2012 Second International Conference on Cloud and Green Computing

The untold story behind the recommendation in micro-blogging network

Tong Man, Hua-Wei Shen, Xue-Qi Cheng
Research Center of Web Data Science and Engineering
Institute of Computing technologies, Chinese Academy of Sciences
BeiJing, China
Email: mantong@software.ict.ac.cn

***Abstract*—Celebrity recommendation is widely-adopted by many micro-blogging services as an important way to enhance users' experience and the visibility of contents submitted. It is critically important for celebrity recommendation to understand which factors and how these factors affect the probability that the recommended celebrities will be accepted. In this paper, taking the Tecent Weibo as a case, we try to empirically tackle the untold story behind the celebrity recommendation in micro-blogging network. We studied the three potential factors, namely the popularity of the recommended celebrities, the structural similarity and the topical similarity between the recommend celebrity and the target user. As shown by experimental results, the popularity of the celebrity and the structural similarity well reflect whether the recommended celebrity is accepted by the target user, whereas the topical similarity is not a good indicator as commonly expected.**

***Keywords*-recommender system; micro-blogging; collaborative filtering;**

I. Introduction

Micro-blogging service is a new form of social network service.In micro-blogging networks, users are allowed to publish small pieces of digital contents, such as brief text messages, pictures, links, videos and other media. Users are also allowed to follow others they are interested in the network. Follow means we add someone to a list and then get his posted messages. Many celebrities, companies and organizations choose micro-blogging network as a channel for publishing news. From the perspective of information diffusion, micro-blogging acts similar as traditional media. From the perspective of linking peoples by follow relationship, micro-blogging could also be looked as an social network. Thus, micro-blogging services are also called as social media.

One of the most notable micro-blogging networks in the world is Twitter, which has received much attentions in both academic and industrial area. In this paper, we use the dataset from one of China's largest micro-blogging service networks, Tencent Weibo.

In micro-blogging network, people get the information mainly from the people he followed. The following relationship is one-way. I.e, the user whom been followed is not required to follow back. The one-way relationships among users is different from some other two-way social network services such as facebook. In a following relationship, the user being followed is called the followee, and the other one who is following is called the follower. In the micro-blogging network, the information flows from the followee to the follower. Without other services provided by the system, the list of followee of one user is his/her sole information source in the network. Sometimes this information source is far more than enough for the users' satisfaction for information requirement. User have to follow more and more people in case of missing any information he maybe interested in. However the more people one user follow, the more unrelated information he will receive. Thus this will forms information overload in micro-blogging services.

Recommender system is the state-of-the-art technology to tackle the information overload problem. In the micro-blogging networks, users have two requirements for recommender systems. The first one is content recommendation. To satisfy users' information demand, system directly recommend the messages to users. The content been recommended should be those messages that users maybe interested in but unaware of. The second one is the link recommendation, or we can call it friend recommendation. Instead of recommending the messages to users directly, we recommend the information sources to them, i.e., recommend users that the target user may follow.

People follow other users in the network for many reasons. Generally speaking, these reasons can be attributed to two kinds of requirements, namely the social requirement and the media requirement. For social requirement, one user may follow the persons who are his friends in real world. For the media requirement, one user may follow the persons who are the potential providers of information. In this paper, we focus on the media requirement of one user's following activity. We use the following activity between the users and celebrities in the network, here celebrity is defined as the celebrity in the real world. In the Tencent Weibo dataset, it is very easy to distinct the normal users from celebrities by the system's labels.

Recommender system in Weibo network recommends celebrities to users by sessions. For each session, three celebrities are presented to the target user. The target user may follow all or part of the three recommended celebrities. However, when none of the three celebrities is followed, we cannot simply say that the target user is not interested

978-0-7695-4864-7/12 $26.00 © 2012 IEEE
DOI 10.1109/CGC.2012.104

in all the three celebrities because the target use may be unaware of the recommendation session at all. Therefore, in this paper, we only consider the sessions, in which at least one celebrity is followed by the target user. In this way, all the celebrities followed by the target user ware taken as positive feedback and the celebrities not being followed can be treated as negative feedback.

In this paper, we focus on analyzing the potential factors which affect the user's feedback to the recommendations. We extracted features for each user and each user-celebrity pair in the micro-blogging network, and analyze how these features affect the user's following activities. We considered three features, namely, popularity, structure similarity and topic similarity. Popularity is the feature of the celebrities, and the two other similarity is the feature of the dyadic user-celebrity pair.

This paper is organized as follow. First we list some related works about micro-blogging networks and recommender systems. Then we introduce the Tencent Weibo dataset, the dataset we use in our paper. Next we analyze how the structure and topic factors affect users' feedback in the recommendation session, and we do some hypothesis testing about the significance of the correlation between the similarity and the feedback of the user. At last we give the conclusion and list some future work to do.

II. Related Works

There are many previous works on micro-blogging networks. The topological and content properties on the whole networks have been analyzed in [1], where the studied network is formed by the twitters and their follower in the twitter network. How to rank the influence of the users and the content in the network are discussed in[6][7], where the influencer in the network is defined and try to find by some algorithms. Information diffusion and link perdition is another two hot topics in this research area. In [5], users were classified into different groups and how the information flow worked between these groups was studied. How the links structure evolved and how to predict the link emerging in the network was studied in[3][4], which both predict the future links in the social networks.

Recommender system is an important technology to overcome the information overload problem[9]. Two state-of-the-art technologies are the content-based recommendation and the collaborative-filtering. Content-based recommendation use characteristics of the users and items, to get the similarity between the items in the waiting list and the items in the user's history list. In one word, these algorithms try to recommend items that are similar to those that a user liked in the past. Collaborative filtering methods are based on analyzing a large amount of user-item interaction data. The basic assumption of collaborative filtering is that the users have the similarly taste in the past will be similarly in the future. User-based and item-based[8] nearest neighbor

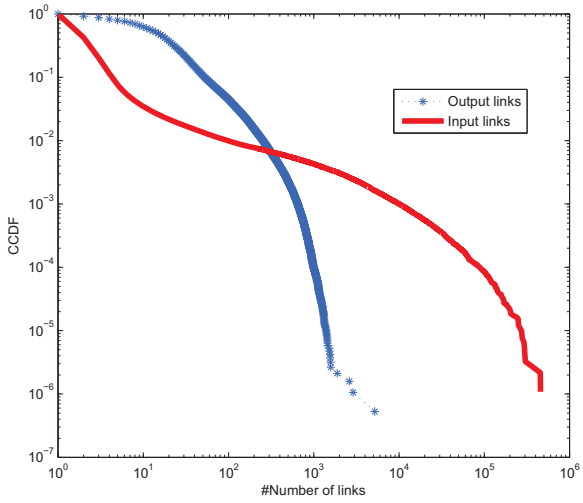

Figure 1. Complementary cumulative distribution function (CCDF) of indegree and outdegree of users in Tencent Weibo.

method are two widely-used methods. The second method is used in this paper.

Sometimes, friend recommending is mixed with the link prediction problem in the social network scenario since that for both of these two problems seem like to predict the links will be formalized in the future. In our viewpoint, the link prediction problem should be treated as a task to get the inherent property of the growing function of the network. However the goal of friend recommendation is to mining the missing links, and recommend the links to the nodes. The recommendation procedure has interposed the growing of the network.

III. Data Analysis

We use the Tencent Weibo dataset, which was launched on KDDCUP 2012, as our experimental dataset. There are 2,314,800 users and 6,095 celebrities, here celebrities could be persons, organizations or groups. We have an initialized social network data as a snapshot at a given time, and a recommending dataset which records users' acceptation and rejection for the recommendations from the system. So we can get positive and negative instances for users' interests in the celebrities. Except for these structure files, we have profile datas that recorded the user's demographical information such as age, gender, etc, and user's content file such as keyword, tags.

We can see the degree distribution of the initial network in the fig.1. For this data is just a sample of the full data set, the distribution of the output and input degrees is showed as a skewed power-law distribution. Then we see the types of the links in the dataset. There are four types of links, user-user, user-celebrity, celebrity-user, celebrity-celebrity,

Table I
LINKS TYPE DISTRIBUTION IN THE SNAPSHOT

Link Type	Numbers	Percentage
User-User	6,184,282	12.21%
User-Celebrity	44,254,247	87.36%
Celebrity-Celebrity	173,716	00.34%
Celebrity-User	42,898	0.08%

Table II
ONE SUCCESSFUL RECOMMENDATION AND THE REPEATLY RECOMMEND TIMES

Repeated Recommendations	Percentage
1 times	73.9%
In 5 times	96.5%
In 10 times	98.9%

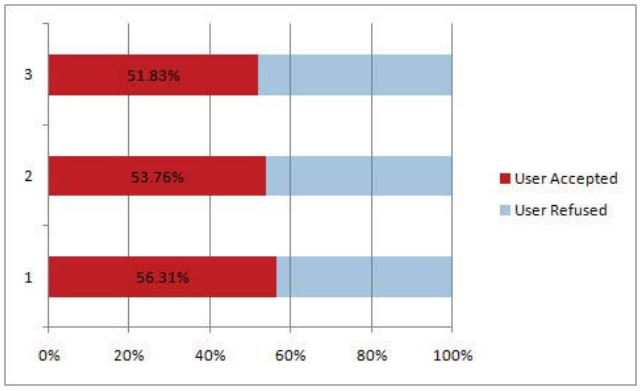

Figure 2. User's adoption of recommendation and the popularity of the celebrity

respectively. In table.1 we list the links constitution of all the 50,655,143 links in the snapshot network. We can see that the most links in the network are user-to-celebrity link.

Then we analysis the user's follow activities in the recommendation record datafile. For each user, Tencent Weibo will recommend some celebrities to users as the candidates, user can accept some of the recommendations and choose someone to follow, or ignore the recommendations. These recommendation are recorded in the data file as sessions. In each session three celebrities was represented to the user, and user feedbacks to each of these recommendation. In the whole file, the proportion of user's positive feedback and negative feedback is 1:13. To make sure of our results are reasonable, we filter out the session that user make no positive feedback.

Celebrities may recommend to users repeatedly. We analysis this activity and found a phenomenon that the acceptation of a recommendation is always done in the early times, which means repeatedly recommendation is useless. There are 5,167,202 successful recommendations in the data, and the rate of the times that one successful recommendation is needed is in the table 2. We can see, almost all(96.5%) the successful recommendation are fulfilled less than 5 duplicated recommendation times.

In this paper, what we want to know is, which and how the factors affect user's reactions for the recommendations of celebrities. Here factors could be tropologic structure-based features, or content-based features. In this paper, we considered three features of one user-celebrity pair, they are popularity, collaborative similarity and topic similarity.

IV. POPULARITY, STRUCTURE, CONTENT

In this section, we considered three features in the user's interaction to the recommendations of celebrities in an useful session. They are popularity, structure similarity and topic similarity, respectively. First we considered the popularity, here we simply use the number of followers of one celebrity as the indicator of the popularity. Then as a graph perspective of the user-celebrity network, we could treat the micro-blogging network as a bipartite graph. So we can use some techniques from collaborating filtering to get the structure similarity of one user and celebrity pair. At last, we can treat both normal users and celebrities as the information source in the network, and then they will generate content in the network. We could get their topic distribution by topic models such as pLSA and LDA. Then we could get their similarity in the space of latent topics. An interesting phenomenon have been found that topic similarity is not significate when user choose the celebrity to follow.

Before we dive into these features, we check for the reciprocity influence for the user in the interaction with recommendations. Here reciprocity influence means that the system recommend one celebrity that have already followed the user. Though this happens rarely, we found this influence is a very strong indicator. In the data session file, these are 61 recommendations that user are been recommended have already been followed by the celebrity, and all these 61 recommendations have been adopted.

A. Popularity

How to estimate the popularity of celebrities in the network actually is a large topic beyond the scope of this paper. To allow propagation of influence along the network, many propagation-based algorithms could be used such as PageRank. In this paper, the popularity of the celebrities be easily estimated by the number of followers. Here we want to know in one recommendation session, if the popularity is an important factor for user to choose one celebrity to follow. In each session, we labeled the three celebrities by 1,2,3 by their popularity rank.

We can see the relationship between the popularity in the recommendation list and the adoption rate of the users by sessions. we can see that user choose the most popularity celebrity to follow more often than the less popularity

celebrity. And this phenomenon have shown that in the user's choice for one celebrity in the recommend list, popularity is one important feature to consider. Notice here the popularity is defined arbitrary by the number of followers, so it is difficult to metric the difference of the popularity across sessions. Hypothesis testing about the significance of the popularity factor in the users' adoption for an recommendation could be better done if we use PageRank or other scaled metrics to judge the popularity.

B. Structure Similarity

To separate the role of user and celebrity in the network, we can treat this micro-blogging network as a bipartite graph. Notice here celebrities could also be the users who have been recommended some celebrities to follow, so some nodes appears in the both set of the bipartite graph. Here first we can treat both user and celebrity as a set of celebrities they followed, so we can get the similarity between a user and a celebrity by Jaccard similarity, which is show as below. Here $\tau(u)$ means the follow list of the users.

$$simi^{Jac}(u,v) = \frac{|\tau(u) \cap \tau(v)|}{|\tau(u) \cup \tau(v)|}$$

Then we considered another similarity which introduced by the idea of collaborative filtering. For each user-celebrity pair, we get the similarity of the active celebrity and all celebrities in the user's follow list. Then we choose the most k similarity celebrities and assemble the celebrities to get the final evaluation. Here the similarity between celebrities is computed based on their follower list. This method is also called item-based k nearest neighbors method. Which is a two-stage procedure showed follow. Here $N(v)$ means the follower list of the celebrities, $S_k(v)$ is the set of the k nearest neighbors of the celebrity.

$$simi^{Jac}(v1,v2) = \frac{|N(v1) \cap N(v2)|}{|N(v1) \cup N(v2)|}$$

$$simi(u,v) = \frac{\sum_{i \subset S_k(v) \cap \tau(u)} simi^{Jac}(i,v)}{|S_k(v) \cap \tau(u)|}$$

Now we check these structure similarities between user celebrity pair and the adoption rate of the users by sessions. First we use the same idea in the last section, then we get the Figure.3 and Figure.4, we can see that the structure similarity could also be an indicator of the user's choice for one celebrity to follow.

Then we do a hypothesis testing about the relationship about user's feedback and the structure similarity between user and celebrity, it can be formalized as a two-sample t-test:

Let $\mu_{positive}$ be the mean similarity value of all the positive feedback pairs, here positive means user accept the recommendation ,and $\mu_{negative}$ be the mean similarity value of all the negative feedback pairs. The null hypothesis is

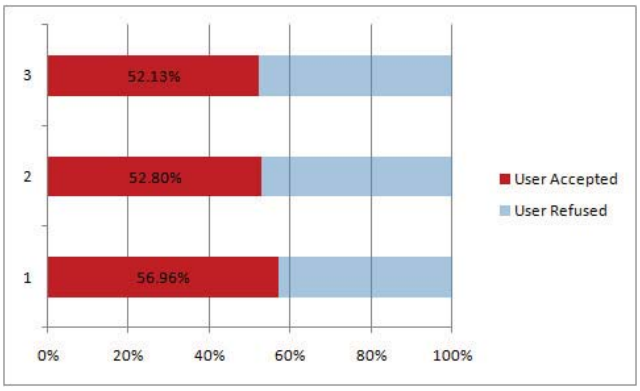

Figure 3. User's adoption of recommendation and the follow list similarity of the user-celebrity pair

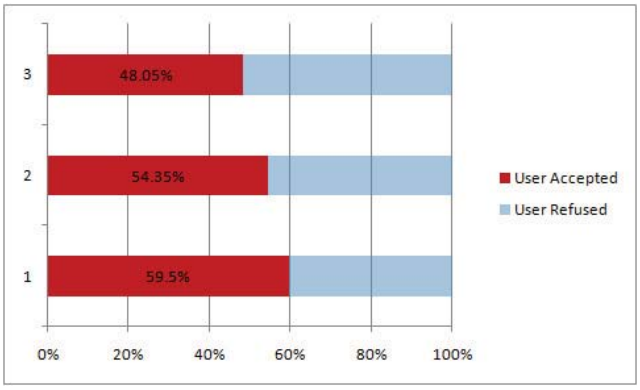

Figure 4. User's adoption of recommendation and the k-nearest neighbor similarity of the user-celebrity pair

$H_0 : \mu_{positive} = \mu_{negative}$, and the alternative hypothesis is $H_0 : \mu_{positive} > \mu_{negative}$.

In the Weibo dataset, we got 1,610,824 positive feedback pairs and 1,389,175 negative feedback pairs. We computed the similarity of all of these pairs both by follow-list similarity and the k-nearest neighbor similarity. Results have show that for both these two similarity, the null hypothesis is rejected at significant level $\alpha = 0.01$.

C. Topic Similarity

In this section we look at relationship between the user's adoption activity and the content similarity of the user-celebrity pair. In the micro-blogging network, users generate a lot of content each day, each user have his own interest or say topical. We can extract some keywords for each user by get rid of some stop words. However, topic interests are not explicitly expressed by twitters, mainly for the sparsity of the keywords. To tackle this problem, some topic modelling is commonly used to get topic distributions for each users. We use Latent Dirichlet Allocation(LDA) model to identify the topics for each users. Just like the procedure of the document

generation, the generative process for generating each user's content as follows:

1.For each user, an topic is picked from his/her topic distribution over topics.

2.Sample a word from the distribution over the words associated with the chosen topic.

3.The process is repeated for all the words in the user's content.

The model have two parameters to be inferred, one is the user-topic distribution and the other is topic-word distribution. To get the content similarity between users, we only need the former one distribution.

After infer all the parameters of the users, we need to compute each two user's topic similarity based on two probability distribution. First we get the distance of two topic distributions by Jenson-Shannon Divergence which is defined as:

$$D_{JS}(i,j) = \frac{1}{2}(D_{KL}(DT_{i.}^{'}||M) + D_{KL}(DT_{j.}^{'}||M))$$

Here M is the average of the two probability distributions. D_{KL} id the Kullback-Leibler Divergence of two probability distribution.Then we define two users' topic similarity as

$$simi(i,j) = 1 - dist(i,j) = 1 - \sqrt{2 * D_{JS}(i,j)} \quad (1)$$

In the topic distillation procedure, we discard the words appeared rarely and the stop words, then we use LDA to get the topic distributions of the users. In the Weibo dataset, we got 947,232 positive feedback pairs and 705,828 negative feedback pairs with useful topic distribution. We computed the similarity of all of these pairs by similarity defined above.

Then we make a hypothesis testing about the relationship about user's feedback and the topic similarity between user and celebrity, it can also be formalized as a two-sample t-test. Results have show the null hypothesis is not rejected at significant level $\alpha = 0.01$.

This sounds interesting as when user choose one to follow in the recommend list, the topic similarity is not an important feature to considered, as least not as important as the structure similarity. The user's adoption activity for one recommendation from the system is not significantly related to the content similarity of the user-celebrity pair.

V. Conclusion And Future work

In this paper, we considered the user's adoption activity for the recommended celebrities in the micro-blogging network, namely, Tencent Weibo. Three factors in one user-celebrity interaction pairs are considered, i.e., popularity, structure similarity and topic similarity. We use the number of followers of one celebrity to define the popularity and the structure similarity is computed by the idea in collaborative filtering. To get the topic similarity, LDA is used to get the topic distribution of the users and celebrities. We found that popularity and the structure similarity are two major factors for users' selection, and the topic similarity is less significant when users choose one celebrity in the recommender list to follow .

More techniques considering the features of one user-celebrity pair could be considered in the future. How to incorporated all these features into a unified prediction framework is our future work. To understand how the structure and content factors affect the user's adoption activities of the recommender system may be useful to make better recommendations for users.

VI. Acknowledgements

This work was funded by the National Natural Science Foundation of China under Grant Nos. 61232010, 60933005, and 61202215. This work was also partly funded by the Beijing Natural Science Foundation under Grant No. 4122077.

References

[1] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 591-600.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993-1022,2003.

[3] Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: predicting and recommending links in social networks. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 635-644.

[4] David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03). ACM, New York, NY, USA, 556-559.

[5] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on twitter. In Proceedings of the 20th international conference on World wide web (WWW '11). ACM, New York, NY, USA, 705-714.

[6] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 65-74.

[7] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining (WSDM '10). ACM, New York, NY, USA, 261-270.

[8] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (WWW '01). ACM, New York, NY, USA, 285-295.

[9] Francesco Ricci and Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer, 2011