A RESOURCE RECOMMENDER SYSTEM BASED ON SOCIAL TAGGING DATA

Zarli Htun[1] and Phyu Phyu Tar[2]

[1]Department of Information Technology, University of Technology, Yatanarpon Cyber City, Pyin Oo Lwin, Myanmar
[2]Department of Information Technology, University of Technology, Yatanarpon Cyber City, Pyin Oo Lwin, Myanmar

ABSTRACT

Recommender systems (RS) are solution to alleviate the information overload problem. Users are suggested with a list of products or items they may interest by analyzing their browsing or purchasing history. These systems generally require user-item rating information to find similar users (neighbors) in order to produce a list of suggested items. However, explicit user-item rating data is difficult to collect in real world applications because most users do not want to give item ratings explicitly. Nowadays, social tagging applications have the content of items as well as users' interest and preferences. Therefore, this paper proposes an alternative approach based on social tagging information to improve the performance of RS The proposed system extracts latent topics from tagging data and uses these topics to build user profile to be used in the system for resource recommendation. The proposed system is tested by using the real world datasets of popular social tagging applications. 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

1. 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 suggests highly relevant items with a given user. The correct suggestion of items is increasingly important because of information overload. Collaborative Filtering (CF) is the dominant technique among recommender algorithms. CF bases on the assumption that previously like-minded users will also share similar tastes in the future. In CF, measuring similarity is important; only the top-k most similar users are allowed to contribute their ratings, and each contribution is weighted according to the specific degree of similarity the neighbor shares with the current user.

The input of CF algorithms is a user-item matrix of ratings values. For a target user u_a, recommender system aims to predict the user's rating value for an item. CF algorithm's output can be the prediction of the active user on this item (how much user u_a will like this item) , called 'rating prediction task' or a list of predicted items in which active user might be interested which is called 'top-N recommendation task'.

The standard CF bases on the users' rating values which users have to provide explicitly to provide the items which they may be interested in. The collaborative filtering approaches based on implicit ratings are arousing more and more attention. A very important advantage of collaborative filtering approaches is that they are applicable to various kinds of application areas and entirely independent with the content types of items.

Collaborative filtering can be of 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 of ratings for items. Memory-based CF search for the users with similar interest (neighbors) and combine different algorithms to produce Top-n list of items of interest based on the formed neighborhoods. Memory-based approaches are easy to implement and popular but they do not always have good results. On the other hand, model-based approaches include several model based learning methods. However, these approaches depend on large amount of user input data to produce output. In real world applications, not everyone is keen to provide data. This led to the development of recommender systems which reduce the dependent on the explicit input data and estimate user interest implicitly.

Social tagging systems have become popular 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. People tag resources for future retrieval and sharing. Tags can convey information about the content and creation of a resource. Tags explain what the resource is about and the characteristics of a resource [9]. Therefore, this metadata could also be used to support the RS process and there are previous works done incorporating social tagging information into RS to improve the performance [2, 3, 4, 5, 6, and 7]. In this paper, we present a resource recommendation method which is based on topics which are derived from tagged resources and tags in a social tagging system.

2. RELATED WORK

Popularity of social tagging systems makes them become the rich source of user's interest and preferences indicators.. Tso-Sutter et.al [4] used tagging information as an additional source to extend the rating data, not to replace the explicit rating information. Gemmis et al. [14] integrated tagging information into content-based recommender systems. Liang et al. [13] also proposed to extend the user -item matrix to user -item -tag matrix to collaborative filtering item recommendations. Sen et al. [15] combined users' explicit ratings with the predicted users' preferences for items based on their inferred preferences for tags. Earlier work didn't consider the tag quality problem. The work of Tso-shuter et al. [4] extended the binary user-item matrix to binary user-item-tag matrix and used the Jaccard similarity measure approach to find neighbors. However, in that work, because of the tag quality problem, tag information failed to be very useful to improve the accuracy of memory based CF approaches. Therefore, effective and efficient ways using tags for recommender systems is still in demand. In the work of Niwa et al. [6], a tf-idf weighted tag based item profiles have been used for web page suggestion. Shepitsen et al. [7] applied hierarchical clustering to tag data from social tagging system to provide resources of interest to user. Besides these memory based CF approaches and content filtering models, in the work of Sen et al. [15], a special tag rating function was used to infer users 'tag preferences. Along with the inferred tag preferences, the click streams and tag search history of each user was used to get user's preferences for items. But their proposed system make use of various kinds of extra information such as click streams and search history and make it difficult for comparing it with others and applying in real-world systems. Bogers and Bosch [11] explored a number of recommender approaches for social bookmarking website users that incorporated social tagging data.

In the proposed system, collaborative tagging information in social tagging system will be explored and derive the hidden topics on collection of resources. Users' interest on these topics is measured based on the users' browsing resources and represented as a profile and build a collaborative filtering RS.

2. SOCIAL TAGGING SYSTEM

Social tagging applications allow users to provide, create, and share more information online. Users create and share information they are interested in the form of textural content, multimedia content and social relationship information [9]. Sharing the textural content online is widely popular and this can be in various forms such as tags, blogs, reviews, micro-blogs, comments, posts, documents and others.

Due to rapid growth in popularity and a high degree of activity by their users, social tagging system has become the rich source of user information which shows their interest and ideas to explore today. For instance; del.icio.us, CiteuLike, Bisonomy, last.fm, flickr.com are popular social tagging systems which users share their interested resources with other users and express their interest online. The increasing popularity of these systems demonstrates that users of these systems are requiring recommender systems in order that they can access the resources they are interested in easily. Accordingly, this paper considers how RS can be employed to this social tagging environment and proposes an RS that use fundamental tagging data available implicitly in these systems. The fundamental data available is tagging data and resource information (descriptions, content, etc). Hence, from the two main tasks of RS, the proposed system focuses on top-N recommendation task which suggests a list of items. The reason is that it will be difficult to evaluate rating predictions in such systems that have no explicit user rating data.

In a social tagging system, there are

$U= \{u_1, u_2, \ldots\ldots\ldots, u_m\}$ 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 resources,
$I= \{i_1, i_2, \ldots\ldots\ldots\ldots, i_n\}$ is the set of 'n' resource items tagged by users.

Users' posts and their tagging data are valuable input to users' interest and preference elicitation. The proposed system explores these valuable data and incorporates it to resource suggestion system to improve the system's performance.

3. PROPOSED RESOURCE RECOMMENDER MODEL

The proposed system works in three steps:

- User preference elicitation
- Neighborhood selection
- Recommendation generation

3.1. User Preference Elicitation

In social tagging systems, users do not provide explicit rating on resource items that they interest. Instead of rating on items, users annotate the resources using keywords called tags. Therefore, tags indicate user's interest and preferences. The proposed system will analyze user's tagging behavior and try to estimate user's interest and preferences. Given a collection of 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.

A popular topic modeling approach, LDA (Latent Dirichlet Allocation) [12] is used to derive latent topics from the collection of resources. 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 θ(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.

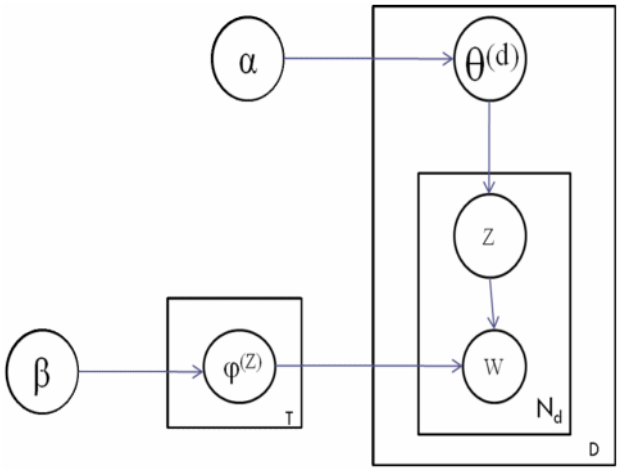

Fig . 1. Probabilistic graphical model of LDA

In social tagging environment, instead of documents (D), users annotate each resource using keywords called tags. Therefore, in order to create topic models using LDA, resources are taken as documents and all of the words in a document (resource) are a set of tags used to describe it by the users. Therefore, each document in social tagging system is a bag of tags used to annotate a resource.

After 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. 2. 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.

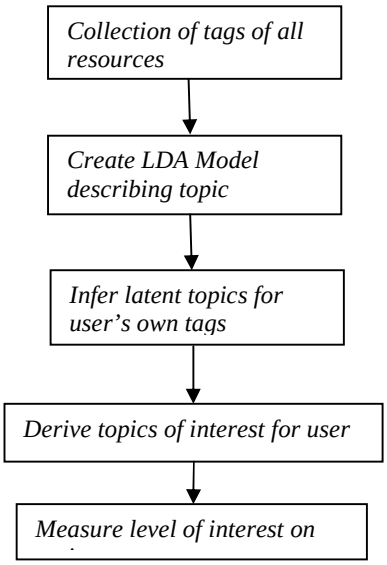

Fig. 2. Topic Modeling Steps

To measure the user's level of interest on a topic, the interest weight of resources for each user is first computed. Interest weight of a resource identifies the user's interest on this resource. Each user 'u' has a set of resources: R (u) = {r1… r_k} of his interest and a set of personal tags: T (u) = {t_1, t2… t_l} which are used to annotate these particular resources by this user. Then, the interest weight of a user for a resource, rs(u,r_j) is

$$rs\,(u, r_j) = \sum_{i \in T_{u,r}} ts\,(u, t_i) \tag{1}$$

Where

rs(u,r_j) = resource score of user 'u' for resource 'r_j',
$T_{u,r}$ = tags used by user 'u' to annotate resource 'r',
ts(u,t_i) = tag score of user 'u' for tag 't_i' which is calculated as

$$ts(u, t_j) = \frac{freq\,(u, t_j)}{\sum_{t_i \in T(u)} freq\,(u, t_i)} \tag{2}$$

Where

ts(u,t_j) = tag score of user 'u' for tag 't_j
freq (u,t_j) = the number of times that user 'u' used tag 't_j'.
Freq (t_l) = total frequency of all tags used by user 'u'.

The assumption is that frequent tags of a user are important to users and its related resources are also important to user and will have high interest weight values.

After calculating resource interest weights, user's interest on latent topics are derived. Each user 'u' has a user profile P represented a vector of his interest topics with its weights,

$$P = \{(b_1, IN\,(u, b1)),\ldots\ldots,(b_k, IN(u, b_k))\} \tag{3}$$

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

$$IN(u, b_k) = Max\{rs(u, r_1), ..., rs(u, r_s)\} \quad (4)$$

where 'r_s' is the resource that belongs to topic ' b_k'. The assumption is that topics related to important resources by the user are also important to user and would have high interest factor values.

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 the system.

3.2. Neighbourhood Selection

Neighbourhood selection is to find users with similar interests for a target user u_i. User topic profiles are matched to measure the interest similarity between users. 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 their resource items to measure similarity between two users. Therefore, user similarity is calculated 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:

$$sim_T(u_a, u_b) = \frac{(T_a \cap T_b)}{|T_a|} \quad (5)$$

(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 7, where I_a represents the item set of user u_a and I_b represents the item set of u_b.

$$sim_R(u_a, u_b) = J(u_a, u_b) = \frac{(I_a \cap I_b)}{|I_a|} \quad (6)$$

(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, Cosine similarity

method is used to measure topic interest similarity called $sim_I(u_a,u_b)$. 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:

$$sim_{a,b} = \frac{\sum_{x \in X(a,b)} (r_{a,x}) * (r_{b,x})}{\sqrt{\sum_{x \in X(a,b)} (r_{a,x})^2 * \sum_{x \in X(a,b)} (r_{b,x})^2}} \quad (7)$$

where $sim_{a,b}$ = similarity between two users, a and b,

X(a,b) = set of topics that both users, a and b, rated, and
$r_{a,x}$, $r_{b,x}$ = rating values for item x by each user a and b respectively.

The final step is to decide which users have the most similar interest with the target user. Since social tagging systems usually have the tagging data as a basic for the user's preference and interest rather than explicit ratings. We investigate how these data can be contributed to RS for resource suggestion. Therefore, to measure the similarity between two users, we test two variations of final similarity calculation method to study more about these implicitly captured users' preferences and interest. The first approach define the similarity sim(u1,u2) by aggregating the three similarity measures above,

$$sim(u1,u2)=sim_R(u1,u2)+sim_T(u1,u2)+sim_I(u1,u2) \quad (8).$$

The second final similarity calculation method is to use resource scores calculated in equation (1) in place of topic interest similarity (sim_I). Since these scores also show user's intensity of interest on his resources, these can be used as user-item rating values. Therefore, the similarity between two users, sim(u1,u2), is calculated as

$$sim(u1,u2) = sim_R(u1,u2)+sim_T(u1,u2)+sim_{RS}(u1,u2) \quad (9)$$

where sim_{RS} is the similarity value calculated by using Pearson Correlation method using resource scores as rating matrix input. The proposed methods are referred to as topic-based method **(TBM)** and resource score-based method **(RSM)** respectively depending on the variations of similarity calculation methods.

3.3 Recommendation Generation

This step chooses resources of similar neighbours for the target user to be recommended. In order to generate recommended list, 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) is the similarity value of user u and his neighbor n.

4. DATASET SPECIFICATIONS

For experimental evaluation, the proposed system and its comparsions are tested with a Delicious.com dataset (HetRec- [11] and a LastFM dataset [11]. These datasets were published for HetRec 2011 Conference and are freely available for research purpose and at the conference website (http://ir.ii.uam.es/hetrec2011/datasets.html). In hetrec-delicious-2k dataset, each user has bookmarks posts, tag assignments, i.e. tuples [user, tag, bookmark], and contact relations

within the dataset social network. Table 1 shows some statistics about the dataset. Delicious.com is a popular social bookmarking service with various kinds of users who have a variety of interests and preferences. As a result, its dataset includes collection of bookmarks and tags which are related to various topics. Therefore, it is suitable for the proposed system which investigate topic-based recommendation task.

Table 1. 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

Another dataset is hetrec-lastfm-2k dataset. Last.fm is a popular social music service. In this dataset, users have listened artists, tag assignments to artists, i.e., tuples [user, tag, artist] and user friend relations. Table 2 shows some statistics about the dataset. The purpose of experiments with last.fm dataset is to study how proposed system performs in a specific domain such as music, movies not only in unspecific domain such as Delicious.com.

Table 2. Data statistics of hetrec-lastfm-2k dataset

Dataset	**Last.fm**
Number of users	1892
Number of Artists	17632
Number of User-artist relations	92834
Number of tags	11946
Number of User-tag-artist	186479
Number of User-user relations	12717

5. EXPERIMENTS AND RESULTS

For experimental evaluation, dataset is divided into two parts: 80% of dataset for training and 20% of dataset for testing. Recall is calculated to measure the performance of recommender. For top-N RS, 'recall' is the number of items in the user's test set that also exists in the top-N recommended items. Therefore, recall is 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.

For performance evaluation, the proposed resource recommender is compared with user-based collaborative filtering system, referred to as UI-IDF-CF, which is one of the approaches investigated by [11]. In UI-IDF-CF, user's items are weighted by inverse user frequencies to recommend resources of interest to social bookmarking users. Another comparison method is tag-based collaborative filtering, referred to as UT-IDF-CF, prensented by [11]. In UT-IDF-CF, user's tags are weighted by inverse user frequencies to recommend bookmark resources to social bookmarking users. Another comparison method is the recommender presented by [16] to alleviate the sparsity problem of recommender system. They use Kullback Libler Divergence method to measure similarity between users. This method is denoted as KL-CF.

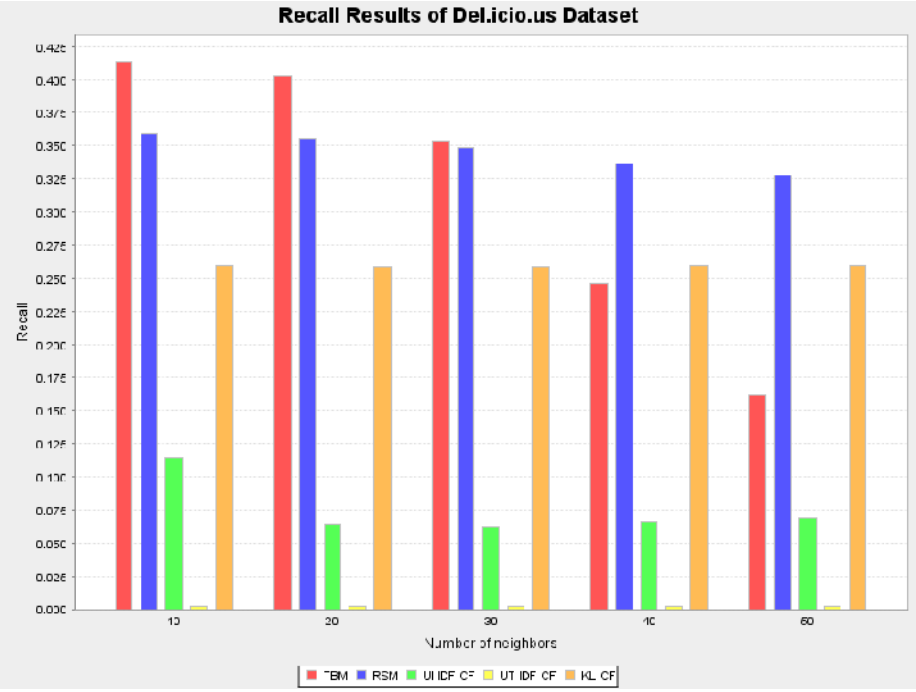

Fig. 3 Recall values of approaches with various number of neighbors (10, 20... 50)

In the performed experiment, the number N of recommended items is set to 100. Numbers of items in the datasets are much larger than the numbers of users. Therefore, setting a small value for N will result poor recall values for all compare methods. And then the system is tested by varying number of neighbors from 10 to 50.

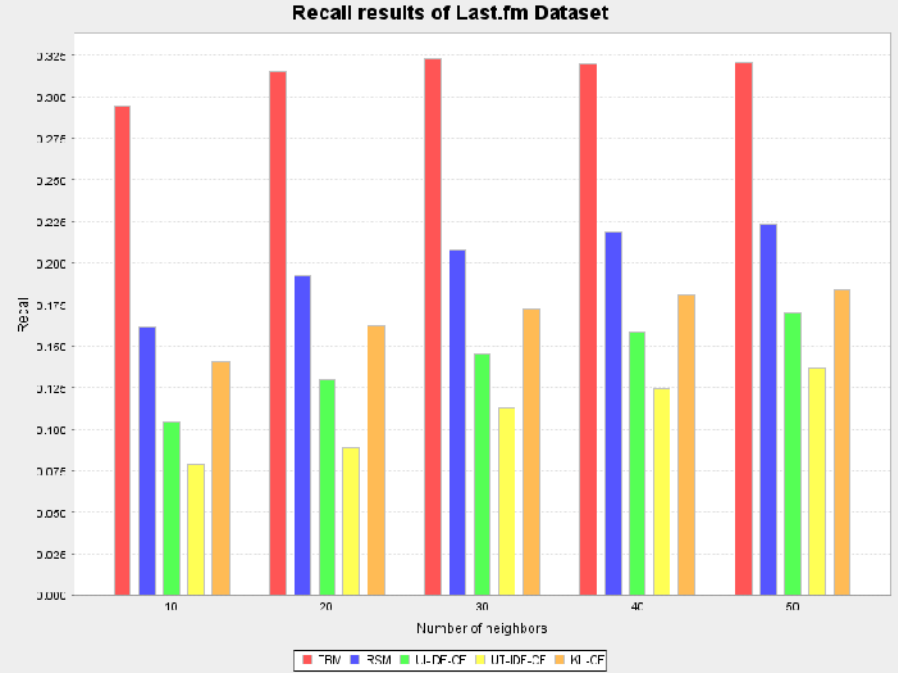

Fig.4 Recall values of approaches with various number of neighbors(10,20,…..,50)

Fig.3 and Fig.4 show the average recall values of our proposed system and comparison methods. When the proposed 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 more than 40 % in average higher than that of comparison methods. If the user has the small number of similar users (neighbors), the derived topics are important to improve the quality of RS' results.

The proposed approaches can perform better than comparison approaches in both datasets. According to the figures, tag-based method, UT-IDF-CF performs worst among all of the methods while KL-based approach can be applied in both dataset. Since Delicious.com dataset is a sparse dataset with unspecific domain, UI-IDF-CF and UT-IDF-CF have poor results with it. With last.fm dataset, UI-IDF-CF and UT-IDF-CF improve their performance and indicate that they are more suitable for specific domains. But their performances are still lower than the proposed approaches.

6. CONCLUSION

In this paper, a recommendation method based on tagging data is presented. The proposed system uses the collaborative tagging information provided by users in a social tagging system and derives user preference topics based on the tagging data by using LDA. Then the proposed system produces the user-topic rating matrix and uses it 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 achieves better results than the other state-of-the-art approaches.

REFERENCES

[1] Adomavicius, G., & Tuzhilin, A. (2005) "Toward 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.

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

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

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

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

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

[7] Shepitsen, A. , Gemmell, J. ,Mobasher, B. and Burke, R., (2008) "Personalized recommendation in social tagging systems using hierarchical clustering", In Proc. Of the 2008 ACM conference on Recommender systems.

[8] McLaughlin, M. R. and Herlocker, J. L., (2004) "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.

[9] Golder, S.A. and Huberman, B. A., (2006) "The structure of collaborative tagging systems", Journal of Information Science, vol. 32, no. 2.

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

[11] Bogers, T. and Bosch, A.V.D. (2009) "Collaborative and Content-based Filtering for Item Recommendation on Social Bookmarking Websites", ACM RecSys '09 workshop on Recommender Systems and the Social Web.

[12] David, M., Blei A. Y. Ng & Michael I. J. (2003) "Latent Dirichlet Allocation", Journal of Machine Learning Research 3.

[13] Liang, H., Xu, Y., Li, Y., Nayak, R., & Weng, L. T. (2009) "Personalized recommender systems integrating social tags and item axonomy", Proceedings of the Joint Conference on Web Intelligence and Intelligent Agent Technology.

[14] Gemmis, M. D., Lops, P., Semeraro, G., & Basile, P. (2008) "Integrating Tags in a Semantic Content-based Recommender", Proceedings of ACM Conference on Recommender Systems, 163-170.

[15] Sen, S., Vig, J., & Riedl, J. (2009) "Tagommenders: Connecting Users to Items through Tags", Proceedings of the 18th International Conference on World Wide Web, 671-680.

[16] Hyunwoo, K., Hyoung-Joo K.(2012) "Improving Recommendation based on Implicit Trust relationships from Tags", The Second International Conference on Computers, Networks, systems, and Industrial Applications, CNSI(2012).

[17] Cantador, I., Brusilovsky, P., & Kuflik, T. (2011) "2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems" (Hetrec 2011), In Proceedings of the 5th ACM conference on Recommender Systems, RecSys 2011, ACM, New York, NY, USA.