Indian Journal of Science and Technology, Vol 9(25), DOI: 10.17485/ijst/2016/v9i25/95143, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Model based Resource Recommender System on Social Tagging Data

V. Vijeya Kaveri* **and V. Maheswari**

Sathyabama University, Chennai - 600119, Tamil Nadu, India; vijerama.kaveri@gmail.com

Abstract

Web (2.0) is the place where people can upload, share and access various sources of information. Web (2.0) has given rise to information overloading problem and knowledge starvation. Recommender Systems (RS) helps in alleviating this overloading problem and gaining the exact information what we need. RS suggest user items or products based on their browsing or purchasing history. RS suggest list of items by identifying similar users with explicit user-item rating. But, in real time applications most users do not rate items. In current web (2.0) social tagging applications help us to find user-item ratings implicitly based on the user's interest and preferences they give for the list of items. In this paper we have proposed a model based resource recommendation on social tagging information which has improved the performance of the RS. In the proposed system the topic is identified from the tagged data, based on the topic user profile is constructed by semantic approach and the recommendation is done for the user.

Keywords: Explicit Rating, Resource Recommendation, Recommender System, Social Tagging, User Profile

1. Introduction

Current Web is a place where lot of information's are shared, uploaded and accessed among the people. The rapid growth of information has lead to the path of overloading and makes the user themselves difficult to gather the knowledge from it. Recommender Systems are the tool/application which help the users to gain knowledge from the information overloading situation. Recommendation is a task where highly relevant items are predicted to the user. Collaborative Filtering (CF) is the most dominant algorithms among the recommending algorithms. This algorithm works in an assumption that previously like-minded users will also share similar taste in future. Similarity measurement plays an important role in CF; the ratings can be contributed only by top-k most similar users and their weights are calculated based on the degree of similarity between the current user and neighborhood. The explicit user-item matrix rating is given as input to CF algorithm. Output of the algorithm is "top-N recommendation task" which gives the 'rating prediction task' or a list of predicted items in which active user might be interested.

Social Tagging System (STS) is also a tool/application in which the user can annotate recourses with tags (Keywords) and categorize content share, upload among them. Tagging helps the user to share or retrieve the resources in future. Generally metadata gives data about data where as in STS tag metadata explains about the resource and the characteristics of the resource[1]. In this paper, we have presented a topic based resource recommendation model derived from tagged resources and tags from a social tagging system.

In[2] the rating data is extended by using tagging information as additional source along with the explicit rating, but over all Recommending System performance is affected by tag quality. In[3] proposed the content-based recommender system by integrating tagging information in it. In[4] proposed collaborative filtering items by extending user-item matrix to user-item-tag matrix. In[5] has proposed social information retrieval based on semantic annotation and ontologies where information's are

*Author for correspondence

extracted from ontologies. In[6] has defined a special tag rating function, combined explicit user ratings with the predicted user preferences for items based on the preferences inferred for tags. In[7] they have suggested the web pages by weighing atf-idf weight based item profiles. In[8] developed a classification tree based on standard long-term HRV for risk assessment in patients suffering from CHF. Using Naïve Bayes Classifier Automatic Classifier is developed. This classifier separates lower risk patients from higher risk ones, using standard long-term Heart Rate Variability (HRV) measures. It achieved the highest performance in terms of accuracy rate and sensitivity. In[9] proposed a STS to provide resources of interest to user by applying hierarchical clustering for tagged data. In[10] explored a number of recommender approaches for social bookmarking website users that incorporated social tagging data. In[11] has proposed Collaborative Filtering RS to extract the hidden topic from the resource collection and interested topics are identified by measuring the similarity[12] of the users based on their browsing resource history. In this paper we have proposed Hierarchical LDA model for extracting hidden topics (tag) and their area of interest are measured by similarity measures form their browsing resource history.

2. Social Tagging System

Online Social Network (OSN) is a place where people can freely share, create and provide information within a community. STS are the tools/applications where users can achieve the same as OSN. Here the information are nothing but resources which be in the form of text content, multimedia content (Image/Audio/Video) and social relationship (FOFA) information[1]. In OSN sharing of textual content has grown more, where the text content can in any one of the forms like tags, review, comments, post, blogs, micro-blogs etc. Due to the active participants of the user in the OSN and overflowing of information, Social Tagging Systems extract the user interested topics from the massive information by the tagged data based on user tag history.

In most of the Social Tagging Systems like CiteuLike, Bibonomy, Flickr.com, Del.icio.us share their interested resources with others and their interest are expressed online. In STS it is difficult to access the resource due to overflowing so it has given a way to Recommending System which can give quick access to the resources by finding their user interest. In this paper we have proposed a RS in Social Tagging environment with implicit rating provided from the tagged data. 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 = {u1, u2,…….,un} is a set of 'n' users.

T= {t1, t2,……,tl} is a set of tags annotated by users to describe resources.

R= {r1, r2,……,rm} is the set of 'm' resource items tagged by users.

3. Proposed Resource Recommender Model

The proposed system works in three steps:

- User Interest Preference Identification.
- Nearest NeighborhoodSelection.
- Recommendation List Generation.

3.1 User Interest Preference Identification

In the proposed system user preferences are identified by topic modeling approach. From the collection of resources latent topics are identified by using Hierarchical LDA (Latent Dirichlet Allocation). This model is approached by arranging the tags into a tree, with the desideratum that more general tags should appear near the root and more specialized tags should appear near the leaves [Hofmann 1999a]. After arranging in desideratum we used probabilistic inference to simultaneously identify the topics and the relationships between them.

Table 1. Variable description for HLDA model

Variable	Description
T	Tag tree
c1,c2,c3 ….. cL	Tags used
Z	Topics (Tags)
D	Documents (Resources)
θ(d)	distribution over tags for a resources
φ(z)	distribution over words for a tag z
W	Collection of words
N	Tags found in each resource
α and β	Dirichlet priors above multinomial distributions

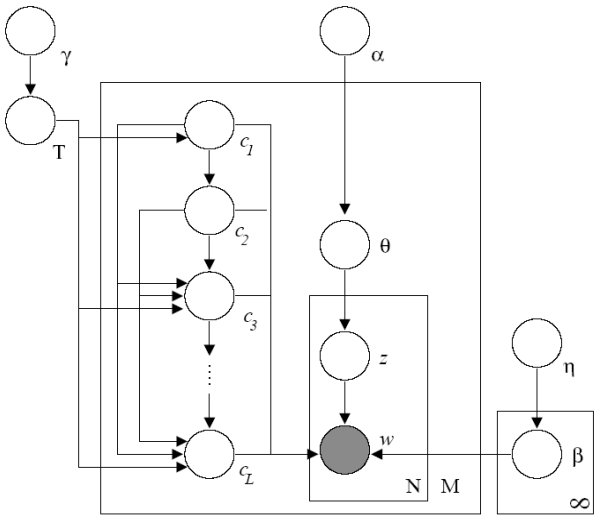

Figure 1. Graphical model for hierarchical LDA.

In STS environment, users annotate each resource usingkeywords called tags instead of Document (D). In HLDA collection of documents are taken as input resource to topic modelling and all the words in a document (resource) are considered as set of tags which is described by the users itself. In this topic model the resources are annotated as with the help of tags identified from each document.

In the proposed HLDA model the user interested topics are extracted as initial phase. After extraction the profile are build on the topic extracted along and user tagging information profile are built on the topics.

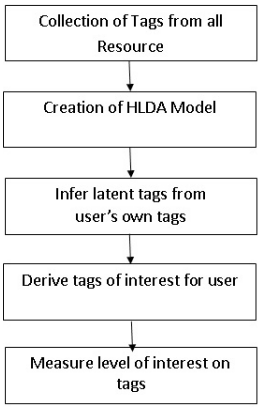

Figure 2. Topic modelling steps.

To measure each user's level of interest on a tag, the interest weight of resource is computed based on the weight the user's interest on this resource is identified .

Each user 'user' has

R (user) = {r1… rm} set of resources of his interest.

T(user) = {t1, t2… tn} set of T personal tags - annotate resources.

rs(user,ri) - interest weight of a user for a resource.

$$rs(user,r_i) = \sum_{j \in Tu,r} ts(user,tj) \quad (1)$$

Where,

rs(user,ri) = resource score of user 'user' for resource 'ri'.

Tuser, r = tags used by user 'user' to annotate resource 'r'.

ts(user,tj) = tag score of user 'user' for tag 'tj' which is calculated as

$$ts(user.t_i) = \frac{freq(user,t_i)}{\sum_{tj \in T(user)} freq(user,t_j)} \quad (2)$$

Where,

ts(user,ti) = tag score of user 'user' for tag 'ti.

freq (user,ti) = No. of times that user 'user' used tag 'ti'.

freq (tn) = total frequency of all tags used by user 'user'.

The assumption made in this model is that higher interest weight values will be obtained for frequent tags used by the user, which shows the importance of the tags as well as the resources related to the tag for the user.

Once the resource interest weights are calculated, latent topics are derived based on user's interest. For each user 'u' user profile UP is build as vector representation of his interest topics and its weights,

UP = {(w1, INF (user, w1)),……..,(wk, INF(user, wk))} (3)

'wk' belongs to set of latent topics and INF (user,wk) is the interest factor of user to this topic. INF factor of user on a topic 'wk' is the maximum of all resource scores of the user related to this topic.

Interest Factor of the user is calculated based on the following formula:

$$INF(user,wk) = \max\{rs(user,r_1),\ldots,rs(user,r_s)\} \quad (4)$$

where 'rs' is the resource that belongs to topic ' wk'. The interest factor values would be high. Since, the

assumption made here is that topics related to important resources by the user are also important to user.

After user profile generation, the input to the system are resulted user interest weight which is consider as implicit user-topic rating matrix.

3.2 Nearest Neighborhood Selection

In Neighbourhood selection Pearson correlation method is used to find users with similar topic interests. User topic profiles are matched to measure the interest similarity between users. In a non-rating environment topic similarity alone cannot give good neighbourhood selection. So in the proposed system the user similarity is calculated based on three similarity measures.

- HLDA Tag similarity.
- HLDA Resource similarity.
- HLDA Interest factor similarity.

For the two users $user_i$ and $user_j$, let Tm and Tn be the sets of tags for each $user_i$ and $user_j$ respectively.

3.2.1 HLDA Tag Similarity

$$Sim_T\left(user_i, user_j\right)=\frac{\left(T_m, T_n\right)}{|T_m|} \quad (5)$$

Equation (5) describes how to measure the HLDA tag usage similarity. The common tags used by the two users, $user_i$ and $user_j$ is used to calculate the tag similarity.

3.2.2 HLDA Resource Item Similarity

$$Sim_R\left(user_i, user_j\right)=J\left(user_i, user_j\right)=\frac{\left(I_i, I_j\right)}{|I_i|} \quad (6)$$

From the Equation (6). HLDA resource item similarity between two users, $user_i$ and $user_j$ is calculated. Here, both of their resource items are considered as two sets and the Jaccard Index is applied between these sets. Equation (7). Describes Jaccard index between two sets. where Ii represents the item set of user ui and Ij represents the item set of uj.

3.2.3 HLDA Topic Interest Similarity

The topic interest similarity is computed based on the User profiles generated between users at the previous phase. In the proposed system Cosine similarity is used to measure the topic interest similarity as SimI (useri, userj).

Let i and j be two users, ri, pbe the rating of user i for topic p and P be the set of topics, rated both by i and j.

Then Pearson correlation coefficient is defined as follows:

$$Sim_{i,j}=\frac{\sum x \in X(i,j)(r_i,x)*(r_j,x)}{\sqrt{\sum x \in X(i,j)(r_i,x)^2 * \sum x \in X(i,j)(r_j,x)^2}} \quad (7)$$

Where,

Sim i,j - similarity between two users, i and j.

X(i,j) - set of topics that both users, i and j rated

ri,x , rj,x - rating values for item x by each user i and j Respectively.

Finally from the similarity measures we have to decide which users have the most similar interest with the target user. Since STS usually use the non-rating variant and it depends only on the tagging data as a basic for the user's preference and interest. We have to investigate how these data can be contributed to RS for resource suggestion.

Therefore, based on the user preferences and interest we have considered two different methods for final similarity calculation. These methods will help to study more about these implicitly captured users' preferences and interest. Based on the variation in the similarity calculation the methods are considered as Topic-Based Method (TBM) and Resource Score-based Method (RSM) respectively.

In TBM method final similarity sim(user1,user2) is calculated by aggregating the three similarity measures above,

sim(user1,user2)=simR(user1,user2)+simT(user1,user2)+simI(user1,user2) (8)

In RSM final similarity is calculated by replacing resource score from equation 1 instead of topic interest similarity SimI. Therefore, the similarity between two users, sim(user1,user2), is calculated as

sim(user1,user2) = simR(user1,user2)+simT(user1,user2)+simRS(user1,user2) (9)

where simRS is the similarity value calculated by using Pearson Correlation method using resource scores as rating matrix input.

3.3 Recommendation List Generation

Recommendation list are generated depending on ranking of an tagged item by choosing the resources of similar neighbor as mentioned by the equation below

$$Rank(user,Ti) = \sum_{x \in Nei(user)} Sim(user,x) \quad (10)$$

Where,

Rank(user,Ti) - Ranking of user and tagged item.

Nei(user) - neighbors of user.

Sim(user,x)- similarity value of user u and his neighborx.

4. Dataset Utilized

The proposed system are tested with two data set of 2K published in HetRec conference available in the website (http://ir.ii.uam.es/hetrec2011/datasets.html). Delicious.com dataset (HetRec[11] and a LastFM dataset[11].

In hetrec-delicious, the dataset is represented as tuples [user, tag, bookmark] and contact relations within the dataset social network. Delicious.com is a popular social bookmarking service with heterogenous user collection with their interest and preferences. The dataset includes bunch of bookmarks and tags of various related topics. This dataset is used in the proposed system for doing the experimental analysis for TBM (Topic-Based Recommendation) method. Table 2 shows the statistics of Delicious.com dataset.

Table 2. 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

In the proposed Resource Score-based Method (RSM) we have done the experimental analysis on a popular social music service last.fm dataset to study about how the proposed system performs in a specific domain such as music, movie etc. The dataset is represented as tuples [user, tag, artist] and user friend relations. Table 3 shows the statistics of last.fm dataset.

5. Experiments and Results

We have utilized 80% of dataset as training set and remaining 20% of dataset for testing purpose. The performance of the recommender system is measured by calculating Recall. Fortop-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,

$$HLDA\,recall = \sum u \mid \frac{HITu}{RELu} \quad (11)$$

Where n is the number of user tested.

We have compared the proposed system with user-basedcollaborative filtering system s UI-IDF-CF and UT-IDF-CF and with KullbackLibler Divergence KL-CF [16].

Table 3. Data statistics of hetrec-lastfm-2k dataset

Dataset	Delicious
Number of users	1892
Number of Artists	17632
Number of User- Artists relations	92834
Number of User-tag-artists	186479
Number of User-user relations	12717

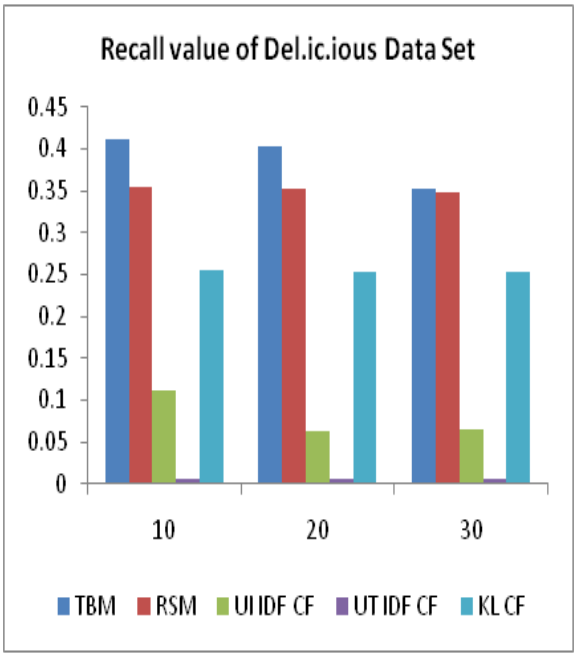

Figure 3. Recall values of approaches with various numbers of neighbours (10, 20 and 30).

Figure 3 and Figure 4 shows that the proposed approaches can perform better than comparison approaches in both datasets.

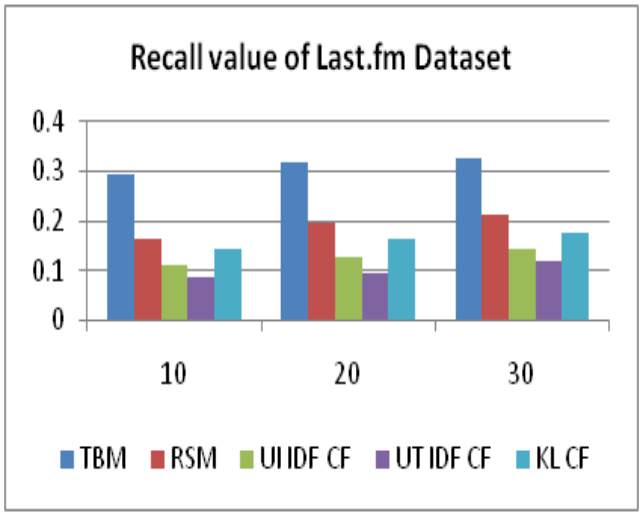

Figure 4. Recall values of approaches with various numbers of neighbours (10, 20 and 30).

6. Conclusion

In this paper we have proposed a recommendation system based on tagging information provided by the user in STS. Where, STS derives the user preferred topics by using HLDA. After topic extraction the implicit rating matrix is generated from a non-rating environment like social book marking as user-topic rating matrix. The top-N recommendation is given to the user by the user-topic rating matrix, which is used in recommender system. From the above, experimental result we conclude that our proposed system achieves better performance than the other systems.

7. Acknowledgements

I would like to express my gratitude to Sathyabama University for providing me with all the resources necessary for the research.

8. References

1. Golder SA, Huberman BA. The structure of collaborative tagging systems. Journal of Information Science. 2006; 32(2).
2. Tso-Sutter KHL, Marinho LB, Schmidt-Thieme L. Tag-aware recommender systems fusion of collaborative filtering algorithms. Proceedings of the 2008 ACM Symposium on Applied Computing; USA. 2008.
3. Gemmis MD, Lops P, Semeraro G, Basile P. Integrating Tags in a Semantic Contentbased Recommender. Proceedings of ACM Conference on Recommender Systems; 2008. p. 163–70.
4. Liang H, Xu Y, Li Y, Nayak R, Weng LT. Personalized recommender systems integrating social tags and item axonomy. Proceedings of the Joint Conference on Web Intelligenceand Intelligent Agent Technology; 2009.
5. Vigneshwari S, Aramudhan M. Social information retrieval based on semantic annotation and hashing upon the multiple ontologies. Indian Journal of Science and Technology. 2015 Jan; 8(2). DOI: 10.17485/ijst/2015/v8i2/57771.
6. Sen S, Vig J, Riedl J. Tagommenders: Connecting users to items through tags. Proceedings of the 18th International Conference on World Wide Web; 2009. p. 671–80.
7. Niwa S, Doi T, Honiden S. Web Page Recommender System Based on Folksonomy Mining. Transactions of Information Processing Society of Japan. 2006; 47(5).
8. Gladence LM, Ravi T, Karthi M. Heart disease prediction using naive bayes classifier-sequential pattern mining. International Journal of Applied Engineering Research. 2014; 9(21):8593–602. ISSN: 0973-4562.
9. Shepitsen A, Gemmell J, Mobasher B, Burke R. Personalized recommendation in social tagging systems using hierarchical clustering. Proc of the 2008 ACM conference on Recommender systems; 2008.
10. Bogers T, Bosch AVD. Collaborative and content-based filtering for item recommendation on social bookmarking websites. ACM RecSys '09 workshop on Recommender Systems and the Social Web; 2009.
11. Htun Z, Tar PP. A resource recommender system based on Social tagging data. MLAIJ. 2014 Sep; 1(1).
12. Shabana AP, Samuel SJ. An analysis and accuracy prediction of heart disease with association rule and other data mining techniques. Journal of Theoretical and Applied Information Technology. 2015; 79(2):254–60.