Capturing Implicit User Influence in Online Social Sharing

Ching-man Au Yeung
NTT Communication Science Laboratories
2-4 Hikaridai Seika-cho Soraku-gun, Kyoto
619-0237, Japan
auyeung@cslab.kecl.ntt.co.jp

Tomoharu Iwata
NTT Communication Science Laboratories
2-4 Hikaridai Seika-cho Soraku-gun, Kyoto
619-0237, Japan
iwata@cslab.kecl.ntt.co.jp

ABSTRACT

Online social sharing sites are becoming very popular nowadays among Web users, who use these sites to share their favourite items and to discover interesting and useful items from other users. While an explicit social network is not necessarily present in these sites, it is still possible that users influence one another in the process of item adoption through various implicit mechanisms. In this paper, we study how we can capture the implicit influences among the users in a social sharing site. We propose a probabilistic model for user adoption behaviour, where we assume that when user adopts an item, he would pick a user and choose from the set of items that this user has adopted. By using the model, we estimate the probability that one user influences another user in the course of item adoption, based on the temporal adoption pattern of the users. We carry out empirical studies of the model on Delicious, a popular social bookmarking site. Experiments show that our model can be used to predict item adoption more accurately than using collaborative filtering techniques. We find that the strength of implicit influence various across different topics. We also show that our method is able to identify the influential users who are more likely to possess items interested by other users. Our model can be used to study the dynamics in a social sharing site and to complement collaborative filtering in recommendation systems.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.5.3 [**Group and Organization Interfaces**]: Web-based interaction

General Terms

Algorithms, Human Factors, Experimentation

Keywords

social sharing, social influence, collaborative tagging

1. INTRODUCTION

Online social sharing sites such as Delicious[1], CiteULike[2] and LibraryThing[3] are becoming very popular nowadays among Web users. Using these Web sites, users can share their favourite items, such as bookmarks, music, books and academic references, with other users. They can also make use of these sites to retrieve and discover new and interesting items introduced by other users. Users can usually assign tags to describe the shared items to facilitate search and retrieval. The user interactions in these sites are also referred to as a form of ballot box communication [28], because users do not really interact with each other by exchange messages, but only implicit interacting with one another by casting their votes or expressing their preferences on the shared items.

The rising popularity of social sharing online has caught the attention of Web information retrieval and data mining researchers, as these sites represent a new source of data that involve user-generated metadata, information about user interests, and collective item adoption patterns. There are studies that focus on leveraging user data in social sharing sites to enhance Web information retrieval [4, 10, 30]. In addition, it has been suggested that social sharing sites are popular particularly because they allow users to discover something that interest them, in addition to retrieving items they need. In other words, exploratory search is what makes social sharing sites attractive to many users [3, 11].

Exploratory search can be enhanced by actively recommending items to users based on their interests expressed in a social sharing site, and many studies have proposed various collaborative filtering methods for recommending potentially interesting items to users in different systems [13, 21, 29]. However, it has also been found that, in addition to the symmetric similarity among users commonly used collaborative filtering, the asymmetric inter-personal influence also plays an important role in understanding user adoption patterns and recommending items to users [25]. In particular, in systems in which explicit social ties exist among the users, degree of influences between users is a good predictor of future item adoption [8, 24].

However, in many online social sharing sites users may

[1]Delicious: `http://www.delicious.com/`
[2]CiteULike: `http://www.citeulike.org/`
[3]LibraryThing: `http://www.librarything.com/`

not necessarily declare explicit social ties among themselves even if such functionality exists. Nevertheless, since items adopted by the users are usually publicly accessible to all other users, there are many different ways by which users may influence one another as time passes. Understanding the implicit influences among users is thus important to the understanding of the dynamics of interactions in a social sharing system. In addition, by capturing the implicit influences among users, better recommendation algorithms can be developed to take into account not only similarity between users but also the influences among them. So far, however, few studies have focused on how implicit influences can be captured in social sharing sites where explicit social ties do not necessarily exist.

In this paper, using the popular social bookmarking site Delicious as our subject, we study how implicit influences among users can be captured by using a probabilistic model of the behaviour of the users. We try to estimate the probability of influence among users by using their temporal adoption patterns. In particular, we assume that while users would adopt some items because they are recently popular in the community, they are also likely to choose to follow some other users and adopt something that they have already adopted. From our experiments on a large dataset collected from Delicious, we find that our model can be used to predict item adoption more accurately than using collaborative filtering techniques. We also find that the strength of implicit influence various across different topics. In addition, we show that our method is able to identify the influential users who are more likely to possess items interested by other users. The results suggest that our model can be used to study the dynamics in a social sharing site and to complement collaborative filtering in recommendation systems.

The remaining of the paper is structured as follows. In the next section, we briefly introduce online social sharing and discuss the possible causes of implicit influence among users. Next, we describe in detail the proposed probabilistic model which we use to capture implicit influence in social sharing. In Section 4 we describe our experiments and analyse the results obtained. Section 5 presents some related work. Finally, Section 6 gives conclusions and mentions possible future research directions.

2. ONLINE SOCIAL SHARING

Online social sharing has attracted much attention in recent years as the Web has been used more and more as a platform for social interactions. In some systems, users contribute materials created by themselves so that they are accessible to other users. Examples are Youtube[4] for video sharing and Flickr[5] for photo sharing. In some other systems, users post something they consider useful or interesting to the system and share with other users. Examples of this kind include Delicious for Web surfers and LibraryThing for book lovers. In many of these social sharing systems, users are able to use some descriptive keywords, also commonly known as tags, to describe the items they have submitted in order to facilitate organisation and retrieval of the items. In some other systems, users can rate the items and share their experience or comments with other users. For example, in MovieLens[6] users can assign ratings to movies and receive recommendation based on collaborative filtering techniques; in Epinions[7], users write reviews of different products and other users can even express whether they think the reviews are useful or not.

In this paper, we focus on a generalisation of online social sharing systems, and consider such a system as one in which users can express their preferences by creating a collection of shared items through adoption over time. More formally, we define a social sharing system $\mathcal{S}$ as follows:

Definition 1. A social sharing system $\mathcal{S}$ is a tuple $\mathcal{S} = \langle U, D, Y, A\rangle$. U is a set of users, D is a set of shared items among the users, and Y is a set of posts. A post $(u, d, t, a) \in Y$ represents the fact that $u \in U$ posts/adopts $d \in D$ at time t, along with certain system-dependent attribute $a \in A$.

This general definition captures user activities in a wide range of existing systems. For example, in Delicious, D is a set of URLs accessible on the Web, and A is a set of tags created by the users. In MovieLens, D is a set of movie titles, and A is a set of possible ratings that users can use to rate the movies.

While users are free to introduce any new items to a social sharing system, users are also likely to discover something interesting among those that have been adopted by other users in the system [3, 11]. Users can either browse the popular items presented by the system, or they can browse the collection of a particular user. In other words, it is very likely that users adopt certain items because they are influenced by other users. The way that one user influences another user can be either explicit or implicit. In some systems, users can subscribe to the collection of other users and be informed of their newly posted items, such that those who 'follow' are more likely to be influenced by those who are 'being followed'. In some other systems, users can form groups of specific interests, and influences occur among users who belong to the same groups.

However, it is not always the case that we can find explicit social ties in a social sharing system. Users may just want to use the system to share or discover interesting items without the intention to create social ties with other users. Even when such functionality is available, influences among users do not have to be confined within the explicit social network. Since in general the activities of all users are publicly accessible, it is possible that in the course of sharing various items over time, a user may be more influenced by one user than by another. There are many different situations in which a user u_a can be influenced by another user u_b. Here are some of the possibilities:

- User u_a bookmarked user u_b's profile page and/or item page, and u_a often visits these pages to look for new items;
- User u_b likes to recommend items to other users, and user u_a tends to adopt items recommended by u_b;
- User u_a has subscribed to a particular RSS feed (e.g. the feed of a tag), and u_b happens to be an active user with respect to that feed;

[4]Youtube: `http://www.youtube.com/`
[5]Flickr: `http://www.flickr.com/`
[6]MovieLens: `http://www.movielens.org/`
[7]Epinions: `http://www.epinions.com/`

- User u_b is an early adopter of items of a particular topic, and u_a is interested in the topic but is a relatively latecomer.

While some of these activities can be known from a system activity log, others cannot be traced. Hence, it is desirable to develop a general framework for capturing implicit influence among users in a social sharing system, so as to acquire a better understanding of the dynamics of user interactions and to facilitate item discovery and recommendation. By identifying users who have greater influences on the others, we can also come to know who make greater impact on or contribution to the system and the user community.

3. MODELLING IMPLICIT INFLUENCE

Implicit social influence among users cannot be directly measured due to the lack of explicit social ties. However, we can estimate such influences based on the activities of the users. In the simplest case, it is possible that user u_a has influence on user u_b if u_b adopts an item that u_a adopted earlier. While this is not a strong evidence as a user can be influenced by any one of the users before him (or not), the probability increases as this pattern happens more frequently among this pair of users. On the other hand, depending on the design of the social sharing system, a user may come to know of a particular item because it is featured on the homepage of the site or because it has attracted much attention recently. In other words, we also have to consider other things that would possibly influence the choices of the users. Since a model of implicit influence depends on the design of the system, we focus on the popular social bookmarking site Delicious in this paper for our study of implicit influences. However, our proposed method can be easily generalised to other systems with different designs.

It should be noted that in studies of adoption patterns in a social network the probability of a user adopting a certain item usually depends on the number of friends of this user who have already adopted the item [1, 9]. However, we believe this is less appropriate in our case in which we focus on systems with no explicit social ties. In particular, adoptions because of influences from other users also tend to be more straightforward in these systems, not to mention the fact that there is no clear definition of a group of neighbours/friends in such a setting. As a result, we choose to infer the degree of influence by focusing on the relative order of adoption of the users.

3.1 The Model

To design our model of user activity in Delicious, we use an approach similar to the one described in [8], in which the authors describe a model of user activity in Wikipedia[8], the collaboratively-written and free online encyclopedia. In our model, a user $u \in U$ chooses to perform an action, i.e. to adopt a particular item, by sampling from a probability distribution. In particular, we consider a total of five different options as listed below:

a. Choose a user and adopt an item that have already been adopted by the user;

b. Choose an item that has recently been adopted by some users;

c. Choose an item from the set of popular items;

d. Randomly pick an item that has been adopted before;

e. Introduce a new item by being the first to adopt this item.

We decide on these options based on our observations of the system of Delicious. In particular, Delicious presents both a list of recently bookmarked items as well as a list of popular items.[9] Thus the options (b) and (c) above correspond to the cases in which users adopt items after browsing these lists. Although the exact algorithm for generating these lists are not known and it has probably undergone various changes since Delicious was first established, we will attempt to use different parameters to approximate these two lists. Moreover, as we do not only want to construct a model for a specific system, we present a probabilistic model that can accommodate various possible system designs that could result in different user behaviour. For example, other systems may choose to present users with a set of all-time popular items, or items that are visited by most users (regardless of whether they adopt the items or not).

In the description below, we mainly consider the temporal adoption pattern of the users, and therefore we temporarily ignore the different attributes associated with each adoption. In other words, we refer to a social sharing system $\mathcal{S}$ as a tuple $\mathcal{S} = \langle U, D, Y \rangle$, and each adoption is represented by $(u, d, t) \in Y$.

What we are interested in is to estimate $P(d|u,t)$, the probability that user u would adopt item d at time t. We assume that when a user tries to adopt some items, he is likely to choose a particular user (e.g. someone he has been following) and then pick an item from the set of items that this user has already adopted. In this way the probability that a user u chooses another user u' in this process corresponds to the extent to which u is influenced by u'. We will describe shortly how this can be extended to accommodate other possibilities as described above.

Firstly, we define $P(d|u,t)$ as follows:

$$P(d|u,t) = \sum_{u' \in U} P(u'|u) G(d|u',t), \quad (1)$$

where $P(u'|u)$ represents the probability that u would pick u' when he attempts to adopt something. Thus, this reflects the degree of implicit influence of u' on u. $G(d|u',t)$ represents the probability that item d is chosen when u' is picked by u at time t. In other words, whether u adopts d depends on the influences of some earlier adopters of this item and the importance of the item with respect to those users. If $P(u'|u) > 0$, we say that u' is an ***influencer*** of u with the degree of influence equals to $P(u'|u)$. Note that we only focus on the times at which the user u adopts something, therefore we have $\sum_d P(d|u,t) = 1$.

While it seems that Equation 1 only models (a) above, it can be easily extended to accommodate all other options. To do this, we only need to introduce some 'virtual users' whose $G(d|u',t)$ are calculated differently. We describe these users in detail as follows.

[8]Wikipedia: `http://www.wikipedia.org`

[9]For example, the recently bookmarked items and popular items under the tag `photography` can be found at `http://delicious.com/recent/photography` and `http://delicious.com/popular/photography` respectively.

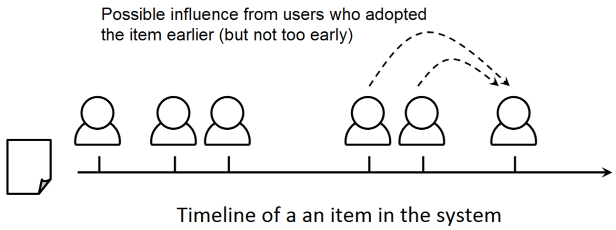


Figure 1: Situation in which a user is considered to be influenced by some other users who adopted an item earlier.

a. Influence of Other Users.

Firstly, we model the case in which a user u picks another user u' and adopts an item from the collection of u'. An aspect that we want to consider in particular is whether a user will only be influenced to adopt items recently adopted by their influencers. In other words, while u' may have a large collection of items, u may tend to adopt items that u' recently adopted, because for example these items are more visible to u. Hence, we can define the probability $G(d|u', t)$ in a way that only items that have adopted by u' in a certain time frame have non-zero probabilities (see Figure 1 for an illustration of this idea):

$$G(d|u', t) = \begin{cases} \frac{1}{H(u',t,\tau_i)} & \text{if } (u', d, t') \in Y \wedge t - t' \leq \tau_i \\ 0 & \text{otherwise ,} \end{cases} \tag{2}$$

where $H(u', t, \tau_i) = |\{d|(u', d, t') \in Y \wedge t - t' \leq \tau_i\}|$ is the number of items adopted by u' in the period $[t - \tau_i, t)$, and $\tau_i \leq t$. In terms of learning the model parameters, this has the effect that u' will not be considered as an influencer of u if u' adopted d at a time much earlier than any time within the specific period. Of course, this can be easily extended to consider the whole timeline by setting $\tau_i = t$. In addition, using the above equation means that when user u' has a lot of items the probability of picking one item from this user will be small. This is reasonable because user u who has chosen u' faces a lot more choices in this case than if he chooses another user who has much fewer items.

b. Recent Items.

Secondly, we use u_r to represent a virtual user who is responsible for 'influencing' ordinary users to adopt recent items (Option b). Thus $P(u_r|r)$ represents the probability that a user would pick an item that was recently adopted by someone. $G(d|u_r, t)$ can then be calculated using the following equation:

$$G(d|u_r, t) = \begin{cases} \frac{1}{M(t,\tau_r)} & \text{if } t - t'_d \leq \tau_r \\ 0 & \text{otherwise ,} \end{cases} \tag{3}$$

where $M(t, \tau_r) = |\{d|(u, d, t') \in Y \wedge t - t' \leq \tau_r\}|$ is the number of unique items being adopted in the period $[t - \tau_r, t)$, and t'_d is the time at which d was last adopted. In other words, the probability is uniformly distributed among all documents that have been adopted once or more in the specific period. By choosing a small value for τ_r (e.g. 1 day), we can then assign probabilities to documents that have recently been adopted.

c. Popular Items.

Likewise, we use u_p to represent a virtual user who is responsible for 'influencing' ordinary users to adopt popular items (Option c). Thus $P(u_p|r)$ reflects how likely a user would adopt an item that is popular. Note that there can be different definitions for 'popular items'. For example, this can refer to items that have recently attracted a lot of attention from users, or it can refer to the all-time popular items determined by the absolute number of times they have been adopted over the entire history. To accommodate different settings, we choose to model the probability $G(d|u_p, t)$ as follows:

$$G(d|u_p, t) = \frac{N(d, t, \tau_p)}{\sum_{d' \in D} N(d', t, \tau_p)} \tag{4}$$

where $N(d, t, \tau_p) = |\{u|(u, d, t') \in Y \wedge t - t' \leq \tau_p\}|$ is the number of users who adopt d in the period $[t - \tau_p, t)$. In other words, an item is more likely to be picked if it has been adopted by more users in the specific period. Hence, to model the case in which users are presented recently popular items, we can choose a certain value for τ_p (e.g. 30 days). We can also choose $\tau_p = t$ to model the case in which the system presents users with all-time popular items.

d. Random Picking.

Besides picking an item from the recent or popular list of items, a user can just randomly pick an item while browsing randomly in the social sharing site. Hence, we introduce u_s as a virtual user who is responsible for 'influencing' users to adopt a random item. Using the definition of the function M above, we define $G(d|u_s, t)$ as follows:

$$G(d|u_s, t) = \frac{1}{M(t, t)}, \tag{5}$$

where $M(t, t)$ returns the number of unique items in the system up to time t.

e. New Items.

At times, users might not be influenced by any of the users or items in the system. Instead, they find something extern to the system (items that have not been adopted by the users of the system before) and introduce them into the system.[10] To model such a case, we introduce a virtual user u_n for the purpose. $P(u_n|u)$ is therefore the probability that user u would introduce a new item to the system. Given the temporal item adoption pattern, we can easily estimate $P(u_n|u)$ using the following equation:

$$P(u_n|u) = \frac{|\{d|(u, d, t') \in Y \wedge t' = l_d\}|}{|\{d|(u, d, t') \in Y\}|}, \tag{6}$$

where l_d is the time at which d is first adopted by some users. In other words, this is simply the ratio of number of times this user is the first to adopt something to the total number of items this user has adopted up to time t. In practise, it would be more realistic to consider the user as the first to adopt an item if he adopts at the first time unit (e.g. the first day on which the item is introduced). Note that in the case of introducing a new item, $G(d|u_n, t)$ is irrelevant

[10] Clearly, a user may still not be influenced by the users or items in the system even though an item appeared in the system before. However, this becomes indistinguishable from the random picking scenario, so we choose to ignore this option in our model.

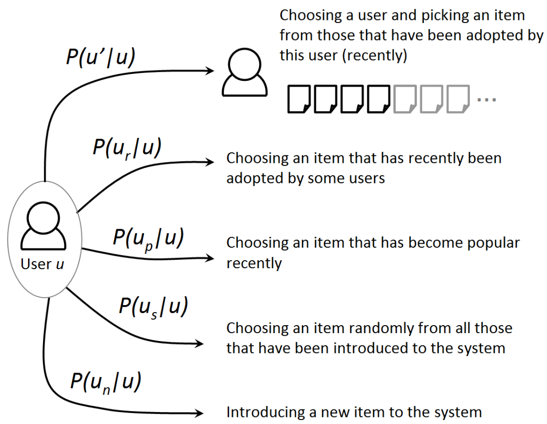


Figure 2: A user selecting from one of the five choices when he attempts to adopt an item.

because there is no documents to pick. Also, for each user u, $P(u_n|u)$ can be estimated easily by counting the number of such occurrences in the training period. Hence, we can treat $P(u_n|u)$ as a constraint while estimating all the other $P(u'|u)$. We will discuss about this in the next section.

As a summary, Figure 2 depicts the different possible actions that a user can take while attempting to adopt some items.

Having defined the probabilities $G(d|u', t)$ for all the real and virtual users, we can then extend Equation 1 as follows,

$$P(d|u,t) = \sum_{u' \in U_A} P(u'|u)G(d|u',t), \quad (7)$$

where $U_A = U \cup \{u_r, u_p, u_s\}$. Note that we assume that every user can only adopt an item once, and therefore a user will never be considered to be influenced by his own earlier adoptions. As a result $P(u|u)$ always equals to zero.

3.2 Estimating the Probabilities

In a history of user activities, we can observe events in the form of (u, d, t). Hence, we can estimate the probabilities $P(u'|u)$ based on a set of training data by maximising the likelihood. Here we describe how this can be done using the EM algorithm. However, it is also possible to use other methods such as Markov chain Monte Carlo techniques.

The log likelihood of a history of user u's activity is given by:

$$\log L_u = \sum_t \sum_{d \in D} \log \sum_{u' \in U_A} P(u'|u)G(d|u',t). \quad (8)$$

We seek to maximise the log likelihood under the constraint that $\sum_{u' \in U_A \cup \{u_n\}} P(u'|u) = 1$. Since for each user u we can obtain $P(u_n|u)$ beforehand, our constraint becomes

$$\sum_{u' \in U_A} P(u'|u) = 1 - P(u_n|u). \quad (9)$$

The conditional expectation of the complete-data log likelihood with the constraint to be maximised is as follows:

$$Q_u = \sum_t \sum_{d \in D} \sum_{u' \in U_A} P(u'|u,d,t) \log P(u'|u)G(d|u',t) + \lambda\Big(\sum_{u' \in U_A} P(u'|u) - (1 - P(u_n|u))\Big), \quad (10)$$

where λ is a Lagrange multiplier. In the E-step, we compute the posterior probability using the Bayes rule:

$$P(u'|u,d,t) = \frac{P(u'|u)G(d|u',t)}{\sum_{u'' \in U_A} P(u''|u)G(d|u'',t)}. \quad (11)$$

In the M-step, we obtain the next estimate of the probabilities $P(u'|u)$ by maximising Equation 10 with respect to $P(u'|u)$ as follows:

$$P(u'|u) \propto \sum_t \sum_{d \in D} P(u'|u,d,t). \quad (12)$$

By iterating the E-step (Equation 11) and the M-step (Equation 12) until convergence, we obtain an estimate for $P(u'|u)$.

3.3 Illustrating Example

In Table 1 we show an illustrating example of estimating implicit influence among users in an online social sharing system. The figure in Table 1(a) depicts the temporal adoption pattern of 4 users involving 2 different items. By setting $\tau_i = 5$ (so that a user's adoption is considered to be influenced by other users' adoption within 5 days), $\tau_r = 1$ and $\tau_p = 10$, we use the above method to estimate the probabilities $P(u'|u)$, which are shown in Table 1(c). We can see that $P(u_n|u_a) = 1$ because u_a did not adopt any item after other users, and all items he adopted were new to the system. In contrast, $P(u_a|u_b) = 0.333$ because u_b adopted two items after u_a. However, we also see that $P(u_s|u_b) = 0.333$ and $P(u_p|u_b) = 0.333$, so the distribution suggests that it is possible that u_b might just have adopted the items because they were popular or because he randomly picked them. Finally, for u_d we can see that the probabilities are no longer evenly distributed. This is because u_d actually picked a less popular item d_y at $t = 8$. Hence, it is estimated that u_d is more likely to be influenced by u_b or to randomly pick an existing item.

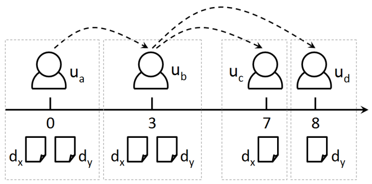
4. EXPERIMENTS

We study whether our proposed model can be used to capture implicit influence among users by performing experiments on a large dataset collected from Delicious. While in Delicious a user can add other users to his own network so that he can be notified of the items adopted by these users, such function was not available in the first few years of the service. The dataset, described in [27], is distributed by the authors for research purpose. It contains the bookmarking activities of over 950,000 users, involving over 50 million items (URLs) and spanning the period between Sep 2003 to Dec 2007.[11]. Since the dataset only contains the date on which a certain item was adopted, the time in all our experiments on the dataset is measured in days.

To avoid data sparsity, instead of working on the whole dataset, we choose to train our model on each of the 50 most

[11] `http://www.dai-labor.de/index.php?id=1726`

(a)



(b)

User	Item	Time
u_a	d_x	0
u_a	d_y	0
u_b	d_x	3
u_b	d_y	3
u_c	d_x	7
u_d	d_y	8

(c)

	Influence $P(u'\|u)$							
	u'							
	u_a	u_b	u_c	u_d	u_r	u_p	u_s	u_n
u_a	-	0.000	0.000	0.000	0.000	0.000	0.000	1.000
u_b	0.333	-	0.000	0.000	0.000	0.333	0.333	0.000
u_c	0.000	0.333	-	0.000	0.000	0.333	0.333	0.000
u_d	0.000	0.479	0.000	-	0.000	0.042	0.479	0.000

Table 1: An illustrating example of item adoption. (a) depicts the the times at which the four different users adopted some items, and also the items they adopted. Here, we set $\tau_i = 5$. The arrows indicate implicit influence deduced from the adoption pattern. (b) shows the tuples corresponding to the pattern in (a). (c) shows the estimated probabilities $P(u'|u)$ based on the adoption pattern.

design	blog	software	web
tools	reference	programming	music
video	webdesign	free	web2.0
art	css	linux	howto
tutorial	news	photography	business
javascript	development	blogs	ajax
google	tips	mac	flash
internet	games	windows	shopping
technology	search	opensource	fun
java	online	security	politics
education	travel	books	science
humor	inspiration	funny	cool
php	food		

Table 2: The 50 most frequently used tags in the Delicious dataset (in descending order of frequency from left to right and top to bottom).

frequently used tags in the dataset (see Table 2).[12] The sub-datasets are obtained by collecting all users who have using the chosen tags and all items that have been assigned the chosen tags. In addition, to reduce the size of the dataset for more efficient processing, we remove users who have only adopted one item.

4.1 Predicting Item Adoption

As a first experiment, we study whether the probabilities $P(u'|u)$ can be used to predict future adoption of the users. In other words, we want to study the accuracy of our model.

Firstly, we split the dataset into two parts by choosing a timestamp $t_x \in [0, T]$ where T is the end of the timeline of the dataset. We train the model using adoption histories of users in the period $[0, t_x]$, and test our model on those in the period after t_x. For a particular user u, using the probabilities $P(u'|u)$ obtained from the training process, we predict the next items that will be adopted by u after time t.

This is done by first calculating the probability $P(d|u, t)$ using Equation 7 and then ranking the items in descending order of their probability of being adopted by the user u. We consider it successful if the next items adopted by the user appears in the top m results of the ranking.[13] To formally describe our performance measure, we first let $top_m(u, t_x)$ be the set of top m results in the ranking, and let D^n_{u,t_x} be the next n items adopted by user u starting from time t_x. We then define ***accuracy***, our performance measure, as below:

$$accuracy(u, t_x, n) = \frac{|top_m(u, t_x) \cap D^n_{u,t_x}|}{n} \quad (13)$$

The value of accuracy ranges from 0 (when not a single item from the next n items adopted by the user appear in our prediction) to 1 (when our prediction contains all the n items).

As for the experiment, we train a separate model for each of the 50 datasets, which correspond to the 50 selected tags. All training periods on average involve 20,000 users and 40,000 items. In each case, we use the first 700, 800, 900 and 1,000 days as training data, and hence we calculate accuracy for $t_x = 700, 800, 900$ and $1,000$. We set $\tau_r = 1$ (considering items adopted in the previous day as recent items) and $\tau_p = 30$ (considering only items that have been adopted in the last 30 days when looking for popular items). In addition, we test our model on different values of τ_i (the 'memory' period of influence), including 30 days, 60 days and 90 days.

To select users for evaluation, we use the following procedure. Firstly, we randomly sample 1,000 users in each case and collect the next 10 items ($n = 10$) they adopt in the testing period. Users and items that did not appear in the training period are ignored. We then use the probabilities $P(u'|u)$ obtained in the training period and Equation 1 to come up with a ranked list of items. In our experiment we set $m = 50$, i.e. we check if the next items adopted by the users appear in top 50 positions in the list. Finally, we calculate the average accuracy of the results using Equation 13.

To gain a better understanding of the performance of the model, we compare our model with a simple k-nearest-neighbour collaborative filtering (C.F.) method. For each user, we consider k other users who are most similar to this user based on the cosine similarity of their adoption histories, which we represent by using vectors of 1 and 0 indicating the items they have adopted. We weight the items by the similarity of the users who have adopted them, and produce a ranking of items in descending order of such scores.

Figure 3 show the accuracy of the model using different parameters and different among of training data. Firstly, we observe our model can be used to predict item adoption at a much higher accuracy when compared with the simple collaborative filtering algorithm. In other words, for a par-

[12]We ignore the tag `imported` because it does not represent a coherent topic and is only used to indicate that a certain URL was imported from an external source.

[13]Instead of only considering the single top-ranked item, this method is chosen because the user can adopt the highly probable items in an order different to the predictions made according to the estimated implicit influences.

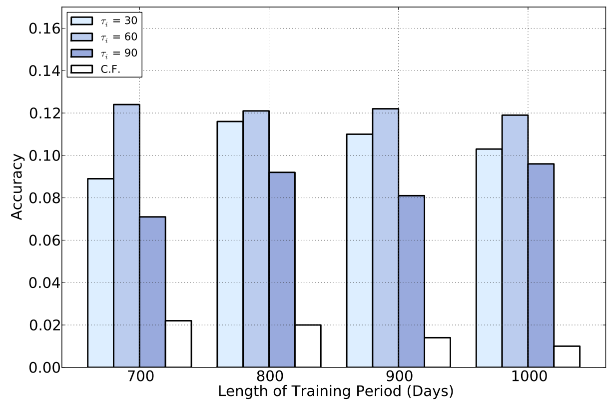


Figure 3: Average accuracy of the model under different conditions across all tags, compared to simple collaborative filtering..

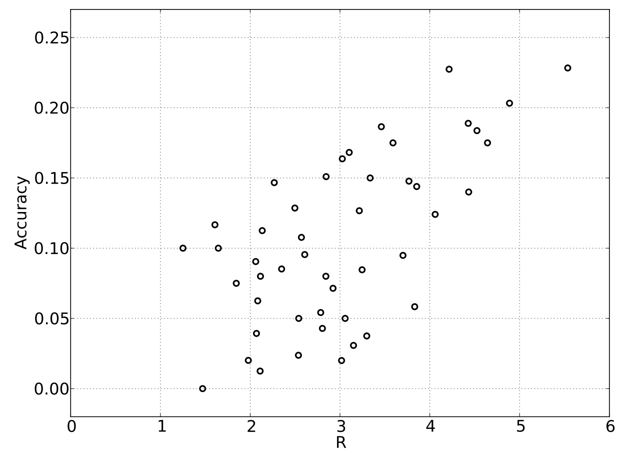


Figure 4: Scatter plot of accuracy against R, the ration of influence from other users to self-influence.

ticular user, other users who have similar adoption histories do not necessarily possess items that are interesting to him. Instead, users who are found to be influential to a particular user form a good basis for predicting item adoption. In addition, collaborative filtering tends to performance poorer as the length of training period increases. This is probably due to the fact that the sparsity of the data increases as more and more new items are introduced into the system. However, although we expect a longer training period will provide more evidence for estimating the implicit influences among users, we do not observer any significant differences in accuracy across different lengths of the training periods. We believe this may also be due to sparsity resulted from the increasing number of users and items.

The above results actually reveal an interesting fact about the importance of considering the temporal adoption pattern when performing adoption prediction and in general item recommendations. Influencers are not very different from neighbours in collaborative filtering because they are all found based on whether two users have adopted the same items. However, in collaborative filtering a neighbour might actually be a follower of the user in question (in which case they would have similar adoption histories), an item that this user adopts at time t is then likely to be adopted by the neighbour only at a time after t, thus leading to the failure of predicting the adoption of this particular item. In contrary, considering the temporal pattern avoids mistaking followers as neighbours in item recommendation.

On the other hand, we observe that accuracy is highest when $\tau_i = 60$. Recall that a later item adoption is considered to be influenced by an earlier adoption by another user only if the time difference between these two adoptions is shorter than the chosen value of τ_i. Hence, our results show that implicit influence between users is subjected to a certain time frame. In other words, a user adopts an item not because of a uniformly distributed influence from all the users who adopted it before him, but only due to the influence from users who recently adopted it. This aspect is probably related to the fact that when a user browse another user's collection of items, he is usually presented with the latest adoption first. From a computational perspective, this is also a desirable characteristic because it means that we do not have to consider the whole timeline when estimating implicit influence among users.

4.2 Self-influence and Predictability

Intuitively, it is reasonable to suggest that the strength of implicit influence among the users would vary across different communities. We can imagine that in some communities users are more independent from each other such that item adoption is less predictable from the relations among the users, and the opposite case can be true in other communities. For the latter it can be expected that prediction of adoption can be more accurate because users are more likely to pick something from their influencers. From our experiment, we also find that there are differences across different topics (tags). To study this, we first introduce R as a ratio defined as follows for a model trained on a particular tag:

$$R = \frac{\sum_{u \in U} \sum_{u' \in U, u \neq u'} P(u'|u)}{\sum_{u \in U} \sum_{u' \in \{u_n, u_s\}} P(u'|u)}. \tag{14}$$

In other words, R reflects whether on average the users who have used the tag are more influenced by other users or are more 'self-influenced' (i.e. tend to introduce items by themselves without paying attention to the items adopted by other users). The larger the value of R, the more likely are the users influenced by other users when adopting an item.

Figure 4 shows a scatter plot of accuracy against the R value for each of the 50 tags, with a training period of 800 days and $\tau_i = 60$. It can be seen that there is a positive correlation between the two variables ($r = 0.69$). In other words, in some topics some users tend to follow the other users, while in some topics users tend to introduce new items or pick items without paying much attention to the collections of other users. For example, the user communities corresponding to the tags `javascript` and `css` score much higher in both R and accuracy than those corresponding to the tags `free` and `howto`. In other words, communities of more specific topics seem to involve stronger implicit influences among the users.

Our discussion here highlights one important aspect of the probabilities of implicit influence estimated using our method. Given any dataset, it is likely that many users may be estimated to have influence on some other users because of their relative order of item adoptions. However, these probabilities are more meaningful when compared to the probabilities of the users being independent from other users. In other words, we should use the R value to get an idea of the strength of implicit influence among users with respect to a certain topic. More generally, the ratio R can be used to determine whether implicit influence is strong enough among users in a social sharing system and whether our approach can be used to complement conventional collaborative filtering techniques in recommendation systems.

4.3 Adoption and Influencers

Besides studying the accuracy of our proposed model, we also want to study whether the influencers identified by our model continue to influence other users in the testing period. While this cannot be directly verified, we can study this indirectly by checking whether items adopted by a user were items adopted by his influencers at an earlier time.

To facilitate our study on this aspect, we introduce a new performance measure here. We define a score S to measure the extent to which user adopts items that have been already adopted by their influencers in the past. Recall that $D_{u,t}$ is the set of items that are adopted by user u at time t. Let I_u be the sequence of influencers of u ordered in descending order of their influence ($P(u'|u)$) on u. We first define a function $q(I_u, d, t)$ as follows:

$$q(I_u, d, t) = \begin{cases} \min_{u' \in I_u}(rank(u')) & \text{if } \exists t' < t, (u', d, t') \in Y \\ \infty & \text{otherwise,} \end{cases} \quad (15)$$

where $rank(u')$ returns the rank of u' according to his influence on the current user u (lower rank means higher influence). In other words, q returns the rank of the most influential user on u who has adopted item d at an earlier time. We then define $S(t)$, which ranges from 0 to 1, as follows:

$$S(t) = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|D_{u,t}|} \sum_{d \in D_{u,t}} \frac{1}{q(I_u, d, t)}. \quad (16)$$

By this definition, S will attain a higher value if it happens that more documents adopted by the users in the testing period have been previously adopted by their influencers. To give an idea of the range of S, $S = 1$ if all documents adopted by the users in the testing period can be found in the collections of their most influential users, and $S = 0.1$ if on average the items adopted are found in the collection of the $10th$ influential users.

We compare our proposed model with other methods. Firstly, we consider again the simple k-nearest-neighbour collaborative filtering method, and generate a ranking of users based on their similarity to the user in question. Secondly, we consider a baseline random method in which a group of k randomly selected users are considered. In our experiment, we set k to be the same as the number of influencers of a user.

To perform our experiment, we use the first 500 days in our datasets as training data, and estimate the influence probabilities among the users. We then treat the next 7 days as a testing period and calculate the averages of $S(t)$ across these 7 days. After that we include these 7 additional days

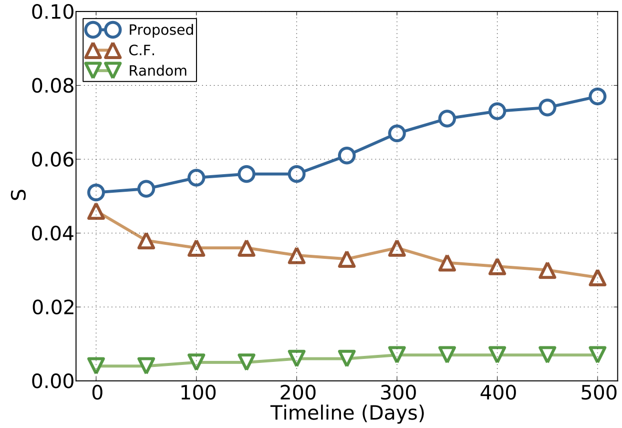


Figure 5: Result of the experiment on how likely users adopt items that have been adopted earlier by their influencers. We only show values of every 50 days here for clarity.

in the training data and repeat the process. We terminate after we have processed 1000 days of adoption history. In this experiment we set $\tau_i = 60$ because it gives the best performance as described above.

Figure 5 shows the results of our experiment. We can see that our proposed model outperform the baseline C.F. method more and more as we have a longer training period. At the end of the timeline, we have S approximately equals to 0.08 for our model, meaning that on average when a user adopts an item, it has been adopted by the 12th most influential user of this user, compared to about 0.03 for C.F., which corresponds to the 33rd most influential user. It shows that instead of adopting something from their similar users, we found that users are more likely to adopt something from their influencers found based on our proposed model.

This result also has implications on recommendation in a social sharing site. In general recommendation systems aims at recommending potentially interesting items to users. However, we can also consider a new recommendation task of recommending users who would possess items that may be interesting, thus allowing a user to choose some other users to follow in order to discover something new and interesting in the future. Our experiment shows that instead of similar users, recommending influential users results in a higher chance of discovering interesting items.

5. RELATED WORKS

Topics related to online social networks, in particular dynamics of information diffusion and social influence, have recently attracted the attention of a lot of researchers from different disciplines. This interest is largely due to the recent surge in the popularity of online social networking sites and the potential of viral marketing in these networks [15, 19]. Research in this area is mainly motivated by the problem of identifying influential users or early adopters for the promotion of certain products among a group of people [5, 20].

While in the past it was difficult to collect data regarding social interaction and information diffusion among a large number of people, the rising popularity of online social networking and sharing sites in recent years have provided an excellent opportunity for the study of real world social network data. Data can be collected easily from social networking sites, some of these even provide API for accessing their public data. As a result, there have been quite a number of studies that analyse social influences and information cascades in specific networks such as an online recommendation network [16], Second Life [2], Flickr [1, 7], and Wikipedia [8].

Several methods have been proposed to model social influences and information diffusion in social networks. In general, in the literature a threshold model is usually considered, in which a user would adopt an item if influence from his neighbours who have adopted the item reaches a certain threshold [9, 12]. There are also different proposals of estimating influences among users in a social network. For example, Song et al. [25], propose to use a Markov chain generated from the relative order of adoption of the users to model information flow in a network, and use this model to perform personalised recommendation. However, they do not consider a restriction of the length of time over which users are considered to influence each other. The authors also proposed another model [24] based on continuous-time Markov chain for studying the diffusion rate of information in a social network. Tang et al. [26] introduce the notion of Topic Affinity Propagation to model social influence in a network with respect to different topics, which are extracted by using topic modelling methods. However, it does not consider the relative order of adoption of the users in the network. In contrast to these previous works, we propose a model to capture implicit influence among users with no explicit social ties. We also consider other factors that would affect the users' choice of adoption.

On the other hand, the results presented in this paper are also highly relevant to research on recommendation systems based on social sharing data. Since users explicitly express their preferences in social sharing sites, these sites offer valuable resources for performing item recommendation to users based on their profiles. For example, Shepitsen et al. [22] generate personalised recommendation by matching user profiles to clusters of tags obtained from a hierarchical clustering process. Bogers et al. [6] and Parra et al [18] present comparative studies of different collaborative filtering techniques on CiteULike. There are also attempts to make use of the explicit social ties in social tagging systems to improve performance of collaborative filtering and recommendation [14, 23]. However, influences among the users, in particular the implicit influences based on the relative order of actions performed by the users, are seldom considered in previous studies of collaborative filtering.

6. CONCLUSION

We study how implicit influence can be captured using a probabilistic model and with some consideration of the design of a social sharing system. By analysing the temporal adoption pattern, our model can be used to estimate the probabilities of influence among users, which can be used to predict future item adoption more accurately than using collaborative filtering. We also find that users are more likely to be influenced by other users' recent activities, such that it is important to focus on a specific time frame when attempting to capture the implicit influence. Overall, our work suggests that while its strength varies across different topics, in general implicit influence is present among users in Delicious and it can be a valuable piece of information that can be used to complement collaborative filtering techniques to generate recommendations to users.

While we make use of the term implicit influence in this paper, the relation among users represented by the probabilities can also be interpreted as one between early adopters and followers/latecomers. An early adopter is more likely to introduce new items to the community or adopt items shortly after they are introduced, and thus have a higher probability of influencing other users. A follower, in contrast, is more likely to adopt some items only after they have become popular, and thus have a lower probability of influencing others users. However, in some cases, an early adopter does not necessarily have great influence on other users. It has been shown that in online social sharing it is not uncommon that some items remain dormant after they are introduced to the system and only gain popularity after a relatively long time [7]. Since our model considers only a limited period instead of the whole adoption history while estimating influence among users, we believe our method should be able to identify users who are truly responsible for the rising popularity of some items. A closely related issue is that if we can identify this kind of users, we can then use them as a basis for estimating the an item's potential of becoming popular in the future. Therefore, we expect to look into this aspect in our future work.

In addition, we see several other possible direction of future research that stems from our current work. In this paper, we work on the datasets of different tags. However, since there exist quite a lot of synonymous tags in Delicious, a single tag might not cover all users and items that are associated with a particular topic. Hence, we plan to employ some methods of clustering or topic modelling in order to generate some comprehensive datasets and study how this would improve the performance of prediction of item adoption. Moreover, while we only consider the patterns of adoption in this work, it has been shown that taking descriptions contributed by users (e.g. tags) into account tend to improve performance in collaborative filtering [17]. Therefore we plan to extend our model by considering also whether two users use similar tags, and see if this will allow us to improve its performance. Finally, we are also interested in applying our model to other social sharing systems and study if there are any differences across systems due to variations in system design or user behaviours.

7. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15, 2008.

[2] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *EC '09: Proceedings of the tenth ACM conference on Electronic commerce*, pages 325–334, 2009.

[3] L. J. Bannon, I. Wagner, C. Gutwin, R. H. R. Harper, and K. Schmidt. Social bookmarking and exploratory search. In *Proceedings of the 10th European Conference on Computer-Supported Cooperative Work*, pages 21–40. Springer, 2007.

[4] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In

WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 501–510, 2007.

[5] F. Bass. A new product growth model for consumer durables. *Management Science*, 15(5):215–227, 1969.

[6] T. Bogers and A. van den Bosch. Recommending scientific articles using citeulike. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 287–290, 2008.

[7] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 721–730, 2009.

[8] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168, 2008.

[9] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Learning influence probabilities in social networks. In *WSDM '10: Proceedings of Third ACM International Conference on Web Search and Data Mining*, 2010.

[10] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proc. of 1st ACM Int'l Conf. on Web Search and Data Mining (WSDM'08)*, pages 195–206. ACM, February 2008.

[11] Y. Kammerer, R. Nairn, P. Pirolli, and E. H. Chi. Signpost from the masses: learning effects in an exploratory social tag search browser. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 625–634, 2009. ACM.

[12] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

[13] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202, 2009.

[14] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202, 2009.

[15] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[16] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 380–389. Springer-Verlag, 2005.

[17] H. Liang, Y. Xu, Y. Li, and R. Nayak. Collaborative Filtering Recommender Systems based on Popular Tags. In *ADCS 2009 : Proceedings of the Fourteenth Australasian Document Computing Symposium*, Springer, 2009.

[18] D. Parra and P. Brusilovsky. Collaborative filtering for social tagging systems: an experiment with citeulike. In *RecSys '09: Proceedings of the third ACM conference on Recommender systems*, pages 237–240, 2009.

[19] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70, 2002.

[20] E. M. Rogers. *Diffusion of Innovations*. Free Press, New York, USA, 5th edition, 2003.

[21] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266, 2008.

[22] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266, 2008.

[23] S. Siersdorfer and S. Sizov. Social recommender systems for web 2.0 folksonomies. In *HT '09: Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 261–270, 2009.

[24] X. Song, Y. Chi, K. Hino, and B. L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 191–200, 2007.

[25] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun. Personalized recommendation driven by information flow. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 509–516, 2006.

[26] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816, 2009.

[27] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Proc. of Mining Social Data Workshop, collocated with ECAI 2008*, pages 26–30, 2008.

[28] M. Xia, Y. Huang, W. Duan, and A. B. Whinston. Ballot box communication in online communities. *Commun. ACM*, 52(9):138–142, 2009.

[29] X. Xin, I. King, H. Deng, and M. R. Lyu. A social recommendation framework based on multi-scale continuous conditional random fields. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1247–1256, 2009.

[30] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116, 2007.