Proc. of the Second Intl. Conf. on Advances In Computing, Communication and Information Technology- CCIT 2014.

ISBN: 978-1-63248-051-4 doi: 10.15224/ 978-1-63248-051-4-91

Friends Recommend in Micro-blog Social Networks

Ruisheng Shi, Shan Feng, Ruifang Liu

***Abstract*—The users of social network sites are always constructing a large network, recommending friends to registered users is a crucial task for these sites. Traditional content- based or collaborative filtering recommend technologies are limited for the task, because the users and items are the same dataset, which have rich attributes and complex social graph. In the paper, we proposed a local random walk and random forests combined method to do friends recommendation for large social networks. The experimental results show that the method has high precision and high recall at the same time.**

***Keywords*—Local Random Walk, Random Forests, Social Network Site, Friends Recommendation**

I. Introduction

All kinds of social network sites(SNS) are becoming more and more popular now, which provide users with a Internet platform of communication and information sharing. Based on the six degrees of separation theory, SNS is trying to build friend circles for users to link each other, to expand the contacts between users through "friends of friends", which is trying to keep the stickiness between users and SNS through expanding social circle to keep stickiness among users. At the same time, the behavior of users will inadvertently affect the behavior of friends, such as sharing, collection, or buy and so on. In SNS, recommending friends to registered users is a crucial task.

Traditional recommendation problem is trying to recommend the most probably interesting items to a user. Two main techniques are often used in recommendation systems [1]. Content-based (CB) techniques recommend solely based on the features of a item which similar to those the user preferred in the past. Purchase activity, for example, CB recommends similar items according to the user's interests. Collaborative filtering (CF) techniques recommend items searched by the other users similar in tastes liked in the past. The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly. Most of the time, CB and CF techniques are combined in order to provide more accurate predictions and overcome the limitations of each technique.

These technologies can be used for the friend recommendation in micro-blog social network. The Track 1 problem of KDD Cup 2012 is to predict which user another user might follow in Tencent Weibo [2], one of the largest micro-blog website in China, The dataset includes user profiles, user actions, item categories, keywords, and social graph. The top 3 winners are all employ model based CF technology.

However, the friend recommendation of micro-blog has its specialty. On the one hand, the items have the same attributes with the users, they are all the users of SNS, there are rich information about the users, such as relationship, tags, tweets, which should be used for the friend recommend. On the other hand, the absent or sparse of the history data of behavior is the biggest shortcoming of CF algorithms.

In the paper, we model the problem as a network link prediction problem. The network nodes are users, which have rich attributes, friend relationships are links. We are trying to predict new relationships of the social network.

Link prediction has attracted much attention recent years [3] and is applicable to a wide variety of application areas, the algorithms can be used to predict the links that may appear in the future of evolving networks. Commonly, two nodes are more likely to be connected if they are more similar, where a latent assumption is that the link itself indicates a similarity between the two nodes. But the sparsity and huge size of the target networks are the difficulties in link prediction, so do in the micro-blog user relationship network. On the other hand, the nodes in social network often have rich attributes, how to use these information to do link prediction is a problem.

In order to improve the efficiency, instead of considering the whole network, we adopt a local random walk algorithm to get a list of latent friends as candidates for each user, we should define the transition probabilities for nodes in a directed graph, because the following relationship in micro-blog does not require confirmation.

In order to improve the effectiveness with the rich attributes of nodes, we extract the features of the target user, the features of candidate users, and the features of pairs of the target user and each candidate user. And then a ranking model is used for the candidates ranking. While training the ranking model, we only consider the followers, following and bi-follow users for each target user.

Ruisheng Shi
Education Ministry Key Laboratory of Trustworthy Distributed Computing and Service, BUPT
School of Humanities, BUPT, Beijing 100876, China

Shan Feng, Ruifang Liu
School of Information and Communication Engineering, BUPT
Beijing 100876, China

Proc. of the Second Intl. Conf. on Advances In Computing, Communication and Information Technology- CCIT 2014.

ISBN: 978-1-63248-051-4 doi: 10.15224/ 978-1-63248-051-4-91

II. Related Works

KDD Cup 2012 Track 1 is to predict which user another user might follow in Tencent Weibo [2]. The dataset is made up of 2,320,895 users, 6,095 items, 73,209,277 training records, and 34,910,937 testing records. There is wealth of auxiliary information on each user and each item. For example, each user possesses age, gender, tags and keywords, as well as a history of followed items, and a history of connections with other users. For each user in the testing dataset, an ordered list of the recommender results is demanded. Mean Average Precision (MAP) is used to evaluate the results. With the history of users accept or reject to the recommend friends, latent factor model [4] based CF technology is wide used by the participants.

To solve these challenges, SJTU team [5] combines two kinds of useful models: feature-based matrix factorization and additive forest. They use feature-based matrix factorization model to incorporate side information such as users' social network, action, keyword/tag and items' taxonomy information. They also develop additive forest models to incorporate users' profile, activity and sequential patterns.

FICO team [6] generates many predictive features as well as latent factor models and interaction terms, and then trains a scorecard model to ensemble these features, interaction terms, and latent factor models.

Shanda team [7] presents a novel approach called context-aware ensemble method. Various features are extracted from the training data and integrated into the proposed models.

In actual, most of the time we could not get the user-item rating matrix or interaction matrix, or the matrix is sparse or poor quality, so the latent factor model based CF technology is not always useful.

Vineeth Rakesh et.al [8] proposed a framework for recommending lists to users by combining several features that jointly capture their personal interests. They develop a ListRec model that leverages the dynamically varying tweet content, the network of twitterers and the popularity of lists to collectively model the users'preference towards social lists.

Ingmar Weber et.al [9] exploited co-following information and hidden correlations to give personalized `"out-of-context" recommendations of Twitter users to follow. The framework is simple and the idea is similar to Amazon's recommendation beyond different domains.

Kristian Slabbekoorn et.al [10] presented a method for Twitter user recommendation based on user relations and taxonomical analysis. This method first ranks users based on user relations obtained from tweet behaviour of each user such as retweet and mention (reply), then picks up users who continuously provide related tweets from the user list.

Marcelo Gabriel et.al [11] proposed a followee recommender system based on both the analysis of the content of micro-blogs to detect users' interests and in the exploration of the topology of the network to find candidate users for recommendation. .

III. Local Random Walk on Directed Graph

Consider an undirected simple network $G=(V,E)$, where V is the set of nodes and E is the set of links, and $|V|=n$. Random walk is a Markov chain describing the sequence of nodes visited by a random walker [12]. This process can be described by the transition probability matrix $P=(p_{ij})_{n\times n}$, with $p_{ij}=\begin{cases} 1/d(i), e_{ij}\in E \\ 0, e_{ij}\notin E \end{cases}$, presenting the probability that a random walker staying at node i will walk to j in the next step, where $d(i)$ denotes the degree of node i. For a directed network, $d(i)$ often means the out-degree.

Given a random walker starting from node i, denoting by $\pi_{ij}(t)$ the probability that this walker locates at node j after t steps, we have $\overrightarrow{\pi_i}(t)=P^T\overrightarrow{\pi_i}(t-1)$, where $\overrightarrow{\pi_i}(0)$ is an $n\times 1$ vector with the i^{th} element equal to 1 and others all equal to 0. While $t\to\infty$, we have $\pi=P^T\pi$.

But for a large social network graph, it is impossible to calculate the steady state distribution of random walk. In 1967, American sociologist Stanley Milgram tested the *small world* phenomenon, as we all know, the phenomenon is ubiquitous in on line social network site. We tested it on micro-blog Sina, we got a data set from Sina Weibo, which contains 3 million users information crawling from Sep. 2013 to March 2014. Started from one user, one step can find 300 friends, two steps can reach 80 thousand users, three steps can reach 10 million users and the result of walking 4 steps almost cover all the network nodes, it is more than one hundred million. From the experiment result shown in Table I, we can see that the result is almost stable while walking four steps. So we will only consider the first 4 steps during the next experiments.

TABLE I. RANDOM WALK TEST

top N friends walk 6 Steps (the number)	50	100	500	2000
Walk 5 steps intersection with 6 steps	49	100	495	1979
Walk 4 steps intersection with 6 steps	48	94	470	1817
Walk 3 steps intersection with 6 steps	45	88	423	1572
Walk 2 steps intersection with 6 steps	32	69	315	823

The relationship of users in micro-blog network is a directed graph, each user is a graph node. Each user of micro-

Proc. of the Second Intl. Conf. on Advances In Computing, Communication and Information Technology- CCIT 2014.

ISBN: 978-1-63248-051-4 doi: 10.15224/ 978-1-63248-051-4-91

blog has three specific attributes: number of friends, number of followers and number of bi-follows. The number of followers, as in-degree, shows the user's popularity. The number of friends, as out-degree, shows the user's activity. The number of bi-follows reflects the reality community.

It is easy to think the random walk should follow the out links, because most of the time we do not care who follow me. But if we define $p_{ij} = 1/friends_i$, we find that the top 2000 users with the highest reach probability in 4 steps, for my account in Sina Weibo, are all star user, because the star users always have high in-degree. Such a recommendation is meaningless.

We thought the user in social network is always interested in two classes of nodes, one is the friends in reality, and the other is the people have a certain influence in social network.

The user will follow the friends in reality, they are trying to build connection on networks. We thought they would have many common friends, we would intersect the nodes bi-follow with user i and the nodes bi-follow with user j, the element number is the number of their common friends, so attractiveness between user i and user j is defined as equation (1).

$$s_{ij} = \frac{common^2}{friends_i * friends_j} \tag{1}$$

The influential people can be described as core degree, they have many followers, and less friends, so the core degree is defined as equation (2).

$$c_j = \log \frac{followers_j}{friends_j^{\,2}} \tag{2}$$

In order to recommend intent friends to common users, we define the transit probability as equation (3).

$$p_{ij} = \gamma(\alpha \times s_{ij} + \beta \times c_j)$$
$$st..\alpha + \beta = 1 \&\& 0 < \beta < 0.1 \&\& \sum_j p_{ij} = 1 \tag{3}$$

By superposing the contribution of each walker, q_{ij} is defined to describe the reach probability from user i to user j in 4 steps, $q_{ij} = \sum_{l=2}^{4} p_{ij}^{l}$. The top k nodes of node j with the highest probability should be recommended to user i as friend candidates.

IV. Random Forests based Ranking Model

In order to use the rich information of users to improve the recommend, ensemble learning methods [13] such as regression, gradient boosting decision trees, and neural networks are always used to build ranking models, while our solution uses random forests to create the final ensemble.

User features selection is shown in table II. There are three types of features selected for each user and their friend candidates. What's more, some pair features between each user and their each candidate are selected, shown in table III.

TABLE II. FEATURE SELECTION OF USERS

C1	Province Number, City Number, Gender, Verify or NOT, Verify Type, Created Time
C2	Followers Number, friends Number, Tweets Number, Bi-follows Number,
C3	Tag based User Group, Content based User Group

The features in C3 are got with a clustering algorithm [14].

TABLE III. FEATURE SELECTION OF PAIRS OF USERS

D1	Common Friends Number, Attractiveness s_{ij}
D2	Similarity of Followers Collection, Similarity of Friends Collection
D3	Similarity of Tags Collection, Similarity of Contents

The similarities in D2 and D3 are calculated with Jaccard Coefficient.

So we have the user vector [c1,c2,c3] and user-user vector [d1,d2,d3]. In order to improve the calculation speed, we only consider the three classes of user relations, the list of followers, the list of friends and the list of bi-follows, quantified with 0, 1 and 2. With these data, we should train a random forest model at first, and then to predict the ranking friends list of a user with the model, class 1 and class 2 are thought positive samples.

Random forests are an ensemble learning method for classification and regression, proposed by Leo Breiman [15], that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. We use an open source toolkit provided by scikit-learn [16].

V. Experimental Results

We crawl a dataset from Sina Weibo with 3 million users information, and choose 35201 users to do friends recommend,

Proc. of the Second Intl. Conf. on Advances In Computing, Communication and Information Technology- CCIT 2014.

ISBN: 978-1-63248-051-4 doi: 10.15224/ 978-1-63248-051-4-91

training dataset include 24740 users and the others are used for testing. The purpose is to predict a ranking friends list for each user, so the followers of a user are negative samples, and the friends and bi-follow users are positive samples.

Table IV shows the result of local random walk. If the top 100 users are recommended as friends, the precision is high, but the recall is low, because the algorithm only uses the users' relationship. The result is used as friend candidates, with the users' rich attributes information, random forests model is used as the ensemble learning method. Table V shows the result of random forests, we can see the precision and recall are improved obviously

TABLE IV. THE RESULT OF RANDOM WALK.

TOP N	Right Precision	Right recall
100	0.720	0.355
200	0.495	0.488
500	0.268	0.660
1000	0.154	0.759
2000	0.083	0.848
3000	0.057	0.935

TABLE V. THE RESULT OF RANDOM FORESTS.

Tree number	Classification precision	Right precision	Right recall
10	0.901	0.929	0.932
20	0.909	0.943	0.930
50	0.912	0.949	0.930
100	0.914	0.951	0.929

The parameters of the random forests model are set as : max_depth=None, min_samples_split=1, max_features=sqrt(total_features), and bootstrap samples are used (bootstrap=True).

VI. Conclusions

In the paper, we proposed a local random walk and random forests combined method to do friends recommendation for large social networks. Different from the KDD CUP 2012 dataset, we crawled 3 million users information from Sina Weibo, without the action history of users about accept or reject a recommendation, which is a sparse matrix, so we do not use CF based technologies. Our method can be used for any kinds of large social networks. First, we adopt a local random walk algorithm to get a collection of friend candidates. And then, random forests model is trained to ranking the friend list with rich users' attributes information. The experimental results show that the method has high precision and high recall at the same time.

Acknowledgment

This work was supported by National Grand Fundamental Research 973 Program of China under Grant No.2013CB329606; National Natural Science Foundation of China under Grant No.91124002; National Science and Technology Support Program of China under Grant No.2013BAH43F00-01; Key Project of Science and Technology in Henan Province (2014) under Grant No.144300510001; Transformation Project of Scientific and Technological Achievements in Henan Province (2014) under Grant No.142201210009; Chinese Universities Scientific Fund (BUPT2014RC0701).

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, 2005.

[2] Yanzhi Niu, Yi Wang, Gordon Sun, Aden Yue, Brian Dalessandro, Claudia Perlich, Ben Hammer, 2012. The Tencent Dataset and KDD-Cup'12. KDD-Cup Workshop

[3] Linyuan Lv and Tao Zhou, Link prediction in complex networks:Asurvey, Physica A, 390(2011)1150–1170

[4] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. Computer, 42, August 2009.

[5] Tianqi Chen,etc. Combining Factorization Model and Additive Forest for Collaborative Followee Recommendation, http://www.kddcup2012.org/workshop

[6] Xing Zhao, Scorecard with Latent Factor Models for User Follow Prediction Problem , http://www.kddcup2012.org/workshop

[7] Yunwen Chen, Context-aware Ensemble of Multifaceted Factorization Models for Recommendation Prediction in Social Networks, , http://www.kddcup2012.org/workshop

[8] Vineeth Rakesh, Dilpreet Singh, Bhanukiran Vinzamuri and Chandan K Reddy, Personalized Recommendation of Twitter Lists using Content and Network Information, Eighth International AAAI Conference on Weblogs and Social Media, 2014, pp416-425

[9] Ingmar Weber and Venkata Rama Kiran Garimella, Using Co-Following for Personalized Out-of-Context Twitter Friend Recommendation, Eighth International AAAI Conference on Weblogs and Social Media,2014, pp654-655

[10] Kristian Slabbekoorn, Tomoya Noro and Takehiro Tokuda, Towards Twitter User Recommendation Based on User Relations and Taxonomical Analysis, , EJC 2013: 115-132

[11] Marcelo Gabriel Armentano, Daniela Godoy, Analía A. Amandi: Followee recommendation based on text analysis of micro-blogging activity. Inf. Syst. 38(8): 1116-1127 (2013)

[12] J. Norris, Markov Chains (Cambridge Univ. Press, 1997).

[13] Cha Zhang and Yunqian Ma, Ensemble machine learning, Springer, 2012.

[14] ShanFeng,Ruifang Liu, Qinlong Wang and Ruisheng Shi, Word Distributed Representation Based Text Clustering, proceeding of CCIS2014, Nov. 27-29, Hong Kong.

[15] Breiman, Leo . "Random Forests". Machine Learning 45 (1): 5–32. 2001

[16] Random Forest, http://scikit-learn.org/stable/modules/ensemble.html