Predicting Social Annotation by Spreading Activation

Abon Chen[1], Hsin-Hsi Chen[2,*], and Polly Huang[1]

[1] Department of Electrical Engineering
[2] Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
{r94921033,hhchen}@ntu.edu.tw, phuang@cc.ee.ntu.edu.tw

Abstract. Social bookmark services like *del.icio.us* enable easy annotation for users to organize their resources. Collaborative tagging provides useful index for information retrieval. However, lack of sufficient tags for the developing documents, in particular for new arrivals, hides important documents from being retrieved at the earlier stages. This paper proposes a spreading activation approach to predict social annotation based on document contents and users' tagging records. Total 28,792 mature documents selected from *del.icio.us* are taken as answer keys. The experimental results show that this approach predicts 71.28% of a 100 users' tag set with only 5 users' tagging records, and 84.76% of a 13-month tag set with only 1-month tagging record under the precision rates of 82.43% and 89.67%, respectively.

Keywords: Collaborative Tagging, Social Annotation, Spreading Activation.

1 Introduction

Collaborative tagging is a very common application on the web. When reading interesting documents, web users are often willing to share their understanding of the documents with others by social annotation. Social bookmark tools [1] facilitate flexible information organization. Consider a social bookmark service *del.icio.us* as an example. It provides online resource organization tools for users, and works as an online collection, just like "my favorites" in a local browser on our own device. Rather than keeping "my favorite" on a local device, it also makes "my favorite" reachable when we surf on Internet. At the same time, we can input our own short description and tags for the resources collected as shown in Figure 1. Initially, tags are only for one's convenience. With the public sharing, users are able to discover and tag their own collection by browsing others.

A resource named an URL receives more and more tags when it is bookmarked by multiple users. Surprisingly, the freedom of annotation does not drive to chaos. Instead, the tag distribution shows a sense of consensus over time, and the stability reveals the collaborative behavior of users [2]. The visualization web service [3] demonstrates the tagging activity in time. The results of social annotation can be

* Corresponding author.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 277–286, 2007.

Fig. 1. Scenario of social annotation

employed to social network analysis [4], semantic web construction [5], enterprise search [6][7], and so on.

Annotation may facilitate recommendation and effective retrieval. However, not all resources can gain the benefits from that. Ill-tagged period of URLs prevents them from being retrieved. The retrieval performance for new-coming URLs degrades inevitably. Thus, how to predict a quality tagging set for a resource is an important issue. Indexing in traditional information retrieval [8] captures content of documents for effective retrieval. It focuses on document contents only. This paper will consider tagging records of users as additional cues. In information retrieval, spreading activation methods have been used to expand search vocabulary and complement the retrieved document [9]. These papers [10][11] adopt this methodology to select useful concepts from outside resources like WordNet and ConceptNet for query expansion. Here, we will employ it to model tag recommendation from users' tagging records.

This paper is organized as follows. Section 2 proposes our methods. Baseline and two alternatives of spreading activation are specified. Section 3 introduces the test material and discusses the experimental results. Section 4 concludes the remarks.

2 Tag Prediction

Given an URL denoting a document and the tagging records T_1, T_2, …, T_h posted by h users, tag prediction aims to recommend suitable tags for this URL. Two possible tasks may be done depending on whether the tagging records are available or not.

(1) Initial tagging: traditional indexing.
(2) History-based tagging: spreading activation.

In the initial tag assignment, terms of larger weights are selected from URL address, document content, outgoing link address and content of the outgoing link, and are regarded as *recommended tags*. Traditional *tf-idf* scheme may be adopted to compute the weight of each term.

In the history-based tag assignments, spreading activation triggered by tagging records is employed. The concept of spreading activation can be explained by a natural phenomenon. When we drop a stone in a pond, oscillation on surface transfers energy to neighborhood, and becomes smaller and smaller in amplitude due to water resistance. In this model, we can imagine a *posted tag* by a user as a stone. Its energy propagates from the most related tags to less relevant ones. A tag has an energy level indicating its relatedness to the posted tag. In general, a user may post more than one tag in a tagging record. In this way, a tag may receive energy contributed from posted tags through different paths. Tags of higher energy are selected and recommended.

In spreading activation, tags are linked as a network. Two tags are linked when they have an association. The degree of the tag association is measured by a weight. A tag t_i may have n outgoing links to tags $t_{i1}, t_{i2}, \ldots, t_{in}$ with weights $w_{i1}, w_{i2}, \ldots, w_{in}$. Assume e_i is an energy level of tag t_i. During spreading activation, t_i will keep portion of energy, say, $\alpha \cdot e_i$ (where α is a decay factor, e.g., 0.8 in our experiments). The fraction of energy, $(1-\alpha) \cdot e_i$, is distributed to the neighbor tags based on their weights. For example, t_{ij} will receive the amount of energy, $(1-\alpha) \cdot e_i \cdot w_{ij}/(w_{i1}+w_{i2}+\ldots+w_{in})$. A tag may have more than one incoming links, so that it may receive energy from different neighbors. We can use mutual information to compute the association of two tags.

An energy spreading matrix $[M]_{n \times n}$ shown as follows defines how much energy is distributed in a spreading activation cycle.

$m_{ij} = \alpha$ when $i=j$;
$m_{ij} = (1-\alpha) \cdot w_{ij}/(w_{i1}+w_{i2}+\ldots+w_{in})$ when $i \neq j$.

Assume $E^i=[e^i_1, e^i_2, \ldots, e^i_n]$ denotes a vector of energy levels $e^i_1, e^i_2, \ldots, e^i_n$ for tags $t_1, t_2, \ldots, t_n$ after i-th user's annotation, but before spreading activation. After one cycle of spreading activation, the new energy vector $E^{i*}=[e^{i*}_1, e^{i*}_2, \ldots, e^{i*}_n]$ is computed by using the formula $E^{i*} = E^i \times M$.

During initial tagging, tags are assigned initial energy E^0. Assume a tagging record $T_i=(b_1, b_2, \ldots, b_n)$ is an n-tuple binary vector. Here b_j is set to 1 when tag t_j is in the i-th user's annotation. In this way, $E^i = E^{(i-1)'} + T_i$, where $E^{(i-1)'}$ denotes the energy vector after $(i-1)$-th automatic tagging.

Spreading activation is triggered by the posted tags, and propagates energy to the relevant tags. Two strategies, called *SA* and *ET*, are shown as follows to control the propagation.

(1) *SA*: Propagation cannot out of a specific number of cycles, e.g., 3.
(2) *ET*: When the energy through a link is below a given threshold, e.g., 0.1, it is too low to be propagated.

After spreading activation, the final energy is stored in each tag. Tags are sorted by the energy, and top-p tags of higher energy are recommended.

Figure 2 illustrates the process of spreading activation by an example. Figure 2(a) shows part of the initial tagging. The arrival of the first user's tagging is depicted by the dark dots in Figure 2(b). The spreading activation goes on by propagating the energy with probability shown in Figure 2(c). The 1st propagation result is specified by Figure 2 (d).

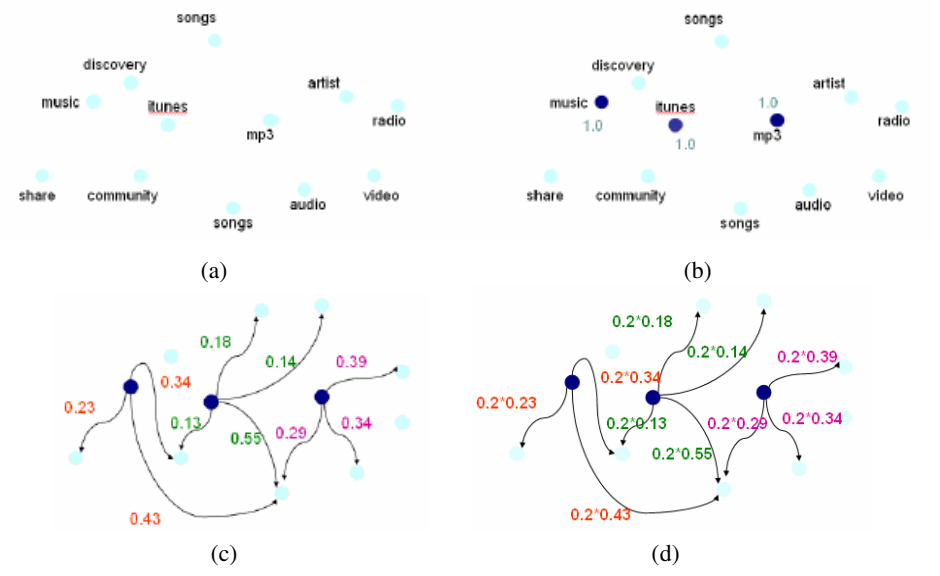

Fig. 2. Illustration of spreading activation process

Algorithms 1 and 2 show two possible implementation of the tag prediction based on these two strategies, respectively.

Algorithm 1. Spreading Activation Method with Limited Cycles

h {total tagging records}
c {maximum number of propagation cycles}
p {total recommendation tags}
n {total number of tags}
E^i {a vector of energy levels for tags}
M {an $n \times n$ energy spreading matrix}
$i = 1$
$E^0 = InitialTagging()$
while $i <= h$ **do**
    $E^i = E^{(i-1)} + T_i$
    $E^i = E^i \times M$
    $j = 1$
    while $j < c$ **do**
        $E^i = E^i \times M$
        $j = j + 1$
    end while
$i = i + 1$
end while
sort the n tags in the descending order of their energy in E^h
return top-p tags

Algorithm 2. Spreading Activation Method with Energy Threshold
h {total tagging records}
t {energy threshold}
n {total number of tags}
E^i {a vector of energy levels for tags}
M {an $n \times n$ energy spreading matrix}
$i = 1$
$E^0 = InitialTagging()$
while $i <= h$ **do**
    $E^i = E^{(i-1)} + T_i$
    $E^i = E^i \times M$
    $block = 0$
    repeat
    $j = 1$
    while $j <= n$ **do**
        if $E^i \times j$-th column of $M > t$
            then $e^i_j = E^i \times j$-th column of M
            else $block = 1$
        $j = j + 1$
    end while
    until $block = 1$
$i = i + 1$
end while
sort the n tags in the descending order of their energy in E^h
return top-p tags

3 Results and Discussion

3.1 Experimental Material

We collected a sample of *del.icio.us* data by crawling its popular feed every 30 minutes during March 27 and April 19, 2007. The data set consists of 2,475,999 taggings made by 10,109 different users on 31,025 different URLs with 125,092 different tags. For evaluation, we extract the mature URLs from the gathered data set by the criteria [5][12], i.e., (1) a mature URL should have its tag distribution remaining stable, and (2) a mature URL should have enough amount of tags applied by users. In this way, we have 28,792 mature URLs for experiments.

3.2 Performance Evaluation

For each URL in the test set, we have its i-th user's tagging record as input at the corresponding suggestion stage. The mature tag set is considered as answer keys in the evaluation. The tags are said to be correctly recommended when they are also listed in the mature tag set. Conventional recall rate and precision rate are adopted to measure the coverage and the quality of recommended tag set. Recall rate is the

number of tags correctly recommended divided by total mature tags. Precision rate is the number of tags correctly recommended divided by total recommended tags.

Figure 3 and Figure 4 show the recall rate and the precision rate of the proposed methods after the 1st, 2nd, 3rd, 4th, and 5th user tagging records have been read. SA_i denotes energy spreads at most i cycles. *ET* means spreading activation controlled by energy threshold. That is, it stops spreading when energy being propagated below a threshold. In the experiments, 0.1 is adopted.

The recall rate of the baseline system stays around 10% even tagging record grows. This is because the tag set proposed by the baseline system is just the union of the initial recommendation and the tagging records collected up to now. Comparatively, the recall rate of all the spreading activation methods improves steadily. *ET* strategy is better than *SA* strategy. With *SA* strategy, even though there is still enough energy for propagation, the spreading is stopped due to the restriction of maximum cycles. Spreading at most 3 cycles outperforms the other four *SA* methods.

Because the baseline system only conservatively includes the tags provided by the users, i.e., it performs without any expansion, its precision rate is higher than its recall rate trivially. The precision rates of *ET* are the best of all the methods. Both the precision rate and the recall rate increase when more user tagging records are posted.

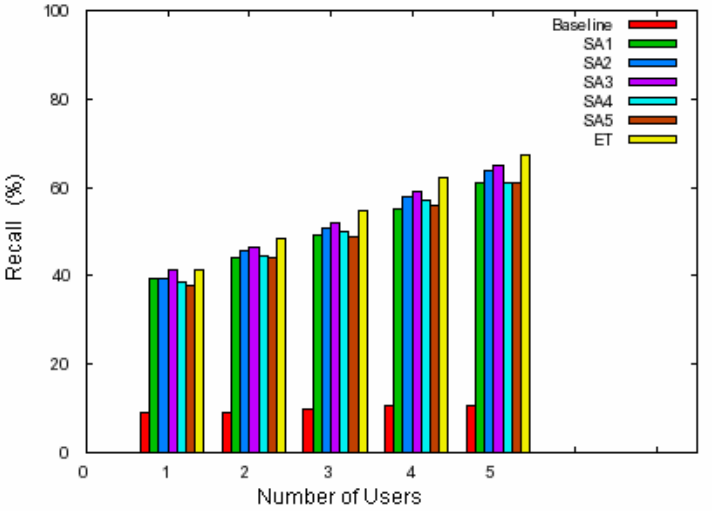

Fig. 3. Comparison of baseline and spreading activation methods from recall perspective

3.3 Coverage over Users and Time

This section discusses how many efforts our system saves from two aspects, i.e., users and time. The spreading activation with energy threshold strategy is the best, so that it is adopted in the latter experiments. In Figures 5 and 6, we assume the mature tag set is achieved when 100 users are involved in social annotation. The upper line and the lower line in Figure 5 show the recall rates of the tag prediction and the social

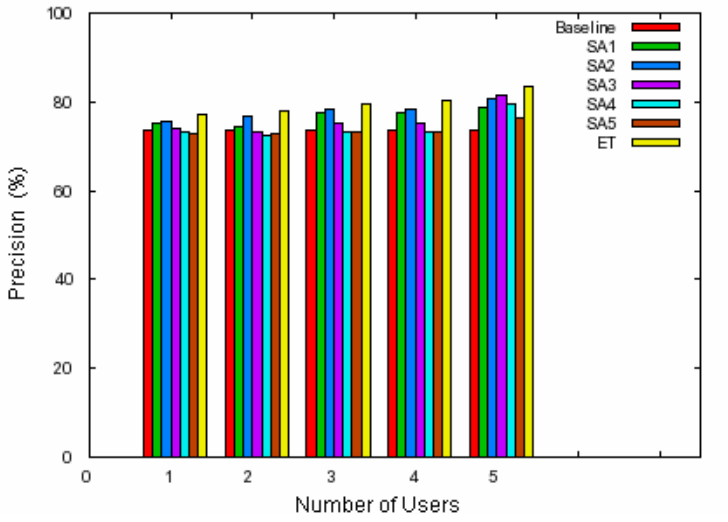

Fig. 4. Comparison of baseline and spreading activation methods from precision perspective

annotation, respectively. The automatic annotation method catches up to the mature tag set much faster than the manual annotation only. For example, the tag records of the first 5 users occupy 20.48% of the mature tag set. In contrast, the spreading activation method can achieve 71.28% of the mature tag set.

The precision rates of manual tagging in Figure 6 are 100%. The precision rates of automatic tagging are also very high, i.e., from 82.43%, 89.67%, …, to 93.42%, under different number of user involvements. That confirms the quality of the recommended tags. From user perspective, the tag prediction method saves 75% of human cost under the recall rate of 71.28% and precision rate of 82.43%.

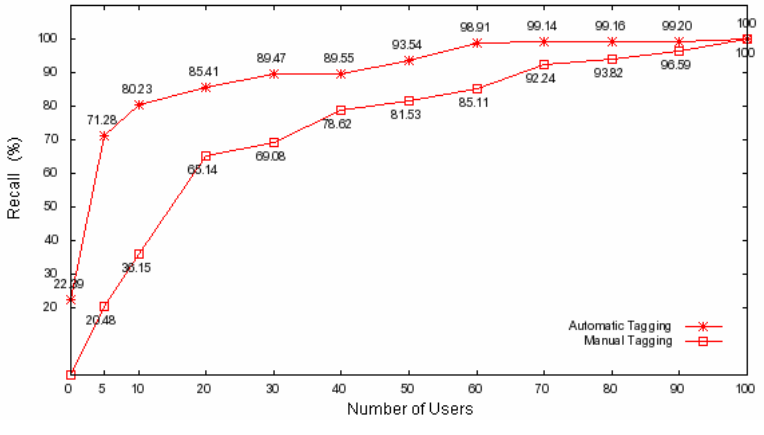

Fig. 5. Recall rate of tagging from user perspective

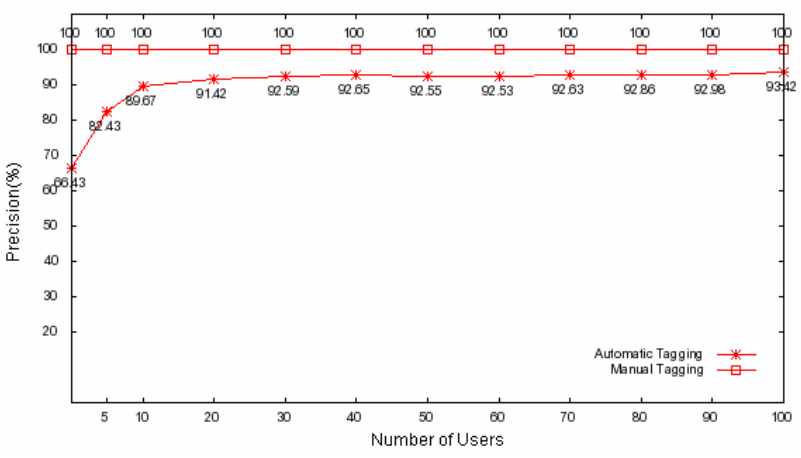

Fig. 6. Precision rate of tagging from user perspective

Figure 7 and Figure 8 show the coverage and the quality of the tag set from the time aspect. Here the developed tag set after 13 months is regarded as mature. The upper line and the lower line of Figure 7 denote the recall rates of the automatic annotation and the manual annotation, respectively. In the first 0.2 month, the corresponding coverage is 70.28% and 20.48%, respectively. After 1 month, the coverage of the spreading activation method increases to 84.76%. It means 12 months can be saved under the coverage of 84.76%. The precision rates shown in Figure 8 ensure the quality of the recommended tags. They are more than 90% after 1-month social annotation.

In summary, the spreading activation method recommends high quality tags without much delay. That makes resources searchable at the earlier stage.

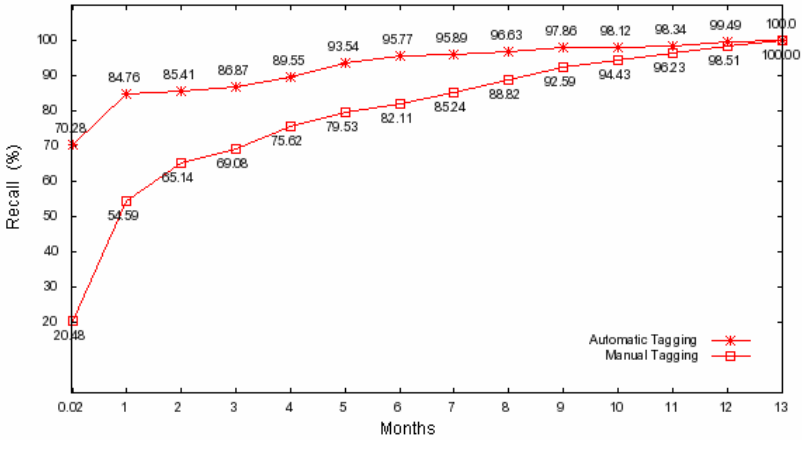

Fig. 7. Recall rate of tagging from time perspective

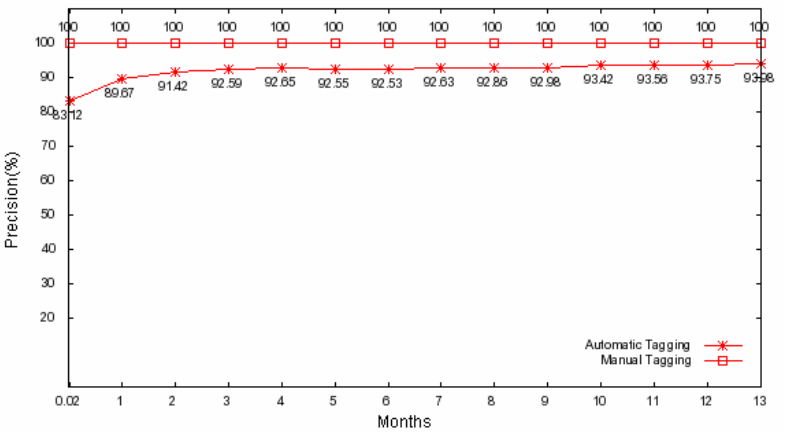

Fig. 8. Precision rate of tagging from time perspective

4 Concluding Remarks

In this paper, a spreading activation method is proposed to predict the tag set of a mature URL based on document content and users' tagging records. The strategies of limited cycles and energy thresholds are explored. The experimental results show that this approach with energy threshold predicts 71.28% of a 100 users' tag set with only 5 users' tagging records, and 84.76% of a 13-month tag set with 1-month tagging record under the precision rates of 82.43% and 89.67%, respectively. Users will benefit from the retrieval performance enhanced by sufficient tags a lot earlier. Currently, only contents of resources and annotation histories are considered. We will investigate more cues like the categorization of resources and the link relationships among resources to predict the social annotation in the future.

Acknowledgments. Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC 96-2752-E-001-001-PAE.

References

1. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social Bookmarking Tools (I): A General Review. D-Lib Magazine 11(4) (2005)
2. Golder, S.A., Huberman, B.A.: Usage Patterns of Collaborative Tagging Systems. Journal of Information Science. 32(2), 198–208 (2006)
3. Russell, T.: Cloudalicious: Folksonomy over Time. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 364–364. ACM Press, New York (2006)
4. Mika, P.: Ontologies are Us: A Unified Model of Social Networks and Semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)
5. Wu, X., Zhang, L., Yu, Y.: Exploring Social Annotations for the Semantic Web. In: Proceedings of the 15th International Conference on World Wide Web, pp. 417–426. ACM Press, New York (2006)

6. Dmitriev, P.A., Eiron, N., Fontoura, M., Shekita, E.: Using Annotations in Enterprise Search. In: Proceedings of the 15th International Conference on World Wide Web, pp. 811–817. ACM Press, New York (2006)
7. Hotho, A., Jaschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
8. Baeza-Yates, R., Riberiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
9. Salton, G., Buckley, C.: On the Use of Spreading Activation Methods in Automatic Information Retrieval. In: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 147–160. ACM Press, New York (1988)
10. Hsu, M.H., Chen, H.H.: Information Retrieval with Commonsense Knowledge. In: Proceedings of 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval, pp. 651–652. ACM Press, New York (2006)
11. Hsu, M.H., Tsai, M.F, Chen, H.H.: Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006. LNCS, vol. 4182, pp. 1–13. Springer, Heidelberg (2006)
12. Halpin, H., Robu, V., Shepherd, H.: The Complex Dynamics of Collaborative Tagging. In: Proceedings of the 16th International Conference on World Wide Web, pp. 211–220. ACM Press, New York (2007)