doi:10.3772/j.issn.1006-6748.2017.02.012

# Mining potential social relationship with active learning in LBSN①

Wang Haiping (王海平)*, Zhang Hong②**, Wang Yong②**, Bing Jia***
(*Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, P.R. China)
(**National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, P.R. China)
(***Henan Worker's Cultural Palace, Zhengzhou 450007, P.R. China)

**Abstract**

Rapid development of local-based social network (LBSN) makes it more convenient for researchers to carry out studies related to social network. Mining potential social relationship in LBSN is the most important one. Traditionally, researchers use topological relation of social network or telecommunication network to mine potential social relationship. But the effect is unsatisfactory as the network can not provide complete information of topological relation. In this work, a new model called PSRMAL is proposed for mining potential social relationships with LBSN. With the model, better performance is obtained and guaranteed, and experiments verify the effectiveness.

**Key words**: data preprocessing, feature fusion, active learning

## 0 Introduction

Local-based social network (LBSN) is a new kind of social network where people could mark their positions information, and it developed rapidly in recent years. But LBSN differs from traditional social network as people could mark their positions information in the network. With the extensive use of smart phones, a large number of local-based social networks like Foursquares and Gowalla have emerged and have been drawing people's attention. Besides, traditional social networks like Facebook and Twitter also add the position information in their products to improve their popularity. In this way, people could publish their statuses in the form of text or picture marked with geographical information in LBSN.

Nowadays, millions of check-ins appear in LBSN every day, which provide sufficient information for the study of social network, including social relationship mining, recommendation of goods and services, community detection, etc. As mining of social relationship is the basis of many studies, it has been drawing wide attention of researchers. Traditionally, researchers use topological relation of social network or telecommunication network to mine potential social relationship[1]. As LBSN could be viewed as the combination of traditional social relationship and marks with position information, potential social relationships could also be mined by traditional methods, however, in which obvious disadvantage exists. As people do not always use a certain local-based social network to communicate with their friends, the relational network extracted from the LBSN could not fully cover their relationships. In other word, features extracted from the existing relational network could not describe the attribute completely. Therefore, researchers have studied mining the potential social relationships by using geographical position information.

Ref. [2] discovered the relation between relationships and geographical position information, and verified the effectiveness to infer potential relationships with geographical information. Ref. [3] defined computation methods to extract features from LBSN and mined potential relationships later. Ref. [4] also studied the problem of inferring links with geographical information to be proved the effective Ref. [5] proposed an entropy-based model (EBM) to infer social connections, and further estimate the strength of social connections with spatial information. However, the studies mentioned above mainly focused on designing features, and less considered about preprocessing data and improving the prediction model. In this paper, a new model called PSRMAL is brought out for mining potential social relationships to predict people's potential social relational network, combined with geographical information extracted from LBSN.

The rest of the paper is organized as follows: In Section 1, a method for designing PSRMAL is ex-

① Supported by the National Natural Science Foundation of China (No. 61501457).
② To whom correspondence should be addressed. E-mail: zhangh@isc.org.cn, wangyong@cert.org.cn
Received on Apr. 20, 2016

plained. The experiment is described in Section 2. Finally, conclusion is given in Section 3.

## 1 Design of PSRMAL

In this section, a method for devising model PSRMAL will be introduced in detail. The model can be viewed as two parts. The first is to extract features from LBSN, and the second is to train the model and further improve its performance. Fig. 1 shows the structure for mining potential social relationships from LBSN, and the process could be divided into four steps, i. e. region partition, feature computation, feature fusion and active learning. The detailed description is as follows.

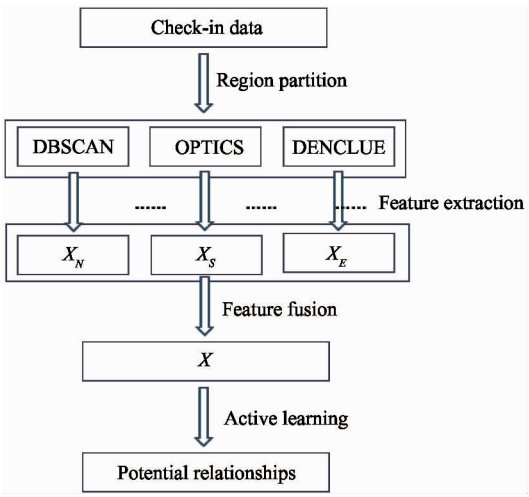



**Fig. 1** The structure for mining potential social relationships

### 1.1 Region partition

As each check-in corresponds to a GPS record, regions could be partitioned by clustering people's GPS records of check-in. Generally, there are four methods for clustering, i. e. partitioning methods, hierarchical methods, density-based methods and grid-based methods. Partitioning, hierarchical and grid-based methods are designed to find spherical-shaped clusters, while positions where people check in are not always in regular shapes. Therefore, density-based methods are used to obtain segment regions in this work. Although density-based cluster is suitable for partitioning regions, it is still insufficient to ensure the rationality. In this paper, partition regions are brought out with three density-based cluster methods, i. e. DBSCAN (density-based spatial clustering of applications with noise), OPTICS (ordering points to identify the clustering structure), and DENCLUE (clustering based on density distribution functions). After region partition, each check-in record corresponds to three region marks to guarantee the rationality of partition.

### 1.2 Feature computation

Generally, people appearing at common positions are likely to be friends[6]. The more frequently they do, the more likely to be. Therefore, three methods that have usually been used for similarity computation in social network are applied, i. e common neighbors, Jaccard index and Cosine index[7]. Here positions are used that people check in to replace the nodes in network. Then features could be computed with three methods and features ComP, JacP and CosP could be obtained respectively. Here ComP denotes the common positions that two persons have checked in, JacP denotes the value that computed with Jaccard index for two persons, and CosP denotes values computed with cosine index. The computational formulas are shown as follows.

$$\mathrm{Com}P_{ij} = \phi_i \cap \phi_j \tag{1}$$

$$\mathrm{Jac}P_{ij} = \phi_i \cap \phi_j / \phi_i \cup \phi_j \tag{2}$$

$$\mathrm{Cos}P_{ij} = \frac{\boldsymbol{P}(i) \cdot \boldsymbol{P}(j)}{\mid \boldsymbol{P}(i) \mid \mid \boldsymbol{P}(j) \mid} \tag{3}$$

### 1.3 Feature fusion with logistic

With the feature definition methods mentioned in subsection 1.2, three different kinds of feature sets $\boldsymbol{X}_N$, $\boldsymbol{X}_S$, $\boldsymbol{X}_E \in \mathbb{R}^{n\times d}$ could be extracted, and $\boldsymbol{X}_N$, $\boldsymbol{X}_S$, $\boldsymbol{X}_E$ denote the feature sets extracted from LBSN with DBSCAN, OPTICS and DENCLUE respectively. Then let $x_{Nij}$ denotes the $i$ pair of persons' position features extracted with $j$ feature computation method, while positions are obtained by DBSCAN. Similarly, let $x_{Sij}$ and $x_{Eij}$ denote features when positions are obtained by OPTICS and DENCLUE correspondingly.

Then the fusion feature could be computed as

$$x_{ij} = \alpha_j x_{Nij} + \beta_j x_{Sij} + \gamma_j x_{Eij} \tag{4}$$

where $\alpha_j$, $\beta_j$ and $\gamma_j$ represent the weights of different features with computation $j$, and $3n$ values will be obtained. To calculate the weighted values, the feature sets are united as

$$\boldsymbol{X}_U = [\boldsymbol{X}_N^T, \boldsymbol{X}_S^T, \boldsymbol{X}_E^T, \boldsymbol{R}^T]^T \in \mathbb{R}^{3n\times d} \tag{5}$$

where $\boldsymbol{R} = [1, 1\cdots 1]^T \in \mathbb{R}^{1\times d}$, and each column of $\boldsymbol{X}_U$ is

$$\boldsymbol{x}_{Ui} = (x_{Ni1}, x_{Ni2}\cdots, x_{Nin}, x_{Si1}, x_{Si2}, \cdots, x_{Sin}, x_{Ei1}, x_{Ei2}, \cdots, x_{Ein}, 1)^T \in \mathbb{R}^{(3n+1)\times 1}$$

Correspondingly, $\boldsymbol{W}$ can be expressed as $\boldsymbol{W} = [\alpha^T, \beta^T, \gamma^T, c]^T \in \mathbb{R}^{(3n+1)\times 1}$, where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2 \cdots \alpha_n]^T \in \mathbb{R}^{n\times 1}$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_n]^T \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{\gamma} = [\gamma_1, \gamma_2\cdots\gamma_n]^T \in \mathbb{R}^{n\times 1}$ are the weight sets of different features, and $c$ is a constant argument.

For convenience, form $\boldsymbol{W} = (w_1, w_2 \cdots w_{3n}, c)^T \in \mathbb{R}^{(3n+1)\times 1}$ is used to denote the weight set, and the

weight sets of different features are $\boldsymbol{\alpha} = [w_1, w_2 \cdots w_n]^T \in \mathbb{R}^{n\times 1}$, $\boldsymbol{\beta} = [w_{n+1}, w_{n+2} \cdots w_{2n}]^T \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{\gamma} = [w_1, w_2 \cdots w_n]^T \in \mathbb{R}^{n\times 1}$.

In this work, logistic regression is applied to calculate the parameters. The probability that two persons have relationship or not can be expressed as

$$p(y = 1 \mid \boldsymbol{x}_U, \boldsymbol{W}) = \frac{1}{1 + e^{-W^T x_U}} \tag{6}$$

$$p(y = 0 \mid \boldsymbol{x}_U, \boldsymbol{W}) = 1 - p(y = 1 \mid \boldsymbol{x}_U, \boldsymbol{W}) = \frac{1}{1 + e^{W^T x_U}} \tag{7}$$

Let $h_W(\boldsymbol{x}_U) = g(\boldsymbol{W}^T \boldsymbol{x}_U)$, and combine the two equations above to obtain $p(y \mid \boldsymbol{x}_U, \boldsymbol{W}) = h_W(\boldsymbol{x}_U)^y (1 - h_W(\boldsymbol{x}_U)^{1-y})$

The likelihood function could be expressed as

$$\boldsymbol{L}(\boldsymbol{W} \mid \boldsymbol{x}_U, \boldsymbol{y}) = \prod_{i=1}^{d} p(y^i \mid \boldsymbol{x}_U^i, \boldsymbol{W}) = \prod_{i=1}^{d} h_W(\boldsymbol{x}_U)^{y^i} (1 - h_W(\boldsymbol{x}_U)^{1-y^i}) \tag{8}$$

And the log-likelihood is

$$l(\boldsymbol{W}) = ln(\boldsymbol{L}(\boldsymbol{W} \mid \boldsymbol{x}_U, y)) = \sum_{i=1}^{d} y^i ln(h_W(\boldsymbol{x}_U^i)) + (1 - y^i) ln(1 - h_W(\boldsymbol{x}_U^i)) \tag{9}$$

When $l(\boldsymbol{W})$ reaches the maximum, $\boldsymbol{W}$ is the weight set. In this study, gradient decent is used to solve it. The elements of $\boldsymbol{W}$ can be got as

$$\begin{aligned}
&\frac{\partial}{\partial w_j} l(\boldsymbol{W}) \\
&= \frac{\partial}{\partial w_j} \sum_{i=1}^{d} y^i ln(h_W(\boldsymbol{x}_U^i)) + (1 - y^i) ln(1 - h_w(\boldsymbol{x}_U^i)) \\
&= \left( \frac{y^i}{h_W(\boldsymbol{x}_U^i)} - (1 - y^i) \frac{1}{1 - h_W(\boldsymbol{x}_U^i)} \right) \frac{\partial}{\partial w_j} h_W(\boldsymbol{x}_U^i) \\
&= \left( \frac{y^i}{g(\boldsymbol{W}^T \boldsymbol{x}_U^i)} - (1 - y^i) \frac{1}{1 - g(\boldsymbol{W}^T \boldsymbol{x}_U^i)} \right) \frac{\partial}{\partial w_j} g(\boldsymbol{W}^T \boldsymbol{x}_U^i) \\
&= \left( \frac{y^i}{g(\boldsymbol{W}^T \boldsymbol{x}_U^i)} - (1 - y^i) \frac{1}{1 - g(\boldsymbol{W}^T \boldsymbol{x}_U^i)} \right) g(\boldsymbol{W}^T \boldsymbol{x}_U^i) \\
&\quad (1 - g(\boldsymbol{W}^T \boldsymbol{x}_u^i)) \frac{\partial W^T x_U^i}{\partial w_j} \\
&= \left( \frac{y^i}{g(\boldsymbol{W}^T \boldsymbol{x}_U^i)} - (1 - y^i) \frac{1}{1 - g(\boldsymbol{W}^T \boldsymbol{x}_U^i)} \right) g(\boldsymbol{W}^T \boldsymbol{x}_U^i) \\
&\quad (1 - g(\boldsymbol{W}^T \boldsymbol{x}_U^i)) \frac{\partial W^T x_U^i}{\partial w_j} \\
&= (y^i (1 - g(\boldsymbol{W}^T \boldsymbol{x}_U^i) - (1 - y^i) g(\boldsymbol{W}^T \boldsymbol{x}_U^i)))) \boldsymbol{x}_{Uj} \\
&= (y^i - h_W(\boldsymbol{x}_U^i)) \boldsymbol{x}_{Uj}
\end{aligned} \tag{10}$$

Then $w_j$ could be got when the partial derivate converges, and the result is

$$w_j := w_j + \eta (y^i - h_W(\boldsymbol{x}_U^i)) x_{Uj}^i \tag{11}$$

### 1.4 Active learning

In this work, active learning is used to improve the performance for mining of potential social relationship. Active learning is proposed relative to passive learning[8-13]. Passive learning refers to selecting samples randomly from the dataset and label them, then trains a model with the labeled samples and classify unlabeled ones with the model. However, there may exist problems such as information redundancy, excessive noise, as the samples for training models are fetched randomly, which would seriously affect the effectiveness of classification[14-19]. Active learning is to divide the sample labeling work into two steps. Firstly, it is to label a few samples as the initial training set, for training a basic classifier, and label the other samples with the classifier. Secondly, it is to select a certain number of samples that are hard to confirm their classes according to the result, and label these ones manually. The new labeled samples to the initial training set are added, and the final classifier with the new training set is trained[20-25]. As the new training set fetched in this way contains more comprehensive information of the dataset, a more robust model would be obtained.

## 2 Experiment

In this section, the performance of PSRMAL is evaluated with a classic algorithm, support vector machine (SVM). Besides, social relationships are also mined with three single features as contrast experiments. Firstly, the features should be extracted with three different methods. Secondly, the features are fused with logistic regression, while the parameter $\boldsymbol{W}$ for fusion is shown in Table1.

Table 1 weight set

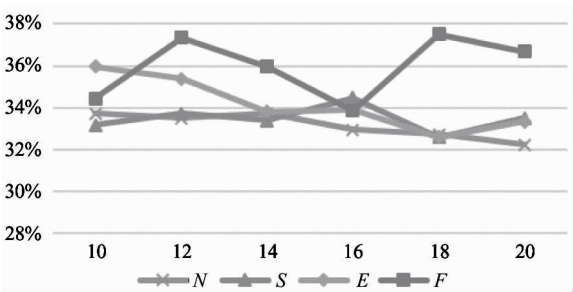
| Parameter | $\boldsymbol{\alpha}$ | $\boldsymbol{\beta}$ | $\boldsymbol{\gamma}$ |
|---|---|---|---|
| | 0.060 | 0.029 | 0.044 |
| | 1.309 | 4.103 | 24.632 |
| | 0.017 | 4.026 | 9.824 |

With the value of weight set $\boldsymbol{W}$, each member of fusion feature $\boldsymbol{X}$ could be obtained as

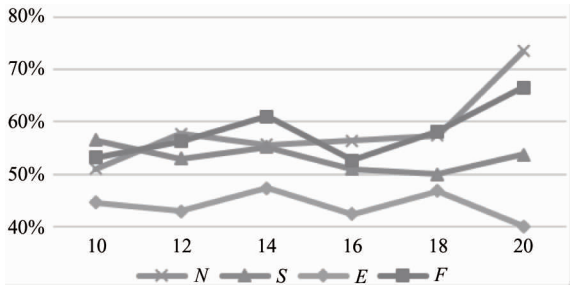
$$X_{ij} = \alpha_j x_{Nij} + \beta_j x_{sij} \tag{12}$$

Lastly, the potential social relationships will be mined by using $\boldsymbol{X}_N$, $\boldsymbol{X}_S$, $\boldsymbol{X}_E$ and $\boldsymbol{X}$ respectively. In Fig. 2(a) ~ (c), it is to express the experiment results of different methods for dividing positions, including DBSCAN, OPTICS, DENCLUE and the fusion. Let $\boldsymbol{N}$ denote the positions divided with DBSCAN, $\boldsymbol{S}$ denote

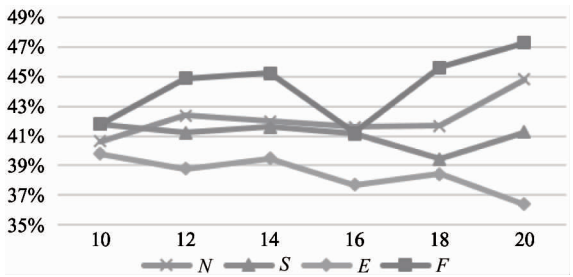
OPTICS, ***E*** denote DENCLUE and ***F*** denote the fusion. As can be seen in the figures, the performance of fusion feature outperforms single features in almost all cases. The fusion is crucial to guarantee the stability of model and achieve high performance. Besides, active learning also contributes to enhance the performance. As the comprehensive assessment, F-measure is more persuasive. The values obtained by using $\boldsymbol{X}_N$, $\boldsymbol{X}_S$, $\boldsymbol{X}_E$ and $\boldsymbol{X}$ are 42.18%, 41.08%, 38.43% and 44.34%, and the F-measure of fusion feature is boosted by 2.16%, 3.26% and 5.92% respectively.




(a) Index of precision




(b) Index of recall




(c) Index of F-measure

**Fig. 2** The experiment results of different methods for dividing positions

## 3 Conclusion and future work

In this study, a new model PSRMAL is proposed for mining potential social relationships with geographical information in LBSN. The importance of region partition is emphasized and the region is segmented with three different cluster methods, in which the rationality of partition is ensured. Then the features are fused with logistical and the performance of model is further improved with active learning mothed. Experiments prove the effectiveness of PSRMAL. In the future, more energy will be put to the efficiency of model to make it more suitable for real-time processing.

**Reference**

[1] Adamic L A, Adar E. Friends and neighbors on the web. *Social Networks*, 2001, 25(3): 211-230

[2] Cho E, Myers S A, Leskovec J. Friendship and mobility user movement in location-based social networks. In: Proceeding of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2011. 1082-1090

[3] Wang D, Pedreschi D, Song C, et al. Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2011. 1100-1108

[4] Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks. In: Proceedings the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2011. 1046-1054

[5] Pham H, Shahabi C, Liu Y. EBM: an entropy-based model to infer social strength from spatiotemporal data. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, USA, 2013. 265-276

[6] Pham H, Hu L, Shahabi C. Towards integrating real-world spatiotemporal data with social networks. In: Proceedings of the 19th ACM SIGSPATIAL, New York, USA, 2011. 453-457

[7] Moyano L G, Thomae O R M, Frias-Martinez E. Uncovering the spatio-temporal structure of social networks using cell phone records. In: Proceedings the 12th International Conference on Data Mining Workshops (ICDMW 2012), Brussels, Belgium, 2012. 242-249

[8] Zhang X Y, Wang S, Yun X. Bidirectional active learning: a two-way exploration into unlabeled and labeled dataset. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(12): 3034-3044

[9] Zhang X Y, Wang S, Zhu X, et al. Update vs. upgrade: modeling with indeterminate multi-class active learning. *Neurocomputing*, 2015, 162: 163-170

[10] Zhang X. Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing*, 2014, 127(3): 200-205

[11] Zhang X Y, Cheng J, Xu C, et al. Multi-view multi-label active learning for image classification. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Cancun, Mexico, 2009. 258-261

[12] Zhang X Y, Xu C, Cheng J, et al. Automatic semantic annotation for video blogs. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hannover, Germany, 2008. 121-124

[13] Zhang X Y, Cheng J, Lu H, et al. Selective sampling based on dynamic certainty propagation for image retriev-

al. In: Proceedings of the Advances in Multimedia Modeling (MMM), Kyoto, Japan, 2008. 425-435
[14] Zhang X Y, Cheng J, Lu H, et al. Weighted co-SVM for image retrieval with MVB strategy. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), San Antonio, USA, 2007. 517-520
[15] Wang S, Zhang X Y, Yun X, et al. Joint recovery and representation learning for robust correlation estimation based on partially observed data. In: Proceedings of the IEEE International Conference on Data Mining Workshop, Atlantic City, USA, 2015. 1-7
[16] Zhang X Y. Preference modeling for personalized retrieval based on browsing history analysis. *IEEJ Transactions on Electrical and Electronic Engineering*, 2013, 8 (S1): 81-87
[17] Zhu X B, Jin X, Zhang X Y, et al. Context-aware local abnormality detection in crowded scene. *Science China Information Sciences* (*SCIS*), 2015, 58(5): 1-11
[18] Zhu G, Wang J, Wu Y, et al. MC-HOG correlation tracking with saliency proposal. In: Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, USA, 2016. 1-7
[19] Zhang Y, Xu C, Zhang X, et al. Personalized retrieval of sports video based on multi-modal analysis and user preference acquisition. *Multimedia Tools and Applications*, 2009, 44(2): 305-330
[20] Zhang X Y, Hou Z, Zhu X, et al. Robust malware detection with dual-lane AdaBoost. In: Proceedings of the IEEE International Conference on Computer Communications, San Francisco, USA, 2016. 1051-1052
[21] Zhang X Y, Zhang K, Yun X, et al. Location-based correlation estimation in social network via collaborative learning. In: Proceedings of the IEEE International Conference on Computer Communications, San Francisco, USA, 2016. 1073-1074
[22] Zhang X Y, Wang S, Zhang L, et al. Ensemble feature selection with discriminative and representative properties for malware detection. In: Proceedings of the IEEE International Conference on Computer Communications, San Francisco, USA, 2016. 674-675
[23] Zhang Y, Zhang X, Xu C, et al. Personalized retrieval of sports video. In: Proceedings of the ACM Multimedia Workshop, Augsburg, Germany, 2007. 313-322
[24] Zhang X Y. Effective search with saliency-based matching and cluster-based browsing. *High Technology Letters*, 2013, 19(1): 105-109
[25] Zhang X Y. Dynamic batch selective sampling based on version space analysis. *High Technology Letters*, 2012, 18(2): 208-213

**Wang Haiping**, born in 1987. He is in pursuit of Ph. D degree, and is currently an engineer in Institute of Information Engineering, Chinese Academy of Sciences. He received his Master degree from College of Information of Renmin University of China in 2012. He also received his B. S. degree from Beijing Technology and Business University in 2009. His research interests include the design of algorithms for parallel processing, big data analysis and text mining.