*I.J. Education and Management Engineering* **2012, 6, 7-14**
Published Online July 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2012.06.02

***Available online at http://www.mecs-press.net/ijeme***

# Tag Recommendation Based on Collaborative Filtering and Text Similarity

[a] Chuanbao Wang, [b] Fang Yuan, [c] Ying Yun

*College of Mathematics and Computer Science, Hebei University Baoding, Hebei, China*

**Abstract**

In current social tagging system, users can freely add tags for the uploaded resources, which caused a problem that many tags could not describe the resource properly and even have some spelling errors. This problem may bring unnecessary troubles for other users who want to search this kind of resource. In this paper, a tag recommendation system based on collaborative filtering and text similarity is presented to solve the problem mentioned above. This system can automatically recommend some relevant tags for the new uploaded resources and thus the users can freely select tags from the system. Experimental results show that the recommended tags can effectively represent the contents of the webpages marked. Compared with the existing tag recommended methods, this method not only improves the accuracy of tags recommended, but also facilitates the webpage sharing and retrieval.

**Index Terms:** Component;Tagging system; Tag; Tag recommended; Webpage



## 1. Introduction

With the development of web 2.0, more and more people begin to upload their favorite articles, pictures and webpages to social networks for sharing with others, so social tagging systems are formed gradually. So how to find the favorite one from the numerous resources becomes an intractable problem. At present, the popular method is using tags for search. Tags can be understood as an object name, which are different from the keywords, they are open and non-hierarchical. They are added by the people who share the resources rather than made by experts in advance [1]. Tags have been applied in many bookmarks system, for example, Delicious, Flickr, YouTube and Furl etc. Because users who share resources can freely add tags, which caused that many tags may not describe the resource properly or even have some spelling errors, the quality of resources sharing is affected. We found some tags added by users from datasets, as shown in Fig. 1, we can see that some tags have spelling mistakes or use both singular and plural form or cannot express any meaning. These tags are not conducive to the management of these resources, nor useful for other users to search the required information.

In order to handle the disadvantage of the current adding tags method, an automatic tag recommendation method is proposed in this paper. The specific recommendation methods depend on the webpages uploaded.

Corresponding author:
E-mail address: [a] fzuwcb@126.com, [b] yuanfang@hbu.cn, [c] yunying_511@126.com

li, yao, service:email:list:archive,nyt, @lat, x, tk, ai, &, music, !read_later, books, book, ~.year:2005, tr, musik, hci, paloalto, ying_not_to_be_an_asshole, !, Israel, phil-wood, bugs, bug

Figure 1. Tags marked by users in webpages

## 2. Related works

The concept of web 2.0 further promotes social network, more and more people begin to study social tags. At present, many systems have generated the appropriate tags automatically for online resources. Reference [2] has proposed a collaborative tag recommendation approach based on HITS algorithm, it uses reward and punishment algorithm to select better tags. Reference [3] has proposed an approach that automatically recommends tags for weblogs. It adopts a hybrid artificial neural network to learn how to predict the best tag set by using the collective intelligence extracted from web 2.0 collaborative tagging as well as word semantics. Reference [4] proposed a content-based tag recommendation, the tags with higher weights in the higher similar webpages are recommended to the user by comparing the web page similarity.

Reference [2] referenced to the content-based tag recommendation, but not gave specific methods and experimental analysis. Reference [3] did not take into account the tag order in the tag recommendation. Reference [4] only considered the words which are the same as tags in the content-based tag recommendation, but some words are filtered out, which can better describe the contents of the pages.

In order to overcome the deficiencies described above we proposed a tag recommendation method based on collaborative filtering and text similarity. The method can automatically recommend tag to the webpages which are marked for the first time or many times. Not only considering the marking sequence of the tags, but also taking account of the words other than the tag set. Many of them can effectively describe the contents of the webpages, so the quality of the tag recommendation is improved.

## 3. Tag Recommendation System

The main function of tag recommendation system is to recommend related tags $t \in T$ for user $u \in U$ who uploads resources $r \in R$. There are many types of resources, including articles, pictures and webpages etc. The following are the definitions of some terms used in this paper: U is the set of users, P stands for the set of webpages, T represents the set of tags and W is the set of words.

The most important work of tag recommendation system is to select high-quality tags. As we all know, the description ability of different tags are different in describing contents of the pages. Aiming at this phenomenon, the different tag recommendation methods are used for the webpages marked for different times.

### *3. 1 Tag Recommendation Based on Collaborative Filtering*

For the webpages whose marking times is more than the threshold ε, the collaborative filtering method is used to recommend tags.

- Search for user u′ that is similar with user u and u≠u′, using a collaborative filtering method. We believe that a user is similar with the one who marked the same webpage. For each tag t marked by similar user u′, the weight is assigned to it according to the tag sequence. The weights are 1.0, 0.8, 0.6, 0.4, 0.2 and 0.1.When the tag sequence is more than 6, the tag weight is set to 0.1. For each tag, calculate the total weight, the formula is defined as Weight ($t_i$) =Freq ($t_i$) + Seq ($t_i$), where Freq ($t_i$) denotes the times that all users use tag $t_i$ to mark webpage p, Seq ($t_i$) denotes the sequence of tag $t_i$.

- The total weights of tags are sorted in descending order. The tags with the same root are combined, using porter stemming algorithm [5]. Finally, the tags with the top n largest weights are recommended to users.

### *3. 2 Tag Recommendation Based on Improved Text Similarity Computing*

For the webpages whose marking times is less than the threshold ε or are marked for the first time, we use the improved cosine similarity formula to search pages to recommend tags.

- The roles of each word are different in describing the contents of the webpages. Some words appear many times in a webpage, but cannot effectively describe it, so we should consider the word discrimination. The words with high discrimination are better in describing the contents of the webpages and representing the types of webpage. The words that only appear in a specific group of webpages are more important than those that appear in a wide range of webpages. The Gini value [6] proposed by Corrado Gini are used to evaluate a word discrimination. The less Gini value, the higher word discrimination. The formula WD(w) is defined as follows:

$$Pro(p,w)=\frac{F(p,w)}{\sum_{w'\in W}F(p,w')},p\in P,w\in W \quad (1)$$

$$WD(w)=Gini(w)=\left|1-\sum_{p\in P}Pro^2(p,w)\right| \quad (2)$$

Where F (p, w) denotes the frequency of word w appearing in the webpage p, Pro(r, w) denotes the proportion of word w to all the words in webpage p.

- In the social tagging system, users can use their favorite tags to freely mark webpages, resulting in that some tags are too specific, and even some tags have spelling errors. These tags cannot effectively describe the contents of webpages, so users cannot find their desired page through these tags. We believe that the pages including more words with higher discrimination are more important. The importance of webpage is defined as:

$$RS(p)=\sum_{w\in W}Pro(p,w)\times WD(w) \quad (3)$$

- Each webpage is represented in the form of vector model [7], such as a vector model: $X=[x_1,x_2,\ldots,x_n]^T$, $xi=[x_{i1},x_{i2},\ldots,x_{im}]$, (i=1,2,…,n), each row vector xi denotes the frequencies of the words in webpage $x_i$. Suppose that there are two row vectors $a=[a_1,a_2,a_3,\ldots,a_n]$, $b=[b_1,b_2,b_3,\ldots,b_n]$, well then the cosine similarity between vectors a and b is defined as follows [8]:

$$cosSim(a,b)=\frac{a\times b}{||a||\times||b||}=\frac{\sum_{i=1}^{n}(a_i\times b_i)}{\sqrt{\sum_{i=1}^{n}a_i^2}\times\sqrt{\sum_{j=1}^{n}b_j^2}} \quad (4)$$

In the calculation of similarity between webpage a and b, we considered the importance of webpages. The (4) is improved as follows:

$$Sim(a,b) = RS(a) \times RS(b) \times \cos Sim(a,b) \quad (5)$$

After the above calculation, we can get the similarity between any two pages. When users upload a new webpage or a webpage which is less marked, firstly, we find the webpages whose marking times is more than the threshold £ and the similarity between those and the current webpage is higher than the threshold δ. Secondly, the total weights of each tag in these pages are calculated using this formula Weight ($t_i$) =Freq ($t_i$) +Seq ($t_i$). The total weights of tags are sorted in descending order and the tags with the same root are combined using Porter Stemming algorithm. Finally, the tags with the top n largest weights are recommended to users.

### *3. 3 Text Similarity Computing GC Algorithm*

Input: designated website
Output: n tags
Algorithm steps:
Begin:
Query the number m of users who have marked the input webpage
If (m > ε) / / ε is the threshold
Add the weight and use frequency of the tag, obtaining each tag's total weight, and the words with the same root are combined into one word. The total weights of tags are sorted in descending order
Return the top n tags with largest total weights
Else
Based on improved the cosine similarity formula, find the webpages whose marking times is more than the threshold £ and the similarity between the those and the current webpage is higher than the threshold δ

1) Extract contents of web, and transform words into lowercase
2) Remove some useless symbols, such as question marks and brackets
3) Use stopwords to filter out webpage contents
4) Calculate the frequency of each word, word probability and Gini coefficient in the corresponding webpage
5) Calculate the importance of each webpage
6) Use (5) to calculates similarity between the webpages and the input one and then save webpages whose marking times is more than the threshold £ and the similarity is higher than the threshold δ
7) Obtain each tag's total weight. The words with the same root are combined into one word and the total weights of tags are sorted in descending order

Return the top n tags with largest total weights
End if
End

## 4. Experimental analysis

### *4. 1 Data sets*

This paper adopts the published datasets by the social bookmark system del.icio.us [9] to evaluate the algorithm performance. This datasets includes 3116 webpages, 22181 tags and 20640 users, and [User], [URL] and [Tag] are the three main components of the website. As shown in Fig. 2, through this page, users can clearly understand the webpage list belonging to the tag [computer] and all the users who have used this tag. After preprocessing, we get 2522 webpages, among these 1123 webpages are used for collaborative filtering and the other 1399 for the text similarity computing.

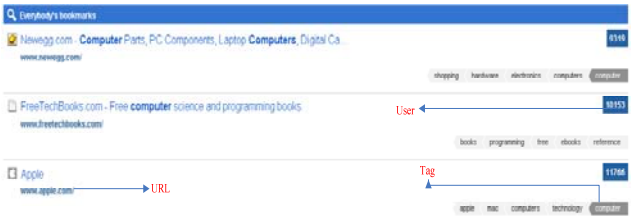


Figure 2. The tag computer's demonstration page

### *4. 2 Preprocessing*

In order to improve the quality of the tag recommendation, we removed the webpages that do not have tags, and also the missing ones caused by error links and time out of servers etc.

For the similarity calculation of webpages, we first filter the useless HTML marks, and delete the words between these marks, such as "script", "style", "cite" etc. However, we retain some words between the marks, e.g., "title", "a", "h1" etc. Second, the symbols like "?", ")", and "<" are removed. Finally, we remove the stopwords, such as "we", "anyone", "other" etc.

### *4. 3 Experimental results*

To evaluate the effectiveness of the proposed collaborative based algorithm on the labeled webpage recommendation, we use the URL http://www.spurl.net as an example. We demonstrate the superiority of our algorithm by the comparison with basic methods.

TABLE I. RECOMMENDED TAGS FOR THE PAGE HTTP://WWW.SPURL.NET

| **Tag Order** | **Freq(t)** | **Seq(t)** | **Freq(t) AND Seq(t)** | **Freq(t) AND Seq(t) AND PS** |
|---|---|---|---|---|
| 1 | bookmarks | bookmarks | bookmarks | bookmarks |
| 2 | del.icio.us | tools | bookmark | tool |
| 3 | bookmark | bookmark | del.icio.us | del.icio.us |
| 4 | bookmarking | del.icio.us | tools | spurl |
| 5 | tools | bookmarking | bookmarking | web |
| 6 | spurl | delicious | spurl | socialsoftware |
| 7 | web | social | web | cool |

In Table I, column 2 lists the recommended tags only considering the using frequency, column 3 lists the ones only considering the being selected orders, while column 4 are the results considering using frequency and selected orders. Finally in the last column we record the results of our method which uses the Porter Stemming algorithm. As the results shown in the table I, our algorithm can remove the tags with the same stem and find some more general tags, which indicates that our method not only can help user to label webpages, but also enable the other users to retrieve webpages using these tags, which indicates that our method not only can help user to label webpages, but also enable the other users to retrieve webpages using these tags.

TABLE II. EXPERIMENTAL RESULTS OF THIS PAPER

| Tag Num / Datsets | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 50 | 0.720 | 0.690 | 0.680 | 0.645 | 0.592 | 0.557 | 0.537 |
| 70 | 0.657 | 0.586 | 0.552 | 0.518 | 0.494 | 0.486 | 0.471 |
| 100 | 0.640 | 0.595 | 0.557 | 0.528 | 0.498 | 0.473 | 0.456 |

TABLE III. EXPERIMENTAL RESULTS OF [4]

| Tag Num / Datsets | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 50 | 0.780 | 0.690 | 0.660 | 0.585 | 0.576 | 0.543 | 0.506 |
| 70 | 0.536 | 0.507 | 0.473 | 0.438 | 0.406 | 0.377 | 0.360 |
| 100 | 0.540 | 0.490 | 0.463 | 0.440 | 0.398 | 0.382 | 0.351 |

To evaluate the effectiveness of the paper improved cosine similarity algorithm, we compare with the [4]. All the results are listed in Table II and Table III. We manually select three datasets with different size and then compare the predicted tags with the true tags. Also the Jaccard coefficient [10] is used to compare the similarity of tags rather than exact matching. Usually, the users will not be patient to give a webpage many tags, so we only recommend seven tags for each webpage. We compute the accuracy of different number of tags. The accuracy is the percentage of the number of correctly matched tags over the number of all recommended tags.

The results in Table III show that only retaining the words belonging to the tag set in the webpage will lead to the low accuracy, since many better words which can represent the contents are removed. However, we can use (5) to deal with this problem. The experimental results note that the proposed improved similarity computation algorithm obtains higher accuracy.

TABLE IV. TAG RECOMMENDED FOR WEB PAGES

| WebPages | Recommended Tags |
|---|---|
| http://slashdot.org/ | font, design, free, typography, web, webdesign, webdev |
| http://www.manybooks.net/ | book, ebook, pda, palm, fre, literature, read |
| http://video.google.com/ | google, search, video, tv, web, tool, television |
| http://www.flickr.com/ | photo, photography, flickr, blog, share, web, community |
| http://www.php.net/ | php, program, web, development, reference, opensource, csie |
| http://www.technorati.com/ | blog, search, web, tag, rss, technorati, news |
| http://news.google.com/ | news, google, daily, search,aggregate, import, web |
| http://maps.pietrosperoni.it/delicious/makemap.html | del.icio.us, mindmap, tool, map, visual, cool, mind |
| http://www.cnn.com/ | news, daily, world, cnn, media, import, us |

Table IV shows more examples of tag recommendation for webpages.

The web http://www.manybooks.net is a free webpage to provide users with e-books. E-books have a lot of formats, the users can choose their favorite format to read, and many e-books can be read in the Pocket PC. The tags we recommend tags for the page is “book”, “ebook”, “ pda”, “palm”, “fre”, “literature”, “read”. Book” and ebook” both denote books, “pda” and “palm” both denote Pocket PC, these tags are effectively describe the contents of the pages, but there are individual tags, such as “fre”, may be the misspelling of “free” and cannot express the meaning of free reading.

The web http://www.php.net is a provider of various versions of php pages. Php is a HTML embedded language and a server-side implementation of the HTML document embedded script Language. The tags we recommend for the page is “php”, “program”, “web”, “development”, “reference”, “opensource’, “csie”. We can see that the all the tags except “csie” can effectively express the contents of the page.

In summary, we can see that tags proposed by this paper can effectively express the majority of contents of the webpages, but there are errors in the individual tags and this is the point we need to improve.

## 5. Conclusions

A tag recommendation method based on collaborative filtering and text similarity is proposed in this paper. The obvious advantage of this method is that it adopts different tag recommendation methods for different webpages. The tags recommended are more popular and with less misspelling, we avoid the situation that the tags with the same root have the same meaning. The drawback is that it can only recommend tags for English webpages, without the analysis of the semantic relationship between tags and the personalized tags recommendation is not realized. The work to be done afterward is to recommend tags for multilingual webpages, analyzes each user’s tagging history to recommend different tags for different users and improve the accuracy of tags.

**Acknowledgements**

This work was supported in part by the Key Research Plan of Education Office of Hebei Province (ZH200804).

**References**

[1] Golder, S., and Huberman, B., "The structure of collaborative tagging systems," Journal of Information Science, vol. 32, pp. 198-208, 2006.
[2] Xu, Z., "Towards the semantic web: Collaborative tag suggestions," In Proceedings of the Collaborative Web Tagging Workshop, pp. 22-26, 2006.
[3] Sigma On Kee Lee, and Andy Hon Wai Chun, "A web 2.0 tag recommendation algorithm using hybrid ANN semantic structures," International Journal of Computers, vol. 1, pp. 49-58, 2007.
[4] Yu-Ta Lu, Shoou-I Yu, and Tsung-Chieh Chang, "A Content-based Method to Enhance Tag Recommendation," In Proceedings of IJCAI'09, pp. 2064-2069, 2009.
[5] Porter MF., "An algorithm for suffix stripping," Program, vol. 14, pp. 130-137, 1980.
[6] Breiman L, Friedman J, and Olshen R, "Classification and regression tress," Monterey: Wadsworth International Group, 1984.
[7] G. Salton, A. Wong and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18 pp. 613-620, 1975.
[8] Jiawei Han, and Micheline Kamber, "Data Mining Concepts and Techniques," China: Machine Press, 2008.
[9] http://130.149.154.91/corpus/delicious/.
[10] S Guha, R Rastogi, and K Shim, "An Efficient Clustering Algorithm for Large Databases," In Proceedings of the ACM SIGMOD international conference on management of data, pp. 73-84, 1998.