u_s(c, s)) \tag{6}$$

5. EVALUATION EXPERIMENT

5.1 Experiment set-up

We conducted evaluation experiments with the data of Hatena bookmark, one of the social bookmark services in Japan. The period of the data is from March, 2005, to October, 2006. We used 268,870 bookmarking events including 7,727 users who saved more than 30 pages and 7,918 pages saved by over 30 users to model transitions. This was because we cannot build accurate transition models for pages with few bookmarking users or users with few bookmarked pages.

To demonstrate the performance of our recommendation algorithm, we divided the dataset into a training set and a test set. We used the training set to build a transition model and the test set to demonstrate the prediction performance. The period of the training set was three months and that of the test set was the one next month. This is based on the consideration that we can model a transition process between users for many pages from the data in three months, because 80% of the bookmarking of a certain web page is done in 1,200 hours (= 50 days). We set the test periods as Oct. 2005, Jan. 2006, Feb. 2006, Jul. 2006, and Oct. 2006, and the training periods as the previous three months for each.

We evaluated the accuracy of prediction by precision. Precision is calculated by eq.7 where U denotes the set of users in the training period, $\hat{s}_c$ denotes the top 10 pages by utility score for each user $c \in C$, and s_c denotes the pages which c bookmarks in the test period.

$$\frac{\sum_{c \in C} |\hat{s}_c \cap s_c|}{|C| \times 10} \tag{7}$$

[7] http : //www.movielens.org/

5.2 Effect of clustering

First we examined the effects of clustering by users and pages in RIF. Figure 6 shows the result of page clustering for RIF on transitions between users (RIFu), and Figure 7 shows the results of user clustering for RIF on transitions between pages (RIFr). In these figures, the x-axis denotes the test period; for instance, we constructed a transition model between users and URLs using the data from August 2005 to September 2005 to predict the user's bookmarking in October 2005, when the corresponding value on the x-axis was 0510. We clustered both users and pages with a group average method based on the Jaccard similarity coefficient [8].

For the prediction by all transitions, precision was highest when the number of the clusters was 3 or 4 and improvement of the precision by clustering was demonstrated. We could find clusters of pages including news articles, pages including engineering topics such as software technology, pages of blog posts, and so on. They are similar to Table 1. User clusters are also the sets of users who bookmark many pages with similar topics such as software technology. Because the pattern of transition between users/pages differs among genres, clustering improves the precision.

On the other hand, RIF with large numbers of clusters of users/pages reduces the accuracy. In the case of page clustering, when the number of clusters became high, some clusters included few pages and thus we could not build appropriate transition models to estimate utilities. The pages to be included for either training period were around 500-1,000 [9]. Some clusters included few users when the number of user clusters was high. Small clusters with less than 10 members appeared when the number of clusters was over 5, so it seems to be hard to build appropriate transition models.

5.3 Effect of combination of transitions

We compared our proposed method with the other existing methods, collaborative filtering (CF) and RIF (either transitions between users or transitions between pages) without clustering.

In CF, we measured the similarity $sim(s_i, s_j)$ between two pages s_i, s_j by the Jaccard coefficient for the bookmarking users and calculate the utility as

$$u_{cf}(c,s) = \frac{1}{|C'|} \sum_{c' \in C'} sim(c,c') \cdot u_{cf}(c',s)$$

where C' denotes the top 50 similar-preference (the Jaccard coefficient is high) users. We extracted the top 10 pages by utility which had not been saved yet for every user and calculated the precision (eq.7). We show the precision of our method and comparison with other techniques in Fig. 8. In comparison with CF, all the methods using RIF give higher precision. CF cannot discriminate between newer and older pages, so CF estimated high utility for pages which were bookmarked early in the training period. Although RIFr does not have a clear sequential order like the innovators in RIFu, Rifr and RIFu both performed with higher precision than CF.

In comparison of RIFu/RIFr with the combination of RIFu and RIFr without clustering, the combined method performed with almost the same precision as RIFu. Both methods estimated high utility for similar pages which appeared late in the training period. Because the effect of RIFr is almost covered by RIFu, the precision of RIFr does not exceed that of RIFu.

When we clustered users/pages (we set the number of clusters as 4 for pages and users because this gave the highest precision in the result of the foregoing paragraph) and combined both transition models, the precision of prediction was higher than the precision of the method without clustering for all periods. This was because the transition model between pages was built more properly by clustering. Although there is a strong tendency for the pages bookmarked by innovators to diffuse (be adopted by other users) immediately based on the transition between users, there are some pages which are bookmarked based not on innovators but on the user's tastes. RIFu tends to predict recent and popular pages with high utility, but the combined method can sometimes predict such pages as well. So the combined method produces the best result.

6. CONCLUSIONS

In this paper, we focused on the temporal sequence of users saving pages and web pages saved by users in a social bookmarking service and proposed a method to predict the pages that a user will save in the future. Our method is inspired by an existing work, which models the sequential order with a continuous time Markov model and estimates the probability that a user will adopt an item by the transition probability. Considering that there are various influences on a user's bookmarking behavior, our method combined two transition models: transitions between users and between web pages. As a result, our method improved the accuracy of bookmarking prediction in comparison with existing methods.

Our utility function (eq. 6) was a simple multiplication without normalization, but it could show the availability of a combination of two transitions. Other factors such as trends and geographical properties may affect the transition process. In future work, we need to try to understand the transition process in more detail and improve the utility function. Then, we will build a more accurate model of the transition of adoption, and the performance will be better.

Our method can be applied to various behavioral datasets. Thus future studies will include a comparative study of different social bookmarking services and other adoption data such as a user log in an online store to investigate the generality of our findings.

7. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6):734–749, 2005.

[2] J. Arguello, J. Elsas, J. Callan, and J. Carbonell. Document representation and query expansion models for blog recommendation. In ICWSM 2008, 2008.

[8] Where s_{c_i}, s_{c_j} denotes the set of pages which user c_i, c_j bookmarks, the Jaccard similarity coefficient between c_i, c_j is $J(c_i, c_j)\frac{|s_{c_i} \cap s_{c_j}|}{|s_{c_i} \cup s_{c_j}|}$.

[9] We eliminated the pages which were not bookmarked by more than 10 users to prevent isolated clusters.

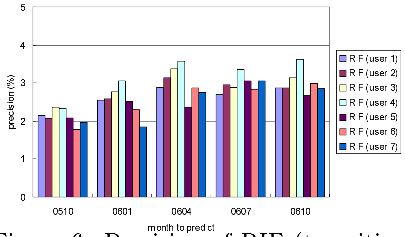

Figure 6: Precision of RIF (transition between users) with clustering. Numbers in parenthesis denote the number of web page clusters; x-axis is the test period.

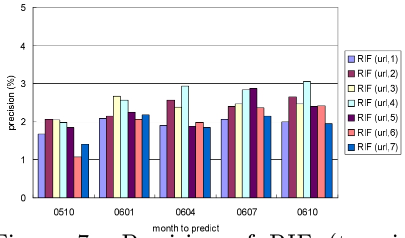

Figure 7: Precision of RIF (transition between web pages) with clustering. Numbers in parentheses denote the number of user clusters.

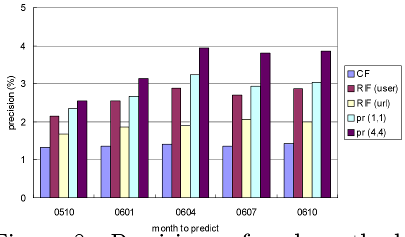

Figure 8: Precisions of each method. CF is collaborative filtering, pr is our combined method. Numbers in parenthesis denote the number of web page clusters and the number of user clusters.

[3] M. Balabanović and Y. Shoham. Fab: Content-based, collaborative recommendation. Communications of the ACM, 40(3):66–72, 1997.

[4] J. Basilico and T. Hofmann. Unifying collaborative and content-based filtering. In Proc. ICML '04, 2004.

[5] G. Begelman, G. Keller., and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In Proc. Collaborative Web Tagging Workshop, WWW 2006, 2006.

[6] D. DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In Proc. ICML '06, 2006.

[7] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12):61–72, 1992.

[8] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2):198–208, 2006.

[9] M. Gori and A. Pucci. Itemrank: A random-walk based scoring algorithm for recommender engines. In Proc. IJCAI-07, 2007.

[10] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Proc. ESWC 2006, 2006.

[11] T. Iwata, K. Saito, and T. Yamada. Modeling user behavior in recommender systems based on maximum entropy. In Proc. WWW 2007, 2007.

[12] R. Jaschke, L. B. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In Proc. PKDD 2007, 2007.

[13] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Proc. 14th ACM SIGKDD, 2008.

[14] R. Li, S. Bao, B. Fei, Z. Su, and Y. Yu. Towards effective browsing of large scale social annotations. In Proc. WWW 2007, 2007.

[15] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1):76–80, 2003.

[16] B. Markines, L. Stoilova, and F. Menczer. Bookmark hierarchies and collaborative recommendation. In Proc. AAAI-06, 2006.

[17] M. G. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In ISWC 2007, 2007.

[18] S.-T. Park and D. M. Pennock. Naive filterbots for robust cold-start recommendations. In Proc. 12th ACM SIGKDD, 2006.

[19] S.-T. Park and D. M. Pennock. Applying collaborative filtering techniques to movie search for better ranking and browsing. In Proc. 13th ACM SIGKDD, 2007.

[20] D. Y. Pavlov and D. M. Pennock. A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains. In Proc. of Neural Information Processing Systems, 2002.

[21] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In Proc. of The Conf. on Computer Supported Cooperative Work, 1994.

[22] E. M. Rogers. Diffusion of Innovations. The Free Press, 1995.

[23] G. Shani, D. Heckerman, and R. I. Brafman. An mdp-based recommender system. Journal of Machine Learning Research, 6:1265–1295, 2005.

[24] X. Song, Y. Chi, K. Hino, and B. L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In Proc. WWW 2007, 2007.