An Automatic Tag Recommendation Algorithm for Micro-Blogging Users

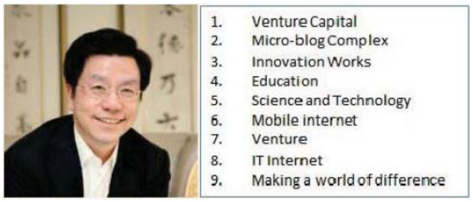
Xiang Wang[1a], Shudong Li[2], Xueqing Zou[1], Ruhua Chen[1], Bin Zhou[1]
[1] School of Computer, National University of Defense Technology, Changsha 410073, China
[2]College of Mathematics and Information Science, Shandong Institute of Business and Technology, Yantai, 264005 China
[a]email: xiangwangcn@nudt.edu.cn

***Abstract*— Online social networks such as Sina Weibo micro-blogging website allow user to annotate himself (or herself) using tags, which describe the characteristics of the user. Many researches have been done on photos, films, commodities but rare researches on users. In this paper, we try to recommend tags for users who can communicate directly with other users. Our method is based on interactive relations between users and is low cost. We present and evaluate tag recommendation method to support the user self-annotation task by recom-mending a set of tags for users. Experiment evaluations on real and large Sina Weibo dataset which contains more than 140 million users and distributed processing framework Hadoop shows that our method can effectively recommend relevant tags to users.**

Keywords-component; Tag Recommendation, User Tags, Micro-blogging, Social Network

INTRODUCTION

Recent years have seen a flourishing of social media like Facebook, Twitter, LinkedIn and so on. In these social media repositories users are allowed to upload personal data and annotate content with tags. These tags are descriptive keywords for the content. With the help of these user labeled tags, user shared content can be better organized and easier acquired.



1. Venture Capital
2. Micro-blog Complex
3. Innovation Works
4. Education
5. Science and Technology
6. Mobile internet
7. Venture
8. IT Internet
9. Making a world of difference

Figure. 1. Tags of Kai-Fu Lee

We take Sina Weibo [1], one of the most popular Chinese microblogging site, as an example to study the value of user labeled tags. Weibo uses a format similar to its American counterpart Twitter with key difference being that it is used almost exclu-sively by Chinese language speakers. Tagging is an important feature of the Web 2.0 and helps to facilitate e.g. browsing and searching. In Weibo, every user can tag him-self to show his personal characteristics. The number of the tags is no more the ten. For example, Kai-Fu Lee, who is one of the most prominent figures in the Chinese internet sector and was the founding president of Google China, labels himself with 10 tags: "IT Internet", "Innovation Works", "Venture Capital", "Education" and so on. Figure 1 shows all tags of Kai-Fu Lee .

Although many people have tagged himself in Weibo but there are still many peo-ple haven't. We have crawled 144,210,854 users and their tags from Weibo for about two months. Figure 2 displays the distribution of the number of tags and users. The y-axis represents the number of users that has a specific number of labels. We find that 78.2% of users have no labels any more and 93.8% of users have less than five labels. It's important to recommend tags to users who have none or a few tags.

User annotations provided by the user himself (herself) reflect self-recognized characteristics and it's important for audiences to find the features and for commercial applications like network marketing, user recommendation, advertisement and so on. Traditional tag recommendation algorithms are based on the content of the object, history tagging logs and so on. In social network, users are connected together with following/follower relationship and interactive relationships like retweet, reply and notify (@username) relationships. Charu C. Aggarwal and Haixun Wang[1] point out that social networks provide a special challenge because the linkage structure provides guidance for mining in a variety of applications. In this paper, our work is based on the strong interactive relationships to perform tag recommendation. These strong iteractive relationships contain retweet relationship and notify relationship.

The rest of this paper is organized as follows: related works will be introduced in Section 2. In Section 3, our algorithm based on the string iteractive relationships for tag recommendation will be described. The experiment results and evaluations will be provided in Section4. We conclude this paper in Section 5.

RELATED WORKS

In a tag system, users can tag resource with keywords or phrases to describe the characteristic of the resource. Tag recommendation is not only useful to improve user experience, but also makes rich annotation available. Golder and Huberman [2] analysis the characteristics of tags for bookmarks. They conclude that tags perform several functions: identifying what (or who) it is about, identifying what it is, identifying who owns it, refining categories and so on. Halpin et. al.[3] examine the dynamics of collaborative tagging systems. They examine the distribution of the frequency of use of tags for popular sites

[1] Sina Weibo website: http://www.weibo.com

978-0-7695-5125-8/13 $26.00 © 2013 IEEE
DOI 10.1109/CSA.2013.100

with a long history (many tags and many users) and find that it can be described by a power law distribution.

Many studies have been conducted in tag recommendation. Tags are recommended based on features of tags such as their quality, co-occurrence, mutual information and object features [4]. In our paper, we utilize the characteristics of social network that it contains rich of links between users. These links can be used to perform tag recommendation and improve the quality of tags recommended.

Hotho et al. [5] proposed FolkRank algorithm which is an adaption of PageRank for tag recommendation. FolkRank generates high quality recommendations results. Krestel et al. [6] employ topic modeling for recommending tags. They use the Latent Dirichlet Allocation algorithm to mine topics in the training corpus, using tags to repre-sent the textual content.

Mishne et al. [7] employ an algorithm which utilizes the social connection of the users to recommend tags for weblogs. The method is based on similar weblogs in other user's website tagged by the same users. Wu et al. [8] utilize the social network and the similar contents of objects to recommend tags. Their system aims to recom-mending tags for Flickr photos. While such personalized schemes have been proven to be useful, some domains of data have limited information about users and their social connections. Börkur et al. [9] analyse a representative snapshot of Flickr. They present and evaluate tag recommendation strategies by recommending a set of tags that can be added to the photo. Liu et al. [10] propose a tag ranking scheme to auto-matically rank the tags associated with a given image according to their relevance to the image content in Flickr. Their method can be used in tag recommendation and is an effectiveness method.

Shepitsen Andriy et al. [11] propose a personalization algorithm for recommendation in folksonomies which relies on hierarchical tag clusters. They use a context-dependent variant of hierarchical agglomerative clustering which takes into account the user's current navigation context in cluster selection. The hierarchical tag clusters are then used for recommendation.

Our method in this paper is to recommend tags for users which is different to many other methods that recommend tags for user's photo, blog or other resources. The tags are considers as the characteristics of a user and can be used in network marketing, user recommendation, advertisement and so on. Our method is based on the communication relationship between users and it's a low cost method compared to methods that is based on content of users.

AUTOMATIC TAG RECOMMENDATION MODEL

In this section, we will introduce our method for tag recommendation. In micro-blogging website Weibo, users are connected together with many relationships. Following/follower relationship is an obvious relationship between users. There are also other relationships between users. A user can notify (@username) others and we name this relationship "notify relationship". The "notify relationship" between users forms notify network. A user can also reply and retweet other users and we call this relationship "reply relationship" and "retweet relationship". In this paper, we don't consider the static following/follower relationship but the interactive retweet and notify relationship, because the interactive relationships explicitly represent that a user strongly influences other user. Reply relationship is also not consider in this paper because of lacking reply data. In future, static following/follower and relpy relationship will be considered.

The interactive retweet and notify relationship links users together and forms an interactive graph. Users influence each other as they communicate with each other. In this paper we suppose that if a user retweet or notify another user , then user influences and they share common interests. User tags shows the characteristics of the user and then the persons who share common interests will probably have the same tags. Tags can propagate from one user to another by the the interactive retweet and notify relationship.

3.1 Formal Description of the Interactive Graph

In Weibo, a user can tag himself (or herself) with any number of labels. That means a user can not tag any more or with many labels that he (or she) wants. Users are connected together with following/follower relationship and some interactive relationships like retweet, reply and notify. In this paper, we don't consider the static following/follower relationship but the interactive relationships, because the interactive relationships explicitly represent that a user influences other user. In future, static following/follower relationship will be considered. We use a weighted directed graph of users $G=(V,E,W)$ to represent relationships between users.

For a specific user u_i , there are no more than ten tags to describe the characteristics. Figure 1 shows the tags of Kai-fu Lee. T_{u_i} is used to describe the set of tags for user u_i. We use $r_{u_i t_j}$ to describe the relevance score of user u_i and tag t_j, where $t_j \in T_{u_i}$. R_{u_i} is the set of relevance score of user u_i and all his (her) tags T_{u_i}. So in the interactive graph $G=(V,E,W)$, a vertex $v_i (v_i \in V)$ can be represented as $v_i = (u_i, T_{u_i}, R_{u_i})$ which show the property of user u_i.

In the interactive graph $G=(V,E,W)$, links between two users is the retweet and notify relationship. If a user u_i retweet or notify another user u_j, then there is a directed edge from user u_j to user u_i. In other words, there is a directed edge $e_{ji}(v_j \rightarrow v_i)$ from vertex v_j to vertex v_i.

In graph $G=(V,E,W)$, the weight w_{ji} of edge $e_{ji}(v_j \rightarrow v_i)$ is calculated by the number of retweet and notify between users. The weight w_{ji} of the edge e_{ji} shows the influence strength from user u_j to u_i. We consider the retweet and notify relationships to be the same relationship. Then weight w_{ji} is calculated as equation (1):

$$w_{ji} = \frac{f_{comm}(u_i,u_j)}{\sum\limits_{u_s \in commSet(u_i)} f_{comm}(u_i,u_s)} \tag{1}$$

where $f_{comm}(u_i,u_j)$ is the total number of retweeting and notifying from user u_i to user u_j. $commSet(u_i)$ is the set of all users that has been retweeted or notified by user u_i. We can find that $0 \le w_{ji} \le 1$ from equation (1).

3.2 Random Walk in the Interactive Graph

In this section, we will introduce our method that tags can propagate from one user to others following the interactive relationships. Every user has a set of tags T_{u_i} and each tag t_i has a relevance score to the user. When user u_i retweets or notifies u_j, then the tags in u_j are probably implicit in user u_i. The influence strength w_{ji} from user u_j to u_i which is calculated in equation (1) is to be the transition probability of the interactive graph. Then a tag $t_s(t_s \in T_{u_j})$ in user u_j will be in user u_i with probability $p_{t_s}(u_j \to u_i)$ which is calculated in equation (2):

$$p_{t_s}(u_j \to u_i) = r_{u_i t_s} \bullet w_{ji} \tag{2}$$

where $r_{u_i t_s}$ is the relevance score of user u_i and tag t_s.

User u_i may communicate with many users and tags in these users can also probably implicit in user u_i. Then the new relevance score $r'_{u_i t_s}$ of a tag t_s to user u_i will be the sum of all users' t_s transferring to user u_i. Let $S_{comm}(u_i)$ be the set of all users that are be retweeted or notified by user u_i, then $r'_{u_i t_s}$ can be calculated in equation (3):

$$r'_{u_i t_s} = r_{u_i t_s} + \sum_{u_m \in S_{comm}(u_i)} p_{t_s}(u_m \to u_i) \tag{3}$$

It's easy to find that sum of relevance score of new tags in user u_i is not equal to one and normalization must be taken. The normalized relevance score $r''_{u_i t_s}$ of tag t_s in user u_i is calculated in equation (4):

$$r''_{u_i t_s} = \frac{r'_{u_i t_s}}{\sum\limits_{t_f \in T'_{u_i}} r'_{u_i t_f}} \tag{4}$$

Where T'_{u_i} is the new set of all tags in user u_i.

EXPERIMENT EVALUATION

4.1 Data Sets and Distributed Experimental Environment

The experiment is conducted in Sina Weibo dataset. The dataset is crawler for about two months which contains 144,210,854 users and 3,052,289,362 relations between users. The relations include both the retweet relationship and notify relationship.

TABLE 1. SINA WEIBO DATASET

Property	Number
users	144,210,854
relations	3,052,289,362

To create a test dataset, we constructed 3 test datasets from the crawled Sina Weibo dataset. We randomly choose 3000 users from the Sina Weibo dataset and each user in the test dataset has more than 8 tags. Each test dataset contains 1000 users whose tags are removed. We use the original tags that are labeled by the users themselves to be the standard results. The tags that are generated from our methods will be checked by the original tags.

We also remove some rare tags in the original dataset. The rare tags are probably caused by mistakes in writing or the tags don't have common meanings. So we re-move the tags whose frequency is less than 20 in the original Sina Weibo dataset.

All our experiments are done on Apache Hadoop [2] which is distributed processing framework of large data sets across clusters of computers. That's because the dataset we use in our experiment is very large that normal server is hard to handle it. Our experimental is conducted in a cluster of computers of hadoop which contains 24 servers. Each server's memory is 32GB and its CPU is 4-core. The disk of each server is 2 TB. Our data is stored in distributed file system HDFS. All our programs in this paper are based on MapReduce [12] programming model which is for processing large data sets with a parallel, distributed algorithm on a cluster.

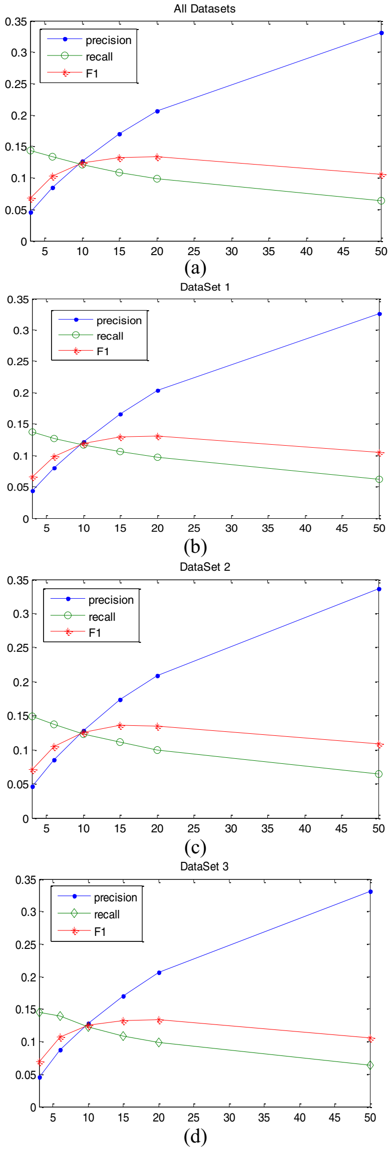
4.2 Experimental Results and Evaluation

To construct the three test datasets, we randomly choose 1000 users and their tags for each. In the test datasets, all users' tags are removed. Our methods are used to recommend tags for user and we will compare our recommend tags with users' real tags to measure performance of our method. Average precision, average recall and average F-measure are used in this paper. Figure 3 shows the performance of our method in the three datasets. All graphs in Figure 3 shows the precision, recall and F1 values with the change of the number of recommended tags. The recommended number is 3, 6, 10, 15, 20, 50. Figure 3 (a) shows the performance in all 3000 users. Figure 3 (b) shows the performance in dataset 1. Figure 3 (c) shows the performance in dataset 2 and Figure 3 (d) shows the performance in dataset 3.

We can find that no matter in dataset 1, dataset 2 or dataset 3, when the number of recommended tags is ten, the precision, recall and F1 values are all the same. This point is the best one for tag recommendation. Why do ten recommended tags to user will get the best performance? We find that in micro-blogging website Sina Weibo, users can only have no more than ten tags. This limit policy influences the best number of recommended tags. In Fig 3, we find that no matter in dataset 1, dataset 2 and dataset 3, the values of precision, recall and F1 are nearly the same. This feature shows that our method's performance will

[2] Apache Hadoop: http://hadoop.apache.org/

change little in the whole dataset. We also can find that the performance of our method is low. It's because that the data of user tags are very sparse and 93.8% of users have less than five labels.



(a)

(b)

(c)

(d)

Figure. 1. Performance with the Number of Recommended Tags

Table 2 shows the precision, recall and f1 values when the number of recommended tags is 10. We can find that there is little difference in different datasets.

TABLE 2. EVALUATED VALUES WITH THE 10 RECOMMENDED TAGS

	Dataset 1	Dataset 2	Dataset 3	All Datasets
Precision	0.1217	0.1286	0.1279	0.1261
Recall	0.1159	0.1235	0.1226	0.1207
F1	0.1849	0.1259	0.1251	0.1233

CONCLUSION

In this paper, we present an automatic tag recommendation method for user in micro-blogging website. Our method is not based on content of user but interactive relations between users. The interactive relations are generated by users retweeting or notifying (@username) each other. The interactive graph is generated because of the interactive relations between users. In the graph, each user contains multi tags. The tags can do random walk in the interactive graph. Our method is based on the random walk algorithms to do multi-tags recommendation. Experiments in real and large Sina Weibo dataset show good performance of our method although the sparseness of user tags data influence the performance.

In the future, other interactive relations such as following/follower relation and reply relation between users will be considered. We will also try to utilize micro-blog content of users for tag recommendation.

ACKNOWLEDGMENT

In this paper, the research was sponsored by National 973 Program (Grant No. 2013CB329604, 2013CB329601, 2013CB329602), NSFC (Grant No. 60933005, 91124002, 61202362), 863 Program (Grant No. 2012AA01A401, 2012AA01A402), National Key Technology R&D Program (Grant No. 2012BAH38B04, 2012BAH38B06).

REFERENCES

[1] C. C. Aggarwal and H. Wang. (2011). *TEXT MINING IN SOCIAL NETWORKS.*

[2] S. Golder and B. A. Huberman, "The structure of collaborative tagging systems," *arXiv preprint cs/0508082: Ithaca: Cornell University Library,* 2005.

[3] H. Halpin, *et al.*, "The complex dynamics of collaborative tagging," presented at the Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, 2007.

[4] M. Gupta, *et al.*, "Survey on social tagging techniques," *SIGKDD Explor. Newsl.,* vol. 12, pp. 58-72, 2010.

[5] A. Hotho, *et al.*, "Information retrieval in folksonomies: Search and ranking," in *The semantic web: research and applications*, ed: Springer, 2006, pp. 411-426.

[6] R. Krestel, *et al.*, "Latent dirichlet allocation for tag recommendation," presented at the Proceedings of the third ACM conference on Recommender systems, New York, New York, USA, 2009.

[7] G. Mishne, "AutoTag: a collaborative approach to automated tag assignment for weblog posts," presented at the Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, 2006.

[8] L. Wu, *et al.*, "Learning to tag," presented at the Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009.

[9] B. Sigurbjörnsson and R. v. Zwol, "Flickr tag recommendation based on collective knowledge," presented at the Proceedings of the 17th international conference on World Wide Web, Beijing, China, 2008.

[10] D. Liu, *et al.*, "Tag ranking," presented at the Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009.

[11] Shepitsen A, Gemmell J, Mobasher B, et al. Personalized recommendation in social tagging systems using hierarchical clustering. Proceedings of the 2008 ACM conference on Recommender systems. ACM, 2008: 259-266.

[12] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM, 2008, 51(1): 107-113.