A Unified Microblog User Similarity Model for Online Friend Recommendation

Shi Feng[1,2], Le Zhang[1], Daling Wang[1,2], and Yifei Zhang[1,2]

[1] School of Information Science and Engineering, Northeastern University
[2] Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education, Shenyang 110819, P.R.China
{fengshi,wangdaling,zhangyifei}@ise.neu.edu.cn, zhang777le@gmail.com

Abstract. Nowadays, people usually like to extend their real-life social relations into the online virtual social networks. With the blooming of Web 2.0 technology, huge number of users aggregate in the microblogging services, such as Twitter and Weibo, to express their opinions, record their personal lives and communicate with each other. How to recommend potential good friends for the target user has been a critical problem for both commercial companies and research communities. The key issue for online friend recommendation is to design an appropriate algorithm for user similarity measurement. In this paper, we propose a novel microblog user similarity model for online friend recommendation by linearly combining multiple similarity measurements of microblogs. Our proposed model can give a more comprehensive understanding of the user relationship in the microblogging space. Extensive experiments on a real-world dataset validate that our proposed model outperforms other baseline algorithms by a large margin.

Keywords: Friend Recommendation, User Similarity, Linear Combination.

1 Introduction

In recently years, people are not satisfied with making friends with their schoolmates, colleagues, neighbors and so on. More and more people are willing to extend their real-life social relations into online virtual social networks. Microblogging services, such as Weibo and Twitter, have become very popular, because it allows users to post a short message named tweet or status for sharing viewpoints and acquiring knowledge in real time. According to statistics, by March 2013, there had been 536 million registered users in Sina Weibo and more than 100 million tweets are generated per day in Weibo. Among huge number of users, how to recommend potential good friends for these users has become a critical issue for both commercial companies and research communities.

Different from some traditional social networks, the characteristics of microblog make it more different for finding appropriate friends for the target users. In microblog, the users can follow someone without his or her permissions. Therefore,

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 286–298, 2014.

the friend links are more casual and informal in microblog than in other online social networks. The users may add a friend link to someone because they share similar hobbies, have similar tags, live nearby, have been to the same places or they have similar opinions and have just discussed about the same trending topics in microblog. These characteristics have posed severe challenges for potential good friend recommendation in microblog.

In this paper, we propose a novel unified microblog user similarity measurement model for online friend recommendation. Our proposed model can integrate multiple similarity measurements of microblog together by linear combination and learn corresponding weight for each measurement. As a result, we can provide a more comprehensive understanding of users' relationship in the microblogging space and recommend potential good friend for the target user. To summarize, the main contributions of our work are as follows.

(1) We leverage the massive real-world microblog data to analyze the characteristics of microblog features and determine which features are critical for friend recommendation.

(2) We proposed a microblog user similarity model for friend recommendation by linearly combining multiple similarity measurements of microblogs.

(3) We conduct extensive experiment on a real-world dataset. The experiment results validate the effectiveness of our proposed model and algorithm.

The structure of the rest of the paper is as follows. Related work is discussed briefly in Section 2. In Section 3, we analyze the crawled dataset and introduce the characteristics of the friend relationship in microblog. In Section 4, we propose the unified microblog user similarity measurement model. Section 5 introduces the settings and details of the experiments. We compare our proposed model with the other baseline models. In Section 6, we make a conclusion and point out the directions for our future work.

2 Related Work

Potential friend recommendation in social network is a hot research topic in the academic area. Guo et al. utilized the tag trees and relationship graph to generate the social network and employed the network topology and user profile to recommend friends [1]. Chin et al. focused on the friend recommendation in Facebook and LinkedIn. They considered the user daily behaviors as good features to indicate potential good friends for the target users [2]. Shen et al. leveraged three dimensions Utilitarian, Hedonic and Social to explore the recommendation model in the mobile terminals [3]. Yu et al. built heterogeneous information networks and transition probability matrix [4]. They conducted random walk on the built graph to recommend friends for the target users. Sharma et al. studied on the value of ego network for friend recommendation [5]. Chu et al. studied on the effect of location information in mobile social network for friend recommendation [6]. The tag-based and content-based friend recommendation algorithms were compared in [7]. The authors described a comprehensive evaluation to highlight the different benefits of tag-based and contend-based recommendation strategies.

Silva et al. analyzed the structure of the social network and utilized the topology of the sub-graphs to recommend potential friends [8]. Their algorithm significantly outperformed the traditional Friend-of-Friend method that is also a topology based algorithm. In [9], Moricz et al. introduced the friend recommendation algorithm People You May Know. Not only the precision, but also the speed of the algorithm was considered for the algorithm in MySpace. Bian et al. presented a collaborative filtering friend recommendation system MatchMaker based on personality matching [10]. The authors collected feedback from users to do personality matching, which provided the users with more contextual information about recommended friends.

Although a lot of papers have been published for friend recommendation, most of the existing literature focused on unique feature for recommendation. Actually, the potential good friends for a target user can be affected by multiple features, which is the basic assumption of this paper.

3 The Characteristics of Friend Relationship in Microblogs

Who is the target user's potential good friend in microblogs? It can be determined by many features because the microblog is full of personal and social relation information. To analyze the characteristics of friend relationship in microblogs, we have crawled huge number of microblog users from Weibo, which is the largest microblogging service in China. The statistics of our crawled dataset is shown bellow.

Table 1. The statistics information of the crawled dataset

Dataset Features	NO. of Features	Percentage of the Whole Dataset
Independent Users	1,459,303	--
Friend Links	3,853,864	--
Bi-directional Friend Links	9,646	--
Users with Tags	1,017,443	69.7%
Users with Location Information	1,292,942	88.6%
Users with Check-in Information	457,520	31.4%
Users with Hot topics	537,741	36.8%

In this paper, we define a friend link from user A to user B if user A follows user B in the microblog, and we say user B is a friend of user A. Due to the characteristics of Weibo, a friend link from user A to B does not necessarily mean that there is a link from user B to A. We have crawled more than 1.4 million microblog users with 3.8 million friend links and there are only about 9,646 links are bi-directional.

Besides the friend relationship, there are a lot of social information in microblogs, such as tags, location and check-in information. The tags are a set of key words that can describe the professions, interests and hobbies of the user. The location information describes the city where the user lives in. The check-in information denotes the GPS location that the user has been to. The hot topics are the trend topics

in microblogs that the user involved in. We compare the similarity of the users using the above features respectively and the results are shown in Table 2.

Table 2. The average similarity between users of friends and strangers

	Tag	Location	Check-in	Hot topic
Friends	4.4×10^{-3}	0.082	0.037	3.0×10^{-4}
Strangers	1.6×10^{-3}	0.038	0.022	2.3×10^{-4}

We can see from Table 2 that in the crawled dataset, the average similarity between friends calculated by specified feature is much higher than the average similarity between strangers. Based on these observations, we assume that tag, location, check-in and hot topic features are good indicators for friend recommendation. In the next section, we will recommend potential good friend for microblog users by linear combination of the multiple similarity measurements.

4 Friend Recommendation by Combining Multiple Measures

In the above section, we observe that tag, location, check-in and hot topic information are good indicators for users' interests, hobbies and professions in microblogs. In this section, we propose a linear combination approach to integrate these different similarity measurements together for friend recommendation. The overall framework of our proposed approach is shown in Figure 1.

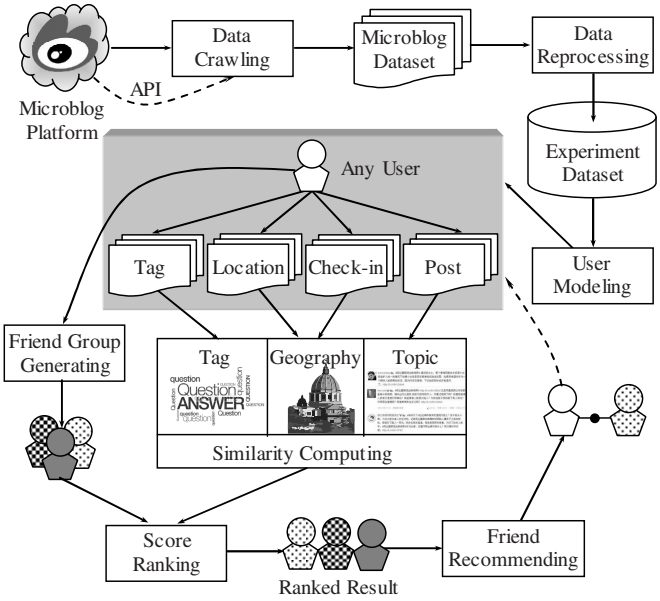

Fig. 1. The overall framework of the proposed unified friend recommendation model

4.1 Candidate Friend Set Generation

Due to the huge number of users in the microblogs, we can not traverse the whole microblogging space to find whether a user is a potential good friend for the target user. The detail of our candidate friend set generation algorithm is shown below. Generally speaking, in line 3-6 we select the friends of the user's friends into the candidate set. In line 7, we add the most popular users that are usually celebrities into the candidate friend set.

Algorithm 1. Generation of user's candidate friend set;

Input: The current friend list of users, the number of followers of users, the number of friends who will be recommended k;

Output: The candidate friends set of target user u;

Description:

1. FOR every target user u
2. Extract u's current friends set $f(u)$, $f(u)=\{u_1,u_2,\ldots u_n\}$.
3. FOR every user u_i IN $f(u)$:
4. Extract u_i's current friends set $f(u_i)$.
5. Generate u's first candidate friends set $f_r(u)$ from the current friends set of u and his friends set, $f_r(u)=\bigcup_{i=1}^{n} f(u_i)-f(u)$.
6. Select top-k users from $f_r(u)$ to build the second candidate friends set $f_{rt}(u)$ according to the number of common friends between the target user u and the candidate friend u_i.
7. Add k most popular users who have most followers and are not the current friends of u into the second candidate friends set to build the final candidate friends set $f_c(u)$ which contains $2k$ users.

4.2 User Tag Similarity

Tags are words or phrases that users utilize to describe online resources. These tags can indicate users' personal interests and hobbies. However, it is usually difficult to directly calculate the similarity between tags because many tags are out of the knowledge base such as WordNet and user-tag vectors are very sparse. Directly using cosine function and tags as vectors may lose many potential information for user similarity calculating. To tackle these challenges, in this paper we employ hierarchical clustering algorithm [11] to build the tags as a tree, based on which the user tag set similarity is calculated. The main steps of our proposed method are discussed as follows.

(1) Given the tag set T that extracted from all the users of the crawled dataset, eliminate the tags that have very low occurrence frequencies in T.
(2) Partition the tags in T based on hierarchical clustering algorithm and build a tag set tree, such as the example in Figure 2. The similarity between tags is calculated by their co-occurrences in T.
(3) Recalculate the personalized tag set similarity based on the tag tree.

If two tags do not co-occur in the T, or they are out of knowledge base such as WordNet, we could not calculate their similarity directly. In Figure 2, the root node contains all the tags in T, and the leaf node contains only one tag. It is obvious that the similarity between the two tags is bigger if there are fewer hops between them in tag tree. When there are the same hops, if the depth of the tags gets bigger, they become more similar with each other. The depth means the number of hops to the root node.

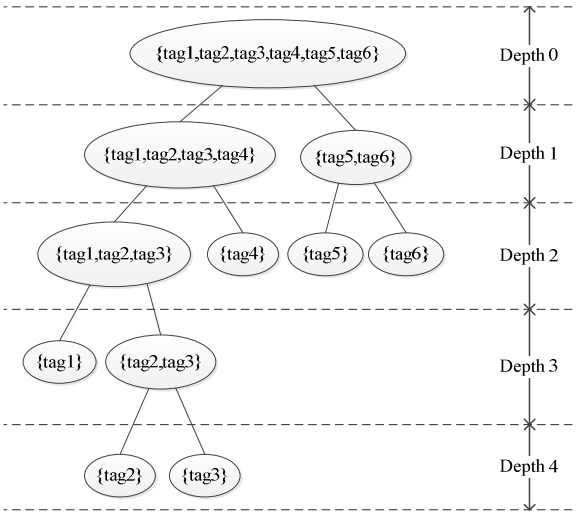

Fig. 2. The example of a tag tree

Given two user u_i, u_j, and their corresponding tag set $T(u_i)=\{t_1,t_2,t_3,\ldots,t_m\}$, $T(u_i)=\{t_1,t_2,t_3,\ldots,t_m\}$, $t_a \in T(u_i)$, $t_b \in T(u_j)$. The similarity between t_a and t_b can be calculated by the tag tree as:

$$sim_{tt}(t_a,t_b)=\frac{1}{\sum_{k,k+1\in SP(a,b)} 2/(d(k)+d(k+1))} \tag{1}$$

where a, b represent the node of t_a and t_b in tag tree respectively; $SP(a,b)$ denotes the nodes that in the shortest path between a and b; $d(k)$ denotes the depth of the node k in the tag tree. Therefore, the personalized tag similarity between u_i and u_j is calculated by the average similarity between $T(u_i)$ and $T(u_j)$ as:

$$sim_{ts}(u_i,u_j)=\frac{1}{|T(u_i)||T(u_j)|}\sum_{t_a\in T(u_i)}\sum_{t_b\in T(u_j)} sim_{tt}(t_a,t_b) \tag{2}$$

We normalized the value of sim_{ts} between users in the candidate friend set for the further calculation.

4.3 User Geography Similarity

Usually, the user's online friendship is a virtual reflection of the real world relationship. We observe that if the users have the same living city and usually come to the similar places, they intend to be friends with each other. In this section, we calculate the geography similarity of microblog users based on their location and check-in information.

Location Similarity. We utilize $sim_{ct}(u_i,u_j)$ to represent the location similarity between user $\underline{u_i}$ and u_j. If two users have the same location value, $sim_{ct}(u_i,u_j)=1$; Otherwise, $sim_{ct}(u_i,u_j)=0$.

Check-in Similarity. The Weibo platform has divided users' check-in information into twelve categories, such as "Train Station", "Library", "School" and so on. Therefore, each user's check-in information is represented by a vector with 12 dimensions, i.e. $chk(u)=\{cp_1,cp_2,\ldots,cp_{12}\}$, where cp_i is the proportion of the number of check-in category i in the latest 50 check-ins. We can use cosine function to calculate their similarity. So the check-in similarity between two users is calculated by $simchk(u_i,u_j)=\cos(\text{chk}(u_i),\text{chk}(u_j))$.

Finally, the geography similarity between users in microblogs is calculated by:

$$sim_{loc}(u_i,u_j)=\gamma\cdot sim_{ct}(u_i,u_j)+(1-\gamma)\cdot sim_{chk}(u_i,u_j) \tag{3}$$

where γ is the weight parameter. We will finally normalize the geography similarity for further similarity linear combination.

4.4 User Hot Topic Similarity

Usually users like to talk about the hot topics in microblog. The hot topic discussion that user takes part in could reflect his/her interests and hobbies. For user u_i, we employ Weibo API to extract the hot topics that u_i takes part in, and the extracted topic set is represented by $TP(u_i)$. The hot topic similarity between u_i and u_j is calculated by Jaccard similarity as:

$$sim_{tp}(u_i,u_j)=Jaccard(TP(u_i),TP(u_j))=\frac{|TP(u_i)\bigcap TP(u_j)|}{|TP(u_i)\bigcup TP(u_j)|} \tag{4}$$

In Formula 4, if two users have discussed more hot topics in common, they will have bigger similarity. We will finally normalize the hot topic similarity for further similarity linear combination.

4.5 A Unified Microblog User Similarity Model

In Section 3, we observe that the tag, location, check-in, and hot topic information are all good indicators for friend recommendation in microblog. In this paper, we propose a unified microblog user similarity model by linearly combining the multiple similarity measurements. Given a user u, and a user u_i from the candidate friend set of u, i.e. $u_i\in f_c(u)$, we have the unified similarity function:

$$sim(u,u_i) = \alpha \cdot sim_{ts}(u,u_i) + \beta \cdot sim_{loc}(u,u_i) + (1-\alpha-\beta) \cdot sim_{tp}(u,u_i) \quad (5)$$

where u_i is a candidate friend of u generated by Algorithm 1; sim_{ts}, sim_{loc}, and sim_{tp} are tag, geography and hot topic similarity respectively; α and β are weight parameters for the linear combination.

Given a target user u, we first employ Algorithm 1 to generate the candidate friend set $f_c(u)$. Then we traverse $f_c(u)$ to calculate the similarity between the candidate friend and the target user u. The top ranked K users will be extracted as recommended friends for the target user u.

5 Experiment

5.1 Experiment Setup

Our experimental dataset are crawled from Sina Weibo platform using API tool [12]. The detail statistics of the crawled dataset is show in Table 1 of Section 3. We conduct experiments using a PC with Inter Core i7, 8 GB memory and Windows 7 as the operation system.

We employ 5-fold cross validation to conduct the experiments. For each target user u, we randomly partition u's current friends and non-friends into 5 groups respectively. We randomly put one group of friends and one group of non-friends together to form a subset of the crawled data. For each run, four of the five subsets are used for training the parameters in Formula 5 and the remaining one subset is used for testing. We utilize Precision, Recall and F-measure to evaluate the performance of the proposed model and algorithms.

5.2 Experiment Results

Firstly, we learn the parameter γ for the Formula 3. In this experiment, we only utilize the geography similarity to recommend friends for the target user. The experiment results are shown from Figure 3 to Figure 5.

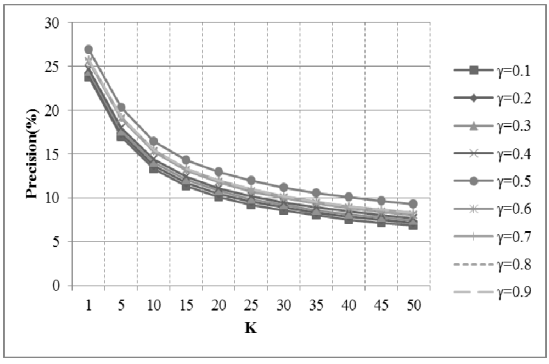

Fig. 3. The result of the parameter selection of the geography similarity model (Precision)

We can see from Figure 5 that when the number of recommended friends K gets bigger, the F-Measure of the model firstly increases dramatically and then gradually decreases. The best performance is achieved when K=10 and for all K settings, we can get the best performance using γ=0.5. Therefore, for the following experiments, we set γ=0.5.

Secondly, we learn the parameter α and β for the Formula 5. In Formula 5, we unified tag, location, check-in and hot-topic information together by linear combining. The experiment result is shown in Figure 6.

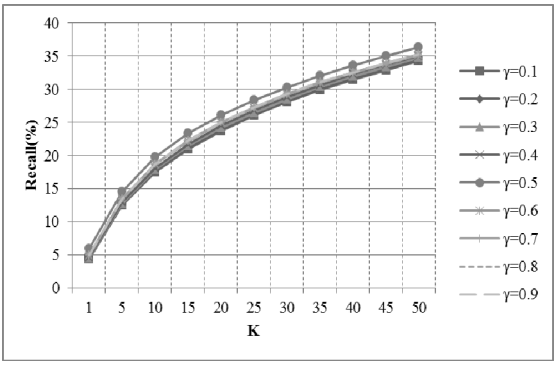

Fig. 4. The result of the parameter selection of the geography similarity model (Recall)

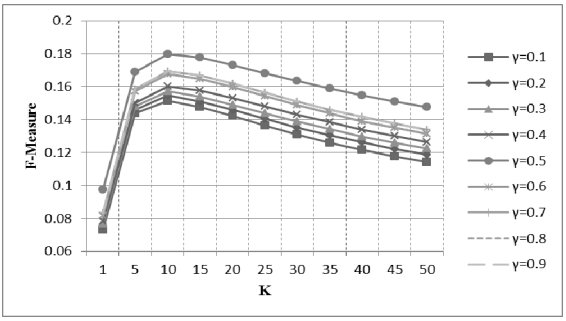

Fig. 5. The result of the parameter selection of the geography similarity model (F-measure)

In Figure 6, when α and β are small, the F-Measure of the unified model is relatively small. When α is fixed, F-Measure grows bigger as β grows. When β is fixed, F-Measure grows bigger as α grows. The best performance is achieved when α=0.4 and β=0.5, which are used as default settings for the following experiments.

To evaluate the effectiveness of the proposed unified model, we compare our method with some other friend recommendation algorithm.

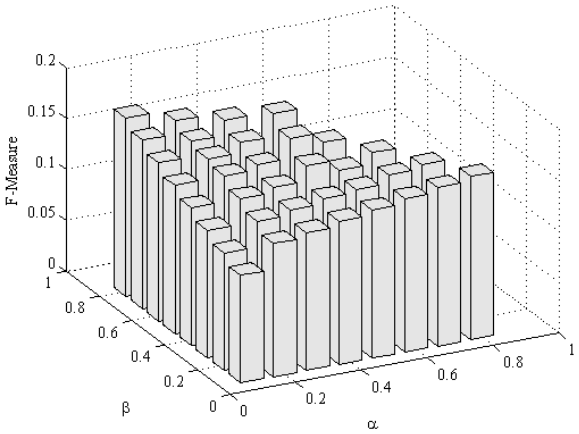

Fig. 6. The result of the parameter selection of the unified friend model

(1) FOF+ (Friends of Friends+). This method utilizes Algorithm 1 in Section 4.1 to generate the friends of the target user.

(2) FC (Follower Count). FC algorithm uses the number of followers to rank the candidate friends. The top K ranked users are extracted as the recommended friends.

(3) FOF+TS (Friends of Friends + Tag Similarity). This method utilizes Algorithm 1 to generate candidate friends set and employ the tag similarity to extract top K ranked users as recommended friends.

(4) FOF+LS (Friends Of Friends + Geography Similarity). This method utilizes Algorithm 1 to generate candidate friends set and employ the geography similarity to extract top K ranked users as recommended friends.

We denote the proposed Unified Microblog Friend Recommendation model as UMFR. We compare UMFR with above baseline methods, and the details are shown in Figure 7, Figure 8, and Figure 9.

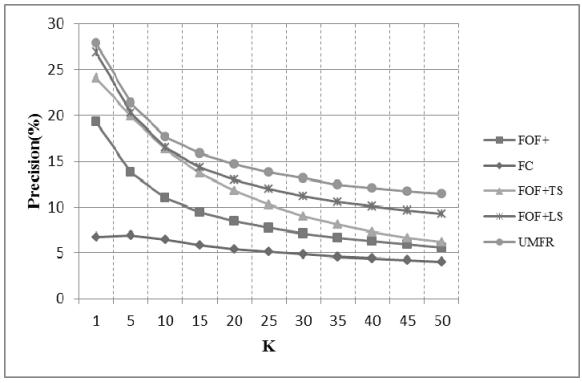

Fig. 7. The result of the friend recommendation algorithms (Precision)

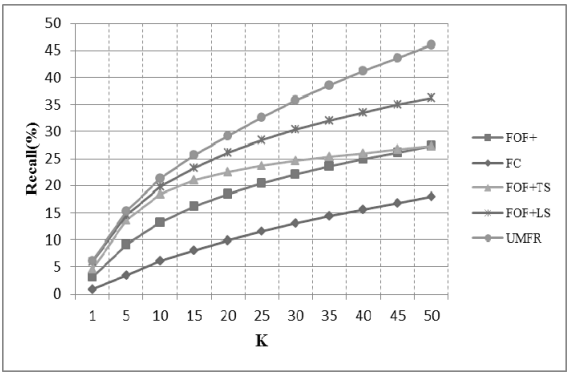

Fig. 8. The result of the friend recommendation algorithms (Recall)

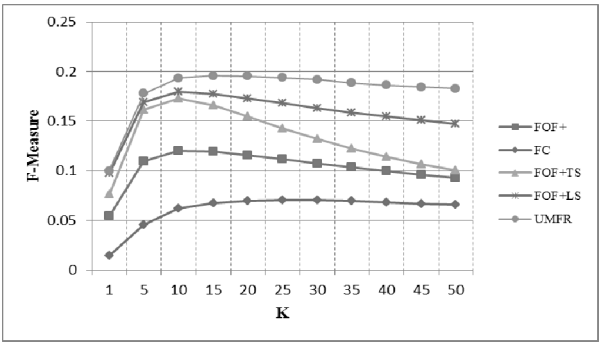

Fig. 9. The result of the friend recommendation algorithms (F-measure)

In Figure 7, the X-axis represents different number of the recommended friends; the Y-axis represents the precision of the algorithms. From Figure 7, we can see that when there are more recommended friends, the precisions of all the algorithms decrease gradually. With the same K setting, when K is small, there is no significant difference between our proposed method and geography based methods. The proposed method significantly outperforms the FOF+ and FC methods. This is because when K is small, the user with similar tags and locations are easily selected from the candidate friend set. The social relations and number of followers do not play a critical role in this case. When K gets bigger, our proposed method's precision has a smoother decline curve. This is because our proposed UMFR model that considering multiple feature measures provides a more comprehensive measurement for user similarities. The precision of FC method does not change much according to the K values, but obviously it has the worse performance of all the algorithms.

In Figure 8, the Y-axis represents the Recall of the algorithms. We can see from Figure 8 that the Recall of UMFR, FOF+LS and FOF+TS are similar to each other. However, when K gets bigger, our proposed method has better and better performance than the other compared algorithms.

In Figure 9, the Y-axis represents the F-Measure of the algorithms. We can see from Figure 9 that as K grows, the F-Measure of all the algorithms firstly get dramatically increases, and then drop down gradually. Our proposed UMFR model significantly outperforms other baseline methods and achieves the best performance when K is between 10 and 15.

6 Conclusions and Future Work

Recently, people are willing to make friends in the online social networks. Because of the multiple features, the traditional friend recommendation algorithms fail to capture the characteristics of the microblogs for user similarity measurement. In this paper, we propose a novel unified microblog user similarity measurement model for online friend recommendation. Our proposed model linearly combines multiple similarity measures of the users, which provides a comprehensive understanding of the user relationship in microblogs. Experiment results show that our proposed method significant outperforms the other baseline methods. Future work includes integrating more microblog features for measuring similarity between users for improving the quality of friend recommendation. We also intend to take the time factor into account, so that we can recommend different potential friends at different time.

Acknowledgements. This work is supported by the State Key Development Program for Basic Research of China (Grant No. 2011CB302200-G), State Key Program of National Natural Science of China (Grant No. 61033007), National Natural Science Foundation of China (Grant No. 61100026, 61370074, 61402091), and Fundamental Research Funds for the Central Universities (N120404007).

References

1. Gou, L., You, F., Guo, J.: SFViz: Interest-based friends exploration and recommendation in social networks. In: Proceedings of the Visual Information Communication-International Symposium (2011)
2. Chin, A., Xu, B., Wang, H.: Who should I add as a friend?: A study of friend recommendations using proximity and homophily. In: Proceedings of the 4th International Workshop on Modeling Social Media (2013)
3. Shen, X., Sun, Y., Wang, N.: Recommendations from friends anytime and anywhere: Toward a model of contextual offer and consumption values. Cyberpsychology, Behavior, and Social Networking 16(5), 349–356 (2013)
4. Yu, X., Pan, A., Tang, L.: Geo-friends recommendation in gps-based cyber-physical social network. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining, pp. 361–368 (2011)
5. Sharma, A., Gemici, M., Cosley, D.: Friends, strangers, and the value of ego networks for recommendation. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp. 721–724 (2013)
6. Chu, C., Wu, W., Wang, C.: Friend recommendation for location-based mobile social networks. In: Proceedings of Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 365–370 (2013)

7. Hannon, J., McCarthy, K., Smyth, B.: Content vs. Tags for Friend Recommendation. In: Research and Development in Intelligent Systems XXIX, pp. 289–302. Springer, London (2012)
8. Silva, N., Tsang, I., Cavalcanti, G., Tsang, I.: A graph-Based friend recommendation system using genetic algorithm. In: Proceedings of IEEE World Congress on Computational Intelligence, pp. 233–239 (2010)
9. Moricz, M., Dosbayev, Y., Berlyant, M.: PYMK: Friend Recommendation at MySpace. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 999–1002 (2010)
10. Bian, L., Holtzman, H.: Online friend recommendation through personality matching and collaborative filtering. In: Proceedings of the Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, pp. 230–235 (2011)
11. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York (1990)
12. Weibo API, `http://api.weibo.com`