A Weighting Scheme for Tag Recommendation in Social Bookmarking Systems

Sanghun Ju and Kyu-Baek Hwang

School of Computing, Soongsil University, Seoul 156-743, Korea
shju@ml.ssu.ac.kr, kbhwang@ssu.ac.kr

Abstract. Social bookmarking is an effective way for sharing knowledge about a vast amount of resources on the World Wide Web. In many social bookmarking systems, users bookmark Web resources with a set of informal tags which they think are appropriate for describing them. Hence, automatic tag recommendation for social bookmarking systems could facilitate and boost the annotation process. For the tag recommendation task, we exploited three kinds of information sources, i.e., resource descriptions, previously annotated tags on the same resource, and previously annotated tags by the same person. A filtering method for removing inappropriate candidates and a weighting scheme for combining information from multiple sources were devised and deployed for ECML PKDD Discovery Challenge 2009. F-measure values of the proposed approach are 0.17975 for task #1 and 0.32039 for task #2, respectively.

Keywords: social bookmarking, folksonomy, tag recommendation

1 Introduction

Social bookmarking systems such as BibSonomy[1] and Delicious[2] have increasingly been used for sharing bookmarking information on the Web resource. Such systems are generally built on a set of collectively-annotated informal tags, comprising a folksonomy. A tag recommendation system could guide users during the bookmarking procedure by providing a suitable set of tags for a given resource. In this paper, we propose a simple but effective approach for tackling the tag recommendation problem. The gist of our method is to appropriately combine different information sources with pre-elimination of barely-used tags.

The candidate tags for recommendation can be extracted from the following information sources. First, resources themselves may have the annotated tags. For example, the title of a journal article is likely to include some of the annotated keywords. Second, the tags previously annotated by other users for the same resource

[1] http://www.bibsonomy.org/
[2] formerly del.icio.us, http://delicious.com/

could be a good candidate set. Third, previously annotated tags for other resources by the same user could also provide some information.[3]

The paper is organized as follows. In Section 2, the proposed tag recommendation method is detailed. Then, Section 3 shows the results of experimental evaluation on the training dataset, confirming the effectiveness of the proposed method. Performance of our method on the test dataset is briefly described in Section 4. Finally, concluding remarks are drawn in Section 5.

2 The Method

In this section, we detail the proposed tag recommendation method. First, the procedure for keyword extraction from resource descriptions with importance estimation and filtering is explained. Then, the keyword extraction and importance estimation method from previously annotated information is described. Finally, tag recommendation by combining multiple information sources is explained.

2.1 Keyword Extraction from Documents (Resource Descriptions)

In our approach, candidate keywords are extracted from the columns *url*, *description*, and *extended description* of the table *bookmark* as well as the columns *journal*, *booktitle*, *description*, and *title* of the table *bibtex*. It should be noted here that the candidates extracted from different fields are processed separately. This means that even the same keywords could have multiple importance values according to the columns from which they are extracted.[4]

In order to estimate the importance of each keyword, its accuracy and frequency ratios are calculated as follows.

D: set of all documents (resources) such as bookmarks or BibTex references.
EC(k, d): extraction count of keyword k in document d.
MC(k, d): matching count of keyword k with one of the tags of document d.[5]
TEC(d): extraction count of all the keywords in document d.

$$\text{Accuracy Ratio, AR}(k) = \sum_{d \in \mathbf{D}} \text{MC}(k, d) \,/\, \sum_{d \in \mathbf{D}} \text{EC}(k, d). \quad (1)$$

$$\text{Frequency Ratio, FR}(k) = \sum_{d \in \mathbf{D}} \text{EC}(k, d) \,/\, \sum_{d \in \mathbf{D}} \text{TEC}(d). \quad (2)$$

The accuracy and frequency ratios of each keyword are calculated across all the documents.

[3] [1] also exploited these kinds of information sources for tag recommendation. We extend this approach by extracting keywords from not only resource title but also other resource descriptions.

[4] It is because average importance values of keywords are different according to extracted columns.

[5] MC(k, d) is equal to EC(k, d) if d is tagged with k, 0 otherwise.

The keywords whose accuracy is lower than average are not considered for recommendation. This elimination procedure is implemented by the following criterion, which also penalizes frequent words.

TMC(d): sum of MC(k, d) across all the keywords in document d.

$$\text{Limit Condition: } \mathrm{AR}(k) / (1 + \mathrm{FR}(k)) > \sum_{d \in \mathbf{D}} \mathrm{TMC}(d) \; / \; \sum_{d \in \mathbf{D}} \mathrm{TEC}(d). \qquad (3)$$

Some keywords with high accuracy ratio values are shown in Table 1. It should be noted that there exist a large amount of keywords having high AR(k) values and the keywords in Table 1 are a sample from them.

Table 1. Example keywords with high accuracy ratios.

Keywords	Extracted columns	Accuracy ratio, AR(k)	Frequency ratio, FR(k)
nejm	*extended description*	1.0000	0.0002579
medscape	*extended description*	1.0000	0.0001146
freebox	*description*	1.0000	0.0000533
harum	*description*	0.9800	0.0000556
ldap	*url*	0.9354	0.0000403
shipyard	*description*	0.9146	0.0002734

The keywords in Table 1 have accuracy ratios much higher than the average, satisfying Equation (3). In Table 2, we present some keywords on the border with respect to Limit Condition.

Table 2. Example keywords on the border in terms of Limit Condition.

Keywords	Extracted columns	Accuracy ratio, AR(k)	Frequency ratio, FR(k)	Difference in Limit Condition	Limit Condition satisfied
netbib	*url*	0.0789	0.0002468	0.0004281	Yes
guide	*url*	0.0778	0.0006510	-0.0007060	No
media	*url*	0.0781	0.0008810	-0.0003974	No
daily	*extended description*	0.0602	0.0002744	0.0005867	Yes
list	*extended description*	0.0601	0.0008056	0.0005053	Yes
engine	*extended description*	0.0590	0.0005598	-0.0006156	No
tool	*description*	0.1279	0.0007647	0.0006509	Yes
ontologies	*description*	0.1271	0.0001312	-0.0000749	No
corpus	*description*	0.1264	0.0000967	-0.0007337	No

The average AR(k) values in *url*, *description*, and *extended description* are 0.07849973, 0.12715830, and 0.05963773, respectively. We also show some example keywords with low accuracy ratios, which do not satisfy Equation (3) as follows.

url: org, co, ac, au, default, main, details, welcome.
description: feeds, economy, review, images, help.
extended description: a, have, that, one, other, are, person, its.

Finally, each extracted keyword, satisfying Equation (3), is stored in ***d-keyword set*** (**DS**). The accuracy weight of each candidate is calculated by multiplying its accuracy ratio and extraction count from the present document as follows.

$$\text{Accuracy Weight from Document Set, } AW_{DS}(k) = EC(k, d) \times AR(k). \quad (4)$$

The accuracy weight, $AW_{DS}(k)$, is calculated when recommending tags for a given document (resource) d.

2.2 Keyword Extraction from Previously-Annotated Information

Candidate keywords could be extracted from the previously annotated tags for the same resource. For the BibTex references, the field *simhash1* of the table *bibtex* is adopted for the semantically-same resource detection. For the bookmarks, a pruning function, which has similar effect of the approach used in [2], was implemented and deployed in our experiments. These candidate keywords are stored in ***r-keyword set*** (**RS**). Their accuracy weight is calculated as follows.

D: set of all documents (resources) satisfying the same document condition with the present document d.
$TC(k, d)$: 1 if document d has keyword k; 0 otherwise.

$$\text{Accuracy Weight from Resource Set, } AW_{RS}(k) = \sum_{d \in \mathbf{D}} TC(k, d). \quad (5)$$

Candidate keywords are also extracted from the previously annotated tags by the same person. These candidate keywords are stored in ***u-keyword set*** (**US**). Their accuracy weight is obtained as follows.

D: set of all documents (resources) which are previously tagged by user u.
$UC(k, d)$: 1 if document d has keyword k; 0 otherwise.

$$\text{Accuracy Weight from User Set, } AW_{US}(k) = \sum_{d \in \mathbf{D}} UC(k, d). \quad (6)$$

2.3 Tag Recommendation by Combining Multiple Information Sources

The last step is to recommend appropriate tags from the three candidate keyword sets, i.e., ***d-keyword set*** (**DS**), ***r-keyword set*** (**RS**), and ***u-keyword set*** (**US**). Given a specific user and a document (resource) for tagging, these three candidate keyword sets are specified with accuracy weight for each candidate. Before unifying these candidates, the accuracy weights are normalized into [0, 1] as follows.

$\mathbf{EK} = \mathbf{DS} \cup \mathbf{RS} \cup \mathbf{US}$ ($ek \in \mathbf{EK}$).
$NW_{DS}(ek) = AW_{DS}(ek) / \sum_{k \in \mathbf{DS}} AW_{DS}(k)$.
$NW_{RS}(ek) = AW_{RS}(ek) / \sum_{k \in \mathbf{RS}} AW_{RS}(k)$.
$NW_{US}(ek) = AW_{US}(ek) / \sum_{k \in \mathbf{US}} AW_{US}(k)$.

We also added tag frequency information, denoting how many times a tag was annotated during the training period. This tag frequency rate is calculated as follows.

$$TFR(ek) = \sum_{d \in \mathbf{D}} TagCount(ek, d) \; / \; \sum_{t \in \mathbf{T}} \sum_{d \in \mathbf{D}} TagCount(t, d); \; 0 \leq TFR(ek) \leq 1,$$

where TagCount(t, d) denotes the number of occurrences of a tag t annotated for a document d. **T** and **D** denote the set of all tags and the set of all documents (resources), respectively.

The above four factors are linearly combined with appropriate coefficients. We have experimented with different coefficient values, trying to obtain nearly optimal results. First, we focused on the fact that the performance of extracted keywords from ***d-keyword set*** (**DS**) is higher than that from ***r-keyword*** or ***u-keyword sets*** (**RS** or **US**). Figure 1 compares the performance using each keyword set on the training dataset when the number of recommended tags is five.

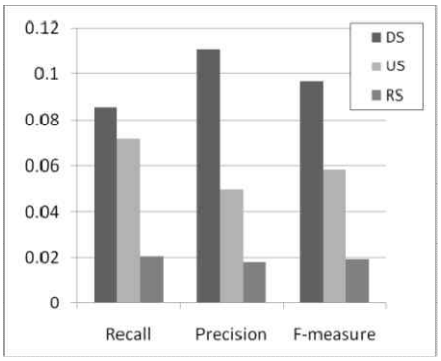

Figure 1. Performance comparison of extracted keywords from different information sources.

Accordingly, we tried high coefficient values on $NW_{DS}(ek)$ and relatively low coefficient values on $NW_{RS}(ek)$ and $NW_{US}(ek)$. However, this scheme does not produce better results than other schemes as shown in Figure 2.

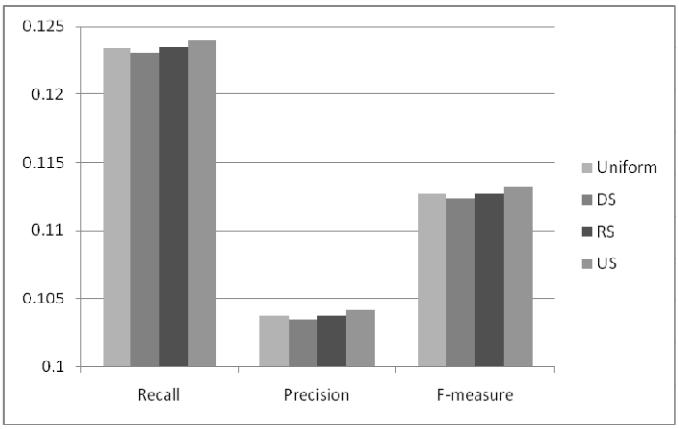

Figure 2. Performance comparison among different weighting schemes.

In Figure 2, Uniform denotes the case of assigning an equal coefficient (0.3) to each keyword set and DS (RS or US) denotes the case of assigning 0.45 to **DS** (**RS** or **US**) and 0.25 to the other keyword sets. TFR(*ek*) was assigned 0.1 or 0.05 in the above cases. On the contrary to our expectation, the weighting scheme assigning high coefficient value to **US** showed the best performance.

The reason for this phenomenon is not clear but one possible clue is that performance of the candidates extracted from **DS** varies much according to extracted data columns. Such keywords are illustrated in Table 3.

Table 3. Variation in accuracy ratio, AR(*k*), of the same keywords extracted from different data columns. Accuracy values higher than the average are represented in bold.

Keywords	*url*	*description*	*extended description*
portal	0.0439	0.1080	**0.1250**
tag	0.0410	0.1194	**0.1237**
tech	0.0318	0.0598	**0.1406**
template	0.0620	**0.1911**	**0.2023**
time	**0.1560**	0.0951	0.0322
youtube	**0.3217**	0.0877	0.0319

In Table 3, it is observed that even the same keyword from **DS** could have extremely different accuracy ratio values. For example, the keyword portal from *extended description* has much higher AR(*k*) value than the average, i.e., about 0.05964. However, the accuracy ratios of the same keyword from *url* or *description* are lower than the averages.

After several trials, we applied the following formula for the recommendation, which has shown fine results on the training dataset.

$$NW_{DS}(ek) \times .2 + NW_{RS}(ek) \times .35 + NW_{US}(ek) \times .4 + TFR(ek) \times .05. \quad (7)$$

3 Experimental Evaluation

To evaluate the proposed approach, we reserved the postings spanning the latest six months from the given training dataset like the real challenge. Hence, the training period is from January 1995 to June 2008 and the validation period is from July to December of 2008. The numbers of postings, resources, and users during these periods are shown in Tables 4 and 5.

Table 4. The Post-Core dataset size

	bookmark	*bibtex*	*tas*	# of users
Training	37037	17267	218682	982
Validation	4231	5585	34933	433

Table 5. The Cleaned Dump dataset size

	bookmark	*bibtex*	*tas*	# of users
Training	212373	122115	1101387	2689
Validation	50631	36809	299717	1292

3.1 Effectiveness of Candidate Elimination

In this subsection, we present the effect of our keyword elimination method (Equation (3)). Note that Limit Condition is applied to the candidate keywords whose accuracy ratio is lower than average with some penalizing effect on frequently-occurred keywords. Figures 3 and 4 show the effect of candidate elimination on the Post-Core and Cleaned Dump datasets, respectively. The results are obtained when the number of recommended tags is five. On the both validation datasets (i.e., Post-Core and Cleaned Dump), the proposed elimination method increases precision and F-measure values regardless of the number of recommended tags (from one to ten, although the results are not shown here). In the case of the Cleaned Dump dataset, recall is also improved by our filtering method.

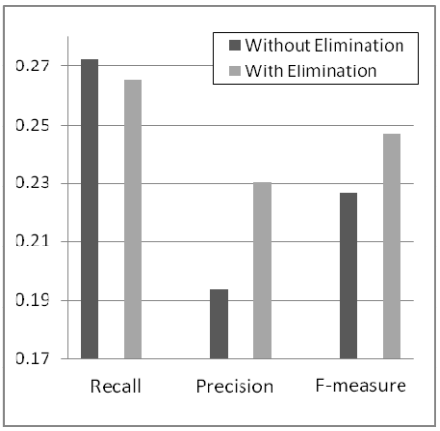

Figure 3. Effect of candidate elimination on the Post-Core dataset

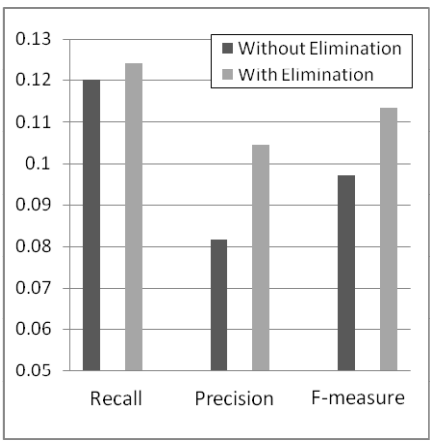

Figure 4. Effect of candidate elimination on the Cleaned Dump dataset

4 Final Results

Here, we append the final results of our method on the test dataset of ECML PKDD Discovery Challenge 2009.

Table 6. The test dataset size

	bookmark	*bibtex*	# of users
Cleaned Dump	16898	26104	1591
Post-Core	431	347	136

Table 7. Final results on the test dataset (Post-Core, Task #1)

# of tags	Recall	Precision	F-measure
1	0.074695721	0.243523557	0.114324737
2	0.121408237	0.213594717	0.154817489
3	0.152896044	0.193533634	0.170831363
4	0.175617505	0.179512968	0.177543872
5	**0.191311486**	**0.169508783**	**0.179751416**
6	0.203439061	0.162068819	0.180412681
7	0.213460494	0.156249045	0.180428119
8	0.22072531	0.151207336	0.179469517
9	0.227309809	0.147113534	0.178623208
10	0.232596191	0.143564862	0.177544378

Table 8. Final results on the test dataset (Cleaned Dump, Task #2)

# of tags	Recall	Precision	F-measure
1	0.142522512	0.42159383	0.213029148
2	0.241682971	0.367609254	0.291633121
3	0.315328224	0.331191088	0.323065052
4	0.367734647	0.295308483	0.327565903
5	**0.406172737**	**0.264524422**	**0.320390826**
6	0.443927734	0.242502142	0.313661833
7	0.47018359	0.221477537	0.301115964
8	0.49385481	0.204859836	0.289591798
9	0.509440246	0.190310422	0.27710381
10	0.520841594	0.176357265	0.263494978

5 Conclusion

We applied a simple weighting scheme for combining different information sources and a candidate filtering method for tag recommendation. The proposed filtering method was shown to improve precision and F-measure for the tag recommendation task in all the cases of our experiments. It has also shown to be effective for improving recall in some cases. Future works include finding more optimal scheme for combining multiple information sources. Evolutionary algorithms would be a suitable methodology for this task.

Acknowledgements

This work was supported in part by the Seoul Development Institute through Seoul R&BD Program (GS070167C093112) and in part by the Ministry of Culture, Sports and Tourism of Korea through CT R&D Program (20912050011098503004).

References

1. Marek Lipczak: Tag Recommendation for Folksonomies Oriented towards Individual Users. Proceedings of the ECML/PKDD 2008 Discovery Challenge Workshop, part of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2008)

2. Marta Tatu, Munirathnam Srikanth, and Thomas D'Silva: RSDC'08: Tag Recommendations using Bookmark Content. Proceedings of the ECML/PKDD 2008 Discovery Challenge Workshop, part of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2008)