Measuring Attention Intensity to Web Pages Based on Specificity of Social Tags

Takayuki Yumoto[1] and Kazutoshi Sumiya[2]

[1] Graduate School of Engineering, University of Hyogo
2167 Shosha, Himeji, Hyogo 671-2280, Japan
yumoto@eng.u-hyogo.ac.jp
[2] School of Human Science and Environment, University of Hyogo
1–1–12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
sumiya@shse.u-hyogo.ac.jp

Abstract. Social bookmarks are used to find Web pages drawing much attention. However, tendency of pages to collect bookmarks is different by their topic. Therefore, the number of bookmarks can be used to know attention intensity to pages but it cannot be used as the metric of the intensity itself. We define the relative quantity of social bookmarks (RQS) for measuring the attention intensity to a Web page. The RQS is calculated using the number of social bookmarks of related pages. Related pages are found using similarity based on specificity of social tags. We define two types of specificity, local specificity, which is the specificity for a user, and global, which is the specificity common in a social bookmark service.

1 Introduction

Recently, social bookmarks are not only used to save private bookmarks on the Web but also for users to be notified of popular or interesting Web pages, therefore, the number of social bookmarks is an important metric. This number, however, depends on not only the quality of Web pages but also users' interests in the social bookmark service. Suppose that there are two pages A and B. The topic of page A is popular and the topic of page B is not popular. In this case, page A tends to gather more social bookmarks than page B. However, if pages A and B have the same number of social bookmarks, the attention intensity to page A is smaller than that of page B. To eliminate this bias of users' interests, we need to compare the number of social bookmarks with those of related pages. We define the relative quantity of social bookmarks (RQS) to measure the attention intensity to Web pages. The RQS is calculated using the numbers of social bookmarks of related pages.

Related pages are found using social tags. In social tags, however, there are synonymity and ambiguity problems. An example of a synonymity problem is that some users use the tag "Programming" (capitalized) and other users use "programming" (not capitalized) to bookmark the same page. An example of an ambiguity problems is that tag "apple" can mean "Apple Computer" or a

M. Yoshikawa et al. (Eds.): DASFAA 2010, LNCS 6193, pp. 264–273, 2010.

fruit. This problem also contains a granularity problem of tags. Suppose that some users use the tag "programming" to bookmark pages about a specific programming language, but others use that tag only to bookmark pages whose topic is common in several programming languages. To bookmark pages about a specific programming language such as perl, these other users would use the tag "perl". In this case, the granularity of concept of "programming" for these users is different.

We focus on the fact that synonymity and ambiguity problems do not occur in tags of one user. To find related pages, we use pages bookmarked with the same tags with which the users bookmark the target page. We also focus on the granularity of the concept of tags, and we propose a method for finding related pages using the specificity of tags. We define two types of specificity, global, which is specificity for a social bookmark service, and local, which is specificity for a user.

The rest of this paper is organized as follows. In Section 2, we describe related work. In Section 3, we explain the specificity of social tags. In Section 4, we define relative quantity of social bookmarks to measure the attention intensity to Web pages. In Section 5, we explain our experiments for specificity and for the RQS. In Section 6, we give the concluding remarks and discuss future work.

2 Related Work

Social bookmarks are often modeled as $(user, page, tags)$ or $(user, page, tags, time)$. There have been many studies on social bookmarks for various purposes. Most research defines some kind of relationship between the one of the elements of the model using the other elements. For example, Krestel and Chen extracted a user graph of social bookmarking data to find spammers[1]. This is an example of the relationship between users. On the other hand, Niwa et al. proposed a Web page recommending system, in which they use tag clustering[2]. This is an example of the relationship between tags. Sugiyama et al. proposed a method for finding related pages using the similarity between pages[3]. To measure the attention intensity to Web pages, we need to obtain pages related to the target page and also define similarity between pages. In general, similarities using social bookmarks can be defined as follows:

$$Sim(o_1, o_2) = Sim(S_1, S_2), \tag{1}$$

where o_i is an element in the social bookmark model and S_i is the other elements. In Sugiyama et al.'s and our research, o_i is a page and S_i is user-tag pairs. These similarities are also based on the similarity measure for sets such as cosine similarity and the Jaccard coefficient. These are generalized as:

$$Sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{Univ(S_1, S_2)} \tag{2}$$

If we change the function $Univ(S_1, S_2)$ in formula (2), this function becomes various similarity measure listed in Table 1. In formula (2), all the elements

Table 1. Universe Function of Similarity between Sets

	$Univ(S_1, S_2)$
Cosine	$\sqrt{\|S_1\|\|S_2\|}$
Jaccard	$\|S_1 \cup S_2\|$
Dice	$(\|S_1\| + \|S_2\|)/2$
Simpson	$\min(\|S_1\|, \|S_2\|)$

$s \in S_1 \cap S_2$ are evenly treated. When we weight s by weighting function f, we can describe formula (2) as follows:

$$Sim(S_1, S_2) = \frac{\sum_{s \in S_1 \cap S_2} f(s)}{Univ(S_1, S_2)} \tag{3}$$

Formula (2) equals formula (3) where $f(s) = 1$. In many studies, this weighting function is defined in various forms. For example, Sugiyama et al. defined it based on corresponding ratios of tags and we define it based on the specificity of tags.

Liang et al. proposed recommendation system based on each user's personal usage of tags and the common usage of tags by many users[4]. Their idea is partly similar to ours but the approaches are different. Though they focused on the tags frequently used by each user, we focused on the tags specifying the bookmarked pages in detail.

Chi et al. focused on the specificity of social tags and reported that this specificity decreases through observing the transition of entropy of social tags[5]. We focus on the concept of this specificity to define similarity between pages.

3 Specificity of Social Tags

3.1 Overview of Specificity

Specificity means the ability of a tag to differenciate the page from a set of pages. For example, if the contents of pages bookmarked using a tag vary, the specificity of the tag is low. On the other hand, if the pages bookmarked using the tag describe narrow topic, the specificity of the tag is high. To calculate the specificity, we analyze the pages that the same user bookmarked using the same tag. We used another method that does not depend on content analysis. Furthermore, we propose two types of specificity, local, which is the specificity for a user, and global, which is the specificity common in a social bookmark service.

3.2 Local Specificity

When the user bookmarks fewer pages using the tag against the number of pages the user bookmarks, the local specificity becomes higher. We define the local specificity of tag t for user u, $sp_l(u, t)$ as follows:

$$sp_l(u,t) = 1 - \frac{|Pages(u,t)|}{|Pages_U(u)|}, \tag{4}$$

where $Pages(u,t)$ is a set of pages that are bookmarked by the user u using the tag t and $Pages_U(u)$ is a set of the pages that are bookmarked by the user u. When user u bookmarks using the tag set T, we define the specificity as follows:

$$sp_l(u,T) = \min_{t' \in T}(sp_l(u,t')) \tag{5}$$

In both cases, the range of local specificity is $[0,1]$.

3.3 Global Specificity

Considering the number of the users who use the tag and the frequency of the tags, we define the global specificity of tag t as follows:

$$sp_g(t) = \min\left(1, \frac{|Users(t)|}{|Pages_T(t)|}\right), \tag{6}$$

where $Users(t)$ is a set of users who use tag t and $Page_T(t)$ is a set of pages bookmarked using tag t. The range of $sp_g(t)$ is $[0,1]$. When the number of pages bookmarked using tag t is large against the number of users who use tag t, the global specificity becomes high. When the number of pages bookmarked using tag t is larger than the number of users who use tag t, we regard the global specificity is high enough and set the value as 1. This specificity is weak at polysemy and synonymity. However, it is used to reduce the effect of the biased usage of user tags.

3.4 Combination of Two Specificities

We discuss the relationship between local and global specificity. If the tendency of local specificity matches that of global specificity, tag's usages are the same. Next, we consider the case when the tendency of local specificity does not match that of the global specificity. Suppose that local specificity is high and the global specificity is low. A user uses general tags for bookmarking a few pages. Therefore, this user seems to be familiar with the topic and the specificity of the tag is high. When the local specificity is low and the global specificity is high, the user uses the tag in his/her own way. In this case, the tag may specify the contents in detail but it does not always specify them. In short, when local and global specificities are high, the specificity of the tag should be high. Otherwise, it should be low. Then, the combined specificity of the tag for the user is defined as follows:

$$sp(u,t) = sp_l(u,t) \times sp_g(t) \tag{7}$$

4 Measuring Attention Intensity to Web Page

4.1 Relative Quantity of Social Bookmarks

Tendency of pages to collect bookmarks is different by their topic. because the number of users who are interested in each topic is different. We need to

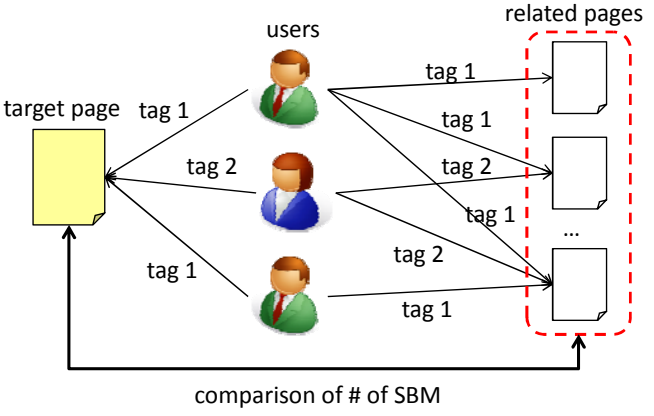

Fig. 1. Schematic of Relative Quantity of Social Bookmarks

normalize the number of social bookmarks of the target page using those of related pages. We define the RQS as the normalized number of social bookmarks. We define RQS of page p, $RQS(p)$ as follows:

$$RQS(p) = BM(p) / \left(\frac{1}{m} \sum_{i=1}^{m} BM(p_i) \right) \tag{8}$$

where $BM(p)$ is the number of social bookmarks of page p, and p_i is a related page of page p. We use the RQS as an estimated value of the attention intensity to the Web page. We show a schematic of this in Figure 1. Related pages are selected from the pages bookmarked with the same tags used by the user to bookmark the target page. RQS is calculated by comparing the numbers of the social bookmarks of the target page and the ones of the related pages.

4.2 Discovering Related Pages

Related pages are found using their similarity to the target page. Similarity is defined using specificity and is based on the Jaccard coefficient.

$$Sim(p_1, p_2) = \frac{\sum_{(u,t) \in UT(p_1) \cap UT(p_2)} sp(u,t)}{|UT(p_1) \cup UT(p_2)|} \tag{9}$$

where $UT(p_i)$ is a set of pairs of users who bookmark page p_i and tags which they use for bookmarking that page. However, if $|UT(p_1) \cap UT(p_2)|$ is large, calculation of $Sim(p_1, p_2)$ requires a large amount of time. Regarding this, there have been studies focused on users who bookmark pages earlier than others. Noll et al. called them discoverers and introduced the concept of discoverers into the HITS algorithm[6] to find experts and spammers among social bookmark users[7]. We use discoverers for approximation and define them as follows:

$$\{u_i|(u_i, p_i, t_i, \tau_i), \tau_i \leq \tau_{i+1}, i = 1, \cdots, n\} \tag{10}$$

where (u_i, p_i, t_i, τ_i) means that user u_i bookmarked page p_i with tags t_i at time τ_i and n is a parameter. We approximate the similarity function (9) using only discoverers to calculate the numerator. We define the approximate similarity function as follows:

$$Sim_n(p_1, p_2) = \frac{\sum_{(d,t) \in UT(p_1,n) \cap UT(p_2)} sp(d,t)}{|UT(p_1) \cup UT(p_2)|} \tag{11}$$

where $UT(p_1, n)$ is a set of user-tag pairs whose users are discoverers and d denotes a discoverer. $|UT(p_1, n)| \leq |UT(p_1)|$ and $sp(d, t) \geq 0$ are always satisfied. Hence, the following formula is also satisfied.

$$Sim_n(p_1, p_2) \leq Sim(p_1, p_2) \tag{12}$$

We use $Sim_n(p_1, p_2)$ instead of $Sim(p_1, p_2)$.

The candidate pages related to the target pages are obtained using those bookmarked by users who bookmarked the target page with the same tags. When we want to obtain the related page candidates of pages bookmarked by many users, however, it requires a large amount of time to obtain the candidates. To avoid this problem, we consider only pages that are bookmarked by the discoverers as the candidates. We developed an algorithm for collecting the candidates of the related pages. If the discoverer d bookmarks page p with the tag t, we collect pages $Pages(d, t)$ bookmarked by d with t. Then, we calculate $sp(d, t)$ of each discoverer d with tag t. We show the pseudo-code as follows:

```
for all (d,t) ∈ UT(p,n) do
  if sp_l(d,t), sp_g(t) is undefined then
    calculate sp_l(d,t) and sp_g(t).
  end if
  for all p' ∈ Pages(d,t) do
    if sim[p'] is undefined then
      sim[p'] ← sp_l(d,t) × sp_g(t)
    else
      sim[p'] ← sim[p'] +sp_l(d,t) × sp_g(t)
    end if
  end for
end for
for all p' in sim do
  sim[p'] ← sim[p'] / |UT(p) ∪ UT(p')|)
end for
```

sim[p'] is an array to reserve the similarity between the pages p and p'.

5 Evaluation

5.1 Evaluation on Specificity

We used Livedoor clip data[1] in December 2008 for the experiments. We found the URLs listed in Table 2 from pages bookmarked by discoverers. The discoverers are defined by formula (10) with $n = 10$. We list the URLs, the number of social bookmarks, and the number of the related page candidates in Table 2. We selected ten pages whose similarity with the target pages were highest from our algorithm using each $sp(d, t)$ function. Pages we could not visit are removed from the experimental targets in advance. All of the related page candidates were sorted in ascending order of their URLs and were presented to the three volunteers who did not know which function each page derived from. The volunteers rated them using the following standard.

- 3: almost the same topic as the target page
- 2: deeply related topic with the target page
- 1: related topic with the target page
- 0: unrelated topic with the target page

We evaluate the similarity ranking using the average of discount cumulated gain(DCG)[8], which is defined as follows:

$$DCG[i] = \begin{cases} G[i], \text{if } i = 1 \\ DCG[i-1] + G[i]/\log i, \text{otherwise} \end{cases} \tag{13}$$

where i is the rank in the similarity ranking and $G[i]$ is the average score rated by the volunteers.

The results are listed in Table 3. The scores in bold are the highest scores for each target page. If $sp(d, t) = 1$, then the similarity function equals the Jaccard coefficient of user-tag pairs. We regard this as the baseline. From the results, the average DCG when only global specificity is used ($sp(d, t) = sp_g(t)$) is highest, and the average DCG when local and global specificity are used ($sp(d, t) = sp_l(d, t) \times sp_g(t)$) is second highest. On the other hand, the average DCG of the case when only local specificity is used ($sp(d, t) = sp_l(d, t)$) is lower than that of the baseline. We also counted the number of URLs whose DCG was higher than the baseline. The number increased the most when only global specificity is used or global and local specificities are used. The number was 8 out of 10. The number when only local specificity was used was 5 out of 10. We discuss the reason local specificity does not contribute to a high DCG. In URL2, the DCG score of local specificity is very low. Therefore, we analyzed the URL2. Most of the tags used to bookmark URL2 were related to movies such as "Movie" but some users used tags related to April Fool. They bookmarked URL2 on April 1. In April 1, this page might have contained contents related to April Fool. On the other hand, we found that most of the related pages of URL2 derived from local specificity contained joke or parody related to April Fool and

[1] `http://clip.livedoor.com/`

Table 2. Pages used for Experiments

ID	URL	#SBM	candidates
URL1	http://codezine.jp/	47	56
URL2	http://eiga.com/	12	28
URL3	http://javascriptist.net/	188	292
URL4	http://lifehacking.jp/2008/03/life-instructions/	45	265
URL5	http://otoko-cooking.com/index.html	77	32
URL6	http://staff.aist.go.jp/toru-nakata/sotsuron.html	299	45
URL7	http://www.asahi.com/	84	54
URL8	http://www.hereticanthem.com/webdesign/295/	144	562
URL9	http://www.iknow.co.jp/	112	18
URL10	http://www.uta-net.com/	22	11

Table 3. Experimental Results for Specificity

$sp(d,t)$	$sp_l(d,t) \times sp_g(t)$	$sp_l(d,t)$	$sp_g(t)$	1
URL1	**5.74**	4.84	**5.74**	5.01
URL2	**6.64**	0.05	**6.64**	0.79
URL3	5.74	**6.21**	6.03	5.95
URL4	6.45	5.97	**6.50**	5.55
URL5	7.78	7.78	7.78	**8.59**
URL6	**4.31**	2.28	**4.31**	3.59
URL7	8.51	**9.01**	7.93	8.45
URL8	**6.69**	2.91	**6.69**	6.66
URL9	**6.66**	6.57	6.58	4.08
URL10	6.99	**7.05**	6.99	6.45
AVG.	6.47	4.97	**6.54**	5.18
# of improved	**8**	5	**8**	-

were published in April 1. Thus, certain users who have different tendencies in tagging can easily affect local specificity. To solve this problem, we need to consider the meaning of the tag.

In addition, the DCG scores were the same in 6 out of 10 tasks between when only global specificity was used when global and local specificities are used. This is because the narrow range distribution of the value of $sp_l(d,t)$ and $sp_l(d,t)$ does not have much effect on the value of $sp(d,t) = sp_l(d,t) \times sp_g(t)$. Hence, we need to analyze the differences in the usage of tags between users and the design of the local specificity function.

5.2 Evaluation of Relative Quantity of Web Pages

To evaluate the RQS as a measure of the attention intensity to Web pages, we use the attention degree of Web pages, which we define as **the attention intensity a user feels when he/she knows the number of social bookmarks**

Table 4. Experimental Results for RQS

ID	#SBM	RQS	User
URL1	47	2.749	0
URL2	12	0.381	-1
URL3	188	20.889	1
URL4	45	1.282	0
URL5	77	4.583	1
URL6	299	25.556	1
URL7	84	3.000	0.667
URL8	144	4.816	1
URL9	112	2.363	0.667
URL10	22	1.424	0
Spearman	0.877	0.922	-

of the target page and its related pages. If the RQS has a strong correlation with the attention degree, the RQS is useful as a measure of the attention intensity to Web pages. The values of the attention degree are obtained from volunteers ratings. Considering the title and the number of social bookmarks of the target pages and those of its related pages when $sp(d,t) = sp_l(d,t) \times sp_g(t)$, the volunteers rated the pages using the following standards:

- 1: Attention degree is high
- 0: Attention degree is medium
- -1: Attention degree is low

If all pages of a blog site have the same title, we use the title of each blog entry instead of the page title. Three volunteers rated each page listed in Table 2 and we used the average of these scores as the attention degree.

If the RQS is useful for measuring the attention intensity to Web page, the Spearman rank correlation coefficient between it and the attention degree should be high. Therefore, we evaluated the RQS using the Spearman rank correlation coefficient between attention degree. To evaluate the RQS, we compared it with the number of social bookmarks. We used Spearman rank correlation coefficient between it and the attention degree as the baseline. The results are shown in Table 4. In Table4, #SBM means the number of social bookmarks and User means the attention degree obtained from user evaluation. Spearman means the Spearman rank correlation coefficient against the attention degree. We found that RQS has a stronger correlation with attention degree than the number of social bookmarks. Therefore, the RQS is more useful than the number of social bookmarks for estimating the attention degree.

6 Conclusions

We proposed a similarity measure based on the specificity of social tags and a method for obtaining related pages using the measure. We defined local and

global specificity and evaluated their effectiveness and the effectiveness of their combination. From the evaluation, we found that global specificity improves the similarity measure but local specificity sometimes makes it worse. One of the reasons seems that certain users who have different tendencies in tagging can easily affect the local specificity. We also define the RQS to measure the attention intensity to the Web pages using related pages. We compared the Spearman rank coefficient between the RQS and attention degree and the one between the number of social bookmarks and attention degree. We found that the Spearman rank coefficient between the RQS and attention degree is higher. For future work, we need to improve local specificity.

Acknowledgment

This work was supported in part by the National Institute of Information and Communications Technology.

References

1. Krestel, R., Chen, L.: Using co-occurence of tags and resources to identify spammers. In: ECML/PKDD Discovery Challenge (RSDC 2008), Workshop at ECML/PKDD 2008 (2008)
2. Niwa, S., Doi, T., Honiden, S.: Web page recommender system based on folksonomy mining. In: ITNG 2006: Third International Conference on Information Technology New Generations, pp. 388–393 (2006)
3. Sugiyama, N., Seki, Y., Aono, M.: A method for finding related pages by users' tagging behavior from social bookmarks (in Japanese). Journal of the DBSJ 7(1), 239–244 (2008)
4. Liang, H., Xu, Y., Li, Y., Nayak, R.: Collaborative filtering recommender systems based on popular tags. In: ADCS 2009: Proceedings of the Fourteenth Australasian Document Computing Symposium (2009)
5. Chi, E.H., Mytkowicz, T.: Understanding the efficiency of social tagging systems using information theory. In: HT 2008: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, pp. 81–88. ACM, New York (2008)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (1999)
7. Noll, M.G., Au Yeung, C.M., Gibbins, N., Meinel, C., Shadbolt, N.: Telling experts from spammers: expertise ranking in folksonomies. In: SIGIR 2009: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 612–619. ACM, New York (2009)
8. Järvelin, K., Kekäläinen, J.: Ir evaluation methods for retrieving highly relevant documents. In: SIGIR 2000: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 41–48. ACM, New York (2000)