

Analyzing Social Bookmarking Systems: A *del.icio.us* Cookbook

Robert Wetzker¹, Carsten Zimmermann² and Christian Bauckhage³

Abstract. Social bookmarking systems have recently gained interest among researchers in the areas of data mining and web intelligence, as they provide a vast amount of user-generated annotations and reflect the interests of millions of people. In this paper, we discuss our initial findings obtained from analyzing a vast corpus of almost 150 million bookmarks found at *del.icio.us*. Apart from investigating bookmarking and tagging patterns in this data, we discuss evidence that social bookmarking systems are vulnerable to spamming and hence need to be preprocessed before any insightful analysis can be carried out. We present a method, which limits the influence of spam in social bookmarking analysis and provide conclusions and directions for future research.

1 Introduction

In the recent past, social data mining has started to gain interest in academia and the practitioner world alike. Here, social bookmarking systems, such as *del.icio.us*, *StumbleUpon* or *CiteULike*, have been very successful in attracting and retaining users. This success initially originated from members' ability to centrally store bookmarks on the web. With the coming of age of these services, however, the users perceived value shifted toward the underlying social effects, such as trend indication, advanced web search or recommendation functionality. For researchers, these services are an invaluable source of information, since they provide a vast amount of user-generated annotations (tags) and reflect the interests of millions of users. The *social* aspect of these systems derives from the fact that resources (in general web pages) are tagged by the community and not by the creator of content alone, as in other services like Flickr or YouTube[9]. This characteristic, called collaborative tagging, provides relevant meta-data[3] and is expected to boost the semantic quality of labels[12].

One of the most popular bookmarking systems is *del.icio.us* which represents a suitable case to analyze the characteristics of social bookmarking communities. This is mainly because of its early acceptance in the market, the vast growth over the past five years and easy data accessibility.

The purpose of this paper is threefold: First, we investigate the underlying bookmarking and tagging dynamics of social bookmarking systems using the example of *del.icio.us*. Second, we discuss evidence that social bookmarking systems are highly vulnerable to spam and hence need to be preprocessed before any sophisticated analysis. For a comprehensive study, we, thirdly, collected a corpus

of 142,341,551 *del.icio.us* bookmarks, which - to the best of our knowledge - is the biggest dataset of its kind analyzed to date.

The remainder of this paper is structured as follows: We start with a brief analysis of related work on social bookmarking systems. This is followed by an introduction into the specifics of *del.icio.us*. Then, we present our method of data mining, and analyze bookmarking and tagging patterns we find in the retrieved corpus. Next, we show that bookmarking systems are vulnerable to different forms of spam which requires filtering, before any sophisticated analysis of the social data can be performed. We present behavioral patterns which characterize spam users and propose a method that limits the influence of spam. Finally, we provide conclusions and areas for future research.

2 Related Work

Research on social web communities including social bookmarking systems has grown in popularity. The authors of [2] provide an overview about the structure of collaborative tagging systems. Based on a small subset of the *del.icio.us* corpus, they investigate what motivates tagging and how tagging habits change over time. The authors of [9] present a taxonomy for the classification of tagging systems based on the design choices such as the tagging rights (who can tag what?) or the type of underlying resources (what can be tagged?). The authors also suggest to classify a tagging system according to the incentives of its users.

The authors of [10] argue that social tagging systems can be described as tripartite graphs, involving users, tags and resources, extending the traditional bipartite ontology model by the user dimension. Based on the tripartite model, the authors show that semantically related tags can be clustered in order to discover emerging ontologies. The integration of collaborative tagging systems with the semantic web concept is also the goal of [11]. The authors combine filter and cluster techniques to extract the semantics emerging from the tagspace. Both, [10] and [11], base parts of their analysis on *del.icio.us* data but only consider data sets with less than 100,000 bookmarks.

Social bookmarking systems also provide a promising source for the detection of trends. The authors of [5] apply the tripartite community model also found in [10] and a diffusion technique similar to Google's PageRank algorithm in order to detect trends in social resource sharing communities. Their algorithm enables the authors to rank items (users, tags, urls) with respect to a given topic preference vector. The authors can thus detect trends by comparing the popularity (rank) of items at different points in time. The proposed method is evaluated on a data set of 17 million *del.icio.us* tag assignments⁴.

⁴ We estimate this to be equivalent to around 5.4 million bookmarks, as in

¹ DAI-Labor, Technische Universität Berlin, Germany, email: robert.wetzker@dai-labor.de

² Cambridge University, Cambridge, UK, email: zimmermann@cantab.net

³ Deutsche Telekom Laboratories, Berlin, Germany, email: christian.bauckhage@telekom.de

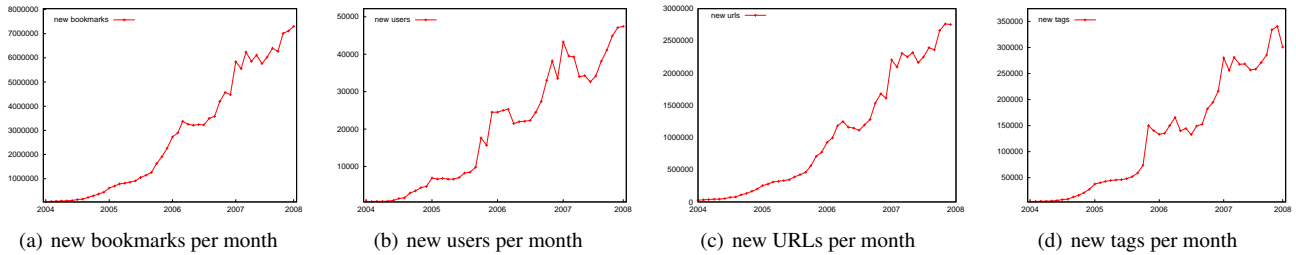


Figure 1. The monthly growth of del.icio.us between 2004 and 2008 by posted bookmarks, new users, new URLs and new tags.

The most complete analysis of the *del.icio.us* bookmarking system is given in [3], where the authors investigate the potential of collaborative bookmarking to enhance web search. Among other results, the authors show that there exists a reasonably high overlap between query terms and the tags found in *del.icio.us*. Furthermore, they report that social bookmarking systems reflect changes within the underlying web structure, such as newly appearing or recently modified web pages, earlier than search engines or directory services, such as the Open Directory.

The effect of spam on the quality of social bookmarking data has been mainly ignored in the past. However, a study on the detection and prevention of spam in tagging systems can be found in [7]. Furthermore, the authors of [8] present their first results on the construction of a spam detection framework for the BibSonomy bookmarking system. The authors test standard machine learning approaches in order to identify the features that characterize social spam.

3 The del.icio.us bookmarking service

As shown in figures 1(a)-1(d), *del.icio.us* is a fast growing social bookmarking service. It allows its users to centrally collect and share their bookmarks, which can refer to any resource on the web as long as this resource can be identified by an URL. When adding a bookmark, users can provide a description, by default the title of the web site, an extended description and tags they consider related. A bookmark can be added by going to the *del.icio.us* web site or, more conveniently, by using one of the many browser plug-ins which make bookmarking to a social bookmarking system as easy as browser wise bookmarking. *del.icio.us* went online in September 2003⁵ and is still constantly growing. According to our dataset there were over 7,305,559 newly added bookmarks and 47,429 newly appearing *del.icio.us* users in December 2007.

There exist two main channels of diffusing information and attention within *del.icio.us*. Users can subscribe to other users' bookmarks or to topics represented by tags. This way, users receive updates in real-time about developments in their friends or colleagues interests or the URLs appearing in a certain domain. The second channel is the *del.icio.us* main page. Once a resource temporarily reaches a certain popularity, it appears on this page and is therefore likely to attract the attention of other users.

Our dataset consists of 142 million bookmarks downloaded from *del.icio.us* between September 19, 2007 and January 22, 2008. As a starting point we downloaded all bookmarks related to the tag "web2.0". From this basis, we extracted all related tags and recur-

⁵ our sample bookmarks were tagged with 3.16 tags on average.
⁵ <http://blog.delicious.com/blog/2007/09/delicious-is-four.html>

Table 1. Corpus details

item	count
users	978,979
bookmarks	142,341,551
assignments	450,113,886
URLs	54,401,067
domains	8,828,058
tags	6,933,179

sively used these for further queries. As a result of this process, we retrieved 45 million unique bookmarks. During our retrieval process, we found that the *del.icio.us* service does not return all relevant bookmarks when queried tag wise. As this limitation does not occur when querying delicious user wise, we additionally downloaded the bookmarks of the most active users within the corpus. For the analysis presented here, we only consider the bookmarks obtained by this user based retrieval. By the time of our analysis, we obtained an overall corpus of 142,341,551 bookmarks from 978,979 users which - to the best of our knowledge - is the biggest dataset of this kind analyzed to date. Details about the corpus are shown in Table 1.

4 Bookmarking patterns

Table 2 lists the most popular URLs in our corpus. The list is thematically split into sites related to social resource or knowledge sharing (1., 2., 5., 6., 7., 8.) and web development related sites (4., 9., 10.). Obviously, the *del.icio.us* community is biased toward web community and web technology related content. This tendency is also reflected by the Top 10 domains given in Table 3. Here, we additionally find news provider and enterprise portals among the Top 10 items, where articles or other content are identified by deep links. Table 3 also emphasizes the fact that enterprise domains tend to have a lower bookmarks per user rate than domains with highly dynamic content (see also Figure 6).

Table 2. Top 10 most frequent URLs in the corpus

	URL	bookmarks
1.	http://www.flickr.com	33,222
2.	http://www.pandora.com	32,634
3.	http://www.netvibes.com	26,743
4.	http://script.aculo.us	26,082
5.	http://slashdot.org	25,272
6.	http://en.wikipedia.org/wiki/Main_Page	23,983
7.	http://www.youtube.com	23,530
8.	http://www.last.fm	22,757
9.	http://oswd.org	20,430
10.	http://www.alvit.de/handbook	20,230

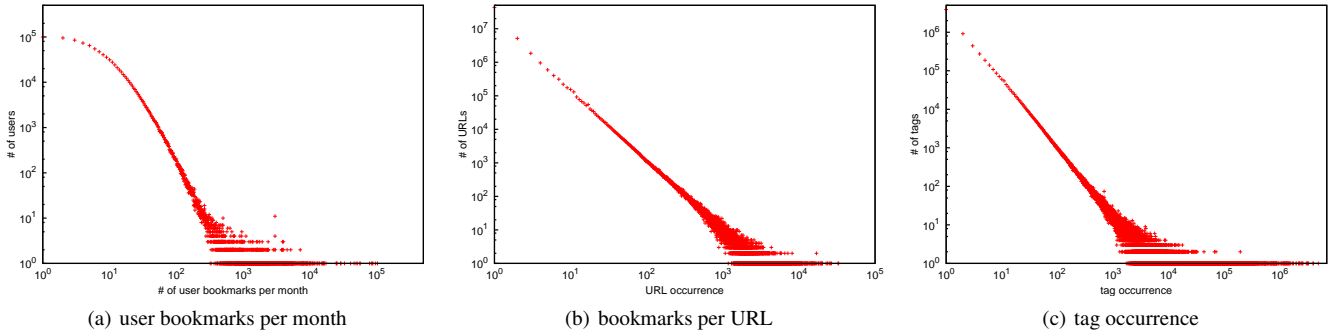


Figure 2. Some power law distributions found on *del.icio.us*.

Table 3. Top 10 most frequent domains in the corpus

	domain	bookmarks	users
1.	http://en.wikipedia.org	919,465	205,639
2.	http://www.youtube.com	915,789	186,326
3.	http://www.flickr.com	535,176	162,363
4.	http://www.nytimes.com	503,776	101,375
5.	http://www.google.com	392,360	156,990
6.	http://lifesacker.com	368,078	90,628
7.	http://www.amazon.com	341,414	91,073
8.	http://news.bbc.co.uk	317,978	75,610
9.	http://www.microsoft.com	290,501	101,947
10.	http://community.livejournal.com	280,020	29,655

As reported by other authors [3, 4], we find the user activity to follow a power law distribution with few users being responsible for a high number of posts as shown in Figure 2(a). The Top 1% of users proliferates 22% of all bookmarks, the Top 10% contribute 62%. These values are above the values reported by [3]. We assume that this difference is due to an increase in spam posts within the recent months (see section 5). Another power law dependency can be found for the occurrence frequencies of URLs where 39% of all bookmarks link to the Top 1% of URLs and 61% to the Top 10% (Figure 2(b)). Furthermore, we find that 80% of all URLs appear only once in the corpus. The URL distribution seems less polluted by spam as users can bookmark an URL only once.

The authors of [2] observe that the *del.icio.us* community pays attention to new URLs only for a very short period of time. As a result, these URLs receive most of their posts very quickly and disappear shortly afterwards. Figure 3 shows the popularity of the most popular URLs in June 2006 that were unknown the month before. As can be seen, each URL peaks within very few days before the number of posts drastically decreases. According to [2], this burst in popularity is likely to be caused by the appearance of an URL on the *del.icio.us* main page triggered by external reasons, such as the appearance of this URL on a widely read blog. Another cause of an initial popularity increase could be the spread of interest within the network of *del.icio.us* users itself.

We find each bookmark to be labeled with 3.16 tags on average. However, the average number of tags assigned to bookmarks varies significantly among users as shown in Figure 4. Moreover, about 7% of all bookmarks are not tagged at all.

The tags assigned to a bookmark can perform different functions, as described in [2]. The authors identify seven tagging purposes the most relevant being the assignment of tags for describing the topic and the type of bookmarked resources. Our analysis underlines these

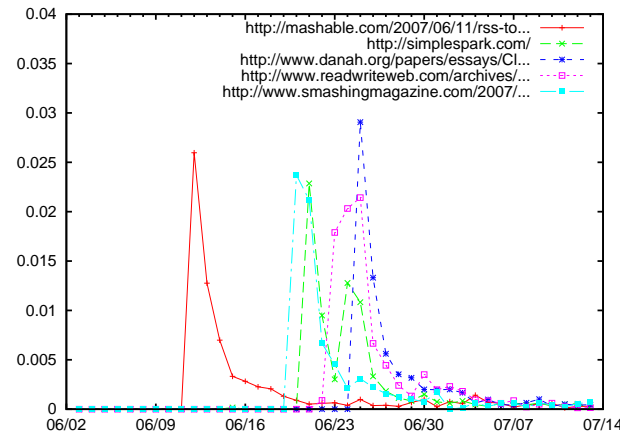


Figure 3. Popularity of 5 sample URLs as percentage of overall bookmarks over time. Most upcoming URLs disappear after peaking.

findings as can be seen from Table 4 which lists the 20 most frequent *del.icio.us* tags. The vocabulary of *del.icio.us* users seems to

Table 4. Top 20 most frequent tags in the corpus

	tag	count		tag	count
1.	design	4,936,513	11.	free	2,501,411
2.	blog	4,027,524	12.	web2.0	2,428,219
3.	software	3,955,838	13.	art	2,303,954
4.	web	3,272,325	14.	linux	2,256,768
5.	tools	3,234,032	15.	css	2,218,035
6.	reference	3,153,890	16.	howto	2,173,611
7.	programming	3,087,505	17.	tutorial	1,980,405
8.	music	2,990,034	18.	news	1,963,509
9.	video	2,603,455	19.	photography	1,766,759
10.	webdesign	2,548,616	20.	business	1,718,118

be highly standardized. Even so, there exist around 7 million tags in our corpus only 700 account for 50% of all assignments. This convergence is likely to be supported by the tag recommendation mechanisms provided by *del.icio.us* which suggests tags based on own or other users previous labels. 55% of all tags were found to appear only once in the data. Figure 2(c) demonstrates the tag distribution.

Tendencies in the *del.icio.us* tag distribution strongly correlate with external events as shown in Figure 5 which presents the dynamics of 5 sample tags in 2007. As can be seen from the time series,

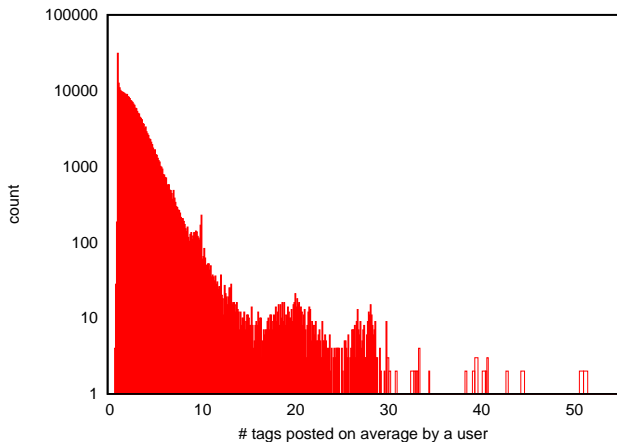


Figure 4. Average number of tags assigned by a user to his bookmarks. Only users with at least 100 bookmarks were considered.

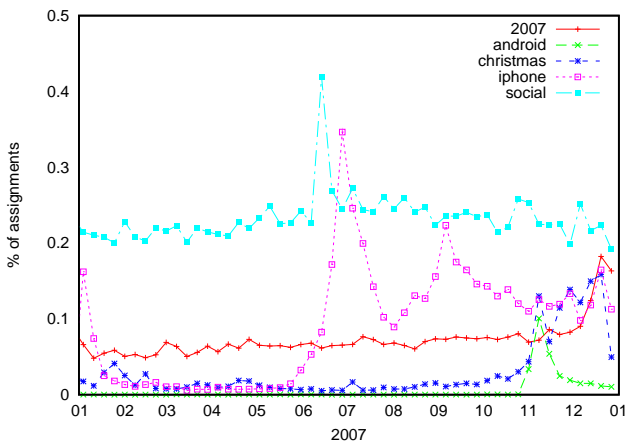


Figure 5. Occurrence of 5 sample tags in 2007 as percentage of overall assignments.

the tagging trends reflect both the upcoming of new technologies, such as Google's 'Android' announced in early November 2007, and periodic events, such as 'Christmas'. The delay between an external event and its echo on *del.icio.us* seems marginal, which was also reported by [3].

5 Social Bookmarking and Spam

An initial analysis of our corpus revealed a frequent occurrence of bookmarks, which presumably are spam and were posted by automated mechanisms. As with applications such as email, the impact of spam is severe for social bookmarking, too. An analysis of the Top 20 most active *del.icio.us* users uncovered 19 users of apparently non human origin posting tens of thousands of URLs pointing to only few domains. These 19 'users' alone account for 1,321,316 bookmarks – around 1% of the corpus. Unfortunately, this result comes to no surprise since *del.icio.us* offers an API for remote postings and URLs that appear on *del.icio.us* have the potential of reaching thousands of users.

The behavior and thus the impact of the spam creators differs. One

user labeled all of his 7,780 bookmarks with the same six tags all referring to the same blog site. Another user, also constantly adding the news from different web portals, labeled each bookmark with more than 100 tags, presumably in order to make the related sources more visible. Furthermore, we encounter bulk uploads, a phenomenon also reported for Flickr by [1], where users upload thousands of bookmarks within minutes and rarely actively contribute again. We also found spammers using multiple user accounts with the most prolific posting at least 538,045 bookmarks to only 5 domains using more than 10 accounts.

Generally, we find spammers to exhibit one or more of the following characteristics:

Very high activity. Automated posting routines may reach much higher participation rates than human users.

Few domains. The URLs posted by spam users are likely to belong to a very small set of domains. Figure 6 plots the number of posts going to a domain versus the number of users bookmarking this domain. As the figure shows, many domains receive a very high number of postings coming from only a few users.

High tagging rate. Some spam users tend to label their bookmarks with an exorbitant number of tags in order to increase their visibility.

Very low tagging rate. Other spam users seem to not care about tagging at all, but constantly upload bookmarks without any tags probably to increase the number of incoming links on their domain(s).

Bulk posts. Bulk uploads are a strong indicator of automated postings. However, automated postings may also appear for human users, e.g. if a user synchronizes his local bookmarks with *del.icio.us* using existing software tools.

Combinations of the above. In most cases, we find a combination of the above characteristics. Figure 7, for example, shows the correlation between the number of bookmarks a user has and the average number of tags he assigns to each bookmark. As can be seen from the figure, some users tend to have very high values in both dimension and can thus easily be identified as spammers.

The existence of spam may result in highly misleading results when bookmarking systems are considered for trend detection or ontology creation based on tag patterns [1, 5, 10]. In these cases, spam fil-

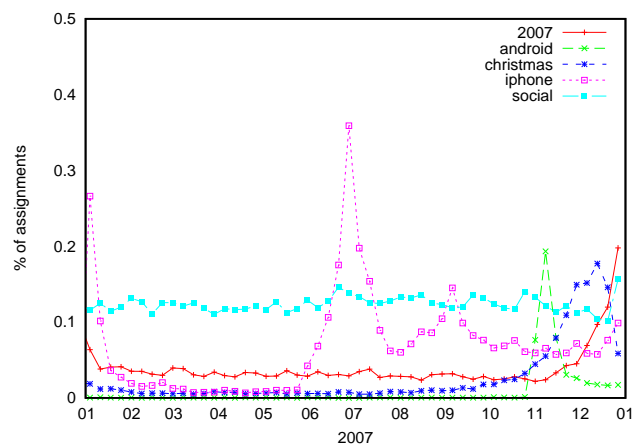


Figure 8. Diffusion of attention in 2007 for the 5 sample tags of Figure 5. The plot shows the number of users assigning a tag for the first time as percentage of the overall first time assignments.

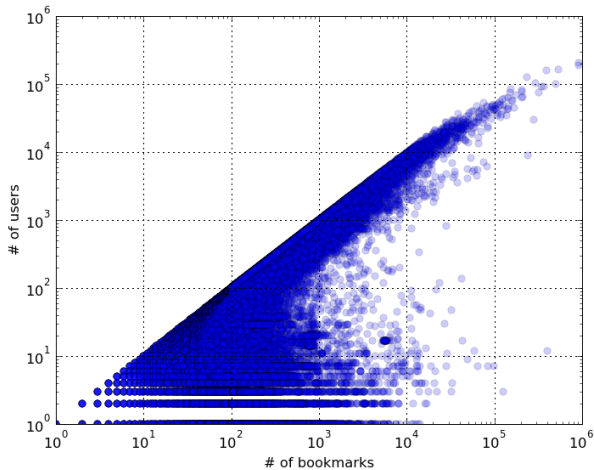


Figure 6. The number of bookmarks compared to the number of users linking to a domain.

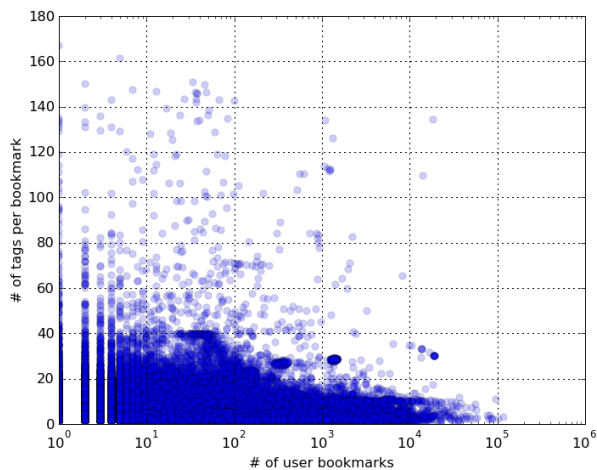


Figure 7. The number of user bookmarks compared to the average number of tags assigned by the user.

tering should precede any sophisticated analysis. However, in some cases, detecting spam may be computationally expensive or ambiguous. We, therefore, propose a new concept called *diffusion of attention* which helps to reduce the influence of spam on the distribution of tags without the actual need of filtering. We define the attention a tag achieves in a certain period of time as the number of users using the tag in this period. The diffusion for a tag is then given as the number of users that assign this tag for the first time. This way, we measure the importance of an item by its capability to attract new users while putting all users on an equal footing. Every user's influence is therefore limited and a trend can only be created by user groups. Figure 8 shows the effect of the concept of *diffusion of attention* on the tags we already plotted in Figure 5. A comparison of both figures reveals that the *diffusion of attention* concept reflects all major trends also obtained by an occurrence based measurement. We find that the measure is able to reflect both seasonal and newly popular tags. The only 'trend' not appearing within the new plot is the peak of the 'social' tag in June 2007. An analysis showed this peak to be caused by the activity of one single user posting 5,666 bookmarks all tagged 'social' within the relevant week. All of these bookmarks link to the same domain. This spam trend does not appear using our *diffusion of attention* measure.

Retaining only combinations of users and tags that appear for the first time in our corpus, reduces the number of total tag assignments from 450 million to 102 million.

6 Conclusions and Future Work

Our analysis shows that social bookmarking provides a valuable source for information retrieval and social data examination. However, we find, that in the case of *del.icio.us* spam highly distorts any analysis. More specifically, among the Top 20 super users of *del.icio.us* we find that 19 users have an apparently automated origin. Therefore, for our future work, we plan to investigate how the spam characteristics we identified can help in the automated spam detection and filtering process. In this context, we also plan to apply

existing spam detection methods known from other domains, such as emails or blogs (e.g. [6]).

Further, we investigate the bookmarking and tagging behavior within the *del.icio.us* community. Here we find, a classical long-tail distribution in the participation rate and the interest levels. Future work could be directed toward the analysis of tagging behavior, identification of user preferences and the topic aware detection of trends.

REFERENCES

- [1] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins, 'Visualizing tags over time', in *WWW '06*, New York, NY, USA, (2006). ACM Press.
- [2] Scott A. Golder and Bernardo A. Huberman, 'Usage patterns of collaborative tagging systems', *Journal of Information Science*, **32**(2), (2006).
- [3] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina, 'Can social bookmarking improve web search?', in *WSDM '08*. ACM, (2008).
- [4] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme, 'Information retrieval in folksonomies: Search and ranking', in *ESWC*, pp. 411–426, (2006).
- [5] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme, 'Trend detection in folksonomies', in *SAMT*, volume 4306 of *Lecture Notes in Computer Science*. Springer, (2006).
- [6] Pranam Kolari, Tim Finin, and Anupam Joshi, 'SVMs for the Blogosphere: Blog Identification and Splog Detection', in *AAAI Spring Sympo. on Comp. Appr. to Analyzing Weblogs*. Computer Science and Electrical Engineering, University of Maryland, Baltimore County, (March 2006).
- [7] Georgia Koutrika, Frans Adje Effendi, Zoltán Gyöngyi, Paul Heymann, and Hector Garcia-Molina, 'Combating spam in tagging systems', in *AIRWeb '07: Proc. of the 3rd int. workshop on Adversarial inf. retrieval on the web*, pp. 57–64, (2007).
- [8] Beate Krause, Andreas Hotho, and Gerd Stumme, 'The anti-social tagger - detecting spam in social bookmarking systems', in *Proc. of the 4th Int. Workshop on Adversarial Information Retrieval on the Web*.
- [9] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis, 'Hit06, tagging paper, taxonomy, flickr, academic article, to read', in *HYPERTEXT '06*, New York, NY, USA, (2006). ACM.
- [10] Peter Mika, 'Ontologies are us: A unified model of social networks and semantics', *J. Web Sem.*, **5**(1), (2007).
- [11] Lucia Specia and Enrico Motta, 'Integrating folksonomies with the semantic web', 624–639, (2007).
- [12] James Surowiecki, *The Wisdom of Crowds*, Doubleday, May 2004.