Link Mining for a Social Bookmarking Web Site

Feilong Chen, Jerry Scripps, Pang-Ning Tan
Computer Science & Engineering
Michigan State University
East Lansing, Michigan, US
{chenfeil, scripps, ptan}@msu.edu

Abstract

Social bookmarking tools enable users to save URLs for future reference, to create tags for annotating Web pages, and to share Web pages they found interesting with others. This paper presents a case study on the application of link mining to a social bookmarking Web site called del.icio.us. We investigated the user bookmarking and tagging behaviors and described several approaches to find surprising patterns in the data. We also examined the characteristics that made certain users more popular than others. Finally, we demonstrated the effectiveness of using social bookmarks and tags for predicting mutual ties between users.

1 Introduction

With the staggering rate at which new content is produced on the Internet, it is becoming increasingly difficult for Web users to keep up to date with the new information. Social bookmarking is a tool that enables Web users to share information they found on the Internet with other users sharing similar interests. It allows users to save and organize their bookmarks on a remote Web server. Users may assign tags to each bookmark to annotate what they perceive to be the content. Some of the popular social bookmarking Web sites include `del.icio.us`, `www.citeulike.org`, and `www.furl.net`.

Social bookmarking is a rich but largely unexplored domain in link mining. Many applications such as Web search and text categorization may benefit from the analysis of social bookmarking data. For example, the number of users who bookmarked a Web page can be used as a metric to measure the authoritativeness of the Web page. The tags used to annotate the bookmarks is another useful data source that can be harnessed to improve Web search [4, 9, 14] or Web page classification [8]. A social bookmarking Web site is also a fertile testbed to investigate many social science phenomena such as information diffusion and social selection.

In this paper, we present a case study on the application of link mining to a popular social bookmarking Web site called `del.icio.us`. We first investigated the user bookmarking and tagging behaviors and described two approaches for finding surprising patterns in the data. One approach involves measuring the deviation of a pattern from its expected frequency while the other is based on performing a temporal analysis on the user's bookmarking history. We then examined characteristics that influence the popularity of a user. At `del.icio.us`, a user becomes a fan of another user by adding the other user to his/her network. The number of fans each user has can be used as a measure of user popularity. Because popular users may influence the bookmarking activities of other users [3], it is useful to understand what makes a user more popular than others. Finally, we studied the linking behavior of users at the social bookmarking Web site. Specifically, we consider links in the form of reciprocal ties, which are pairs of users who are mutual fans of each other. Our goal is to predict the formation of such ties based on the user bookmarking and tagging activities. We developed a technique for predicting links based on a kernel alignment approach [6] and obtained very promising results.

2 Preliminaries

We begin with a brief discussion of the terminology used in this paper. While our terminology is based on the features available at `del.icio.us`, it is also applicable to other social bookmarking web sites.

- **User:** A registered visitor of the social bookmarking Web site. Let $\mathcal{U}$ be the set of all users.
- **Bookmark:** A shortcut to a Web page or URL. Let $\mathcal{B}$ be the set of all bookmarks.
- **Fan:** A directional link from one user to another. If a

user X adds another user Y to his/her network, then X becomes a fan of Y.

- **Tag:** A keyword or text description assigned by a user to a bookmark. Let $\mathcal{T}$ be the set of tags used for all the bookmarks in $\mathcal{B}$.
- **Post:** A 4-tuple (b, u, t, τ), where $b \in \mathcal{B}$, $u \in \mathcal{U}$, $t \subseteq \mathcal{T}$, and τ is the timestamp. Let Π denote the set of all posts, also known as the posting history.
- **Reciprocal Tie:** A mutual tie between two users. A reciprocal tie exists between X and Y if X is a fan of Y and vice-versa.

A social bookmarking network can be represented as a graph $\mathcal{G} = (V, E)$, where V is the set of all nodes (users) and $E \subseteq V \times V$ is the set of links created from the reciprocal ties between users. We define the popularity of bookmarks, users, and tags in the following way.

- **Bookmark Popularity:** The number of users who saved a given bookmark.
- **User Popularity:** The number of fans associated with a user.
- **Tag Popularity:** The number of posts that contain a given tag.

Our tag popularity measure is somewhat different than the one used by `del.icio.us`, which is based on the number of unique bookmarks associated with a given tag. There are two advantages for using our measure. First, it is more informative because it takes into account both the number of bookmarks and number of users who have used the tag. Second, our measure is more resilient to spam tags, which are the tags used to mislead users about the content of a Web page or to attract users to a spam Web site. If tag popularity is measured using the number of bookmarks only (instead of number of posts), it would be easier to promote a spam tag into a popular tag.

3 Analysis of del.icio.us Data

Let $\mathcal{D} = \langle \mathcal{B}, \mathcal{U}, \mathcal{T}, E, \Pi \rangle$ be a collection of social bookmarking data, where $\mathcal{B}$ is the set of bookmarks; $\mathcal{U}$ is the set of users; $\mathcal{T}$ is the set of tags; E is the set of links in the social bookmarking network, i.e., $E : V \times V \rightarrow \{0, 1\}$, and Π is the posting history. We wrote a crawler program to collect the data from the `del.icio.us` web site. The crawler was deployed for a three month period (from February 2008 to April 2008) to retrieve the posting history of $330,000$ users.

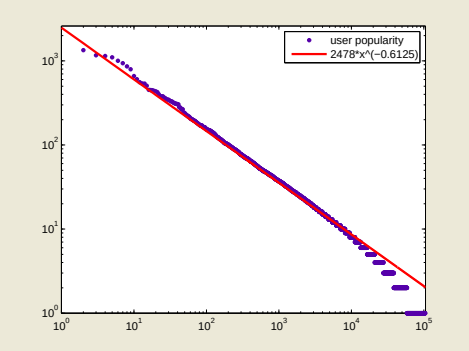

Figure 1. Log-log plot of the user popularity versus the rank of a user. The most popular user has 2244 fans. About 62% of the users have no fans.

3.1 Frequency Distribution Analysis

This section analyzes the popularity of users, bookmarks, and tags based on their frequency distributions. For user popularity, we observe that the number of fans follows a power law distribution [7]. Figure 1 shows the log-log plot of the number of fans (f) versus user popularity rank (r). The least square fit of the curve is given by $f(r) \propto r^{-0.61}$ with a root mean squared error of 0.157. Although the fit is obtained using only a sample of the bookmarks and users in `del.icio.us`, we observe that the exponential factor does not change considerably when the sample size is varied. Note that only 38% of the users have at least one fan.

Table 1 shows the ten most popular users, along with the number of bookmarks they have saved and the top three tags they have used most frequently. At first glance, it appears that the popularity of users is somewhat correlated with their number of bookmarks. In fact, for seven out of the ten most popular users, the number of bookmarks they saved is $2 \sim 8$ times larger than their number of fans. However, we found that the average number of bookmarks to number of fans ratio is larger (≈ 75) for less popular users, since many of them have hundreds of bookmarks but only a few fans. Some popular users such as `adobe`, `ambermac` and `jgwalls` also have more fans than bookmarks. This suggests that the quality of the bookmarks is just as important as the quantity when determining popular users.

Table 2 shows the ten most popular bookmarks, including the number of users who saved them and their five most frequently assigned tags. Most of the popular bookmarks are links to software and other social media Web sites. Unlike the distribution for user popularity, the bookmark popularity does not appear to follow the power law distribution[1].

[1] We have omitted the plot due to lack of space.

Table 1. List of ten most popular users.

User	Identification	# Fans	# Bookmarks	Top 3 Tags
adobe	Adobe Systems Inc.	**2244**	**1193**	Adobe, Macromedia, Photoshop
twit	TWiT Netcast Network	1341	3422	system:unfiled, 67,mbwideas
merlinmann	blogger	1144	4327	43folders,quick_post,system:unfiled
joshua	blogger	1099	9967	tl, undescribed, gis
steverubel	blogger	1034	5142	Blogs, Marketing, rss
wfryer	blogger	1003	4999	InternetSafety, science, EducationReform
regine	blogger	955	4722	gogogo,fun, art
willrich	blogger	950	1569	tools, social, blogging
jgwalls	unknown	**863**	**320**	bit200f06, bit300f07,bit300f06
jonhicks	designing company	795	1567	sidenotes, osx, design
ambermac	web strategist	**682**	**408**	citynews, natn, commandn

Table 2. List of ten most popular bookmarks.

URL	# Users	Top tags
http://www.netvibes.com/	30239	web2.0, netvibes, rss, aggregator, portal
http://digg.com/	24504	news, digg, social, web2.0,community
http://www.alvit.de/handbook/	22066	webdesign,css,reference,web,development
http://www.mininova.org/	22066	webdesign, css, reference, web, development
http://www.lifehacker.com/	21992	blog, lifehacks,productivity, technology, tech
http://kuler.adobe.com/	21871	color, design,webdesign,adobe, tools
http://www.sxc.hu/	21650	photos, images,stock, free, photography
http://www.dafont.com/	21641	fonts, typography, free, design, font
http://www.zamzar.com/	20280	converter, conversion, tools, file, online
http://www.imdb.com/	19941	movies, film, database,cinema, reference

Our data set contains nearly a million tags. The list of most popular tags is shown in Table 3. The frequency distribution of tag popularity is highly skewed—while the most popular tag has been used in more than 3 billion posts, the majority of the tags have been used only a few times. The average tag popularity is $67,537$, but only 35 tags have been used more than this many times. Figure 2 shows a tag cloud displaying the most popular tags at the web site, where popularity is given by the number of bookmarks a tag is associated with. Despite the extensive overlap between the popular tags in the tag cloud and the list shown in Table 3, some of the tags in our list do not appear in the tag cloud—e.g., `isbn`, `asin`, `nsid`, and `watched`—and vice-versa. One possible explanation is that our sample is only a subset of the `del.icio.us` data. Another possible explanation is because of the difference in the tag popularity measure.

3.2 Temporal Analysis of Posting History

We have also performed temporal analysis to examine the bookmarking behavior of the users. We restrict our analysis to users who started bookmarking after December 2004 and have at least 30 months of posting history. There are 5015 users who satisfy these requirements. A collection of time series are then generated, which correspond to the number of bookmarks saved every month by each user. The time series are clustered to identify groups of users with similar bookmarking behavior.

To account for the different lengths and starting periods of the time series, we compute the pairwise similarity between users by using the dynamic time warping approach [11]. Clustering is then performed on the similarity matrix by applying the kmeans [13] algorithm. Figure 3.2 shows the centroids of the clusters we had found. The cluster $c4$ contains users who are extremely active at the beginning, but become less active after the first few months. Cluster $c6$ behaves similarly, except the initial number of posts is not as high as that for cluster $c4$. Both clusters represent quite a significant fraction of users in our data set. There are several potential explanations for the observed behavior. Some users may lose interest after the first few months of testing the system while others may have added all their important bookmarks during the first few months and no longer have many new bookmarks to add thereafter. In contrast, clusters $c1$ and $c3$ have low activities at the beginning but their users become more active later. The difference between the

Table 3. List of 35 most popular tags.

Tags	# Posts	Tags	# Posts	Tags	# Posts	Tags	# Posts
isbn	3191826937	software	160252	webdesign	112846	free	79657
asin	316172324	web	152037	art	98455	tutorial	77556
date	180486826	music	150082	zip	96730	wc	76197
nsid	48973658	reference	146354	css	96517	photography	75066
watched	20071012	programming	141390	business	91199	javascript	73918
congrelicious	3703832	blog	139152	google	85908	development	73129
year	236262	video	123624	tips	84797	politics	72964
design	223971	howto	120505	linux	84357	news	71848
tools	166238	web2.0	120280	blogs	82345		

Figure 2. Tag cloud from http://del.icio.us/tag/.

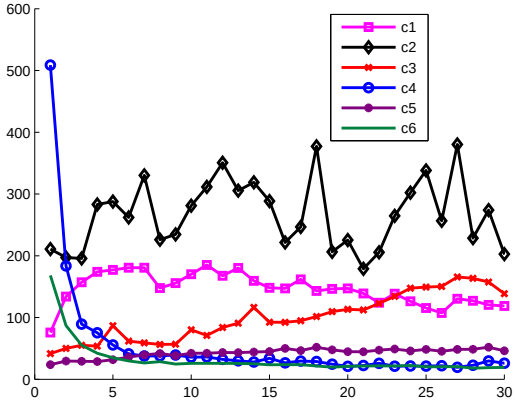

Figure 3. Clustering of User Time Series

two is that the activities of users in $c1$ stabilize after the first few months, whereas the bookmarking activities of users in $c3$ keep growing gradually. Cluster $c2$ represents a group of very active users, whose activities are consistently high during the first 30 months of their posting history. Finally, cluster $c5$ contains users with very few bookmarking activities.

3.3 Discovery of Surprising Patterns

This section describes two approaches for finding surprising patterns in the `del.icio.us` data. The first approach is based on analyzing the frequency distributions while the other is based on performing temporal analysis on the posting history.

For the first approach, we fit the frequency (f) versus rank (r) plot with a known parametric distribution (e.g., power law). The empirical distribution is then used to determine the expected behavior of each observation (user, bookmark, or tag) in the data. A surprising pattern is defined as an observation whose frequency deviates significantly from its expected value. For example, Figure 5 shows the frequency of each user being an *early adopter*[2] of a bookmark versus the user's rank in popularity. After fitting the data to a power law distribution, we obtain $f(r) = 41.81r^{-0.4229}$. We compute the "surprisingness" score (σ) of a user as follows: $\sigma = |\frac{f'-f(r)}{f(r)}|$, where f' is the observed frequency and $f(r)$ is the expected frequency obtained from the empirical power law distribution. Table 4 shows the list of users with highest surprisingness scores, along with the ranks of their user popularity. The list allows us to find users who are often one of the first ten bookmarkers of a URL but yet

[2](A user is considered an early adopter if he/she is among the first ten users who bookmarked a given URL.

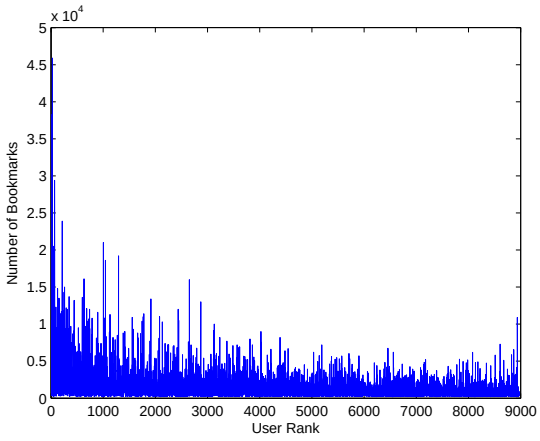

Figure 4. Distribution of the number of bookmarks each user has saved, sorted by the user popularity in decreasing order.

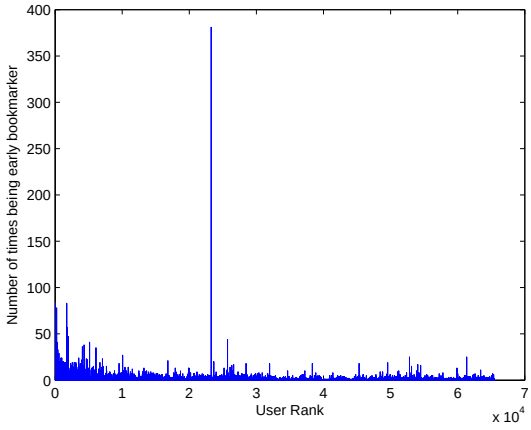

Figure 5. Distribution of the number of times a user is an early adopter.

are not as popular as other users. The user with highest surprisingness score corresponds to the one with highest peak in Figure 5.

Table 4. List of users with largest surprisingness scores (σ).

Rank of User Popularity	σ
23266	896
25703	144
61354	85
10067	79
52819	72

For the second approach, we create a similarity time series for each pair of users. Specifically, we compute the Jaccard similarity [13] of the bookmarks added each month by the users. We expect the monthly similarity values to be low for users who do not collaborate with each other. If the average value of the time series is higher than expected, we suspect potential collusion between the users. In our experiment, we selected a subset of users with posting history no less than 25 months since January 2005. There are 8326 users who fit the criteria. The average similarity value of their time series is about 0.000005 and the maximum similarity is 0.04. Our analysis does not indicate any strong evidence of collaboration among the users. However, after artificially adding a few common bookmarks each month to five randomly selected users, we were able to easily detect such collaboration since their average similarity values exceed 0.5. This approach therefore shows promise in detecting potential collusion in a social bookmarking web site.

4 What makes a user popular?

In `del.icio.us`, popular users may potentially impact the bookmarking activities of their fans. To determine what makes a user more popular than others, we investigated the following user characteristics:

- **Bookmarking activities:** Do popular users bookmark a large number of Web pages? Figure 4 shows the distribution of the number of bookmarks for each user ranked by user popularity. The number of bookmarks generally decreases with decreasing user popularity.
- **Early adoption:** Do popular users tend to be one of the first ten users to bookmark a Web page? Figure 5 shows the frequency each user being an early adopter.
- **Tag Popularity:** Do popular users tend to use popular tags? To answer this question, we examined the tags used by the top $4,000$ users and compute the average popularity of their top 5 tags. Figure 6 shows there is no obvious relationship between the use of popular tags among popular users.

Finally, we employ a decision tree classifier to determine the importance of these factors. We labeled the top 2,000 users with largest number of fans as popular users and the bottom 7,000 users with least number of fans as unpopular users[3]. The F_1 measure obtained using 10-fold cross validation is 0.3418 ± 0.0472, which is significantly better than random guessing (0.2222). Figure 7 shows the decision tree obtained from the data.

[3]The thresholds are obtained by examining the elbow of the frequency versus rank curve for user popularity.

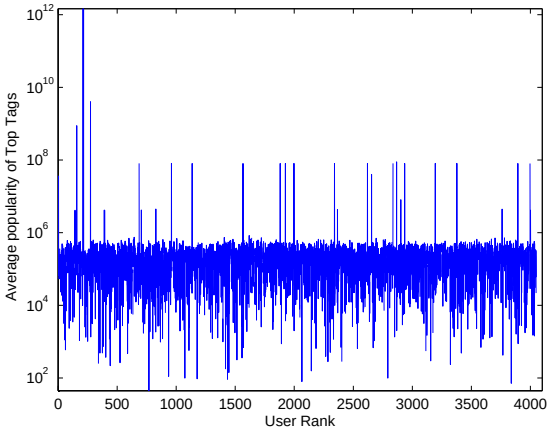

Figure 6. Distribution of the average popularity of a user's five most frequently used tags. The users are sorted by their popularity in descending order.

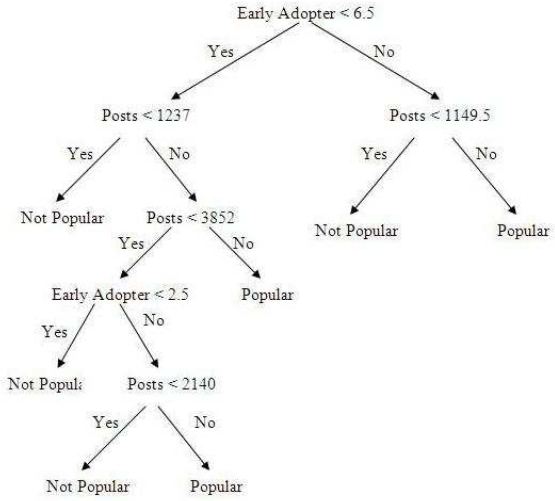

Figure 7. The tree learned for user popularity.

5 Link Prediction in Social Bookmarking

This section presents our approach for predicting links in a social bookmarking network. Although the link prediction problem was previously studied by Shen et al. [12] for discovering friendship relations in a blogosphere network, their approach is not directly applicable to our data since we have two types of features, namely, bookmarks and tags.

We formalize the link prediction problem as follows. Suppose we are given a set of users $\mathcal{U}$, their posting history Π, and a partial link relation $E : V \times V \rightarrow \{-1, 0, 1\}$, where 1 represents a link (mutual tie), -1 represents no link,and 0 represents unknown (or missing) link. The objective of link prediction is to infer the unknown links in the network based on their common bookmarks and tags.

5.1 Algorithm

This section presents a link prediction approach based on kernel target alignment framework [6]. First, we compute the pairwise similarity between users based on the similarity of their bookmarks (K_b) and the similarity of their tags (K_t). Obviously, we can apply a classifier such as support vector machine (SVM) [5] to directly predict the missing links using either K_b or K_t as the input kernel matrix. Instead, our goal here is to combine evidence from both similarity matrices in a principled way to improve link prediction. A straightforward approach would be to take the average of their similarity values, i.e., $\overline{K} = (K_b + K_t)/2$, but this approach does not take into consideration the relative importance of the bookmarks and the tags in predicting links.

Let A be the adjacency matrix induced by the link relation E, i.e., $A(i, j)$ is equal to 1 if the node pair is linked, -1 if they are unlinked, and 0 if unknown. Our objective is to learn a set of weights α and β that maximize the degree of alignment between the adjacency matrix A with the bookmark similarity (K_b) and tag similarity (K_t) matrices.

$$\max_{\alpha,\beta} \frac{\langle A, \alpha K_b + \beta K_t \rangle_F}{\sqrt{\langle A, A \rangle_F \langle \alpha K_b + \beta K_t, \alpha K_b + \beta K_t \rangle_F}} \quad (1)$$

where $\langle P, Q \rangle_F = \sum_{i,j} P(i, j)Q(i, j)$. We simplify the computation by first obtaining the rank-k approximations of the bookmark and tag similarity matrices. Let $X_1 = \sum_i \alpha_i v_i v_i'$ and $X_2 = \sum_i \beta_i w_i w_i'$, where v_i's and w_i's are the top k eigenvectors of K_b and K_t, respectively.

Our objective function in (1) reduces to the following expression:

$$\max_{\alpha,\beta} \frac{\sum_i \alpha_i \langle v_i v_i', A \rangle_F + \sum_j \beta_j \langle w_j w_j', A \rangle_F}{m\sqrt{\sum_i \alpha_i^2 + \sum_j \beta_j^2}}, \quad (2)$$

where m is the number of 1's in A. This is also equivalent to maximizing the following objective function:

$$\max_{\alpha,\beta} \quad \sum_i \alpha_i \langle v_i v_i', A \rangle_F + \sum_j \beta_j \langle w_j w_j', A \rangle_F - \lambda \left(\sum_i \alpha_i^2 + \sum_j \beta_j^2 - 1 \right)$$

Table 5. Link prediction results.

Data	Algorithms	F1	Recall	Precision
Tags only	SVM	0.2656 ± 0.0780	0.6999 ± 0.2830	0.1639 ± 0.0805
	Alignment + SVM	0.2693 ± 0.0682	0.5629 ± 0.1178	0.1770 ± 0.0788
Bookmarks only	SVM	0.2851 ± 0.0013	0.9599 ± 0.0124	0.1866 ± 0.0152
	Alignment + SVM	0.3064 ± 0.0057	0.6826 ± 0.0150	0.1976 ± 0.0036
Tags + Bookmarks	SVM	0.4661 ± 0.1095	0.7436 ± 0.2460	0.3221 ± 0.0171
	Alignment + SVM	0.5167 ± 0.0058	0.5192 ± 0.0073	0.5143 ± 0.0059

which can be solved in closed form as follows:

$$\alpha_i = \frac{\langle v_i v_i', A \rangle_F}{\sqrt{\sum_i \langle v_i v_i', A \rangle_F^2 + \sum_j \langle w_j w_j', A \rangle_F^2}} \quad (3)$$

$$\beta_j = \frac{\langle w_j w_j', A \rangle_F}{\sqrt{\sum_i \langle w_i w_i', A \rangle_F^2 + \sum_j \langle w_j w_j', A \rangle_F^2}} \quad (4)$$

After alignment, we can apply SVM on the aligned kernel matrix, $\sum_i \alpha_i v_i v_i' + \sum_j \beta_j v_j v_j'$, to predict the missing links.

5.2 Experimental Results

We conducted our experiment on 9868 users who have at least 3 friends. Let N_+ to be the number of linked pairs and N_- be the number of unlinked pairs. To overcome the problem of the skewness in the distribution ($N_+ \ll N_-$), we selected all the linked pairs and a random sample of the unlinked pairs ($N_- = 5 \times N_+$) to form our training and test sets. We employed support vector machine (SVM) as our classification algorithm. The results reported here are based on 10-fold cross validation on the reduced data set.

Table 5 summarizes the results of our experiments. First, using both bookmarks and tags clearly improve the effectiveness of link prediction compared to using bookmarks or tags alone. Second, our results also suggest that applying the kernel alignment approach would improve the performance of SVM classifier.

6 Conclusion

This paper presents a case study on the application of link mining to a popular social bookmarking Web site called `del.icio.us`. First, we investigated the user bookmarking and tagging behaviors by analyzing their frequency distributions. Temporal analysis is also performed to identify users with similar bookmarking patterns. We then illustrate two approaches for finding surprising patterns in the social bookmarking data. We also examined the characteristics that make a user becomes popular using a decision tree classifier. Finally, we developed a kernel alignment framework to predict mutual ties in a social bookmarking network and obtained very promising results. The breadth of analysis performed in this case study shows that social bookmarking is indeed a rich domain for applying link mining. The domain also presents interesting research problems such as how to identify potential collusion between users or tag spam [10, 1, 2] in social bookmarking data.

References

[1] Del.icio.us spam, social technology and trust. http://chimprawk.blogspot.com/, Sept 2006.

[2] Manipulating `del.icio.us` with spam. March 2006.

[3] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proc. of WSDM'08*, pages 207–218, Palo Alto, California, 2008.

[4] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proc of WWW 2007*, pages 501–510, Banff, Canada, 2007.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[6] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel target alignment. In *NIPS*, pages 367–373, 2001.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc of SIGCOMM '99*, pages 251–262, Cambridge, MA, 1999.

[8] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[9] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proc of WSDM '08*, pages 195–206, Palo Alto, California, USA, 2008.

[10] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems. In *Proc of AIRWeb '07*, pages 57–64, Banff, Canada, 2007.

[11] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389–1409, September 1981.

[12] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Latent friend mining from blog data. In *Proc of ICDM '06*, pages 552–561, Hong Kong, 2006.

[13] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[14] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *Proc of JCDL '07*, pages 107–116, Vancouver, BC, Canada, 2007.