Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weighted hierarchical geographic information description model for social relation estimation

Kai Zhang [a,b], Xiaochun Yun [a,b,c], Xiao-Yu Zhang [a,*], Xiaobin Zhu [d], Chao Li [c,*], Shupeng Wang [a]

[a] Institute of Information Engineering, Chinese Academy of Sciences, 89 Minzhuang Road, Haidian District, Beijing 100093, China
[b] University of Chinese Academy of Sciences, 19 Yuquan Road, Shijingshan District, Beijing 100049, China
[c] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China
[d] Beijing Technology and Business University, 11 Fucheng Road, Haidian District, Beijing 100048, China

ARTICLE INFO

Article history:
Received 3 November 2015
Received in revised form
23 April 2016
Accepted 4 August 2016
Communicated by: Jinhui Tang

Keywords:
Social network
Social relation estimation
Weight computation
Geographic feature extraction
Hierarchical description
Semi-supervised learning

ABSTRACT

Social relation estimation has been attracting researchers' attention worldwide, and rapid development of LBSN (Location-Based Social Network) provides researchers an additional resource to estimate users' social relations. Previous works have fulfilled the social relation estimation with spatial information extracted from LBSN, while ignored or paid a little attention to the property of location. In this paper, a hierarchical grid based method is proposed to define location ID, and location's property is taken full advantage of when extracting features, in which way to exploit users' spatial information more sufficiently. Besides, in order to train a robust estimation model, we design the model based on semi-supervised learning. Our careful consideration of the above issues ultimately leads to a general framework that outperforms competitors, and experiments prove the effectiveness finally.

1. Introduction

Social relation estimation is an important task in the field of social network study, which is not only an interesting and challenging topic, but also the foundation of many applications, such as communication detection and recommender system [1,2]. For instance, people prefer to buy products which are recommended by their closely related friends or relatives, and the communications are also usually initiated among people who know each other. As a consequence, social relation estimation is of great significance and has drawn extensive attention.

Traditionally, social relation estimation is usually casted as the study of topological relation in a graph model. Researchers would train estimation model by using the observed links, and infer the remaining part of the network by recovering the missing links [3,4]. The method has been applied in other fields such as biology and complex network [5,6], but there exit two obvious disadvantages when estimating social relation with this traditional method. Firstly, the observed links are far from sufficient for reliable estimation in most cases. Secondly, related people do not necessarily contact within only one social network, which limits the estimation range. With the rapid development and the wide spread application of LBSN such as Foursquare, Twitter et al., people could constantly check in wherever they visit as a fashion way of sharing locations, and millions of check-in tags emerge every day. Therefore, we propose a framework with users' spatial information to fulfill the social relation estimation. We recognize that this is not the first paper to estimate social relation with spatial data, but there are important differences versus past works. Studies [7,8] have proved the correlation between users' social relations and their spatial locations. Other previous works [9–11] mainly focused on studying the methods of feature extraction, and proved the effectiveness with universal supervised or unsupervised learning methods. However, they ignored or did not pay enough attention to issues as follows:

- The definition of location's ID. The determination of scale of map division is a critical and non-trivial factor. As GPS could provide users' geographic information in high resolution, researchers usually acquire location's ID according to GPS information. In previous works, the map is divided into regions under an empirically specified scale. But if regions are divided in

* Corresponding authors.
E-mail addresses: zhangkai@iie.ac.cn (K. Zhang), yunxiaochun@iie.ac.cn (X. Yun), zhangxiaoyu@iie.ac.cn (X.-Y. Zhang), brucezhucas@gmail.com (X. Zhu), lc_lichao@126.com (C. Li), wangshupeng@iie.ac.cn (S. Wang).

http://dx.doi.org/10.1016/j.neucom.2016.08.030
0925-2312/

larger scale, it may lead to a high recall and a low precision. That is because a location in larger scale contains more check-ins, yielding more pairs of related users. On the contrary, if regions are divided in smaller scale, the precision will be high while the recall will be low. Therefore, it is hard to ensure the stable performance of estimation model under this situation.

- The property of location. When estimating social relation, the contribution of different locations varies. The more popular a location is, the less contribution it does. For instance, hundreds of customers would visit a coffee house each day, and many of them would happen to check in at the same time. As a result, we could not draw any conclusion about their social relations just according to this information. On the contrary, if a user appears in another's home, we would be convinced of their intimate social relation. However, the property of location has not received enough attention in previous works.
- The imbalance of data distribution. In a common social relation network, a few pairs of users own social relations. The ratio of users who own social relations and those do not is pretty small (usually less than 1%). The imbalanced data usually leads to lack of positive samples, yielding that the learned estimation model is not strong. Previous works reduced the ratio by decreasing negative samples before classification, but it would not help to get enough positive samples for training a strong estimation model.

In this paper, motivated by feature extraction works [12,13], we combine the process of feature extraction and classification, and propose a more robust model called Weighted Hierarchical Geographic Information Description Model (WHGIDM for short) for social relation estimation. Our key contributions are summarized below.

- WHGIDM takes the effect of check-in data's sparsity into account and divides the regions hierarchically.
- WHGIDM puts more emphasis on location's property and sets weight for each location when extracting features.
- WHGIDM makes full use of semi-supervised learning, and especially combines with the advantages of hierarchical ID definition method, in which way to expand training set and reduce the adverse impact caused by the imbalance of data distribution.
- Finally, we evaluate WHGIDM by using real-world datasets collected by location-based social networks, Gowalla and Brightkite, and we experimentally compare our model with state-of-the-art approaches. The results show that the WHGIDM's performance outperforms competitors'.

The rest of this paper is organized as follow: In Section 2, we formally define the problem and describe it in mathematic. In Section 3, we explain the methods for extracting features. Section 4 describes the algorithm for designing WHGIDM. Section 5 demonstrates experiment results. Section 6 introduces related works and Section 7 concludes this paper.

2. Problem formulation

In this part, we will briefly formalize the problem of social relation estimation. Generally, the estimation problem can be modeled as a classification problem. More specially, a binary classification model can be implemented to indicate the existence of relation between two users.

In this paper, we define social relation as: if two users are neighborhoods in the telecommunication network or friendship network, we assume the two users own social relation.

Let $U = (u_1, u_2, \ldots, u_m)$ denote users' ID set and e_{ij} denote the relation between user i and user j. If there is a social relation between the two users, the value of e_{ij}would be set as 1, otherwise 0. As there are $d = m(m-1)/2$ pairs of users' relation, i.e. $m(m-1)/2$samples, then we could get the social relation set $E = (e_1, \ldots, e_{ij}, \ldots, e_d)$, and the relation network could be described as an undirected graph $G = (V, E)$.

Like classification, the estimation also needs two attribute sets, feature set and class label set. Assuming the number of methods for extracting features is n, then we could obtain the feature set $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and the class label set $\boldsymbol{Y} = \boldsymbol{E} \in \mathbb{R}^{1\times d}$. Here each column of $\boldsymbol{X}$ represents features belonging to a pair of users, and each row of $\boldsymbol{X}$ represents the features extracted in a certain method; Each element of $\boldsymbol{Y}$ represents a pair of users' social relation correspondingly. As each pair of users could be viewed as a sample, then samples with $y = 1$ would be viewed as positive ones while the others negative. Then our work is to build a model with part of the d samples, and to estimate the remaining ones' class labels.

3. Method for hierarchical feature extraction

3.1. Location ID definition

Location ID means the mark of location. The most convenient way to define location ID is using its name like Starbucks, Sofitel et al. However, there are millions of locations in a common location-based social network, yielding only a few interactions between users in space would be observed in this way. Therefore, meshing is used to define location ID when studying users' social relations with LBSN data. As explained in Section 1, positions in large scale or small scale have their own advantages and disadvantages, so it is not proper to divide the map in single scale, and the combination of features extracted in different scales is crucial to achieve high performance.

In this paper, we exploit three scales to divide the Austin and Tokyo maps into grids with latitude and longitude values, including 0.1 degree, 0.01 degree and 0.001 degree. We define the girds obtained in 0.1 degree as large scale. Similarly, the girds in 0.01 degree are taken as middle scale and 0.001 degree as small scale. In the real world, 0.1 degree is equivalent to nearly one kilometer, which would be the size of a school, a community or a public park. 0.01 degree is equated with a few hundreds of meters that would be the size of a shopping place, a building in a school or a company. 0.001 degree is equal to dozens of yards, which may be the scope of a house or a restaurant. Then each check-in owns three location IDs in three different scales, and we could obtain three kinds of users' spatial interaction features.

3.2. Calculation of location's weight

Users' check-in data could reveal their tracks, and users appearing in the same locations are more likely in social relations. As a result, we exploit the similarity of users' behaviors extracted from spatial information where users check in to estimate their social relations.

As mentioned in the Section 1, different locations contribute differently to social relation estimation. Generally, locations with less users checking in would contribute more. However, note that although locations with less users checking in would contribute more for relation estimation, locations with only one user or even no user checking in do not help at all, as these locations would not reveal users' interaction in space. In this paper, we would pay more attention to location's weight when extracting features.

Then we'll describe the property in mathematic. Let D_k denote the weight of location k. And let $d(i,l_k)$ represent whether user i appears in location k. If appears, $d(i,l_k)=1$, otherwise $d(i,l_k)=0$. As for the locations where no user or just one user has checked in as

Please cite this article as: K. Zhang, et al., Weighted hierarchical geographic information description model for social relation estimation, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.030