International Journal of Hybrid Information Technology
Vol. 9, No. 5 (2016), pp. 61-70
http://dx.doi.org/10.14257/ijhit.2016.9.5.06

Tag-Based Interest-Matching Users Discovery Approach in Online Social Network

Sheng Bin[1], Gengxin Sun[2], Peijian Zhang[2] and Yixin Zhou[2]

[1]*Software Technical College of Qingdao University, Qingdao, China*
[2]*International College of Qingdao University, Qingdao, China*
binsheng@qdu.edu.cn

Abstract

With the popularity of online social networks, such as Facebook, YouTube and Douban, have become a channel for fast information propagation and influence. The problem of discovering common interests shared by groups of users has become one of the central researches because it helps to connect users with common interests and can eventually influence as many users with common interests in the online social network as possible. Unfortunately, most of existing related works have only focused on the network topologies and so unable to identify the common interests of users who have no online connections. In an online social network, users tend to use descriptive tags to annotate the contents that they are interested in. User-generated tags are consistent with the content they are attached to. Thus, patterns of frequent co-occurrences of user tags can be used to characterize and capture topics of user interests. In this paper, we propose a novel common interests discovery approach based on user-generated tags to discover the set of interest-matching users whose interests are similar. The experiment results show that our approach can effectively discover user communities with common interests no matter if they have any online connections.

Keywords*: Online social networks, tag, interest-matching, common interests*

1. Introduction

With the recent rapid growth of Online Social Networks (OSNs), such as Facebook, YouTube and Douban, has made them become one of the most important channels for fast information propagation and influence. In online social networks, users self-organize into different communities to share the common interests and contents, so discovering common interests shared by users is a fundamental problem in online social networks. It is the essential function of identifying user communities of the same interests, detecting the domain experts in different subjects and recommending personalized relevant contents.

Currently, there are two kinds of existing approaches to discover common shared interests in online social networks. One is use-centric which focuses on detecting common interests based on the social connections among users, the other one is object-centric, which detects common interests based on the objects fetched by users in the same social community. Most of the existing user-centric approaches mainly depend on the network topologies and ignore the implicit factors. Schwartz [1] and Hasan [2] analyzed user's social network connections to discover users with particular interests or expertise for a given user. However, for some online social networks such as Douban, social connections among users in networks are hard to identify. Different from this kind of approaches, our approach aims to find the users who share the same interests no matter whether they are connected with each other in a network or not. For the object-centric approaches, Sripanidkulchai [3] and Guo [4] explored the common interests among users based on the objects they fetched in peer-to-peer networks. However, without other information of the objects, it cannot differentiate the various common interests on the same object.

ISSN: 1738-9968 IJHIT
Copyright © 2016 SERSC

Furthermore, not all of objects are popular. So it is difficult to discover interest-matching users on them. Our approach focuses on detecting interest-matching users by taking advantage of user tags. Through user tags, the related objects and the users would cluster under the same topic, it would remove the limitation of the object-centric approach.

Tag techniques have been widely used in different online social networks, but so far there have been few researches on retrieving user interests from tags in online social networks. Golder found that in a given online social network the proportion of frequencies of tags tend to stabilize with time due to the collaborative tagging by all users [5]. Halpin proposed that the distribution of frequency of tags for online social network follows the power law and a model of collaborative tagging to explain how power law distribution could generate [6]. Brooks clustered blog articles that share the same tag, and analyzed the effectiveness of tags for blog classification. They found that the average pairwise cosine similarity of articles in tag-based clusters is only a little higher than that of randomly clustered articles [7]. Different from above works, our approach is based on the co-occurrence of multiple tags, thus can identify common interests and cluster similar objects more accurately.

2. Data Collection and Pre-Processing

Douban is a popular Chinese online social network and media comments platform, whose users can indicate their preference and comments on particular media items, such as books, movies and music. Our data used in this paper is a partial dump of the Douban database representing activity during a limited period of time. In Douban, a media item has some keywords which can be used to describe sententiously the content and categorize the media item. A user can create tags for a media item that he wants to share with other users. The tags can later be used for searching, sharing and categorizing the media items. Users can add their own tags to the same media item independently, called collaborative tagging.

In our dataset, there are 2.9 million tags saved by 32451 users on 115460 media items. After crawling user tags, we used our own stop word dictionary to filter out all stop words in user tags. Then, we normalized the remaining tags and keywords with the Porter stemming algorithm. After normalization, the vocabulary of tags contains 278150 distinct tags, while the vocabulary of media item keywords contains 4572215 unique words.

After data pre-processing, some simple statistical characteristics are firstly analyzed. The distribution of frequencies that the media items were accessed in our dataset is shown Figure 1.

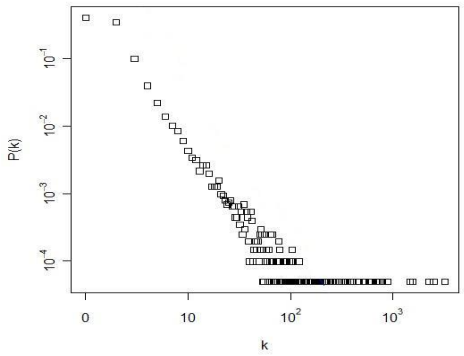

Figure 1. The Distribution of the Media Items Accessed Frequency

 Copyright © 2016 SERSC

In Figure 1, the points are nearly in a straight line in the log-log scale, which indicates that the distribution follows the power law. It means that most media items are rarely accessed, which only a small number of the media items are frequently accessed. It is consistent with the Zipf-like distribution of Web object popularity [8].

Figure 2 shows the distribution of the keywords appearance in our dataset. The long tail of this distribution in the log-log scale means that most keywords are less appearing while a few keywords are highly appearing.

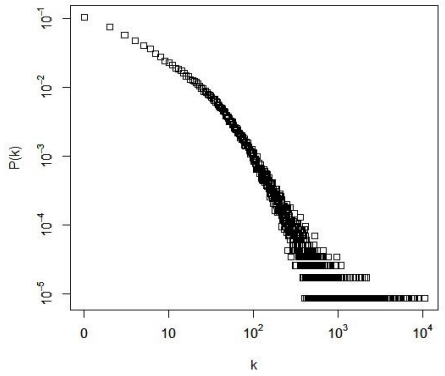

Figure 2. The Distribution of the Keywords Appearance Frequency

Figure 3 shows the distribution of tag frequencies in our dataset. From Figure 3 we can see that the use of tags also follows power law distribution, which means the selection of tags is highly concentrated. The most popular tag was used more than 160000 times by different users' altogether.

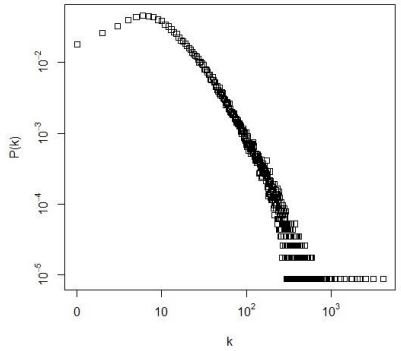

Figure 3. The Distribution of the Tag Frequency

Above analysis results show that the top popular tags connect closely most of the users and media items, which motivate us to utilize tags to discover common interests among users in Douban, where most users are inactive and media items are unpopular.

3. Discovering Common Interests with Tags

In this section, vector space model (VSM) [9] is used to describe a media item, which is a standard technique in information retrieval. A media item can be represented with two vectors, one in the space of all tags and the other one in the space of all media item keywords.

Copyright © 2016 SERSC

In VSM, a corpus with t terms and d documents can be represented by a term-document matrix $A = (a_{ij}) \in R^{t \times d}$. Each column vector $a_j (1 \le j \le d)$ corresponds to a document j. Weight a_{ij} represents the importance of term i in document j.Let f_{ij} be the frequency of term i in document j.The tf -based weight of a term i in document j can be defined as:

$$a_{ij}^{tf} = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{t} f_{kj}^2}} \tag{1}$$

The $tf \times idf$ -based weight of a term i in document j can be defined as:

$$a_{ij}^{tf \times idf} = \frac{b_{ij}}{\sqrt{\sum_{k=1}^{t} b_{kj}^2}} \tag{2}$$

where b_{ij} is defined as

$$b_{ij} = f_{ij} \cdot \log\left(\frac{d}{D_i}\right) \tag{3}$$

D_i is the number of documents that contain term i, and $\log(d/D_i)$ is called inverse document frequency (idf).

3.1. The Vocabulary of Tags and Keywords

Keywords of a book in Douban, which is about Linux operating systems are shown in Table 1.

Table 1. The tf and tf×Idf Keywords and User-Generated Tags of a Book

Top tf keywords	computer, programing, linux, operating system, software, network
Top $tf \times idf$ keywords	domain, server, linux, function, operating system, computer
Tog Tags	linux, dns, open source, unix, operating system, gcc

We show the top-6 keywords using both tf and $tf \times idf$ approaches. Along with them, top-6 tags that have been attached to this book by all users are also listed. From the compare of keywords with tags in Table 1, we can see that the tags and keywords express the same content of the book. Both tf and $tf \times idf$ keywords contain terms such “linux”, “operating system”, “computer”, and so on. On the other hand, user-generated tags have a higher-level abstraction and extended meaning on the content, thus, because of its higher-level abstraction, the tags are closer to the people’s understanding of the content than the keywords. For example, “open source” and “gcc” together carry the main difference of Linux from other operating systems. Moreover, the terms such as “computer”, “network”

 Copyright © 2016 SERSC

are in fact unrelated to the true purpose of this book, these keywords will not make any sense in finding user' interesting book. This simple example shows that intuitively, tags are more appropriate to represent human being's interests about media content.

Before using tags for capturing common interest, it is necessary to examine the vocabulary of the user-generated tags as compared with the vocabulary of keywords. Given a book, we are interested in seeing if the most important characteristics of the book have all been covered by the vocabulary of user-generated tags.

The coverage of user-generated tags for the tf keywords of 8000 random books in our dataset are shown as Figure 4. The cumulative distribution function (CDF) of the percentage of the missed keywords by the tags is plotted.

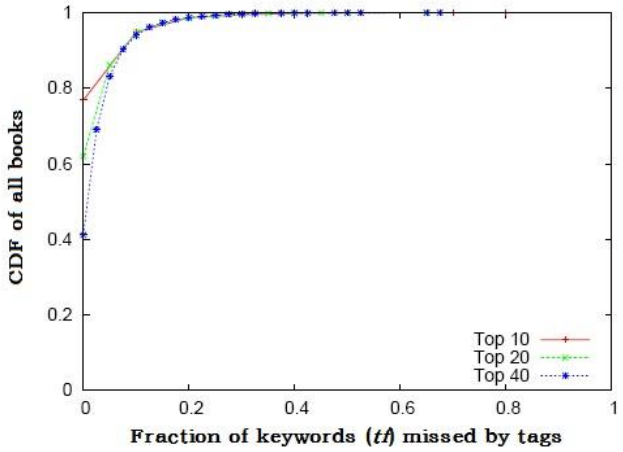

Figure 4. Tag Coverage for tf Keywords

The three plots in Figure 4 are for different amounts of top keywords, when Top-10 tf keywords chosen, 75% of all books are fully covered by user tags. Cumulatively, the cases where the set of user tags missed at most 2 keywords accounts for 98.2% of all sampled books. Similar conclusion can be drawn for Top-20 plot and Top-40 plot. Overall, the cases where user tags missed at most 15% of the keywords accounts for more than 98% of all sampled books. From Figure 4, we can see that the vocabulary of user-generated tags can cover the main keywords of books.

3.2. Tag Match Ration

In this section, statistical analysis about the correlation between the tags of a media item and its content is presented. The total number of occurrences of a tag in our dataset is acted as the weight to characterize the importance of the tag. In an online social network like Douban, most users have the motivation to use descriptive tags for summarizing and sharing with other users. So for a given set of tags for a media item, the matching on a popular tag is more significant than the matching on an unpopular tag.

Let $T = t_i$ be the set of tags attached to a given media item U by all the users. Let $w(t)$ be the weight of tag t .The tag match ration $e(T, U)$ represents the ratio of tags of this media item that can be matched by the content, it is defined by the following equation:

$$e(T, U) = \frac{\Sigma_{k|t_k \in U} w(t_k)}{\Sigma_i w(t_i)} \tag{4}$$

Copyright © 2016 SERSC

where the numerator measures the total weight of the tags that have also appeared in the keyword set of U .

The distribution of tag match ratio for media items in our dataset is shown in Figure 5.

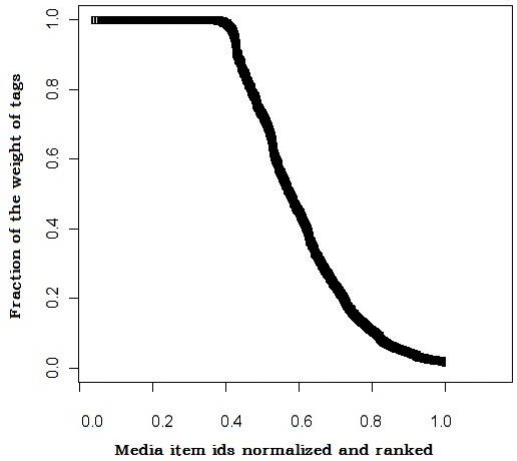

Figure 5. The Distribution of Tag Match Ratio

As shown in Figure 7, the tag match ratio of nearly 50% of all media items in our dataset is 1, meaning for these media items can be covered by all tags. More than 60% of all media items have a tag match ratio greater than 0.5, while only 5% of all media items have no matched Tags.

3.3. Relationship between User's Common Interest and Tags

In online social network, the keywords that can categorize the media items reflect the interest of the user. On one hand, if a user repeatedly access the similar media items, then we can say that the user have interest on the content. On the other hand, we have shown that in most cases, user-generated tags capture the content of a media item, and tags can be more concise and closer to the users' understanding. Based on above reasons, we would believe that tags can be used to represent the content of media items and hence the interest of users. When multiple tags are frequently used together, they compose a topic of interest.

Our ultimate aims are to find the sets of tags that are shared by many users on many media items. If a set of tags are frequently used by many users, then we can think that these users spontaneously form a community of interest, even though they may not have any online connections in the online social network. The tags represent the common topics of interests of these users and the media items tagged by these users represent the commonly interested contents to this community. Therefore, the assignment of discovering common interest for users is to extract frequently used tags and cluster the media items and users based on the identified tags.

4. Discovering Interest-Matching Users with Tags

Our Tag-based interest-matching user's discovery approach have two steps: find topics of interests and clustering.

In the step of finding topics of interests, a set of media items with keywords are given, all topics of interests need be found out. Each topic of interests is a set of tags with the number of their co-occurrences exceeding a given threshold. The problem in nature is to

 Copyright © 2016 SERSC

find the frequent tag patterns. The frequent pattern discovery problems have been studied in other domains. Among them, association rules have been explored for many years and efficient solutions have been developed [10-11]. The basic idea of association rules algorithms is to discover frequent item patterns for a set of transactions and then derive the implication relationship among item sets for transactions [12].

In the step of clustering, each topic is collected, and the media items and users are inserted into two clusters. The clustering algorithm for a given set T of topics and a given set P of media items is shown below:

1: for all topic $T \in \mathrm{T}$ do

2: $T.user \leftarrow \phi$

3: $T.media \leftarrow \phi$

4: end for

5: for all media item $P \in \mathrm{P}$ do

6: for all topic $T \in \mathrm{P}$ do

7: $T.user \leftarrow T.user \cup \{P.user\}$

8: $T.media \leftarrow T.media \cup \{P.media\}$

9: end for

10: end for

The output of the clustering algorithm is two clusters identified by topics: one cluster contains all the media items that have been saved with all the tags in the topic, another cluster contains all the users who have been used all the tags in the topic.

The metric to evaluate the tag-based common interest discovery approach is whether similar contents can be well clustered under the topics. In our experiment, 500 interest topics, each consisting of more than 30 media items that share 5-6 co-occurring user tags is randomly selected. For each interest topic, we compute the average cosine similarity of all media item pairs in the cluster, which is called intra-topic similarity. Then 10000 topic-pairs among these 500 interest topics are randomly select, and compute the average pairwise similarity between every two topics, which is called inter-topic similarity. The inter-topic similarity between this topic and all other topics among these 10000 topic pairs are averaged, and are compared with its intra-topic similarity.

The comparison between the intra-topic and the inter-topic cosine similarity for each interest topic in selected topic with the keywords set is shown as Figure 6.

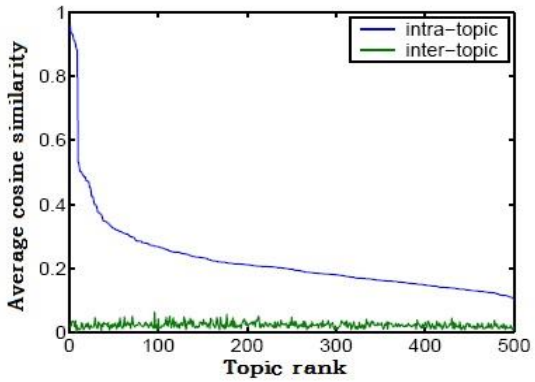

Figure 6. Keyword-Based Cosine Similarity of Interest Topics

Copyright © 2016 SERSC

From Figure 6 we can see that for all interest topics, the intra-topic cosine similarity is obviously higher than the inter-topic cosine similarity.

Corresponding to Figure 6, the comparison of the tag-based intra-topic and inter-topic cosine similarity for each interest topic is shown as Figure 7.

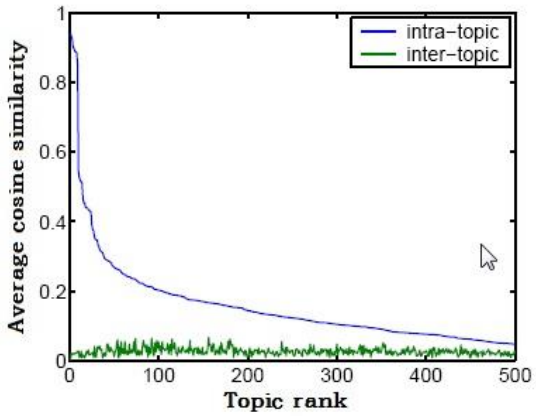

Figure 7. Tag-Based Cosine Similarity of Interest Topics

From Figure 6 and Figure 7 we can see that the tag-based cosine similarity is quite close to keyword based cosine similarity, which indicates that tags really capture the main concepts of media items. The experiment results can prove that tag-based clustering is highly accurate and the common interest captured by a number of co-occurring tags is highly focused.

5. Conclusion

In this paper, a tag-based common interest discovery approach in online social networks has been proposed. The user-generated tags are effective to represent user interests because these tags can more concisely and closely reflect understanding. So the consensus among users for the content of a given object can be reached more likely by tags than by keywords. It is proved by experiments that our approach is very effective to discover common interest topics in online social networks such as Douban, without any information on the online connections among users.

Acknowledgments

This work is supported by the Humanity and Social Science Youth foundation of Ministry of Education of China. This research is also supported by the Social Science Foundation of Qingdao, China (grant no. QDSKL150437) and the key research in Statistics Foundation of Shandong Provincial Bureau of Statistics (No. KT15168, No. KT15172).

References

[1] M. F. Schwartz and D. C. M, "Discovering Shared Interests Using Graph Analysis", Communications of the ACM, vol. 8, no. 36, **(1993)**.
[2] N. Ali-Hasan and L. Adamic, "Expressing Social Relationships on the Blog through Links and Comments", Proceedings of the International Conference on Weblogs and Social Media, **(2007)**.
[3] K. Sripanidkulchai, B. Maggs and H. Zhang, "Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems", Proceedings of INFOCOMM, **(2003)**.
[4] L. Guo, S. Jiang, L. Xiao and X. Zhang, "Fast and Low-Cost Search Schemes By Exploiting Localities", In P2p Networks, Journal of Parallel and Distributed Computing, vol. 6, no. 65, **(2005)**.

 Copyright © 2016 SERSC

[5] S. A. Golder and B. A. Huberman, "Usage Patterns of Collaborative Tagging System", Journal of Information Science, vol. 2, no. 32, **(2006)**.
[6] H. Halpin, V. Robu and H. Shepherd, "The Complex Dynamics of Collaborative Tagging", Proceedings of ACM WWW, **(2007)**.
[7] C. H. Brooks and N. Montanez, "Improved Annotation of Blogosphere via Autotagging and Hierarchical Clustering", Proceedings of ACM WWW, **(2006)**.
[8] L. Breslau, P. Cao, L. Fan, G. Philips and S. Shenker, "Web Caching and Zipf-Like Distributions: Evidence and Implications", Proceedings of INFOCOM, **(1999)**.
[9] P. Glenisson, P. Antal, J. Mathys, Y. Moreau and B. DeMoor, "Evaluation of the Vector Space Representation In Text-Based Gene Clustering", Proceedings of Pacific Symposium on Biocomputing, **(2003)**.
[10] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules Between Sets of Items In Large Databases", Proceedings of ACM SIGMOD, **(1993)**.
[11] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", Proceedings of VLDB, **(1994)**.
[12] A. Plangprasopchok and K. Lerman, "Exploiting Social Annotation For Automatic Resource Discovery", Proceedings of AAAI workshop on Information Integration from the Web, **(2007)**.

Authors

Sheng Bin, She received her Ph. D. degree in Computer Science from Shandong University of Science and Technology, China in 2009. She is currently a lecturer in the School of Software Technology at Qingdao University, China. Her main research interests include embedded system, operating system, complex networks, cloud computing and data mining.

Gengxin Sun, He received his Ph.D. degree in Computer Science from Qingdao University, China in 2013. He is currently an Associate Professor in the School of Computer Science and Engineering at Qingdao University. His main research interests include embedded system, operating system, complex networks, web information retrieval and data mining.

Peijian Zhang, He received his Ph.D. degree in Computer Science from Beijing University of Posts and Telecommunications, China. He is currently a lecturer in the School of Computer Science and Engineering at Qingdao University. His main research interests include complex networks, web information retrieval and data mining.

Yixin Zhou, She received her Ph.D. degree in Computer Science from Qingdao University, China. She is currently an Associate Professor in the School of Computer Science and Engineering at Qingdao University. Her main research interests include complex networks, web information retrieval and data mining.

Copyright © 2016 SERSC

Copyright © 2016 SERSC