International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:8, 2010

Collaborative and Content-based Recommender System for Social Bookmarking Website

Cheng-Lung Huang Cheng-Wei Lin

***Abstract*—**This study proposes a new recommender system based on the collaborative folksonomy. The purpose of the proposed system is to recommend Internet resources (such as books, articles, documents, pictures, audio and video) to users. The proposed method includes four steps: creating the user profile based on the tags, grouping the similar users into clusters using an agglomerative hierarchical clustering, finding similar resources based on the user's past collections by using content-based filtering, and recommending similar items to the target user. This study examines the system's performance for the dataset collected from "del.icio.us," which is a famous social bookmarking website. Experimental results show that the proposed tag-based collaborative and content-based filtering hybridized recommender system is promising and effectiveness in the folksonomy-based bookmarking website.

***Keywords*—**Collaborative recommendation, Folksonomy, Social tagging

I. Introduction

RECENTLY the new Web 2.0 websites providing interactive information sharing and user-centered collaboration are growing rapidly on the World Wide Web. Examples of Web 2.0 website include social networking sites, social bookmarking sites, blogs, photograph sharing sites, video sharing sites, wikis and etc. Folksonomy is one of the characteristics in Web 2.0, and it is also known as collaborative tagging or social tagging, which allows users to collaboratively create and manage tags to categorize and classify users' collections or contents. Collaborative tagging in Web 2.0 is becoming widely used as an important tool to classify dynamic content for searching and sharing [1].

Currently, researches have shown that social tagging can be used to classifying blogs [2], to enhancement information retrieval [3-4] and to improve recommender systems [1]. Recommender systems are developed to deal with information overload and provide personalized recommendations, content and services to users [5-6]. These software systems have been applied in many areas including e-commerce, news, advertisement, document management and e-learning. Using tags can release the limitation of the cold-start and sparsity problems in the collaborative filtering based recommender systems [1]. User's tag information represents part of this user's preference or interest in the social bookmarking website. This triggers our research to develop a new tag-based recommender system.

Our investigated system incorporates the tag-based collaborative filtering and content-based filtering approaches. The two-stage recommender system groups similar users into clusters using clustering algorithm in the collaborative filtering stage, and then recommends similar items to the target user according to the target user's past collections in the content-based filtering stage.

The rest of this paper is organized as follows. Section 2 introduces the related works. Section 3 describes the framework of the proposed methodology. Section 4 demonstrates the empirical experiment. Section 5 provides conclusions.

C.-L. Huang is with the Department of Information Management, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, ROC. (phone: +886-7-6011000 Ext 4127; fax: +886-7-6011042; e-mail: clhuang@ccms.nkfust.edu.tw).

C. W. Lin was with the Department of Information Management, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, ROC.

II. Related works

A. Recommender systems

Recommender systems use a specific type of information filtering (IF) technique to recommend information items which are likely of interest to the user. Examples of these information items are blogs, commercial products, movies, music, news, photographs, and etc. Recommender systems make recommendations using three basic steps: acquiring preferences from the user's input data, computing recommendations using proper techniques, and presenting the recommendations to users [7]. The recommendation techniques include the content-based filtering approach (CBF) [8], the collaborative filtering approach (CF) [9] and hybrid-based recommender systems [10].

B. Folksonomy and social resources sharing systems

Folksonomies became very popular on the web as part of social software applications such as social bookmarking and photograph annotation. The important factors of folksonomy system are that the overall costs for users in terms of time and effort are far lower than systems that rely on complex hierarchal classification and categorization schemes [11].

Social resource sharing systems are web-based systems that allow users to upload their resources, and to label them with arbitrary tags [12]. Users, resource items, and tags are three important roles in this kind of systems. Users label the resource item using social tags as shown in Figure 1. These systems can be categorized according to what kinds of resources are supported, such as bookmarks, bibliographic references, photos,

merchandizes, or video. “Delicious” (del.icio.us) is a social bookmarking web service for storing, sharing, and discovering web bookmarks. Delicious uses a non-hierarchical classification system in which users can tag each of their bookmarks with freely chosen keyword terms

C. Tag-based collaborative recommender systems

People tag resources for future retrieval and sharing [13]. Tags can convey information about the content and creation of a resource [14]. Tags identify what the resource is about and the characteristics of a resource [15]. The tags collected by the user represent part of this user’s preference or interest in the social bookmarking website. This study models the user’s preference by using the tag-based information.

Kim et al. [1] used tag-based user profile in the collaborative filtering based recommender systems to release the limitation of the cold-start and sparsity problems. Unlike previous researches, this study constructs a two-stage recommender approach that hybridizes the collaborative filtering and content-based filtering.

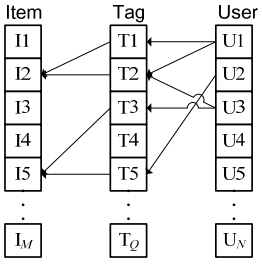


Fig. 1 Relationships among resource item, tag, and user in the social resources sharing systems

III. Proposed Methodology

The proposed system includes the following four steps: building the tag-based user and item profile, collaborative filtering for users, content-based filtering for resource items, and resource items recommendation.

A. Tag-based user profile and resource item profile

(1) Tag-based user profile

The user collects social resource items and labels these items with tags in the bookmarking websites. Since the tags collected by the user represent part of this user’s preference or interests in the social bookmarking website, the preference of a user can be modeled by analyzing the user’s tag information. The user’s tag information includes the tag name and the corresponding number of items collected by this user for each tag. For example, a certain user from del.icio.us collected 94 resource items of “javascript” and 73 resource items of “ajax.” We can notice that this user’s preference indicated by his/her tag information is major in “programming.”

(2) Tag-based resource item profile

Each collected resource item is associated with tags defined by users who are interested in the same resource item. The resource item’s tag information includes the tag name and its corresponding frequency tagged by these users. For example, a certain resource item (msn.com) has been tagged 101 times using “MSN,” 80 times using “news,” and 48 times using “search” as shown in Figure 2. From the item tagging information, we can identify the classification of this resource item defined by all users who is interested in this resource item.

B. Collaborative filtering for users

(1) Tag frequency–inverse user frequency (TF-IUF) for users

The tag-based user profile can be transform into a vector of TF-IUF (tag frequency–inverse user frequency) which is modified from the TF-IDF (term frequency–inverse document frequency) for representing the document description. The weight of tag *j* in user *u*’s tag collection is defined as:

$$UserTagW_{u,j} = UserTagTF_{u,j} \times UserTagIUF_j \quad (1)$$

$UserTagTF_{u,j}$ is the local weigh of tag frequency and it is defined as:

$$UserTagTF_{u,j} = \frac{freq_{u,j}}{\max_u freq_{u,j}} \quad (2)$$

where $freq_{u,j}$ is the number of occurrence of tag *j* in user *u*’s tag collection; $\max_u freq_{u,j}$ is the maximum number of occurrence in user *u*’s tag collection. This equation normalizes or scales the tag occurrence.

The global weight, $UserTagIUF_j$, which represents the relative important among user *u*’s tag collection, is defined as:

$$UserTagIUF_j = \log(\#Users / \#UsersCollectingTag_j) \quad (3)$$

where #Users is the total number of users in the training set; $\#UsersCollectingTag_j$ is the total number of users who collect tag *j* (in the training set).

This model incorporates local and global information. The $UserTagTF$ accounts for local information and $UserTagIUF$ is the inverse tag importance which represents the global probability of a certain tag for a user.

(2) Similarity between users

The cosine similarity between the user *u* and user *v* can be defined as the inner product of the two users’ tag weights:

$$UserSim_{u,v} = \frac{\sum_{j=1}^{N} UserTagW_{u,j} \cdot UserTagW_{u,j}}{\sqrt{\sum_{j=1}^{N} (UserTagW_{u,j})^2 \cdot \sum_{i=1}^{N} (UserTagW_{u,j})^2}} \quad (4)$$

where *N* is the number of common tags that is collected by user *u* and *v*.

Fig. 2 Resource item's tag information (e.g. msn.com)

(3) Users clustering

The purpose of this step is to cluster users based on their TF-IUF tag profile. Clustering is an unsupervised data segmentation technique for grouping a set of data objects into classes of similar data objects. Some popular clustering methods can be adopted such as partitioning methods (e.g., k-means clustering), hierarchical methods, grid-based methods, model-based methods and density-based methods [16]. This study used the hierarchical methods clustering approach, which can be easily performed clustering based on the cosine similarity matrix among users obtained from the previous step.

C. Content-based filtering for resource items

Users in the same cluster have similar preferences. The content-based filtering based on the resource item's tag information is applied in each cluster. The purpose of this step is to find the similar resource items which the user may interest and then recommend these similar resource items to the target user.

(1) Tag frequency for the resource item

User defines a resource item using tags. The tag i's normalized frequency for item q represented as $ItemTagTF_{q,i}$ is defined as follows.

$$ItemTagTF_{q,i} = \frac{freq_{q,i}}{\max_q freq_{q,i}} \quad (5)$$

where $freq_{q,i}$ is the number of occurrence of tag i that defines item q; $\max_q freq_{q,i}$ is the maximum number of occurrence of tags that define item q.

(2) Tag frequency threshold for frequent tags

The number of tags that defines an item may be very large, and it is not necessary to use all tags. We use the normalized tag frequency (item-tag-TF) threshold to filter out unnecessary tags for an item. That is, if the item-tag-TF threshold is set to 30%, we select the tags whose TF values are greater than or equal to 0.3. For example, in Table 1, the msn, news, search, email, and hotmail are selected because their TF are greater than 0.3.

The merits of item-tag-TF threshold filter are as follows. (i) The infrequent tags can be filtered out, thus the computation time for similarity calculation is reduced. (ii) Since some incorrect or improper tags which are rarely used by users may exist, the item-tag-TF threshold can filter out these incorrect or improper tags.

(3) Tag weight for the resource item

The tag i's relative importance among collected tags in a cluster represented as $ItemTagIIF$ (inverse item frequency) is defined as:

$$ItemTagIIF_i = \log(\#Items / \#ItemsDefinedByTag_i) \quad (6)$$

where $\#Items$ is the total number of items in a cluster; $\#ItemsDefinedByTag_i$ is the total number of items (in a cluster) defined by tag i.

The weight of tag i for resource item q is defined as:

$$ItemTagW_{q,i} = ItemTagTF_{q,i} \times ItemTagIIF_i \quad (7)$$

(4) Similarity between resource items

The tag-based cosine similarity between resource item q and item r is calculated as the inner product of the item tag weights:

$$ItemSim_{q,r} = \frac{\sum_{i=1}^{M} ItemTagW_{q,r} \cdot ItemTagW_{q,r}}{\sqrt{\sum_{i=1}^{M} (ItemTagW_{q,r})^2 \cdot \sum_{i=1}^{M} (ItemTagW_{q,r})^2}} \quad (8)$$

where M is the number of common tags which label both resource item q and r.

TABLE I TAG NAME, FREQUENCY AND TF VALUE FOR RESOURCE ITEM "MSN.COM"

Tag	*freq.*	*TF*	Tag	*freq.*	*TF*
msn	101	**1.00**	imported	24	0.24
news	80	**0.79**	searchengine	13	0.13
search	48	**0.48**	e-mail	10	0.10
email	45	**0.45**	MSN.com	6	0.06
hotmail	35	**0.35**	Portal	6	0.06
Microsoft	25	0.25	Daily	5	0.05

D. Resource item recommendation

To recommend items to the target user, the content-based filtering is applied in this study. The k-nearest similar items for each item collected by a certain target user were recommended according to the item similarity matrix. That is, given m items collected by the target user, and k (maximum) similar items found from the item similarity matrix, $k \times m$ similar resource items are recommended to the target users. Note that the total number of recommended resource items for a target user may not be the same as that of the other target users, because the number of collected items of a target user and the number of similar items found are not the same as that of the other target users.

IV. Experiments and Results

The proposed recommender system (as shown in Figure 3) was developed using C#.NET language under the platform of WINDOWS XP operating system and Microsoft SQL Server 2008. Experiments was performed on the computer of Intel Pentium M 740 1.73GHz Processor and 4 GB RAM.

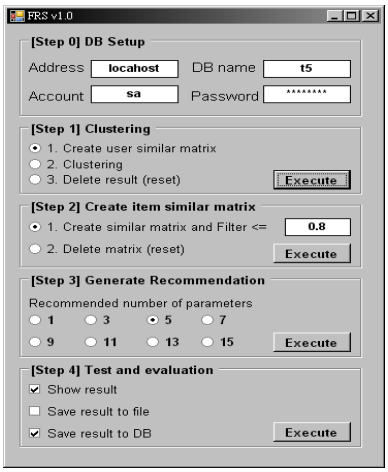


Fig. 3 A prototype of the proposed recommender system

A. Experimental design

(1) Data set

The experimental dataset was collected from the del.icio.us website, which is a popular website that helps users share their favorite information items (links). We crawled the del.icio.us to randomly collect the newly active users. The dataset contains 34,613 bookmarked items which are labeled using 45,784 tags by 473 users. The user profile of a specific user includes resource item, tag, and the number of collected item that is defined by each tag. Item profile for a specific user's collected item includes the name of the tag that label the item, and the number of occurrence of the tag that defines the item by all users.

(2) Training and test set

For each user's collection, the collected items were sorted according to the time-horizon. And then each user's collections were divided into two parts according to the time-horizon: 50% of items for training and 50% of items for test. Each user takes turn as the target user. The number of hit items is counted if the recommended items hit the target user's collected items in the test set. The average performance for target users is computed to evaluate the reliability of the proposed recommender system.

(3) Evaluation measures

Three performance measures are considered in evaluating the effectiveness of the proposed method: recall, precision, and F1-measure [17]. The precision is the ratio of target user's hit items from the recommended items. The recall is the ratio that the recommended items successfully predict (hit) the target user's collected items (test set). F1-measure combines recall and precision with an equal weight in Eq. (11).

$$Precision = \frac{\text{Number of hit items}}{\text{Number of recommended items}} \quad (9)$$

$$Recall = \frac{\text{Number of hit items}}{\text{Number of items in the test set}} \quad (10)$$

$$F1 = \frac{recall \times precision}{(recall + precision)/2} \quad (11)$$

(4)User clustering

The default number of clusters is set to nine. Most people can't be grouped together by clustering, because their tags are quite different. This study did not include the users whose size of cluster member was less than five, since recommending to users in a cluster with small member size was not effectiveness. This resulted in 183 users with nine clusters included in this study.

B. Experimental Results

(1) Comparison with random recommendation models

The item-tag-TF threshold of our proposed model was set to 20% and the recommendation size was set to five. To identify the relative performance improvement of our proposed system against that of the random recommendation (no-model approach), we conducted a performance comparison with random recommendation under the same recommendation size. The random recommendation which randomly recommends resource items to the user was conducted via two ways: (i) the model of random recommendation with clustering randomly recommended items to target users in the same cluster; (ii) the model of random recommendation without considering randomly recommended items (from all items) to target users. That is, the former one used the cluster information while the second one did not use the cluster information.

Figure 4 shows the precision, recall and F1-measure for the proposed system and random models. We found that our proposed model was obviously better than the random recommendation models. Our model had an improvement 2.5 times F1-measure of the random recommendation with clustering.

(2) Recommendation size and item-tag-TF threshold

A proper recommendation size should be studied in considering the precision, recall, and especially the practices in application domain. In this study, the recommendation size of 1, 3, 5, 7, 9, 11, 13, and 15 were conducted for a particular cluster. Their average precision, recall and F1-measure are illustrated in Figure 5. We found that F1-measure was slightly improved in increasing recommendation size; however, it is not adequately to recommend too many items to user to reduce information overloading. Thus the recommendation size of five was adequate for this experimental dataset.

For F1-measure and precision, the item-tag-TF threshold of 20% was better than that of 80% and 100% under various recommendation sizes. Though the recall of item-tag-TF

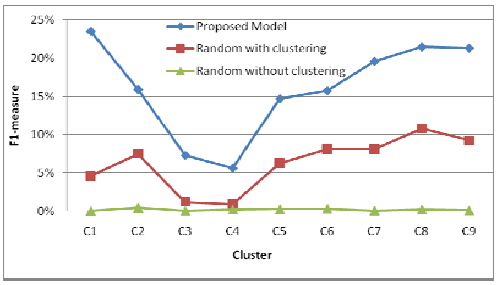
threshold of 20% was not better than that of 80% and 100%, this study selected item-tag-TF threshold of 20%. The computational time for item-tag-TF threshold of 20% can be reduced, as the number of tags used in the user's preference profile was smaller than that of 80% and 100%.

V. Conclusions and Future Works

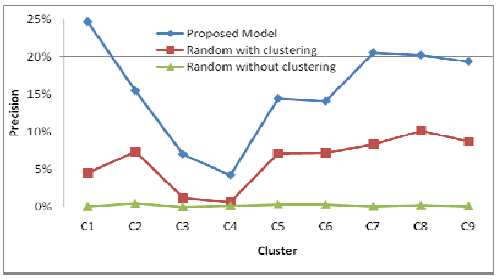
This study used a two-stage recommendation approach. First, the collaborative filtering finds similar user group by clustering algorithm using the tag-based user preference profile. And then the content-based filtering recommends resource items to target users by analyzing the tag-based content of the target user's collected items. From the experimental results of the dataset from del.icio.us website, we found the proposed hybrid recommender system were promising and effectiveness.

This study has demonstrated that the tag information can be used to represent users' preferences in the social bookmarking website; and the proposed recommender system can successfully suggest social resource items to user based on the user's tag-based preferences. The proposed model can be adapted in application areas of tagging, such as books, articles, documents, pictures, audio and video.

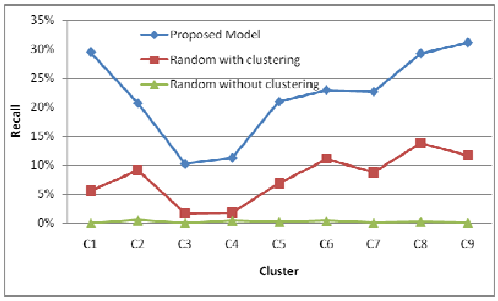
In an environment in which the user gradually changes interests, the tag data close to the current temporal period are usually more important than that temporally far from the current period. This is called the time decay in the recommender systems [18]. This implies that in the social tagging system, the newly tagged items by the user are more important for this user currently. Our future work will incorporate the recency consideration into our tag-based recommender systems.



(a) F1-measure

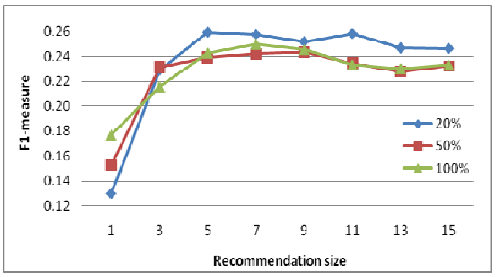


(b) Precision

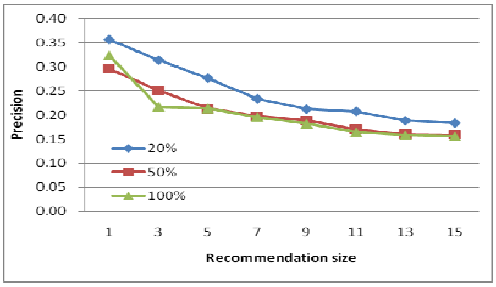


(c) Recall

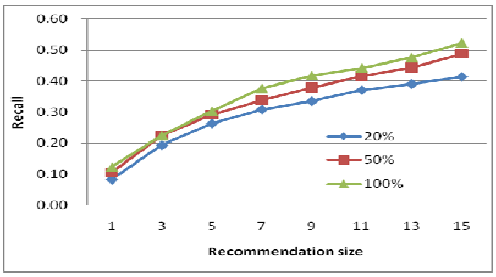
Fig. 4 Performance comparison of the proposed system vs. random models



(a) F1-measure



(b) Precision



(c) Recall

Fig. 5 Performances with various recommendation sizes and item-tag-TF thresholds

References

[1] H.-N. Kim, A.-T. Ji, I. Ha and G.-S. Jo, "Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation," *Electronic Commerce Research and Applications*, vol. 9, Issue 1, pp. 73-83, January-February 2010.

[2] C. H. Brooks and N. Montanez, "An analysis of the effectiveness of tagging in blogs," in: *Proceedings of the 2005 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*. Stanford, CA. March 2006.
[3] P. J. Morrison, "Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web," *Information Processing & Management*, vol. 44, Issue 4, pp. 1562-1579, July 2008.
[4] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, Issue 2, pp. 101-114, June 2008.
[5] G. Adomavicius and A. Tuzhilin, "Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp.734-749, 2005.
[6] A.M. Rashid, I. Albert, D. Cosley, S.K. Lam, S.M. McNee, J.A. Konstan and J. Riedl, "Getting to know you: learning new user preferences in recommender systems," in: *International Conference on the Intelligent User Interfaces*, San Francisco, CA, 2002, pp.127-134.
[7] K. Wei, J. Huang and S. Fu, "A survey of e-commerce recommender systems," in: *IEEE International Conference on Service Systems and Service Management*, Chengdu, China, 2007, pp.1-5.
[8] K. Lang, "Newsweeder: Learning to filter netnews," in: *Proceedings of the Machine Learning conference*, Tahoe City, CA, USA, 1995, pp. 331-339.
[9] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in: *Proceedings of the Computer Supported Cooperative Work Conference*, Chapel Hill, NC, 1994, pp. 175-186.
[10] M. Balabanovic and Y. Shoham, "Fab: content-based collaborative recommendation," *Communications of the ACM*, vol. 40, no. 3, pp.66-72, 1997.
[11] A. Mathes, "Folksonomies-cooperative classification and communication through shared metadata," Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, Technical Report LIS590CMC, December 2004.
[12] R. Jäschke, A. Hotho, C. Schmitz, B. Ganter and G. Stumme, "Discovering shared conceptualizations in folksonomies," Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 6, No. 1, pp.38-53, February 2008.
[13] C. Marlow, M. Naaman, D. Boyd and M. Davis, "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead." in: *Proceedings of Hypertext*, New York: ACM Press, 2006.
[14] M. Memmel, M. Kockler and R. Schirru, "Providing multi source tag recommendations in a social resource sharing platform," *Journal of Universal Computer Science*, vol. 15, no. 3, pp.678-691, 2009.
[15] S.A. Golder and B. A. Huberman, "The structure of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198-208, 2006.
[16] J. Han and M. Kamber, *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, USA, 2006.
[17] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers, Norwell, MA, 1997.
[18] C.-L. Huang and W.-L. Huang, "Handling sequential pattern decay: Developing a two-stage collaborative recommender system," *Electronic Commerce Research and Applications*, vol. 8, Issue 3, pp.117-129, May-June 2009.