Topic-based Resource Recommendation

Zarli Htun, and Phyu Phyu Tar

***Abstract*—**Recommender systems are used to alleviate the information overload problem. They provide users with products or items that they may interest by analyzing users' past interests and preferences. Standard recommender systems need user-item rating data to know the user's interest to provide recommendations. But in real world systems, not every user wants to provide rating information explicitly. Social tagging systems have become important information source to describe the content of items as well as users' interest and preferences. Therefore, we propose a topic-based resource recommendation method based on social tagging information by considering how tagging data can be incorporated into recommender system to improve recommendation. The proposed system extracts latent topics from tagging data and uses these topics to build user profile used in the system for resource recommendation. We tested proposed system using the real world dataset. The experimental results show that the proposed system outperforms the other state-of-the-art approaches.

***Keywords*—**Recommender system, social tagging, user profile, explicit rating, resource recommendation

I. Introduction

With the rapidly growing amount of information available on the World Wide Web, it becomes necessary to have tools to help users to select the relevant part of online information. Recommendation is a task that recommends highly relevant items with a given user. The correct recommendation is increasingly important because of information overload. It is impossible for a user to search all items to discover interesting items which are matched with the user's preference because the number of existing items is too large. Typically, in a recommender system, we have a set of users and a set of items. Each user rates a (small) subset of the set of all items with some numeric score, e.g. on a scale from 1 to 5. The recommender system has to predict the unknown rating for source user on a non-rated target item based on the known ratings. Collaborative Filtering is the most popular method and widely implemented technique in recommender systems [13].

Collaborative-filtering-based systems compare the preferences of a user (such as explicit item ratings) with the records of preferences of other users to find users who have similar preferences to the user. These records are then used to predict the preference value of the user on a particular item (rating prediction) or to recommend the top N items which could be of interest to the user (top N recommendation).

Collaborative filtering has two types: (1) memory-based approaches, (2) model-based approaches. [13] The memory-based approaches use either user-based approaches or item-based approaches for prediction (recommendation) of ratings for items. Memory-based approaches are easy to implement and popular but they do not guarantee good prediction results. On the other hand, model-based approaches include several model based learning methods. However, the above CF-based approaches assume users are independent and identically distributed and ignore additional information such as the content of item and social connections of users while performing the recommendation task.

Social tagging systems have gained popularity on the web. In social tagging systems, users can manage their resources easily and annotate them with their keywords called tags and categorize content and share them with other users. It is very helpful in searching information of interest. People tag resources for future retrieval and sharing [10]. Tags can convey information about the content and creation of a resource [11]. Tags identify what the resource is about and the characteristics of a resource [12]. Therefore, this metadata could also be used to support the recommendation process and there are previous works done incorporating social tagging information into recommender system to improve recommendation performance [2, 3, 4, 5, 6, 7, and 8]. In this paper, how the social tagging information can be used to improve the performance of traditional collaborative filtering algorithm. This paper presents a resource recommendation method which is based on topics which are derived from tagged resources and tags in social bookmarking system.

II. Related work

Popularity of social tagging systems makes them become the rich source of user's interest and preferences indicators. Although tag information has been incorporated into recommender systems in some ways, not so much work has been done on the item recommendation. Tso-Sutter [5] discussed about using the tag information to do item recommendation. In Tso-Sutter's work, the tag information was converted into two 2-dimensional relationships, user-tag and tag-item, and was used as a supplementary source to extend the rating data. In the work of Niwa et al. [7], a tf-idf weighted tag based item profiles have been used for web page

Zarli Htun is a Ph.D student in Department of ICT at University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Mandalay Division, and The Republic of the Union of Myanmar. (e-mail: zarlihtun@gmail.com).

Phyu Phyu Tar is with University of Technology(Yatanarpon Cyber City), Myanmar. She is now with the Department of Information Technology of UTYCC. (e-mail: thitagu07@gmail.com).

http://dx.doi.org/10.15242/IIE.E0314129

recommendation. Shepitsen et al. [8] applied hierarchical clustering to tag data from social tagging system to provide recommendation of resources.

In the proposed system, collaborative tagging information in social bookmarking website will be explored to derive the hidden topics on collection of resources. After determining these hidden topics, users' interest on these topics is measured based on the users' tagging behavior and build a collaborative filtering recommender system that provide a top N list of resources.

III. Recommender System

Recommender systems are used to alleviate the information overload problem of World Wide Web by providing users with information of interest. A recommender system assumes a set of users $U=\{u_1,\ldots u_N\}$ and a set of items $I=\{i_1,\ldots\ldots i_M\}$. Each user u rates a set of items by $r_{u,i}$. $r_{u,i}$ can be any real number, but more often ratings are integers, e.g., in the range [1,5]. The basic task of recommendation is as follows. Given a user $u \in U$ and an item $i \in I$ for which $r_{u,i}$ is unknown, predict a rating for user 'u' on item 'I'.

User-based collaborative filtering for top-N recommendation [9] first finds the top-K similar users to the source user. To measure the similarity between users, there are various similarity calculating methods such as Pearson Correlation similarity, Jaccard coefficient similarity, and Cosine similarity that can be used. The list of items rated by similar users is aggregated. Then we find the top-N highly ranked items in this aggregated list and return them as the top-N recommended items. In the aggregated list, the aggregated rating of each item i would be

$$\hat{r}_{c_{u,i}} = \frac{\sum_{v \in N_u, i \in I_v} sim_{u,v} \times r_{v,i}}{\sum_{N_u, i \in I_v} sim_{u,v}} \quad (1)$$

In (1), is the predicted rating of item i for the source user u using CF. Top-N recommended items are the items with the top-N highest values of . $sim_{u,v}$ is the similarity value between user u and v. Similarity value can be calculated by using various similarity calculation methods such cosine similarity, Pearson correlation similarity, etc.

In item-based collaborative filtering recommender system, for each item rated by the user, the set of top K similar items to that item is searched and these items are aggregated to compute the set of top-N recommended items. Ranking score in the set of K similar items for all items rated by user as is computed as follows:

$$s_i = \sum_{j \in I'_u, i \in N_j} sim_{i,j} \quad (2)$$

In (2), s_i is the score computed for each item i similar to one of the items in user's rated items. The top N items with highest values of s_i will be returned as the top-N recommended items. $sim_{i,j}$ measures the similarity of items i and j.

IV. Social Bookmarking System

Social bookmarking system is a web-based resource sharing system in which users store and share their bookmarks online. Social bookmarking websites have seen a rapid growth in popularity and a high degree of activity by their users. In social bookmarking systems, users store the bookmarks of web resources they interest and describe these bookmarks with their keywords. For instance; del.icio.us, CiteuLike, Bisonomy are popular social bookmarking services. Among them, "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.

In social bookmarking system, the following data exists:

$U=\{u_1, u_2,\ldots\ldots\ldots,u_m\}$ is a set of 'm' users,

$T=\{t_1, t_2,\ldots\ldots\ldots\ldots,t_l\}$ is a set of tags annotated by users to describe bookmarked resources,

$I=\{i_1, i_2,\ldots\ldots\ldots\ldots,i_n\}$ is the set of 'n' resource items tagged by users.

V. Proposed Topic-Based Recommender Model

A. User Profile Generation

Traditional collaborative filtering recommender system bases on users' rating data which reflect the user's preferences and interest on items. But in social bookmarking systems, users do not provide explicit rating on bookmarked resource items that they interest. Instead of rating on items, users annotate the bookmark resources using their own keywords called tags. Therefore, tags show user's interest and preferences on the bookmarked resources. The proposed system will analyze user's tag usage and try to estimate user's interest and preferences. Given a collection of bookmark resources, the proposed system generate implied topics (latent topics) on the given resources and based on these resulted topics, user's interests on these latent topics are estimated to create a user interest profile.

We use the LDA (Latent Dirichlet Allocation) which is a popular topic modeling approach to derive latent topics from the collection of bookmark resources. Latent topic models have been successfully applied as an unsupervised topic discovery technique in large document collections. In topic modeling, a document is transformed into a bag of words, in which all of the words of a document are collected and the frequency of the occurrence is recorded. In LDA, documents are represented as a mixture of implied (or latent) topics, where each topic can be described as a distribution of words.

Fig. 1 illustrates the LAD process in plate notation. In this the generative model, z and d variables identify topics and documents, while $\theta(d)$ is the distribution over topics for a

document d and φ(z) is the distribution over words for a topic z. These distributions can be used to generate documents in the form of a collection of words (w). D is the number of documents, T is the number of topics in the corpus and N_d the topics found in each document. Hyperparameters α and β identify the Dirichlet priors of the above multinomial distributions respectively. These hyperparemeters can be changed in order to control the smoothing of the distributions.

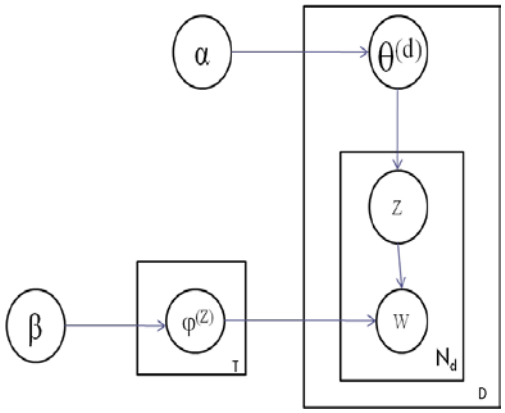

Fig. 1 Probabilistic graphical model of LDA

In SBS, Instead of documents (D), users annotate each bookmark resources using their own keyword called tags. Therefore, in order to create topic models using LDA, bookmarked resources are taken as documents and all of the words in a document (bookmark resource) are a set of tags used to describe it by the users. Therefore, each document in SBS is a bag of tags used to annotate a resource.

Once the LDA model is generated, it is used to infer the mixture of topics that the user interests. This process in it is entirety is shown as a block diagram in Fig. 3. Based on the resulted latent topics and user's tagging information, user profile based on topics is built. Map user's tags with latent topics and assign weights.

To measure the user's interest on a topic, we first compute the tag scores of a user. Each user 'u' has a set of personal tags: T (u) = $\{t_1,t_2,\ldots\ldots,t_l\}$ which are used to annotate the particular resources by this user. Therefore, the tag score of a user for a tag, tw(u,t_j) is

$$tw(u,t_j)\frac{freq(u,t_j)}{\sum_{t_l \in T(u)} freq(u,t_l)} \quad (3)$$

where freq(u,t_j) is the number of times that user 'u' used tag 't_j'. Freq(t_l) is total frequency of all tags used by user 'u'.

Each user 'u' has a user profile P with a vector of his interest topics with its weights,

$$P = \{(b_1, IN(u,b_1)),\ldots\ldots,(b_k, IN(u, b_k))\} \quad (4)$$

where 'b_k' belongs to set of latent topics and IN(u,b_k) is the interest weight of user to this topic. Interest weight of user on a topic 'b_k' is the maximum of all tag scores of the user related to this topic. It is described according to the following formula:

$$IN(u,b_k) = Max\{ts(u,t_1),\ldots\ldots,ts(u,t_j)\} \quad (5)$$

where 't_j' is the tag that belongs to topic word distribution of topic ' b_k' The assumption is that topics related to important tags by the user are also important to user and would have high interest weight values.

In our experiments, we use a dataset from Delicious.com social bookmarking website and 50 topics is extracted from a corpus of 69226 documents.

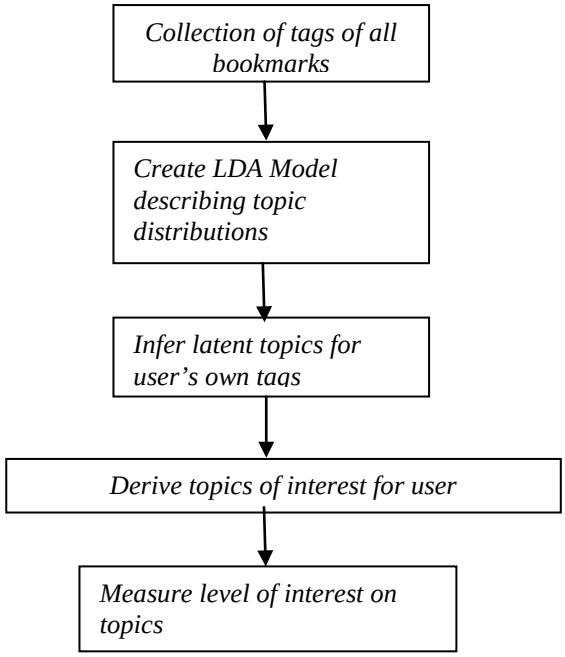

Fig. 2 Topic Modeling Steps

Fig. 3 Example topics from Delicious .com dataset

After user profile generation, the resulted user interest weights on topics are regarded as implicit user-item rating matrix (user-topic, here) and used as input to recommender system.

B. User Profile Generation

Neighborhood formation is to generate a set of users with similar topic interests for a target user u_i. The "K-Nearest-Neighbors" technique is used to the top K neighbors with shortest distance from the target user. The distance or similarity measure can be calculated through various kinds of proximity computing approaches such as cosine similarity and Pearson correlation. In the proposed system, Pearson

correlation method is used to measure the topic interest similarity between two users.

In such a non-rating environment, depending on topics only is not enough. Therefore the proposed system also considered user's tagging behavior and bookmarking behavior to measure similarity between two users.

Therefore, neighbors are selected based on three similarity measures: tag usage similarity, resource item similarity and interest factor similarity.

For the two users u_a and u_b, let T_a and T_b be the sets of tags for each user u_a and u_b respectively.

(1)Tag Usage similarity

Tag usage similarity is measured based on the common tags used by the two users, u_a and u_b. It is described in formula as follows:

$$sim_T(u_a, u_b) = \frac{(T_a \cap T_b)}{|T_a|} \quad (6)$$

(2) Resource Item Similarity

To compute the resource item similarity between two users u_a and u_b, both of their resource items are considered as two sets, and the Jaccard Index is applied between these sets. The Jaccard index is a well known statistic, widely used to compare the similarity between two sets. This formula is presented below in equation 2, where I_a represents the item set of user u_a and I_b represents the item set of u_b.

$$sim_R(u_a, u_b) = J(u_a, u_b) = \frac{(I_a \cap I_b)}{|I_a|} \quad (7)$$

(3) Topic Interest Similarity

User profiles generated at the previous phase are used to compute the topic interest similarity between users. Since topic interest values in user profile can be regarded as the rating values in user-item (user-topic, here) matrix. Therefore, the similarity between two users' rating behavior can be calculated by using various similarity measures. In the proposed system, Pearson correlation method is used to measure topic interest similarity called simI(u_a,u_b). Let a and b be two users, $r_{a,p}$ be the rating of user a for topic p and P be the set of topics, rated both by a and b. Then Pearson correlation coefficient is defined as follows:

$$\text{sim}(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \overline{r_a})(r_{b,p} - \overline{r_b})}{\sqrt{\sum_{p \in P}(r_{a,p} - \overline{r_a})^2}\sqrt{\sum_{p \in P}(r_{b,p} - \overline{r_b})^2}} \quad (8)$$

Finally, the similarity sim(u1,u2) is measured by aggregating the three similarity measures above,

sim(u1,u2)=simR(u1,u2)+simT(u1,u2)+simI(u1,u2) (9)

C. Recommendation Method

After choosing neighbor users with similar topic interests, the resource recommendation module subsequently chooses which resource items of neighbor users to be recommended to user. In order to generate a ranked list of items, the rank of an item is computed according to the following equation,

$$Rank(u,i) = \sum_{n \in Nei(u)} sim(u,n) \quad (10)$$

where Nei(u) is neighbors of user u produced from neighborhood formation phase. sim(u,n)=similarity value of user u and his neighbor n.

VI. Dataset Specifications

For experimental evaluation, we use the version of the Delicious.com dataset published for the HetRec 2011 (hetrec2011-delicious-2k). This dataset was obtained from Delicious social bookmarking system. Its users are interconnected in a social network generated from Delicious "mutual fan" relations. Each user has bookmarks, tag assignments, i.e. tuples [user, tag, bookmark], and contact relations within the dataset social network. Each bookmark has a title and a URL. Table 1 shows some statistics about the dataset.

TABLE I.
DATA STATISTICS OF HETREC-DELICIOUS-2K DATASET

Dataset	Delicious
Number of users	1867
Number of Items	69226
Number of User-items relations	104799
Number of tags	53388
Number of User-tag-items	437593
Number of User-user relations	15328

VII. Experimental Results

An experimental study was performed to evaluate the performance of our topic-based resource recommender. The recommender was tested for various numbers of neighbor users. We use 80% of dataset for training and 20% of dataset for testing. For top-N recommendation, the quality was measured by looking at the number of hits, i.e., the number of items in the test set that also exists in the top-N recommended items. Therefore, 'recall' is measured as evaluation metric, since precision is difficult to evaluate in such no-rating environment like social bookmarking. We define recall as the ratio of hit set (HIT) size to the relevant set(REL) size (test set). Therefore, for all n tested users, the average of recall is:

$$recall = \frac{\sum_u \frac{|HIT_u|}{|REL_u|}}{n}, \quad (11)$$

where n is the number of users tested.

A recall value of 1.0 shows that the recommendation algorithm was able to retrieve all relevant items, whereas a recall vale of 0.0 shows that the recommendation algorithm was not able to recommend any of the relevant items.

For performance evaluation, the proposed system is compared with traditional user-based collaborative filtering (CF) and tag-vector similarity-based recommender (TVS) [6].

In the performed experiment, the number N of recommended items is set to 100. The dataset contains significantly more items than users which is different from many other data sets, so using a small value for N will produce generally poor results for all compare methods. And then the system is tested with various neighborhood sizes from 10 to 50 by an interval of 10.

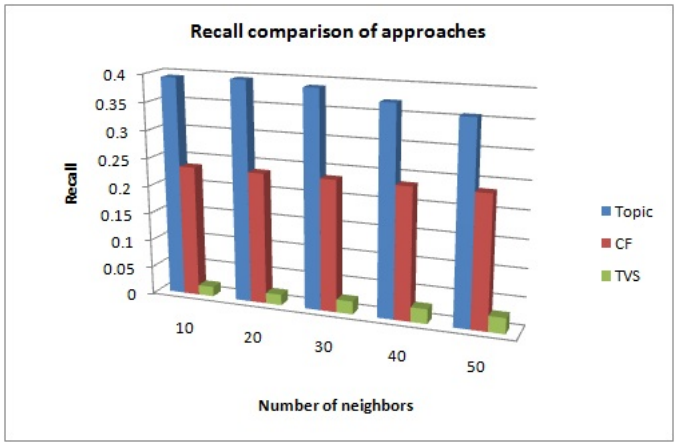

Fig. 4 Recall values of approaches with various number of neighbors (10, 20… 50)

Fig. 4 shows the average recall values of our proposed system and comparison methods. When the recommender system deploys the topic-based profiling, the performance of the system is higher than that of other two systems. Recall values of Topic-based system are 40 % in average higher than that of CF and TVS. If the user has the small number of similar users (neighbors), the derived topics are important to improve the quality of recommendation results.

VIII. Conclusion

In this paper, a recommendation method based on folksonomic data is presented. The proposed system uses the collaborative tagging information provided by users in a social bookmarking system and derives user preference topics based on the tagging data by using LDA. With reference to the resulted topics, the proposed system generates the user-topic rating matrix which is used as implicit rating matrix in a non-rating environment like social bookmarking. The resulted rating matrix is used in recommender system to provide top-N recommendations to users. The experimental results show that the proposed system outperforms the other state-of-the-art approaches.

References

[1] G. Adomavicius and A. Tuzhilin , "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extentions. IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.6, pp.734-749.
http://dx.doi.org/10.1109/TKDE.2005.99

[2] C. Basu, H. Hirsh, W. Cohen, and C. Nevill-Manning. "Techical Paper Recommendation : A Study in combining multiple information sources", JAIR, 1: 231-252, 2001.

[3] C.L. Huang, C.W Lin, "Collaborative and Content-Based Recommender System for Social Bookmarking Website", World Academy of Science, Engineering and Technology,2010.

[4] 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.
http://dx.doi.org/10.1016/j.elerap.2009.08.004

[5] K.H.L. Tso-Sutter, L.B. Marinho, and L.Schmidt-Thieme. "Tag-aware Recommender Systems Fusion of Collaborative Filtering Algorithms", Proceedings of the 2008 ACM symposium on Applied computing, ACM, USA, 2008.

[6] A.Bellogín, I.Cantador, and P.Castells, "A Study of Heterogeneity in Recommendations for a Social Music Service", HetRec 2010.

[7] S.Niwa,T. Doi and S. Hon'Iden, "Web Page Recommender System Based on Folksonomy Mining", Transactions of Information Processing Society of Japan, 47(5):2006.

[8] A. Shepitsen, J. Gemmell, B. Mobasher and R. Burke, "Personalized recommendation in social tagging systems using hierarchical clustering", In Proc. Of the 2008 ACM conference on Recommender systems, 2008.
http://dx.doi.org/10.1145/1454008.1454048

[9] M. R. McLaughlin and J. L. Herlocker. "A collaborative filtering algorithm and evaluation metric that accurately model the user experience", In SIGIR '04: Proceedings of the 27th international ACM SIGIR conference on Information Retrieval, New York, NY, USA, 2004.

[10] 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.

[11] M.hbcgt 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, 2009.

[12] S.A. Golder and B. A. Huberman, "The structure of collaborative tagging systems," Journal of Information Science, vol. 32, no. 2, 2006.
http://dx.doi.org/10.1177/0165551506062337

[13] P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.),"Collaborative Filtering Recommender Systems", The Adaptive Web, LNCS 4321, pp. 291-324, 2007.

[14] B. Liu, Web Data Mining, Springer 2007.

[15] F. Ricci et al. (eds.), Recommender Systems Handbook, DOI 10.1007/978-0-387-85820-3_1, Springer 2011.
http://dx.doi.org/10.1007/978-0-387-85820-3_1

[16] David M. Blei Andrew Y. Ng Michael I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research 3 (2003).

[17] http:// delicious.com/.