An Approach to Social Relationship Ranking on Internet-based Social Platforms by Tempo-spatial Data Mining using Location Prediction Technique

Chao Ma[1], Yinda Wang[2], Haowen Liu[3,], Hao Gui[4], Weiping Zhu[5], Xiaochuan Shi[6], Xuhui Li[7]*

[1] International School of Software, Wuhan University, Wuhan, Hubei, P.R.China, chaoma@whu.edu.cn
[2] International School of Software, Wuhan University, Wuhan, Hubei, P.R.China, ydwang9367@whu.edu.cn
[3] International School of Software, Wuhan University, Wuhan, Hubei, P.R.China, hwenliu@whu.edu.cn
[4] International School of Software, Wuhan University, Wuhan, Hubei, P.R.China, hgui@whu.edu.cn
[5] International School of Software, Wuhan University, Wuhan, Hubei, P.R.China, cswpzhu@whu.edu.cn
[6] International School of Software, Wuhan University, Wuhan, Hubei, P.R.China, shixiaochuan@whu.edu.cn
[7] School of Information Management, Wuhan University, Wuhan, Hubei, P.R.China, lixuhui@whu.edu.cn
**Corresponding Author: hwenLiu@whu.edu.cn*

Keywords: social relationship ranking; Internet-based social platform; tempo-spatial data mining; location prediction

Abstract

During the last decade, we have witnessed the prosperity of the Internet-based social platforms and mobile social applications such as Facebook, Twitter, etc. Meanwhile, due to the popularity of mobile terminals such as smart phones and variety of PADs, it is feasible to obtain relatively accurate tempo-spatial data from mobile terminal holders when they visit and upload geo-tagged messages or pictures to Internet-based social platform. Therefore, it is observed that the volume of tempo-spatial social data posted on social platforms keeps increasing. This brings us more opportunities to mine the semantic information according to the analysis on the tempo-spatial social information collected from Internet-based social platforms. In this paper, we present an approach to social relationship ranking by mining the tempo-spatial social data. To deal with the sparsity of raw tempo-spatial social data, the location prediction technique is employed. According to the comparison between the social relationship ranking method with location prediction and its version without location prediction, it shows that the former outperforms the latter substantially. Finally, we make several helpful observations about the social relationship ranking method when location prediction technique is adopted.

1. INTRODUCTION

Compared with the traditional personal computers, intelligent mobile terminals such as smart phones and variety of PADs, due to their portability and convenience, are becoming more and more popular, especially among young end-users. According to the latest statistic data released by International Telecommunication Union (ITU), there are nearly 7 billion mobile terminals in use worldwide, which are almost four times more than personal computers by 2014. The survey conducted by China Internet Network Information Center shows that by the end of December 2014 there are more than 600 million users surfing on Internet while more than 80% of them access Internet via mobile terminals. And due to the fact that the production and sale of smart mobile terminals keep increasing, we can foresee that the aforementioned statistics about smart mobile terminals will become higher and higher. Meanwhile, as the continuous advancement of wireless communication technology and sensor technology, smart mobile terminals act not only as voice communication tools, but also personalized intelligent interface for environmental information acquisition & processing.

Nowadays, end-users especially the young users are very glad to use smart mobile terminals for visiting Internet-based social platforms due to the easy access and no physical place limitation. They can interact with their on-line friends via text, voice, pictures with/without geo-tags in an instant and convenient manner. In this way, it greatly enhances the efficiency of social interactions. And it also brings us more opportunities to provide attractive services such as potential friend recommendation service by analyzing their geo-tagged social data. So far, the tempo-spatial social data is widely utilized in variety of application fields such as recommendation systems [1, 2], social network modelling [8, 9, 10, 14] and other related systems [11].

One feature, which is becoming more and more obvious on Internet-based social networks, is the volume of tempo-spatial social data keeping increasing. In our previous work [16], it is shown that online social network users are glad to share their geo-tagged activities with their friends. And it is observed that people shared more and more geo-tagged activities than ever before. Obviously, this makes the volume of tempos-spatial social data being an important portion of the social data on Internet-based platforms. Therefore, to provide attractive and

This work is supported in part by Fundamental Research Funds for the Central Universities under grant No. 2042014kf0056, Science Foundation of Hubei Province under grant ZRY1708, NSF of China under grant No.61272110.

978-1-4673-7278-7/15 $31.00 © 2015 IEEE
DOI 10.1109/BigDataCongress.2015.56

personalized services such as potential friend recommendation, online marketing, etc., it is required to learn more semantic knowledge by mining the raw social data.

In this paper, we aim to propose an approach for social relationship ranking by mining tempo-spatial information collected from Internet-based social platforms. Here the tempo-spatial social data is the end-users' historical location records attached with timestamps. However, according to our observation in the previous work [16], it is found that the historical location records are relatively sparse which leads to the low social relationship prediction accuracy. To cope with the sparsity problem of tempo-spatial social data, we plan to employ the location prediction technique in our social relationship ranking method.

The rest of this paper is organized as followed: related works are introduced and analyzed in Section 2. In Section 3, the problem to be studied is defined to help us clearly understand the research goal. The methodology part of the social relationship ranking approach is illustrated in detail in Section 4. The experimental results and analytics are provided in Section 5. Finally, the conclusion and future works are presented in Section 6.

2. RELATED WORK

During the last decade, the prosperity and popularity of Internet-based social networks such as Facebook, Twitter, have drawn attentions from scientific researchers. Early research works such as [4, 12, 13] which mainly focus on semantic knowledge mining including structural property mining and evolution pattern recognition of online social networks. However, due to the lack of theoretical support, these research works cannot be easily applied to semantic feature extraction in realistic and large-scale social network modeling.

As the efficiency to acquire interaction data among end-users in online social networks being enhanced, inferring social link with rich semantic information becomes the reality. The new challenge brought by this research topic is how to design more effective and efficient online social network modeling technique. The representative works of this research topic include [15, 14, 7]. In the research work [15], it proposes the method using dual un-weighted link to denote social relationships among users and try to speculate potential social network topology by analyzing email sending/receiving history. It relies heavily on the email sending/receiving frequency between a pair of users in the social network. Then, it further determines whether there is a social link between two users by using predefined threshold. According to the experimental results and comprehensive analysis, it shows that the prediction accuracy of potential social links has an apparent correlation with the selection of predefined threshold. Most of the early research works focusing on online social network modeling have conducted similar research as done in [15].

However, the drawbacks of these works are also obvious:

- The email records are highly private which makes the raw data collection task being difficult to be completed;
- They simply regard the social links among end-users as equivalent. But the truth is that social links of different pairs of end-users are different in various aspects such as the strength of social links. For example, one social link is detected between end-users A and B. From A's point of view, B is his/her best friend, but B does not necessarily regard A as his/her best friend. Therefore, to model the social network in a more precise manner, the social link between two end-users should be bi-directed and has independent semantic features (such as social relationship strength, etc.).

Besides the binary social network modeling methods, alternatives are presented recently. Authors of [14] present a supervised method for social network modeling which constructs the social links with weights (i.e. strength). The model proposed is able to conduct the classification among end-users according to their social interaction data. In the research work of [7], it proposes a symbol weighted social network modeling technique to measure the strength of social links. To improve the measurement accuracy, a machine-learning framework is introduced for extracting the most representative interaction data.

Different from the above-mentioned research works, the research team of [6] tries to predict the existence of social relationships among end-users according to the location data collected from their mobile phones. Experiments conducted show that the accuracy for predicting the existence of social relationships can amazingly reach 95%. However, to achieve such high prediction accuracy, two prerequisites must be met. The first one is that all volunteers in experiments must continuously record their locations and upload all location records to the server periodically. The second one is that most of end-users' mobility pattern should be highly predicable. And these two prerequisites limit the application of this prediction strategy of [6] in real scenarios.

One fact we have noticed is that most of the datasets utilized in similar research works are elaborately collected and well organized. But in real scenarios, the fact or the challenge is that the volume of geo-tagged records collected from Internet-based social platforms could be very large, usually containing millions of geo-tagged records. The more challenging issue is that the average time gap between two adjacent location records is much longer than that in datasets containing frequently collected location records such as in the Wifi localization scenario. In this paper, we call such phenomenon as location record sparsity problem. To cope with this problem, we tend to employ the location prediction technique [21] which is capable of predicting the next location according to the historical location records. Hopefully, this will help us alleviate the location record sparsity problem. Different from above-mentioned works in social network modeling, in this paper we aim to propose the social relationship ranking method instead of measuring the strength of social relationships. Therefore, in the experimental part, we care more about the ranking sequence rather than the quantified values of social relationships.

3. PROBLEM DEFINITION

The definition of the problem we aim to address in this paper is given as follows:

Let a set of users be denoted as U = {U_1, …, U_i, …, U_j, …, $U_{|U|}$} who have posted geo-tagged information within the specific region R on social platforms. By analyzing the social data posted by all users, their tempo-spatial information (i.e. location records with timestamps) could be extracted. Each location record encompasses the latitude and the longitude with the unique timestamp. For the user U_i, his historical trajectory is denoted as L_i={$L_{i,1}$, …, $L_{i,x}$, …, $L_{i,Ri}$} ($L_{i,x}$=($Lat_{i,x}$, $Lon_{i,x}$) where $Lat_{i,x}$ and $Lon_{i,x}$ are the latitude and longitude of the x-th location record $L_{i,x}$), all location records of user U_i are ranked in the time ascending order according to their timestamps (i.e. for each location record $L_{i,x}$ is placed before the location record $L_{i,x+1}$). Furthermore, the timestamp attached to the location record $L_{i,x}$ is denoted as $T_{i,x}$ where for all the $1 \le x \le R_i-1$ the inequation $T_{i,x}<T_{i,x+1}$ is always satisfied.

Since we define the social relationship as bi-directed, only when two users are in the friend list of each other, one social relationship between them will be recorded. The social relationship from the user U_i to U_j is denoted as $S_{i,j}$ (note that $S_{i,j} \neq S_{j,i}$). The value of S(i, j) is measured by the method M. For all the non-zero S(i, j), we rank them according to the value of all S(i, j) in the descending order which is denoted as SR^M. Our goal for designing method M is to maximize the matching degree of SR^M and ground truth. The evaluation metric, which quantifies the matching degree, is defined by Eq. 4 in section 5.

4. METHODOLOGY

In part A of this section, we firstly introduce the location prediction technique employed for alleviating the location record sparsity problem. Then, in part B of this section, the social relationship ranking method of which the early work is presented in [16] will be illustrated in detail.

A. Location Prediction Technique

Location prediction is widely used in the participatory sensing systems for predicting the next place that certain mobile node will visit. The study [17] on human mobility pattern shows that the human movements present the highly repeated pattern. It is observed that most of the time people may visit several specific places with much higher probability than other places. On the basis of this observation, location predictors can be designed and utilized for predicting the place that certain mobile node may visit at next time point.

In this paper, to cope with the location record sparsity problem we have noticed in the tempo-spatial information collected from Internet-based social platforms, we employ the location prediction technique before we conduct the social relationship ranking. To the best of our knowledge, this is the first paper presenting the idea for applying location prediction technique in social relationship ranking.

To search for approaches to the location prediction problem, many researchers conducted a bunch of experiments and made a lot of fundamental observations in [17, 18, 19, 20]. In the research work of [17], it is found that human trajectory can be described by certain simple and repeated pattern according to the analysis on trajectories collected from 100000 volunteers over 6 months. Authors of [18] prove that human mobility pattern is predictable. And by applying the concept of "entropy", they propose the method to make the decision that whether human trajectory is predictable or not. By conducting the experiments on historical trajectories, the potential predictability of user trajectories could be higher than 90%.

On the basis of aforementioned theoretical and experimental support, different solutions to location prediction problem are presented, such as [21, 22, 23, 24, 25]. So far, the most widely used location prediction technique is Markov-Model-based location predictor. The basic idea of Markov-Model-based location predictor assumes that the next location can be predicted based on the current context. For the Order-K (denoted as O(K) for short) Markov predictor, the context is the sequence of the K most recent locations in the historical trajectory. By initializing the transition probability matrix, the O(K) Markov predictor can start to work. And every time no matter the O(K) Markov predictor makes right, wrong or no prediction, the transition probability matrix will be updated until the prediction process is terminated. Besides Markov predictors, LZ-tree based predictors are also proposed. However, in the research work of [21], Markov predictors outperform LZ-tree based predictors. Moreover, the O(2) Markov predictor which reaches the prediction accuracy as high as 72% in average is the best of all. Therefore, we employ the O(K) Markov predictor to at least alleviate the location record sparsity problem.

To illustrate how Markov predictor works, we can consider a user whose historical trajectory is $L=a_1 \ldots a_i \ldots a_j \ldots a_n$ where a_i is one of the area that this user has ever visited during determined time period. Please note that a_i and a_j could be the same area but with different visiting time. $L(i, j)=a_i a_{i+1} \ldots a_j$ is defined as the substring of L where $1 \le i \le j \le n$. The location of the user is regarded as the random variable X. $X(i, j)=X_i X_{i+1} \ldots X_j$ represents the sequence of random variants X_i, X_{i+1}, …, X_j where $1 \le i \le j \le n$. The current context is defined as c=L(n-k+1, n). Let R be the set of all possible areas which may be visited by all users. If the following equation is always true for all $a_i \in R$ and $i \in \{1, 2, \ldots, n\}$:

$$\begin{aligned} &P(X_{n+1}=a \mid X(1,n)=L) \\ &=P(X_{n+1}=a \mid X(n-k+1,n)=c) \\ &=P(X_{i+k+1}=a \mid X(i+1,i+k)=c) \end{aligned}$$

where the notation $P(X_{n+1}=a \mid X(1,n)=L)$ is the probability that at the next time point n+1 the user's location is a when the current context is X(1,n)=L. The first and second lines of the equation indicate the assumption for O(K) Markov predictor that the probability the user visited the location a only depends on the k most recent locations. The second and third lines of the equation indicate the assumption that the transition probability is the same when the context is the same.

The transition probability matrix denoted as TPM for short should record all such probabilities. With TPM and current context c, by checking all entries of TPM, it is easy to figure out the most probably location to be visited. And of course, the entry with highest probability indicates the most possible next location. To initialize TPM, we utilize the following equation to estimate the probability to visit location a as the next location when historical trajectory is L:

$$\hat{P}(X_{n+1} = a \mid L) = \frac{N(ca, L)}{N(c, L)}$$

where N(ca, L) and N(c,L) represent the number that substrings ca and c occur in the historical trajectory L. By analyzing this equation, we can list the cases that O(K) Markov predictor can not make prediction:

- The context c has never occurred in L ever before. Because it makes N(ca, L) and N(c,L) being 0 and 1 respectively. And the estimate transition probability is 0 for all locations in R.
- The length of the historical trajectory L is shorter than K. And it makes both of N(ca, L) and N(c,L) being 0. Hence, it is impossible to calculate the estimate transition probability. However, compared with the first case, this case only happens at the starting stage of O(K) Markov predictor.

To cope with the first no prediction case, one strategy is proposed in [21] which is called as "fallback". The idea is that for the O(K) Markov predictor when the current context c is never seen in L before, then the first location symbol of c is removed from c of which the remaining part forms the new context c' containing K-1 location symbols. And now the new context c' is used in the O(K-1) Markov predictor for next location prediction. If the O(K-1) Markov predictor still can not make prediction, the above steps will be conducted again and the new context c'' containing K-2 location symbols will be used in the O(K-2) Markov predictor until the prediction is made. The so-called "fallback" strategy is able to improve the prediction accuracy for around 3% while extra space and calculation costs become much heavy. Hence, in this paper, we do not adopt "fallback" strategy for Markov predictor.

During the execution of the O(K) Markov predictor, the transition probability matrix TPM keeps being updated. Different from classical Markov predictor, it is impossible for us to confirm whether the predicted location is right or not. The assumption is that if the social ranking method reaches better results by using Markov predictor, it shows the effectiveness of Markov predictor; otherwise we think it does not work for alleviating location record sparsity problem.

Another barrier to adopt Markov predictor for improving the social relationship ranking is that for Markov predictor each location symbol in historical trajectory L or predicted locations have no absolute time which is one of the key features for social relationship ranking. In the Markov predictor used in this paper, we let t_a be the timestamp of the location record a. If a is the predicted location by using the context c and historical trajectory L, the timestamp t_a of location a is calculated according to the following equation:

$$t_a = t_c + \frac{ATG(ca, L)}{N(ca, L)}$$

where t_c is the timestamp of the last symbol location of current context c and ATG(ca,L) is the average time gap between the last symbol location of all context c in L and the symbol location a followed by the context c.

One note is that in the Markov predictor employed in this paper we can not verify whether the predicted location is right or not. But by observing the social relationship ranking results, we can infer the effectiveness of the Markov predictor. Another issue about the Markov predictor used here is that the predicted locations will not be used for updating the transition probability matrix TPM due to the above reason.

B. Social Relationship Ranking

The main contribution of this paper is to propose the idea of introducing location prediction technique for coping with the location record sparsity problem in social relationship ranking. Therefore, the design of the social relationship ranking method is partially borrowed from our previous research work of [16] but with some modifications in this paper.

Before conducting the social relationship ranking, it is critical to quantify the social relationships (also known as social relationship measurement) just like the search engine which will calculate a value for each web page before ranking them. The rule for measuring the social relationship is that the social relationship tends to be strong if users have more social interactions.

Therefore, we can measure the potential social relationship between users U_i and U_j where $1 \le i, j \le |U|$ as designed in Eq. 1:

$$S(i, j) = Tmp(i, j) * \alpha + Spa(i, j) * \beta \qquad (1)$$

where variables α and β are utilized to optimize the final result S(i, j) for improving the accuracy of social relationship measurement, Tmp(i, j) and Spa(i, j) will be explained later. And to reduce parameters in Eq. 1, we make the constraint α+β=1.

As shown in Eq. 1, S(i, j) is consisting with two parts. The first part Tmp(i, j) is designed for inferring the possible interaction between users U_i and U_j from the temporal point of view. More specifically, we make a rule that the shorter the time gap between two location records from two different users is, the more possible there used to be real social interaction between these two users. Therefore, Tmp(i,j) is defined by Eq. 2 which is described as follows:

$$Tmp(i, j) = \sum_{1 \le i, j \le |U|, i \ne j} \left(\frac{1}{|T_{i,x} - T_{j,y}|}\right) \qquad (2)$$

where $T_{i,x}$ and $T_{i,y}$ are timestamps of two location records in historical trajectories of users U_i and U_j.

Similar to the definition of Tmp(i, j), Spa(i, j) is utilized for inferring the possible interaction between users U_i and U_j from the spatial point of view. More specifically, we make a similar rule that the shorter distance between two location records from two different users is, the more possible there used to be real social interaction between these two users. Therefore, Spa(i, j) is defined by Eq. 3 which is described as follows.

$$Spa(i, j) = \sum_{1 \le i, j \le |U|, i \ne j} \left(\frac{1}{dis(S_{i,x}, S_{j,y})}\right) \qquad (3)$$

where the function dis($S_{i,x}$ $S_{j,y}$) is the physical distance between two location records $L_{i,x}$ and $L_{j,y}$ in historical trajectories of users U_i and U_j.

To reduce the unnecessary calculation task, one constraint is made that only when the corresponding two locations are within the same "region" Tmp(i, j) and Spa(i, j) will be calculated according to Eq. 2 and Eq. 3. All "regions" have the same size. And each "region" is a rectangle of which the horizontal sides are the equal division on latitude while the vertical sides are the equal division on longitude. In the evaluation section, we conduct experiments with different "region" sizes to validate the effectiveness of social relationship ranking method with the Markov predictor.

5. EVALUATION

In this section, we first introduce the details about the data preparation work. And then in part B of this section, the evaluation metric is defined and explained. Finally, in part C, experiments with different configurations are conducted to evaluate the effectiveness of the proposed social relationship ranking method with Markov location predictor.

A. Data Preparation

To prepare the data for social relationship ranking purpose, we conduct following 2 steps: raw data collection and raw data pre-processing.

To collect social data from Internet-based social platforms, there are two classic ways. One way is to use web spider to download and parse related web pages for collecting useful information. The other way is to utilize the official APIs provided by Internet-based social platforms. The main advantage of web spider way is that it is able to reach high collection efficiency if well designed and implemented. The problem of web spider way is that it is not easy to identify useful information from dynamic web pages especially when those web pages are full of unrelated information such as advertisements. And after the raw data is collected by web spider, the cost for data pre-processing maybe much more than that for raw data collection. Different from web spider way, the official API-based way has the advantage that it is easy to identify the useful social data. Moreover, the social data collected via this way is of high-quality and requires few pre-processing efforts. The main disadvantage of official API-based way is that the access to official APIs is limited for preventing malicious attack purpose such as DoS attacks. Obviously, by using the official API-based way, it is not easy to achieve very high data collection efficiency. But by considering all relevant factors of web spider way and official API-based way, we make the decision to choose the official API-based way for the raw data collection.

Then, the next task for data collection is to determine the data source. We investigate 3 popular Internet-based social platforms in China including Sina Micro-blog, Tencent Micro-blog and Renren as the data source candidates. For these 3 Internet-based social platforms, we have following obvervations:

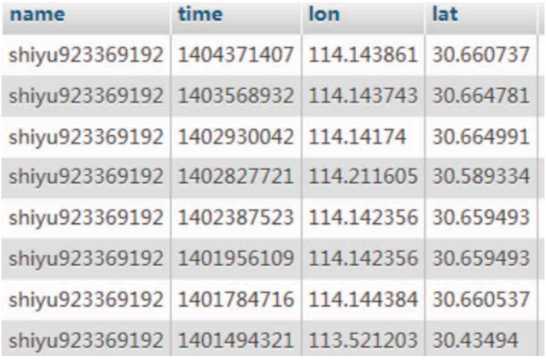
- Sina Micro-blog: user tempo-spatial information is intensive and can be obtained via official APIs, but social relationships cannot be conveniently collected due to the limitation of corresponding APIs;
- Tencent Micro-blog: user tempo-spatial information is relatively intensive and can be obtained via official APIs, and social relationships can also be conveniently collected by using corresponding APIs;
- Renren: user tempo-spatial information is intensive and can be obtained via official APIs, but social relationships cannot be conveniently collected due to the limitation on aacessing corresponding APIs.

According to the above observations, we finally select Tencent Micro-blog as the data source due to the reason that it provides comprehensive supports to efficiently obtain various kinds of social data including the tempo-spatial information of end-users without rigorous limitations. Before you can actually collect data, it is required to get the authorization from Tencent Micro-blog on which Oauth is used as the authorization management mechanism. After the access authorization is successfully obtained, the raw data collection can start.

The raw data collection is conducted as described below:

1. Select a region with longitude and latitude constraints. In this paper, the region is restricted within and around the campus of Wuhan University;
2. Repeatedly call the location record request API with the above region as the input parameter until no new location records are returned. This is because every time this API just returned part of location records within this region.
3. Traverse the location record list we get by executing step 2 and list all unique user IDs.
4. Traverse the unique user ID list and for each user ID extract the number that he/she "@" each of his/her friends. The more that he/she "@" one of his/her friends, the higher the rank of social relationship from his/her to that friend is. Hence, we take this value as the ground truth for validating the proposed social relationship ranking method.

Each location record collected encompasses 4 elements including the user ID ("name" column as shown in Figure 1), timestamp ("time" column as shown in Figure 1), longitude and latitude ("lon" and "lat" columns respectively). All social data we collected are permanently stored as database tables. Part of the raw location record database table is shown in Figure 1.



name	time	lon	lat
shiyu923369192	1404371407	114.143861	30.660737
shiyu923369192	1403568932	114.143743	30.664781
shiyu923369192	1402930042	114.14174	30.664991
shiyu923369192	1402827721	114.211605	30.589334
shiyu923369192	1402387523	114.142356	30.659493
shiyu923369192	1401956109	114.142356	30.659493
shiyu923369192	1401784716	114.144384	30.660537
shiyu923369192	1401494321	113.521203	30.43494

Figure 1. Raw Location Record

So far, we have finished the raw data collection work. But in the raw location record list it may contain incomplete or abnormal records. For instance, there may be some missing values on the latitude, longitude or timestamp. And some exceptional location records such as the latitude or longitude values being obviously out of the normal range may also be included in the raw dataset. Therefore, to guarantee the data quality, the pre-processing is necessary by eliminating incomplete or abnormal location records from the raw data.

Finally, after successfully conducting the raw data collection and pre-processing tasks, the dataset for validating the proposed social relationship ranking method contains: 962 users and 167514 location records over 4 months while there are 1203 social relationships among 962 users. The median value of location record number per user is 180 while the average value of location record number per user is 174. The detail about the dataset we collected is shown in Figure 2 as followed. The vertical axis of Figure 2 denotes the number of users who have location records within the certain range. It is obviously observed that users having location records within the range 180-210 are the most which reaches nearly 180 among 962 users in total.

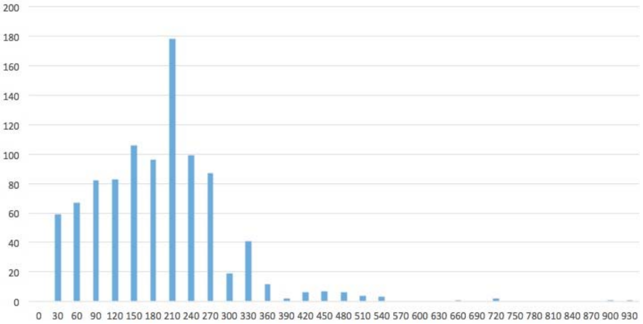


Figure 2. Statistics on Dataset

B. Evaluation Metric

To evaluate the effectiveness of the proposed social relationship ranking method, a metric called *R*anking *A*ccuracy of *S*ocial *R*elationship (denoted as *RASR*) is defined as shown in Eq. 4.

We let the friend list of the user U_i be $FL_i=\{fl_{i,1},\ldots, fl_{i,k},\ldots, fl_{i,fi}\}$ where FL_i is the subset of U and U_i is not contained by FL_i. All elements of FL_i are sequenced in descending order according to the number that U_i and each of his/her friends "@" each other. And FL_i is regarded as the ground truth for the ranking results on social relationships between user U_i and all his/her friends.

The estimated ranking list of user U_i derived from the social relationship ranking method M is denoted as $EL_i=\{el_{i,1},\ldots, el_{i,k},\ldots, el_{i,ei}\}$ where EL_i is the subset of U and U_i is not contained by EL_i. All elements of EL_i are sequenced in descending order according to the qualified value of social relationships measured by Eq. 1 between U_i and each of his/her friends.

The *RASR* of user U_i to all his/her friends is calculated as shown in Eq. 4.

$$RASR_i = \frac{1}{\sum_{y=1}^{|FL_i|} y} \sum_{j=1}^{|FL_i|} \frac{|FL_i| - j + 1)}{|j' - j| + 1} \qquad (4)$$

where j is the index of the j-th user in FL_i and j' is the index of the same user in EL_i. Obviously, the value of $RASR_i$ is bounded by the interval [0,1]. In this paper, to validate the effectiveness of the proposed social relationship ranking method in the quantitative manner, we use the cumulative distribution function (CDF) of RASR across all users.

C. Experimental Results

The statistic results of all experiments with different configurations (i.e. different Markov predictors and region sizes) are listed in Table 1. The first columns of Table 1 lists the social relationship ranking names of which SRR is used to denote the one we proposed in part B of the design section. And in this paper, 4 different versions of the social relationship ranking methods: SRR (without location prediction), SRR+O(1) MP (denote Markov Predictor for short), , SRR+O(2) MP and SRR+O(3) MP. The second column is the region size used by Eq. 2 and 3 for inferring possible interaction purpose. The value in the second column ranges from 10^{-1} to 10^{-6} which means the smallest unit for the side length of each region is from 10^{-1} to 10^{-6} of latitude in vertical and longitude in horizontal. Obviously, the size of the region becomes smaller and smaller. The third column is used to record the number of regions according to corresponding region size. For different Markov predictors with same region size, the number of regions keeps the same. The fourth column records the number of predicted locations by specific Markov predictor with certain region size. The last column of Table 1 lists the median RASR value across all users, which is regarded as the important metric for evaluating the effectiveness of the proposed social relationship ranking method.

Table 1. Statistic Results of Experiments

Ranking Method	Region Size	No. of Regions	Predicted Locations	Median of RASR
SRR	NA	NA	NA	0.172
SRR+ O(1) MP	10^{-1}	4740	130343	0.384
	10^{-2}	9860	131965	0.417
	10^{-3}	17630	132922	0.368
	10^{-4}	20570	130383	0.353
	10^{-5}	23200	129809	0.357
	10^{-6}	24050	129546	0.361
SRR+ O(2) MP	10^{-1}	4740	134012	0.323
	10^{-2}	9860	134673	0.358
	10^{-3}	17630	134775	0.346
	10^{-4}	20570	131601	0.313
	10^{-5}	23200	130043	0.301
	10^{-6}	24050	129315	0.286
SRR+ O(3) MP	10^{-1}	4740	133621	0.266
	10^{-2}	9860	133773	0.283
	10^{-3}	17630	135529	0.269
	10^{-4}	20570	131438	0.263
	10^{-5}	23200	129482	0.244
	10^{-6}	24050	128774	0.212

By analyzing the statistic results in Table 1, we make following helpful observations:

- By introducing location prediction technique into social relationship ranking method, the performance is apparently improved from 0.172 (SRR) to 0.417 (SPP+O(1) MP with the region size 10^{-1}) which is more than the double of SRR. And it is found that all ranking methods with location prediction get better median RASR values with more or less improvements. This indicates that the location prediction technique is indeed helpful for enhancing the performance of the social relationship ranking method based on tempo-spatial social data mining. Although, undoubtedly, some wrong predictions will be also made which may lower down the performance of SRR method.
- When the region size increases, the number of region increases also. But the newly added regions are not increased as more as initially. This phenomenon may show that location records tend to cluster within few places. And another interesting issue is that the performance of SRR with O(1), O(2) and O(3) MPs reach the highest point when the region size is 10^{-2}. We guess the reason is when the region size is too big the predicted location are of low spatial accuracy and when the region size is too small the possibility that Markov predictor makes no prediction may increases. Therefore, selecting an appropriate region size does have matter for tuning the Markov predictor-based social relationship ranking methods.
- It shows that the social relationship ranking method with lower order Markov predictor outperforms that with higher order Markov predictor. We infer that this is due to the fact that the location records of users in our dataset are quite sparse over time. And for high order Markov predictor it is difficult to detect multiple long contexts, which causes the bad performance of high order Markov predictor.

6. CONDLUSION

In this paper, to cope with the location record sparsity problem, we propose a novel idea for introducing the location prediction technique into social relationship ranking method. By conducting a set of experiments and comparing the results of different social relationship ranking methods, the effectiveness of O(K) Markov predictor for enhancing the performance of social relationship ranking method is confirmed. And on the basis of analysis on experimental results, several helpful observations are made which provide valuable clues for us in the future to further improve the overall performance of location prediction based social relationship ranking method.

Our future works will include following tasks.

- Since the dataset we use in this paper is of small-scale, the processing efficiency is not taken into account. However, if we would like to apply the proposed method in real scenarios, the efficiency issue will become a key point. Therefore, our first task n the future is to design and implement a parallel version of the location prediction based social relationship ranking method for greatly enhancing its processing efficiency.
- As shown in the experimental results, there is still big room for further improving the performance of location prediction based social relationship ranking. Therefore, the second task is to tune the Markov predictors and try to enhance the prediction accuracy of potential locations. Hopefully, it may help us increase the median RASR value of the proposed social relationship ranking method.

ACKNOWLEDGMENTS

We would like to thank Zihan Jiao for his excellent work on data preparation, Ting Wang for her help on providing improvement suggestions for the data analysis process and Jianxin Ren for his help on polishing this paper.

REFERENCES

[1] H. Yin, Y. Sun, B. Cui, Z. Hu and L. Chen. LCARS: A Location-Content-Aware Recommender System. In: Proceedings of KDD 2013. pp. 221-230

[2] W. Zhang, J. Wang and W. Feng., Combining Latent Factor Model with Location Features for Event-Based Group Recommendation. In: Proceedings of KDD 2013. pp. 910-918

[3] Kate Ching-Ju Lin , Chun-Wei Chen, Cheng-Fu Chou (2012) Preference-aware content dissemination in opportunistic mobile social networks. In: Proceedings of the 31st IEEE International Conference on Computer Communications, Orlando, FL, USA, INFOCOM'12, pp 1960-1968

[4] Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Groupformation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '06, pp 44–54

[5] Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks. KDD 2011, San Diego, CA, USA, 2011

[6] Eagle N, Pentland A, Lazer D (2009) Inferring social network structure using mobile phone data. Proc Natl Acad Sci USA 106(36):15274–15278

[7] Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. In: Proceedings of the 19th international conference on World wide web, ACM, New York, NY, USA, WWW '10, pp 641–650

[8] Wellman (2001) Computer networks as social networks, Science 2001: 293 (5537), 2031-2034

[9] Haibo Hu, Xiaofan Wang (2009) Evolution of a large online social network, Physics Letters A, Volume 373, Issues 12-13, 16 March 2009, Pages 1105-1110

[10] N. Eagle et al. (2009) Inferring Social Network Structure using Mobile Phone Data. In: Proc. of the National Academy of Sciences (PNAS), 106(36), pp. 15274-15278, 2009

[11] Y. Zheng et al. (2010) GeoLife: A Collaborative Social Networking Service among User, location and trajectory. IEEE Data Engineering Bulletin 33(2), 32-40, 2010

[12] Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '08, pp 160–168

[13] Singla P, Richardson M (2008) Yes, there is a correlation: from social networks to personal behavior on the web. In: Proceedings of the 17th international conference on World Wide Web, ACM, New York, NY, USA, WWW '08, pp 655–664

[14] Xiang R, Neville J, Rogati M (2010) Modeling relationship strength in online social networks. In: Proceedings of the 19th international conference on World Wide Web, ACM, New York, NY, USA, WWW '10, pp 981–990

[15] De Choudhury M, Mason WA, Hofman JM, Watts DJ (2010) Inferring relevant social networks from interpersonal communication. In: Proceedings of the 19th international conference on World Wide Web, ACM, New York, NY, USA, WWW '10, pp 301–310

[16] Chao Ma, Hao Gui, Haowen Liu, Weiping Zhu, Lv Xie. Inferring social relationship in mobile social networks using tempo-spatial information. In: Proceedings of International Conference on Software Intelligence Technologies and Applications, Hsinchu, Taiwan, December 2014, pp. 116-122

[17] Gonzalez M C, Hidalgo C A, Barabasi A L. Understanding individual human mobility patterns. Nature, 2008, 453(7196): 779-782

[18] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility. Science, 2010, 327(5968): 1018-1021

[19] Jiang B, Yin J, Zhao S. Characterizing the human mobility pattern in a large street network. Physical Review E, 2009, 80(2): 021136

[20] Qin S M, Verkasalo H, Mohtaschemi M, et al. Patterns, entropy, and predictability of human mobility and life. PloS one, 2012, 7(12): e51353

[21] Song L, Kotz D, Jain R, et al. Evaluating location predictors with extensive Wi-Fi mobility data// Proceedings IEEE INFOCOM 2004, the Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies. Hong Kong, China, 2004, 2: 1414-1424

[22] Scellato S, Musolesi M, Mascolo C, et al. Nextplace: a spatio-temporal prediction framework for pervasive systems. Pervasive Computing. San Francisco, USA : Springer Berlin Heidelberg, 2011: 152-169

[23] M AS., AN M A. A novel approach for protein subcellular location prediction using amino acid exposure. BMC bioinformatics, 2013, 14(1): 342

[24] Yavaş G, Katsaros D, Ulusoy Ö, et al. A data mining approach for location prediction in mobile environments. Data & Knowledge Engineering, 2005, 54(2): 121-146

[25] Akoush S, Sameh A. The use of bayesian learning of neural networks for mobile user position prediction. In: Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications. Rio de Janeiro, Brazil, 2007: 441-446