Detecting Trends in Social Bookmarking Systems using a Probabilistic Generative Model and Smoothing

R. Wetzker, T. Plumbaum and A. Korth
DAI-Labor, Technische Universität Berlin
Berlin, Germany
www.dai-labor.de

C. Bauckhage, T. Alpcan and F. Metze
Deutsche Telekom Laboratories
Berlin, Germany
www.laboratories.telekom.com

Abstract

We propose a method for the detection of trends in social bookmarking systems. Compared to other work in this emerging field, our approach has a more sound statistical basis. In order to cope with the problem of vanishing probabilities due to data sparsity, we apply smoothing and show that it allows for an easy calibration of our trend detector resulting in better generalization and scalability. We test our approach on a collection of $105,000,000$ *bookmarks collected from the del.icio.us bookmarking service. To our knowledge, this is the largest corpus of a real world bookmarking service analyzed in this context. The results show that our method outperforms previously proposed methods and successfully detects trends in the data.*

1 Introduction

Social bookmarking systems, such as *del.icio.us*, *StumbleUpon*, or *CiteULike* have recently attracted the interest of the research community since they provide a vast amount of user-generated annotations (tags) and reflect the interest of millions of users. The *social* aspect of these services derives from the fact that resources (usually web pages) are tagged by the community as a whole and not only by the creator of content alone as it is the case for services like Flickr or YouTube [8]. Collaborative tagging, as considered in this paper, has been shown to provide relevant meta-data [6] and is expected to boost the semantic quality of labels [10].

Trend detection in on-line communities creates new opportunities in areas such as product tracking or marketing. So far, existing studies in this direction mainly focused on the identification of opinion leaders and the detection of trends in the blogosphere [1, 3]. However, bookmarking systems, too, have recently attracted interest in this regard [4, 7]. We shall discuss these works in detail in section 4 of this paper. First, however, we introduce a model of social bookmarking systems based on a set of bipartite graphs. Then, we describe how trends in social bookmarking systems can be detected by means of a probabilistic generative model with corresponding smoothing priors. In our experimental evaluation, we consider a corpus of 105 million *del.icio.us* bookmarks, discuss the trend detection capabilities of our approach and conclude by comparing our findings to the trends detected by other measures.

2 The model

According to [7], a social bookmarking system can be described using a tripartite graph whose vertex set is partitioned into three disjoint sets

$$U = \{u_1, \ldots, u_k\}$$
$$T = \{t_1, \ldots, t_l\}$$
$$R = \{r_1, \ldots, r_m\}$$

which correspond to the sets of users, tags and resources (URLs), respectively. [7] also defines relations $Y \subseteq U \times T \times R$, so that the tripartite hypergraph is given by $G = (V, E)$, where $V = U \cup T \cup R$ and $E = \{\{u, t, r\} | (u, t, r) \in Y\}$. We simplify this structure to three bipartite graphs UR, UT, RT which model the link structure between each pair of sets separately and thus resemble the approach in [9]. Due to symmetry and without loss of generality, we restrict our discussion to UT and look for trends apparent from the tagging behavior of *del.icio.us* users. In this case, we have the graph

$G_{UT} = (V_{UT}, E_{UT})$ with

$$V_{UT} = U \cup T$$
$$E_{UT} = \{\{u,t\} | (u,t) \in U \times T\}.$$

2.1 The dataset

Our dataset consists of 105 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 the corresponding collection, we extracted all related tags and recursively used them 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 by tag. We therefore additionally downloaded the bookmarks of the 400,000 most active users. This second crawl resulted in a corpus of 105,258,294 bookmarks and 331,449,289 tag assignments which, to the best of our knowledge, is the biggest dataset of this kind analyzed to date.

2.2 Avoiding spam

Authors of related work weight the edges in the graph that relates users and tags by the number of times a user chooses a tag [4, 5, 8]. This, however, does not take into account that most collaborative tagging systems are vulnerable to spam. After an initial analysis of our dataset, we found some users of *del.icio.us* to be bots. The behavior of these automatic "users" varies but is generally characterized by a very high rate of participation and anomalous tag assignments [11]. In order to limit the influence of spam, we apply the *diffusion-of-attention* concept presented in [11]. We, therefore, prescind from count based edge weights in our graph G_{UT} but set all weights to 1. In this way, we measure the importance of an item by its capability to attract new users while treating all users equally. Removing from our corpus all combinations of users and tags that occur more than once, reduces the number of total tag assignments from 331 million to 70 million.

3 Trend detection

We consider an item, i.e. a tag, to signify a trend, if it attracts significantly more new users in a currently monitored period of time than it did in past periods. For each period, we therefore count the number of users that assign a tag for the first time. This count equals the number of new edges E_{UT} in that period. Since we consider trends as statistical anomalies, we measure their significance using a probabilistic generative model. For this purpose, we compare the item distributions D_0 and D_1 resulting from the counts over two consecutive periods t_0 and t_1 and assume that all items in D_1 were generated independently at random. If the probability of an item i at time t_0 is assumed to be $p_0(i)$, the probability of observing item i at a frequency of $f_1(i)$ in the period t_1 is given by

$$p(f_1(i)) = \binom{n_1}{f_1(i)} p_0(i)^{f_1(i)} (1 - p_0(i))^{n_1 - f_1(i)} \quad (1)$$

where n_1 denotes the number of all observations in t_1. In order to avoid problems related to items of zero counts, i.e. items that did not occur in t_0 but do occur in t_1, we choose a beta distribution

$$F(p_0; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_0^{\alpha - 1} (1 - p_0)^{\beta - 1} \quad (2)$$

as a prior, where $\alpha = \epsilon$ and $\beta = (Z - 1)\epsilon$. Here, Z denotes the number of existing items and ϵ the assumed prior frequency of each item.

Combining the prior and the observations at t_0, the $p_0(i)$ in equation (1) becomes

$$p_0(i) = \frac{f_0(i) + \epsilon}{n_0 + \epsilon Z} \quad (3)$$

Using the prior has thus the same effect as *additive smoothing* well known from language modeling (e.g. [2]). Moreover, the additive parameter ϵ allows for controlling the influence of the prior distribution with respect to the observations at t_0 and thus for controlling the degree of smoothing. As the resulting bias toward rare items may be interpreted as a trend itself, we also apply smoothing to D_1 in order to neutralize this effect.

Finally, we calculate a score for all items for which $p_1(i) > p_0(i)$. It is given by

$$score(i) = -log(p(f_1(i)) \quad (4)$$

so that we can consider the items with the highest score to be most significant.

As the exact calculation of $p(f_1(i))$ is computational expensive, we estimate the binomial probability mass function by a Gaussian of the type $\mathcal{N}(n_1 p_0(i), n_1 p_0(i)(1 - p_0(i)))$.

		$\epsilon = 0.2$		$\epsilon = 1$		$\epsilon = 100$	
		tag	score (f_0/f_1)	tag	score (f_0/f_1)	tag	score (f_0/f_1)
Oct'07	1	inrainbows	51747 (0/148)	leopard	65453 (102/3203)	leopard	44647 (102/3203)
	2	leopard	44284 (102/3203)	radiohead	12617 (81/1231)	radiohead	7483 (81/1231)
	3	bit200f07	42420 (0/134)	opensocial	11477 (5/422)	halloween	7168 (476/2380)
	4	twine	31743 (1/285)	twine	9750 (1/285)	imap	5596 (207/1459)
	5	decenturl	23626 (0/100)	prism	9307 (27/663)	prism	3425 (27/663)
Nov'07	1	android	290314 (22/3556)	android	360132 (22/3556)	android	102030 (22/3556)
	2	kindle	130602 (3/930)	kindle	78912 (3/903)	opensocial	18865 (422/3366)
	3	gos	78083 (2/579)	gos	37159 (2/579)	kindle	8470 (3/903)
	4	gracenote	47711 (0/136)	opensocial	17378 (422/3366)	thanksgiving	5300 (158/1223)
	5	dalvik	20434 (0/89)	quicklook	9113 (14/480)	vector	4945 (1818/4030)
Dec'07	1	simpledb	1955161 (0/826)	simpledb	117574 (0/826)	simpledb	8127 (0/826)
	2	knol	937602 (0/572)	knol	56484 (0/572)	christmas	6180 (2219/4648)
	3	remastersys	39233 (0/117)	alphabetize	5078 (2/203)	2007	5314 (1030/2803)
	4	blazeds	20708 (0/85)	alphabetizer	4404 (2/189)	wii	4376 (993/2531)
	5	diso	19745 (0/83)	christmas	4353 (2219/4648)	knol	4080 (0/572)

Table 1. Top 5 tag trends for different months and different values of ϵ.

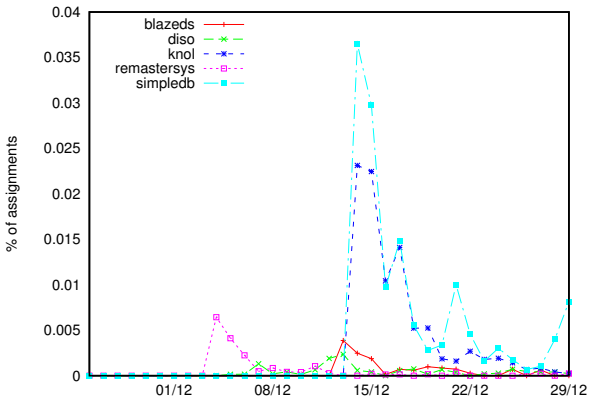

Figure 1. Performance over time for the Top 5 tags of December 2007($\epsilon = 0.2$).

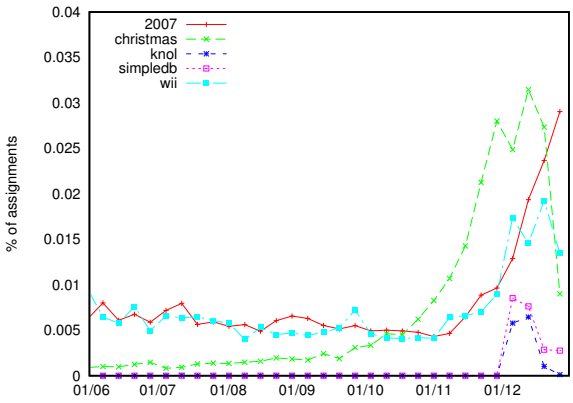

Figure 2. Performance over time for the Top 5 tags of December 2007($\epsilon = 100$).

4 Experiments

In our experimental evaluation, we generate monthly snapshots of the *del.icio.us* corpus. For each snapshot, we determine the number of new edges per tag in G_{UT}. We then measure the evolution of a tag from one month to the next by calculating its score using equation (4). Table 1 presents the Top 5 tag trends for different months and different values of ϵ. It shows that our measure can detect new events, such as the announcement of Google's "Android" or of Apple's "Mac OS X Leopard", and, at the same time, seasonal events, such as "christmas" and "halloween". Furthermore, we observe that the parameter ϵ lets us control the importance of the relative versus the absolute development of an item. Choosing larger values for ϵ favors popular items which are already well spread throughout the bookmarking system, whereas upcoming, often shortly living, tags benefit from a small ϵ. To illustrate this effect, we plot the importance over time for the Top 5 tags of December 2007 for different ϵ values (see Figures 1 and 2). The tags in Fig. 1 peak for a few days before users lose interest. Whereas, choosing a large ϵ allows for the detection of seasonal trends, such as "christmas" or "2007", even though these tags were already popular within the previous months.

4.1 Comparison to other measures

In order to visualize tag trends over time, Dubinko et al.[4] propose a measure which they term *interestingness* and define as $int(i,t) = f_t(i)/(C + f(i))$, where $f(i)$ is the overall occurrence of an item and $f_t(i)$ its occurrence in interval t. C is a regularization constant that increases robustness against scarce observations. Tab. 2 shows the tags of high-

est *interestingness* for November 2007 with $C = 50$. Our experiments highlight, that the interesting tags identified by the measure overlap with the top tags presented above. Moreover, we find the measure to be computational inexpensive. However, in contrast to our method, it only performs well if the overall number of items remains almost constant over all intervals. This, however, limits its general applicability for fast growing communities, such as *del.icio.us*.

	tag	int (f_0/f_1)
1	android	0.98 (22/3556)
2	kindle	0.94 (3/903)
3	gos	0.92 (2/579)
4	hackintosh	0.89 (43/723)
5	quicklook	0.88 (14/480)

Table 2. Top 5 tag trends in November 2007 according to *interestingness* ($C = 50$).

Hotho et al. [7], too, attempt to identify trends from social bookmarking services. They present a method related to the PageRank algorithm that ranks items (tags and URLs) with respect to their importance in relation to a given preference vector. In order to track the impact of an item in a given period, they introduce a measure called *popularity change* defined as: $pc_{t_0 \to t_1}(i) = (\frac{r_0}{n_0} - \frac{r_1}{n_1}) \log(\frac{n_1}{r_1})$ where r_t is the rank of item i at time t and n_t the total number of ranks. Although we do not consider a tripartite link structure, for our evaluations we investigated how the *popularity change* measure would perform when items where ranked simply by their occurrence. We found, that, in this case, *popularity change* strongly favors items that where previously unknown, whereas items above a certain popularity tend to be ignored by the measure. We attribute this to the fact that many items appear at most once or twice within a month. A small increase in occurrence therefore results in a much better rank for formerly unknown items. As an example, Tab. 3 shows the Top 5 upcoming items for November 2007.

5 Conclusions

We presented a new statistical model for the emerging problem of trend detection from social bookmarking services. Our experiments on a large corpus show that our method successfully detects latent trends. As compared to other trend detectors proposed in the literature, our principled use of priors results in increased robustness and scalability.

	tag	pc (f_0/f_1)
1	gravenote	2.19 (0/136)
2	dalvik	2.03 (0/89)
3	23andme	2.02 (0/86)
4	blog-directories	1.97 (0/76)
5	goating	1.93 (0/68)

Table 3. Top 5 tag trends in November 2007 according to *popularity change*.

References

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the Influential Bloggers in a Community. In *Proc. Int. Conf. on Web Search and Web Data Mining*, pages 207–218. ACM, 2008.

[2] S. F. Chen and J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proc. Ann. Meet. of the Ass. for Comp. Ling.*, pages 310–318, 1996.

[3] Y. Chi, B. L. Tseng, and J. Tatemura. Eigentrend: Trend Analysis in the Blogosphere Based on Singular Value Decompositions. In *Proc. Int. Conf. on Information and Knowledge Management*, pages 68–77. ACM, 2006.

[4] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing Tags over Time. In *Proc. Int. Conf. on World Wide Web*, pages 193–202, 2006.

[5] S. A. Golder and B. A. Huberman. Usage Patterns of Collaborative Tagging Systems. *J. of Information Science*, 32(2):198–208, 2006.

[6] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can Social Bookmarking Improve Web Search? In *Proc. Int. Conf. on Web Search and Web Data Mining*, pages 195–206. ACM, 2008.

[7] A. Hotho, R. Jschke, C. Schmitz, and G. Stumme. Trend Detection in Folksonomies. In *Proc. SAMT*, volume 4306 of *LNCS*, pages 56–70. Springer, 2006.

[8] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proc. Conf. on Hypertext and Hypermedia*, pages 31–40. ACM, 2006.

[9] P. Mika. Ontologies are us: A unified model of social networks and semantics. *J. Web Sem.*, 5(1):5–15, 2007.

[10] J. Surowiecki. *The Wisdom of Crowds*. Doubleday, May 2004.

[11] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Mining Social Data, Workshop Proc.* ECAI 2008, July 2008.