Towards Effective Recommendation in a Social Annotation System Through Group Extraction

Yanhui GU†, Zhenglu YANG†, and Masaru KITSUREGAWA†

† Institute of Industrial Science, the University of Tokyo
4–6–1 Komaba, Meguro–ku, Tokyo 153–8505 Japan
E-mail: †{guyanhui,yangzl,kitsure}@tkl.iis.u-tokyo.ac.jp

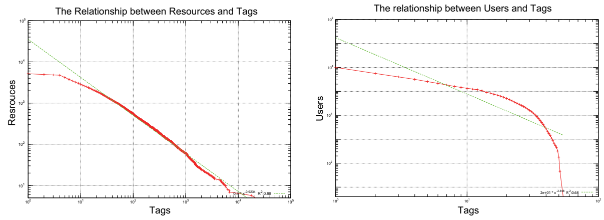
Abstract With the recent information explosion, social websites have become popular in many applications where abundant social data is available. Many social annotation services allow users to annotate various resources with tags, which can facilitate users finding preferred resources. However, in social annotation based recommendation researches, obtaining the proper relationship between user, resource and tag is still a challenge. In this paper, we judiciously extract affinity relationship from between tags and resources and between tags and users. The key idea is to obtain the implicit relationship groups among users, resources and tags and then fuse them to generate recommendation.

Key words relationship groups, bipartite graph, tag clustering, tag-based recommendation

1. Introduction

With the development of Web 2.0 application services, tag-based services, e.g., Delicious, Flickr have undergone tremendous growth in the past several years. A social recommender system uses social information with the traditional recommender systems to make the recommendation results more accurate and serve the users' demand. Different from the traditional recommender system, A tag-based social recommender system has three entities that are considered by the recommender: user, resource, and tag. The user prefers some resources and annotates some words on them. In some cases, the degree in which a user prefers a resource can hardly measured. Therefore, we only obtain the information on the interaction among user, resource, and tag forming a three-dimensional (3D) matrix. However, there is a limitation when few users annotate some resources. An experiment on detecting the pairwise relationship between tags and resources and between tags and users is shown in Fig 1 on MedWorm dataset which will be introduced in experimental evaluation section. As illustrated in the figure, several resources are annotated frequently with many tags, whereas a large number of resources are not. The same powerlaw distribution can also be found from between tags and users. Therefore, determining the relationship among users, resources, and tags is difficult because of the sparse data.

On the other hand, since tags are the personalized keywords annotated on resources by users, they are unsupervised. i.e., a variety of tags that can be redundant, ambiguous or entirely idiosyncratic. Tag redundancy, in which several tags have the same meaning, can obfuscate the similarity among resources. Redundant tags can hinder algorithms that depend on identifying similarities between resources. In addition, tag ambiguity, in which a single tag has many meanings, can falsely give the impression that resources are similar when they are in fact unrelated. Because of this complicated information space by redundancy and ambiguity, a good recommender system is needed to aid the user when they interact with the system. Take, for example, consider the user Tom, who annotated several Chicago Bulls with the ambiguous tag "bulls". The recommender system cannot recommend "Milwaukee Bucks" or "Houston Rockets" to him. In addition, other Chicago Bulls fans may have annotated alternative tags: "Chicago" or "Jordan", etc. These resources may not have been annotated with "bulls", but they should still be make available to Tom. Typically, a social annotation recommender system should have the ability to deal with the three aspects: users, resources, and tags. Consider again on the basketball fan, Tom, the recommender



(a)Tags and Resources (b)Tags and Users

Fig. 1 Power-Law Distribution betweenTags and Resources and between Tags and Users (Dataset: Medworm).

system should have the ability to recommend "Milwaukee Bucks" or "Houston Rockets" to him. So a primary concern of recommendation mechanism in social annotation system is to present users with avenues for navigation that are most relevant to their information needs. So far, we can see that we have to face several problems:

(1) The annotation data may not properly capture the interests of the user because it only has the binary relationship.

(2) Due to the ambiguity of the annotation [18], we cannot easily distinct the topics which the tags present.

(3) In the annotation services, tags and resources follow the power law distribution, which indicates that the data is very sparse. All of these problems hinder the applicability of the traditional collaborate filtering algorithms.

In this paper, we judiciously extract information from between tags and resources and between tags and users then form several communities where we call **Groups**. Through groups, users can obtain their desire information more accurately. We extract two kinds of groups based on different affinity relationships, Topic-Groups based on the bipartite relationship between tags and resources; and Interest-Groups based on the bipartite relationship between tags and users. To obtain these latent relationships, it is better to organize the different social relational groups. We regard that, by fusing these two kinds of groups, we can obtain the more accurate recommendation.

Some researchers have worked on integrating the social relationship into the recommender system [4], [10], yet they just considered the single relationship, such as friendship data [5]. Prior approaches are lack of exploration of other possible ways of illustrating the latent relationship among users, tags and resources more effectively. As the social relationship is inherently in the bipartite graph [12], [17] of tags and resources and tags and users, we extract such relationships through this graph.

The contributions of our paper are as follows:

- We address the problem of extracting the group information from the tags based on the bipartite graph between tags and resources or between tags and users to obtain the latent social relationship. We call these groups as Topic Groups and Interest Groups.

- We propose the group formulation approach of clustering tags based on such group information. Through this, we can deal with the redundancy of the tags.

The remainder of this paper is organized as follows. We introduce the preliminaries in Section 2. The group extraction and organization solutions are presented in Sections 3. Section 4 introduces the related work and Section 5 concludes the paper.

2. Preliminaries

2. 1 Social Tagging System Model

In this paper, our work is to deal with tagging data. A typical social tagging system has three types of objects, users, tags and resources which are interrelated with one another. Social tagging data can be viewed as a set of triples [6], [7]. Each triple (u, t, r) represents user u annotate a tag t on a resource r. A social tagging system can be described as a four-tuple, there exists a set of users, U; a set of tags, T; a set of resources, R; and a set of annotations, A. We denote the data in the social tagging system as D and define it as: $D =< U, T, R, A >$. The annotations, A, are represented as a set of triples containing a user, tag and resource defined as: $A \sqsubseteq \{< u, t, r >: u \in U, t \in T, r \in R\}$. Therefore a social tagging system can be viewed as a tripartite hypergraph [13] with users, tags and resources which are represented as nodes and the annotations are represented as hyper-edges connecting one user, tag and resource.

2. 2 Standard Recommendation Model in Social Tagging System

Standard social tagging system may vary in the ways of their ability of handling recommendation. In previous researches, the possible approaches include recency, authority, linkage, popularity or vector space models. In this paper, we conduct our work on the vector space model, which is derived from the information retrieval principle. Under such scheme, each user, u, is modeled as a vector over the set of tags, where $w(t_i)$ denotes the weight, in each dimension corresponds to the relationship of a tag t_i with this user, u, $\vec{u} =< w(t_1), w(t_2), ..., w(t_{|T|}) >$. Likewise each resource, r, can be modeled as a vector over the set of tags, $\vec{r} =< v(t_1), v(t_2), ..., v(t_{|T|}) >$. Some work [9], [15] use the tag frequency, $F_t =| a =< u, r, t >\in A : u \in U |$, to calculate the weight of the vector. After that, similarity calculation techniques such as the Jaccard similarity coefficient or Cosine similarity are applied to obtain the similarity scores between users or resources via various recommendation strategies, such as user-based or resource-based recommender system. In this paper, we mainly adopt the binary expression of tags, i.e. the user annotation activities. We take the occurrence into consideration only. That means if a user, u, annotate a tag, t, on a resource, r, w will be "1" in this 3D matrix, "0" otherwise.

3. Group Extraction and Fusion Strategy on Social Recommender System

The framework of our **Group Extraction and Group Fusion Social Recommender System** is illustrated in Fig 2. It mainly includes three steps to obtain the final

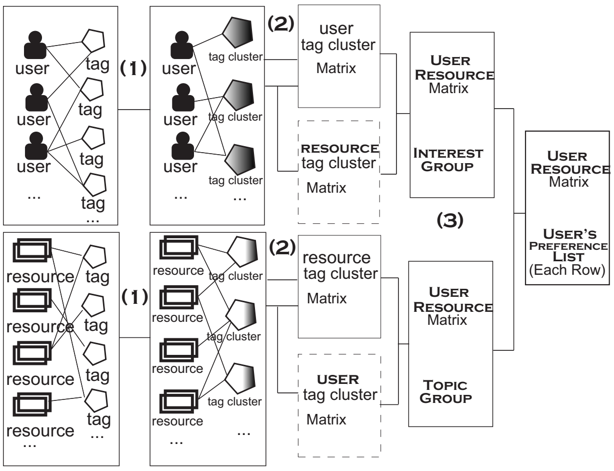


Fig. 2 The framework of Group Extraction and Group Fusion Social Recommender System.(Three Steps:(1) Tag Clustering for tackle tag redundancy; (2) Group Extraction to organize two different groups; (3) Group Fusion for the generate the Users' Preference List on resources.)

recommendation-user's preference list. The following, we will introduce our approach step by step.

3. 1 Affinity Graph-based Tag Clustering based on Bipartite Graph

The first step is tag clustering which can tackle tag redundancy. After clustering, community relevant partitions are created and tags are categorized into clusters which can relieve the redundancy. In addition, ambiguous tags can also be abstracted into their desired meaning. In this step, the tag-resource or tag-user relationships can be represented by a bipartite graph as modeled in Fig 3. We illustrate such procedure by using a tag-user model. Based on the bipartite graph between users and tags, we can build a user-tag matrix. Each tag vector over users denotes that whether this tag has been annotated by a user. After that, we apply cosine similarity computation on this matrix. So far, we get the tag similarity matrix which is also illustrated in Fig 3. Upon such similarity matrix, we obtain an affinity graph of tags where the edge indicates the similarities of pair tags. After affinity graph of tags has been built, we employ the Hierarchical Agglomerative Clustering algorithm [15] for tag clustering. The input to the clustering module is a set of tags, an adjustive factor step and a division coefficient. Tags are represented as a vector of weights over a set of resources or a set of users. With the length limitation of this paper, we do not introduce how the Hierarchical Agglomerative Clustering works. After this step, we obtained a set of tag clusters based on users pairs. Likewise, we can also obtain cluster sets based on resources pairs.

3. 2 Group Extraction

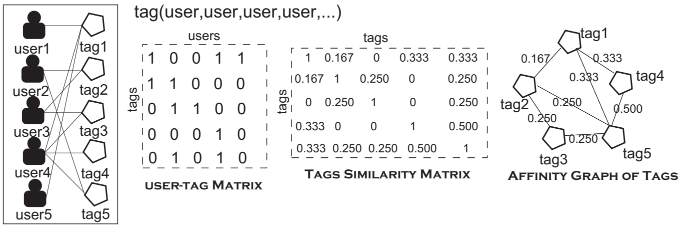


Fig. 3 The illustration on affinity graph generation. (An example on tag-user model.)

By clustering the tags, each user or resource can be re-presented by a vector over tag clusters. For instance, after performing this procedure by clustering the tags based on tag-resource matrix, we obtain two vectors $\vec{u} =< w(Rtc_1), w(Rtc_2), ..., w(Rtc_{|TC|}) >$ and $\vec{r} =< v(Rtc_1), v(Rtc_2), ..., v(Rtc_{|TC|}) >$, where $w(Rtc_i)$ and $v(Rtc_i)$ denote the different weight over two different types of clusters. Furthermore, we compute the similarity between **u**, **r** to form a matrix $\mathbf{UR}_R$ whose element represents the relationship between users and resources at the granularity level of tag aggregations. We can also call Topic-Group matrix, that is:

$$\mathbf{UR}_R = \begin{pmatrix} t_1(r_1) & t_1(r_2) & ... & t_1(r_{|R|}) \\ t_2(r_1) & t_2(r_2) & ... & t_2(r_{|R|}) \\ ... & ... & ... & ... \\ t_{|TC|}(r_1) & t_{|TC|}(r_2) & ... & t_{|TC|}(r_{|R|}) \end{pmatrix} \quad (1)$$

Likewise, we can also obtain the matrix $\mathbf{UR}_U$, which means the matrix is extracted from the information of tags and users. The procedure is Step (2) in Fig 2.

3. 3 Group Fusion

By far, we obtain two matrixes, $\mathbf{UR}_R$ and $\mathbf{UR}_U$. We term $\mathbf{UR}_R$ which is focused on topic group and $\mathbf{UR}_U$ on interest group. Now we fuse these two group information by a tunable factor $\lambda \in [0, 1]$, that is:

$$\mathbf{UR} = (1 - \lambda)\mathbf{UR}_R + \lambda\mathbf{UR}_U$$

If $\lambda = 0$, which means such fusion is weighted on topic group only while $\lambda = 1$ indicates interest group only. In this paper, we empirically investigate the impact of choosing different λ on recommendation.

3. 4 The Recommender Generation

The final step is to generate the recommendation. The matrix **UR** denotes the user-resource preference information. For each user vector over resources that means the affinity relationship between this user and those resources. For example, $\vec{u} =< w_{r_1}, w_{r_2}, ..., w_{r_{|R|}} >$ denotes the weight what a user prefer a resource. For a giving user, we sort the value in each vector and push top-N values as the recommendation.

4. Related Work

We review the related literature of group extraction and fusion to improve the standard recommender system from three perspectives:

- **Tag Clustering**

The efficiency of tag clustering is the ability of aggregate tags into topic domains. [3], [14]demonstrated how tag clusters serving as coherent topics can aid in the social recommendation of search and navigation. In [8] topic relevant partitions are created by clustering resources rather than tags. By clustering resources, it improves recommendation by distinguishing between alternative meanings of a query. While in [1], clusters of resources are shown to improve recommendation by categorizing the resources into topic domains. Our work is orthogonal to such works but different that our approach is to cluster tags. We believe that tags can hold richer social information rather than resources can.

- **Information Extraction from Bipartite Graph**

We studied several works which are related to our proposal. [17] using bipartite graph to do neighborhood formation and anomaly detection. [12] also conduct the personalized query recommendation based on the Query-URL bipartite graph. In social tagging system, because of the triple relationships, it is difficult to extract the social information directly. i.e., We cannot apply their strategies to our recommender systems. Therefore, in this paper, we extract our group information based on the bipartite graph which are tags-resources and tags-users respectively.

- **Fusion of Social Relationship**

Recently, with the increasing development of social websites and appearance of social data, researchers have begun to pay attention to the social data and explored its usage in recommender systems. [4] used social network data for neighborhood generation. [11] adopted Random Walk with Restart to model the social tagging in a music track recommendation system. In addition, [10]proposed an online social recommender system attempting to use more social information for recommendation generation. All the work show that, their fusion social information can benefit the recommender system. However, their work mainly focus on friendship, i.e. the similarity between users'. Compared to friendship, other community relationships, **group** in this paper, contains more information about users' activities [2], [16]. Our proposal is orthogonal but fuse the two different groups to obtain better recommendation.

5. Conclusion and Future Work

In this paper, we presented an approach on Group extraction and Group fusion in the social tagging system. In traditional tagging system, the recommendation generated only based the similarities between the pairs of every two objects. But in this paper, we find that, by clustering the tags based on the bipartite graph of tags and resources, tags and users, groups can be detected. After the groups were extracted from the tag information, we conducted the preferences which user preferred on resources by going through such groups. Our adjustable factor λ control the effect of two kinds of groups. However, our tag clustering was yet based on the bipartite graph only. The weight is only based the occurrence between the tags and resources, or users pairs. An efficient tag clustering strategy can also promote the accuracy of the group extraction for the better precision of our approach.

For the next step, we are interested in future exploring the effect of latent social relationship on recommender system, for example, the time series. We can see that the triple of users, tags and resources and their annotation relationships combining by users' activities form a big graph. Regarding such consideration, we believe that the graph will evaluate [19] as time goes by. So, time information fusion can be promoted for better precision of the social recommender system.

References

[1] Hao Chen and Susan Dumais. Bringing order to the web: automatically categorizing search results. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152, New York, NY, USA, 2000. ACM.

[2] Jilin Chen, Werner Geyer, Casey Dugan, Michael J. Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson, and Saul Greenberg, editors, *CHI*, pages 201–210. ACM, 2009.

[3] J. Gemmell, A. Shepitsen, M. Mobasher, and R. Burke. Personalization in folksonomies based on tag clustering. In *Proceedings of the 6th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems*, July 2008.

[4] Georg Groh. Recommendations in taste related domains: Collaborative filtering vs. social filtering. In *In Proc ACM Group07*, pages 127–136, 2007.

[5] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, New York, NY, USA, 2005. ACM.

[6] Ziyu Guan, Jiajun Bu, Qiaozhu Mei, Chun Chen, and Can Wang. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *SIGIR*, pages 540–547, 2009.

[7] Ziyu Guan, Can Wang, Jiajun Bu, Chun Chen, Kun Yang, Deng Cai, and Xiaofei He. Document recommendation in social tagging services. In *WWW*, pages 391–400, 2010.

[8] Conor Hayes and Paolo Avesani. Using tags and clustering to identify topic-relevant blogs. In *International Conference on Weblogs and Social Media*, March 2007.

[9] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Folkrank: A ranking algorithm for folk-

sonomies. In *Proc. FGIR 2006*, 2006.

[10] Hans G. K. Hummel, Bert Van Den Berg, Adriana J. Berlanga, Hendrik Drachsler, Jose Janssen, Rob Nadolski, and Rob Koper. Combining social-based and information-based approaches for personalized recommendation on sequencing learning activities. *Int. J. Learn. Technol.*, 3(2):152–168, 2007.

[11] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M. Jose. On social networks and collaborative recommendation. In *SIGIR*, pages 195–202, 2009.

[12] Lin Li, Zhenglu Yang, Ling Liu, and Masaru Kitsuregawa. Query-url bipartite based approach to personalized query recommendation. In *AAAI*, pages 1189–1194, 2008.

[13] Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, *ISWC 2005*, volume 3729 of *LNCS*, pages 522–536, Berlin Heidelberg, November 2005. Springer-Verlag.

[14] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266, New York, NY, USA, 2008. ACM.

[15] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin D. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys*, pages 259–266, 2008.

[16] Ellen Spertus, Mehran Sahami, and Orkut Buyukkokten. Evaluating similarity measures: a large-scale study in the orkut social network. In *KDD*, pages 678–684, 2005.

[17] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, pages 418–425, 2005.

[18] Kilian Q. Weinberger, Malcolm Slaney, and Roelof van Zwol. Resolving tag ambiguity. In *ACM Multimedia*, pages 111–120, 2008.

[19] Zhenglu Yang, Jeffrey Xu Yu, Zheng Liu, and Masaru Kitsuregawa. Fires on the web: Towards efficient exploring historical web graphs. In *DASFAA (1)*, pages 612–626, 2010.