Tag Recommendation by Link Prediction Based on Supervised Machine Learning

Manisha Pujari and **Rushed Kanawati**
LIPN CNRS UMR-7030
Université Paris Nord
93430-Villetaneuse, France
firstname.lastname@lipn.univ-paris13.fr

Abstract

In this work, we explore applying a link prediction approach to tag recommendation in broad folksonomies. The original idea of the approach is to mine the dynamic of the tagging activity in order to compute the most suitable tag for a given user and a given resource. The tagging history of each user is modeled by a temporal sequence of bipartite graphs linking tags to resources. Given a target user and a target resource, we first compute a set of similar users. The tagging history of the identified set of users is merged in one temporal sequence on bipartite graphs. The obtained sequence is used to learn a model of link prediction in bipartite graphs. The learned model is then applied to predict tags to be linked to the target resource and a list of top similar resources. We get hence several ranked lists tags, one list for each considered resource. These ranked lists are then merged, applying classical preference merging methods in order to obtain the final output: a list of ranked tags that will be recommended to the user. We show through experiments conducted on real datasets extracted for the CiteULike folksonomy the soundness of the proposed approach.

Introduction

Social tagging sites such as Delicious (for web site sharing), CiteULike and Bibsonomy (two sites for sharing bibliographical data) have become a major tools of sharing resources on the Web. In such sites, called also broad folksonomies, users annotate resources (new or existing ones) by a set of user-defined words called *tags*. The most important feature of a folksonomy that makes it different from any other resource sharing applications, is the freedom given to users to select their own tags for annotating resources. This feature gives an advantage of eased cost factor but at the same time, leads to various problems. One key issue to handle is the *tag ambiguity* problem. This refers to the situation of having the same tag being used to index semantically different resources by different users or even by a same user but at different points of time. It may also refer to a condition where similar resources can be indexed by different tags by different users. This witnessed phenomena limits the utility of tags as a mean for sharing new resources. One widely studied approach to cope with this problem is tag recommendation.

Different approaches for tag recommendation computation has been proposed in the scientific literature. Some make use of resources contents (Mrosek et al. 2009). Others relay mainly on analyzing the topological features of the graph induced from the ternary relation linking users to resources they annotate (Jäschke et al. 2008). However, according to our knowledge, no prior work has proposed to mine the evolution of the folksonomy graph in order to compute appropriate tags to recommend.

In this work, we describe a new approach for tag recommendation that we call LiPTaR[1]. The original idea of the approach is to mine the dynamic of the tagging activity in order to compute the most suitable tag for a given user and a given resource. The tagging history of each user is modeled by a temporal sequence of bipartite graphs linking tags to resources. Given a target user and a target resource, we first compute a set of similar users. The tagging history of the identified set of users is merged in one temporal sequence of bipartite graphs. The obtained sequence is used to learn a model of link prediction in bipartite graphs. The learned model is then applied to predict tags to be linked to target resource.

The reminder of this paper is organized as follows. In section , we present a short state of the art of tag recommenders in folksonomy. The LiPTaR approach is presented in section 3. Experimentations and preliminary results are given in section 4. Finally we conclude in section 5.

Related work

Various approaches have been proposed for tag recommendation which can be broadly categorized into two main classes:

- *Content-based approaches* that involve extraction of tags from the content of the resources or titles of the resources. They are efficient in recommending very relevant tags but without taking into account users's choices. These methods cannot be efficient when the resources do not provide a rich content of information.
- *Topology-based approaches* that find tags to recommend by analyzing the graphical structure linking users, tags

[1] Link Prediction for Tag Recommendation

and resources. In this case recommended tags are mostly those, that has already been used in the system.

One first content-based approach has been proposed in (Mrosek et al. 2009) where recommended tags are generated from the content of resources to be tagged. Individual scores are computed based on different informations provided by each resource and then an aggregated global score is calculated for each tag. Tags with top five highest scores are then recommended. Another content-based approach is proposed in (Lu et al. 2009). This is based on the observation that similar web pages usually have same tags. So, each web page can share tags with similar ones. The propagation of a tag depends on its weight in the originating web page and the similarity between the sending and receiving web pages. The similarity metric between two web pages is defined as a linear combination of four types of cosine similarities, taking into account both tag information and page content. In (Lipczak 2009) authors propose yet another content-based approach where recommendation computation is made in a three-step process: tags are first extracted from resource titles. The set of potential recommendations is then extended by related tags proposed by a lexicon based on co-occurrence of tags within resource posts. In the third and the final step tags are filtered by user's *personomy*: a set of tags previously used by the user.

In the category of topology-based approaches, one of the prominent work is given in (Jäschke et al. 2008). Here, authors compare a number of recommendation techniques such as collaborative filtering, *PageRank* and its modified version for folksonomy known as *FolkRank*. They show that the *FolkRank* based recommender outperforms the other two other approaches. They propose two tag recommendation algorithms: an adaptation of user-based collaborative filtering and a graph based recommender built on the top of *FolkRank*. Tests were performed on the dense core of folksonomy, so it may not be very representative. Moreover, they do not take into account the dynamic nature of a folksonomy.

Another work is given in (Zhang, Zhou, and Zhang 2009) in which authors propose a tag recommendation algorithm based on an integrated diffusion on user-item-tag tripartite graphs. Authors propose an algorithm using both user-resource relations and the collaborative tagging information. They emphasize on the fact that two resources, sharing many common tags, have greater probability of being closely related in content. They conclude that the use of tag information can significantly improve the accuracy, diversification and novelty of recommendation. Another graph-based method is FolkDiffusion(Liu, Chi, and Sun 2010) which uses the concept of heat diffusion to rank tags. This method can suggest user and resource specific tags without having topic drift. It uses a graph having users, resources and tags with edge weights representing the relatedness among the three entities. It uses the concept of physical phenomenon of flow of heat from high to low temperature. The user and resource for which tag suggestion is to be made are given a temperature more than zero. All other tags, users and resources are given a temperature equal to zero. The heat is then assumed to flow from target user and target resource to all other nodes according to the edges between them. After a certain number of iterations the heat value on tags show their relatedness to target user and target resource and are accordingly selected for recommendation.

LiPTaR system

The `LiPTaR` approach is based on link prediction in folksonomy graphs. The system takes as input a target user u_t and a target resource r_t. The goal is to compute a list of tags best suited for the user u_t to annotate resource r_t.

Fig. 1 illustrates the general outlines of the tag recommendation cycle applied in our system. The cycle is structured in three main steps :

1. First, the system computes a set of k most similar users $\mathcal{U}_s$ based on their *similarity* to u_t. Many user similarity metrics can be used for this purpose. In the current prototype the top k similar users have been found by application of k-nearest neighbors with a similarity metrics based on both resources and tags used by a user. Another important aspect of this system is that while computing similarity, it takes into account the users's time of activity. So the similar users found have at least one year of activity time common with the target user u_t. Here we explore the idea that users active during same period of time may have common interests and choices.

2. Each user $u \in \mathcal{U}_s$ is associated with a sequence of temporal bipartite graphs relating resources added by user u to tags used by him at various point of time. These graphs are combined to create a single resource-tag bipartite graph for training (G_{learn}). During this process, only graphs corresponding to a time within the duration of training period are used. We have :

$$G_{learn} = \bigcup_{u \in \mathcal{U}_s} \bigcup_{i=t_0}^{t_{learn}} G_i \tag{1}$$

Similarly, G_{label} and G_{test} are generated to be used for examples labeling and validation correspondingly. A couple of nodes (resource-tag pair) that are not linked in G_{learn} but both belonging to the same connected component represent an example (in terms of supervised learning convention). For each such couple of nodes, we compute a set of topological attributes that characterize their roles in the network as well as their *similarity*. The class label for them is obtained by checking whether the couple of nodes is indeed connected in G_{label}. If such a connection is found then it is labelled positive in the supervised learning task. If not it is labelled negative. All the training examples thus found are used by a supervised machine learning algorithm to learn a classification model. This model is then used to predict links in the validation graph G_{test} in order to find probable links between target resource r_t and different tags during validation time period. It does the same for each of the similar resources. At this point, we make an assumption that the tags used by the similar users, for resources that are somehow similar to the target resource, can also be useful for recommendation. In the end, we obtain one or more lists of tags for annotating the resource r_t and other similar resources.

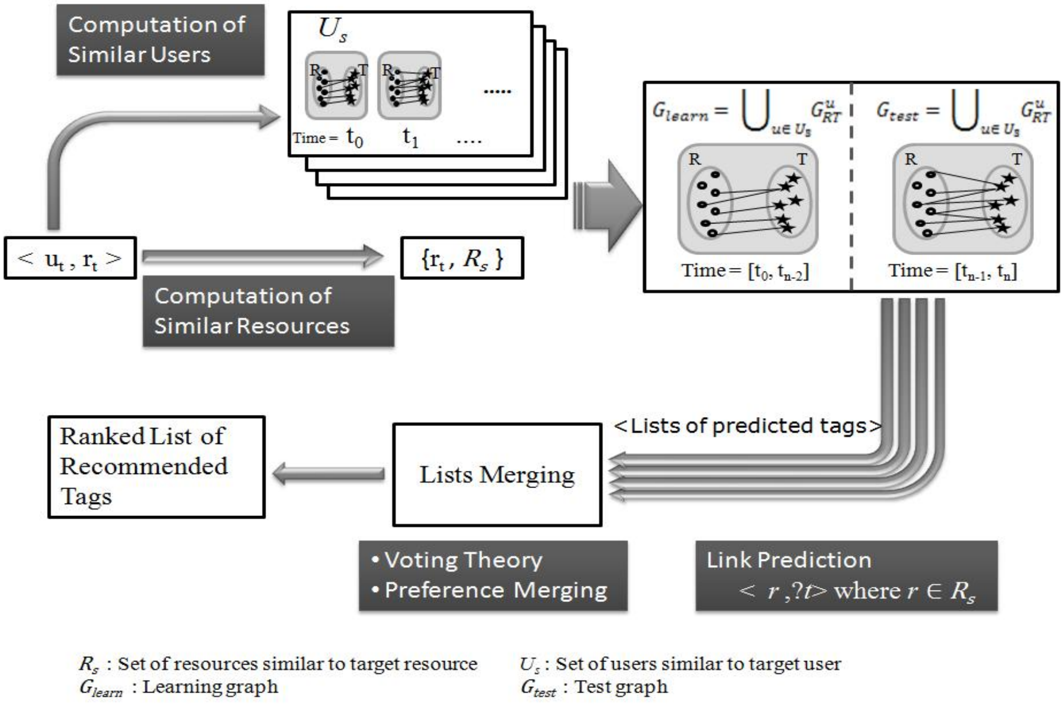
Computation of Similar Users
U_s
R T
R T
Time = t_0 t_1 ....
.....
$G_{learn} = \bigcup_{u \in U_s} G^u_{RT}$
$G_{test} = \bigcup_{u \in U_s} G^u_{RT}$
R T
R T
Time = $[t_0, t_{n-2}]$
Time = $[t_{n-1}, t_n]$
$< u_t , r_t >$
$\{r_t , R_s\}$
Computation of Similar Resources
<Lists of predicted tags>
Ranked List of Recommended Tags
Lists Merging
• Voting Theory
• Preference Merging
Link Prediction
$< r ,?t>$ where $r \in R_s$
R_s : Set of resources similar to target resource
U_s : Set of users similar to target user
G_{learn} : Learning graph
G_{test} : Test graph

Figure 1: LiPTaR work cycle

3. At the end of step 2, we get one or more ranked lists of tags, obtained for r_t and/or a set of similar resources using the data related to retrieved similar users.These lists include both already used tags and predicted tags. We apply a suitable ranked list aggregation approach (Dwork et al. 2001) to merge these lists.

To sum up, the `LiPTaR` approach is conceived as a framework offering three main hotspots to be adapted: a) the user and resource similarity metrics, b) the link prediction approach to be applied to infer tags for recommendation, from the point of view of each retrieved similar user and c) the rank aggregation method to be applied to merge all obtained list of tags computed in step b).

Experiments

We experimented our system on data extracted from CiteULike[2] which is a bibliographic reference sharing website. Like any other folksonomy, users can share their resources with other users and annotate them using their own tags. The dataset covers a time period from year 2004 to year 2010. The total number of data entries are $10, 504, 915$. After pre-processing we get a tripartite graph with $71, 464$ users, $2, 402, 913$ resources and $489, 682$ tags. We use only meaningful tags, discarding the system generated ones. We found that there are $397, 252$ resources without a tag which counts for 16.53% of total resources.

The inputs for our tag recommendation system are a user (target user) and a resource (target resource). We apply a combination of Jaccard's similarity coefficient based on both tags and resources for computing similarity between users. As mentioned before, these users also have some common time of activity. We make use of a modified version of link prediction approach proposed in (Benchettara, Kanawati, and Rouveirol 2010) for predicting new links in the bipartite graph linking resources and tags used by top-k similar users. For computation of resource similarity we use the same Jaccard's coefficient but only based on tags.

At present we are using the following topological measures: product of coefficient of clustering, product of degree centrality, preferential attachment (Barabasi et al. 2002), an indirect computation of number of common neighbors with respect to tag and with respect to resource, shortest path length, measure of Adamic Adar (Adamic, Buyukkokten, and Adar 2003). Finally we use local Kemeny optimal method (Dwork et al. 2001) for list merging which gives us an optimized aggregation and is computationally efficient.

We use data from period of $2005 - 2007$ for training the link prediction model. We use a boosted decision tree classifier (the Adaboost using $J48$ classifier) in the Orange[3] platform. Validation is done on examples constructed from data of period 2006-2008 to predict the tags used in period 2009. (We discarded the data of 2004 and 2010 as they were not complete.) The performance of the system is measured in

[2] http://www.citeulike.org/

[3] http://orange.biolab.si/

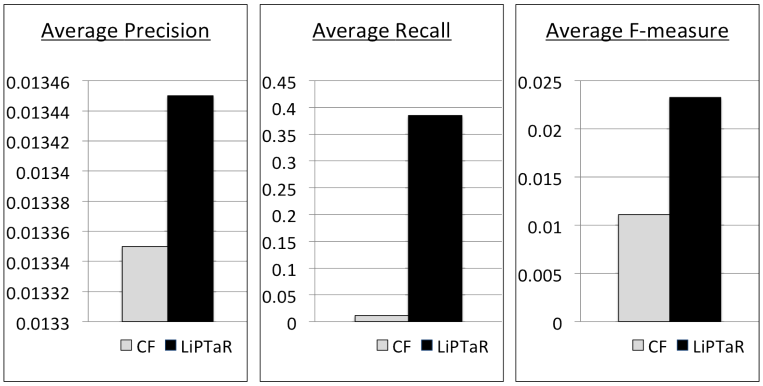

Figure 2: Preliminary Results

terms of precision, recall and F-measure. We experimented on 31 users and a varying number of resources for each of them. The average precision is found to be approximately 0.01345, average recall is 0.385 and average F-measure is 0.02326. The average precision and F-measure may seem to be low. However, low values are due to the fact that we have not restricted the number of predicted tags to be used for recommendation. The result is also affected by the data sparsity of the large scale dataset we are using.

To make a comparison with a classical approach, we experimented with a basic method of tag-based collaborative filtering. The input graph is the union of the temporal resource-tag graphs for top k similar users. We make a prediction of tags for resources used by the target user in validation period of 2009. This prediction is made on the basis of target user's history and the choice of similar users. Using this approach, for the same number of target users and target resources, the average precision is found to be approximately 0.01335, average recall is 0.011 and average F-measure is 0.01113. Fig. 2 shows a comparison between the two approaches. Our approach seems to give a better result as compared to collaborative filtering method which encourages us to continue our experiment further.

Conclusion

We propose, in this paper a new approach of tag recommendation in folksonomy based on a method for link prediction in the bipartite graphs. The approach includes decomposition of a tripartite graph representing a folksonomy, into three bipartite graphs. The proposed approach is implemented as a framework structured around three main hotspots: 1) the similarity metrics to be used for retrieving similar users and similar resources, 2) the link prediction approach to apply and 3) the ranked list merging method to use. Results obtained from applying first implementation of this framework to a real world dataset extracted from a broad folksonomy (where tagging is oriented towards sharing resource within a community) argue for the validity of the approach. Further experiments are required in order to evaluate effects of using more elaborate similarity metrics for retrieving similar users and similar resources. Evaluating different rank aggregation results as well as applying the framework to different types of folksonomies including narrow ones (where tagging is mainly motivated by personal usage).

References

Adamic, L. A.; Buyukkokten, O.; and Adar, E. 2003. A social network caught in web. In *First Monday*, number 6.

Barabasi, A.-L.; Jeong, H.; Néda, Z.; Ravasz, E.; Schubert, A.; and Vicsek, T. 2002. Evolution of the social network of scientific collaboration. *Physica A* 311(3-4):590–614.

Benchettara, N.; Kanawati, R.; and Rouveirol, C. 2010. Supervised machine learning applied to link prediction in bipartite social networks. In *ASONAM'10*, 326–330.

Dwork, C.; Kumar, R.; Naor, M.; and D.Sivakumar. 2001. Rank aggregation methods for web. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, 613–622. Hong Kong: ACM.

Jäschke, R.; Marinho, L. B.; Hotho, A.; Schmidt-Thieme, L.; and Stumme, G. 2008. Tag recommendations in social bookmarking systems. *AI Commun.* 21(4):231–247.

Lipczak, M. 2009. Tag recommendation for folksonomies oriented towards individual users. In *ECML PKDD Discovery Challenge 2009, CEUR Workshop Proceedings Vol. 497*, 189–199.

Liu, Z.; Chi, C.; and Sun, M. 2010. Folkdiffusion: A graph-based tag suggestion method for folksonomies. In *Information Retrieval Technology*, 231–240. Springer Berlin / Heidelberg.

Lu, Y.-T.; Yu, S.-I.; Chang, T.-C.; and jen Hsu, J. Y. 2009. A content-based method to enhance tag recommendation. In Boutilier, C., ed., *IJCAI*, 2064–2069.

Mrosek, J.; Bussmann, S.; Albers, H.; Posdziech, K.; Hengefeld, B.; Opperman, N.; Robert, S.; and Spira, G. 2009. Content-and graph-based tag recommendation: Two variations. In *ECML PKDD Discovery Challenge 2009, CEUR Workshop Proceedings Vol. 497*, 189–199.

Zhang, Z.-K.; Zhou, T.; and Zhang, Y.-C. 2009. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs. *CoRR* abs/0904.1989.