Location-Interest-Aware Community Detection for Mobile Social Networks Based on Auto Encoder

Ming Chen, Wenzhong Li(✉), Sanglu Lu(✉), and Daoxu Chen

State Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, China
mchenflood@gmail.com, {lwz,sanglu}@nju.edu.cn

Abstract. Community detection partitions users in social networks into sub-groups according to structural or behavioral similarities, which had been widely adopted by a lot of applications such as friend recommendation, precision marketing, etc. In this paper, we propose a location-interest-aware community detection approach for mobile social networks. Specifically, we develop a spatial-temporal topic model to describe users' location interest, and introduce an auto encoder mechanism to represent users' location features and social network features as low-dimensional vectors, based on which a community detection algorithm is applied to divide users into sub-graphs. We conduct extensive experiments based on a real-world mobile social network dataset, which demonstrate that the proposed community detection approach outperforms the baseline algorithms in a variety of performance metrics.

Keywords: Location based service · Mobile social network · Interest community detection

1 Introduction

Community detection is an important topic in mobile social network, which could help us to understand mobile social network structures and provide good location based services. Community detection in mobile social network has a broad range of applications such as Ad pushing, crowd-sourcing, marketing etc. There are two types of community detection algorithms in mobile social network. The first type is graph-based community detection algorithms such as Newman algorithm [11], Louvain algorithm [2] and Label Propagation Algorithm (LPA) [14]. These state-of-the-art graph-based community detection algorithms could find cohesive subgroups in a network based on graph theory, while ignore the characteristics of node in the network. The second type is semantic-based community detection method [1] can explore overlapping semantic communities from the perspective of characteristics analysis. In mobile social network, users' location interest is very important for location-based services which indicate users' location preference.

C. Douligeris et al. (Eds.): KSEM 2019, LNAI 11775, pp. 176–184, 2019.
https://doi.org/10.1007/978-3-030-29551-6_16

Many existing community detection approaches [6,8,9] did not take mobile users' location interest into account, that cannot support personalized location-based services.

We take mobile users' location interest into account which could support personalized location-based services. In this paper, we propose an location-interest-aware community detection approach for mobile social networks. Specially, we develop a spatial-temporal topic model to describe users' location interest, and introduce an auto encoder mechanism to represent users' location features and social network features as low-dimensional vectors, based on which a community detection algorithm is applied to divide users into sub-graphs. We conduct extensive experiments based on a real-world mobile social network dataset, which demonstrate that the proposed community detection approach outperforms the baseline algorithms in a variety of performance metrics.

Our main contributions are summarized as follows.

- A novel idea of community detection approach based on location interest mining.
- A spatial-temporal topic model to extract users' interests.
- Comprehensive feature representations.
- An auto encoder based location-interest-aware community detection approach.
- Experiments based on real-world dataset.

2 Related Work

In this section, we review related work of community detection models. The input of the community detection algorithm is a graph (a set of nodes representing individuals, they are connected by edges), and the output is a list of node groups representing the communities. Usually everyone belongs to a community. The community detection algorithm in social networks can be divided into graph-based methods and semantic-based methods. Graph-based community detection methods, such as Newman algorithm [11], Louvain algorithm [2] and Label Propagation Algorithm (LPA) [14], these graph-based community detection algorithms [11,12], which only use graph theory, have obvious limitations. For example, they ignore the characteristics of user interest which can lead to the lack of interest cohesiveness in the community detection results. Semantic-based community detection methods [1,14] can explore overlapping semantic communities from the perspective of characteristics analysis. Community detection can also transform users into vectors, use unsupervised clustering algorithms for community segmentation, and classify similar users into the same cluster. Clustering algorithms could be K-means [7], DBSCAN [5], spectral clustering [13], etc. The more similar or closer the users in the same cluster, the better the result is. Many community detection approaches did not take mobile users' location interest into account, that cannot support personalized location-based services.

3 Our Work

The solution framework shows in Fig. 1, including the following subsection: data cleaning, extract location interest feature by LDA, extract location distribution features, extract user relation features, user representation by auto encoder, interest community clustering and visualization.

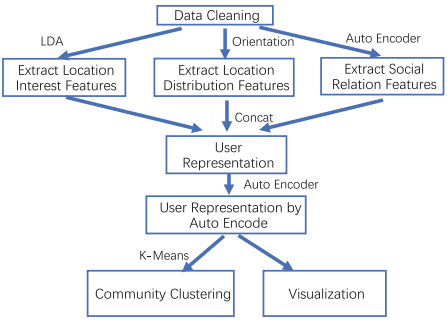

Fig. 1. The solution framework

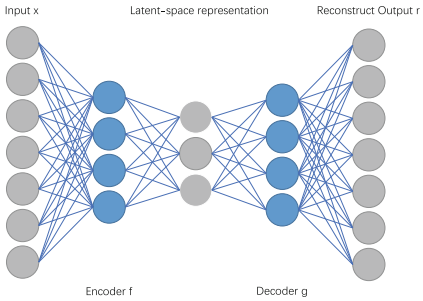

Fig. 2. Auto encoder architecture

3.1 Data Cleaning

We use the mobile social network dataset from Gowalla [3], a location-based social networking service through which users share their locations by checking-in. The Gowalla dataset contains a total of 6,442,890 check-ins of 196,591 users over the period of Feb. 2009 - Oct. 2010.

In our work, we extract users from San Francisco. Since some of the users rarely check-in, we filter out users with fewer than 10 check-in locations. Because some locations have few associated check-ins, we filter out locations with fewer than 10 check-ins. After filtering out the unqualified users and locations, the dataset has 1,995 users, 3,251 locations, and 106,098 check-ins.

3.2 Extract Location Interest Features by LDA

As we found in [10], check-ins at a location at different time represent different degrees or different types of location interest. In order to extract users' interests, we consider both time and location factors. We propose a temporal-spacial LDA (Latent Dirichlet Allocation) model to extract interests. Specifically, we treat the combination time and location of a check-in event as a word, and view the historical trajectory of a user as a document.

We apply LDA to all users' document, and convert their check-in trajectories into interest distribution represented by K topics. The user distribution on each location are shown in the user-location matrix, which is obtained by counting the trajectories. The users' interest distribution is shown in user-interest matrix,

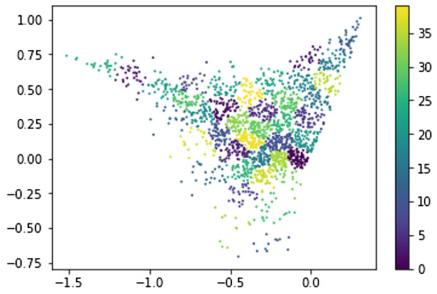

Fig. 3. Community visualization, Q = 0.379313

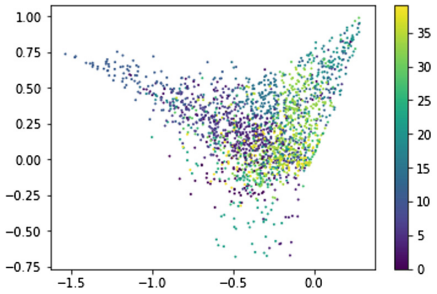

Fig. 4. Community visualization at low dimension, Q = 0.720976

where the number in the i-th row and j-th column represents the interest degree of the i-th user in the j-th topic. The interest distribution at each location is shown in the interest-location matrix. These latter two matrices come from the LDA process which use Gibbs sampling [4] as inference technique. To determine the best number of K for LDA, we use cross validation to determine the likelihood. It is observed that the maximum likelihood achieved for $K = 40$. Therefore 40 topics are selected in our experiments. In our experiment after 30,000 iterations the LDA model converges.

3.3 Extract Location Distribution Features

Here we use the way of dividing the map into $N * M$ grids, extracting the user's position distribution to form a position vector. The dimension of each user location vector is $N * M$. Here we set it to $N * M = 6 * 6$.

3.4 Extract User Relation Features

The representation of the user relationship can be represented by the adjacency matrix. In this case, if the number of users is N_u, the user relationship can be represented by an adjacency matrix of $N_u * N_u$. Each user's friend relationship vector dimension is N_u. Since the dimension of the user relationship is very high here, the dimensionality reduction is performed by auto encoder (Fig. 4).

The algorithm principle of user friend relationship vector dimension reduction is shown in Fig. 2. Auto encoder is neural network that intended to replicate the output of the inputs. It works by compressing the input into a latent-space representation, and then reconstructing the output of this representation.

We set the user's friend relationship vector by the auto encoder. We set the size of input layer to 1995, and set the size of hidden layer to 40, with the structure of $1995 - 128 - 64 - 40$. The network convert user vector from 1995 dimensions to 40 dimensions.

3.5 User Representation

The representation of a user is based on the interest vector, location distribution vector, and user relationship vector derived from last three steps.

3.6 User Representation by Auto Encoder

Here we use auto encoder to convert the user vector to low dimension for clustering. The algorithm principle of the user vector dimension reduction is shown in the Fig. 2.

3.7 Community Clustering and Visualization

After the user vector is encoded by auto encoder, we then cluster the users according to the K-Means algorithm. After clustering, We plot the user vector of 2 dimensions, and we can see the visual location community. Different colors represent different clustering groups, which intuitively visualizes the distance relationship between users, as shown in Fig. 3.

4 Experiment

4.1 Experimental Setting

Dataset. We use the mobile social network dataset from Gowalla [3]. Gowalla is a location-based social networking service where users share their locations by checking-in. The Gowalla dataset contains a total of 6,442,890 check-ins of 196,591 users over the period of Feb. 2009 - Oct. 2010.

Default Parameters. Each user and location have at least 10 check-ins. The number of LDA topics is 40. The vector dimension of the user's geographic location distribution is $6 * 6 = 36$ dimension. The user friend relationship vector is converted to 40 dimensions by auto encoder. The user vector is convert to 116-dimensional via auto encoder. The K-Means parameter K is 120.

4.2 Baseline Algorithms

We compare the proposed method with the following algorithm.

- Graph model based algorithm-FastNewman [11], LPA [14].
- Spectral clustering based on interest. The user's interest vector is extracted using the topic model algorithm LDA, and then the user interest vector is classified by spectral clustering [13].
- Geographic location based algorithm. Each user's most frequent geographic location is used as the classification basis.

4.3 Metric

We use an improved community module metric for performance evaluation. It includes original community modularity, community location similarity, and community interest similarity scores.

The original metric of community modularity [12], which measures the quality of community clustering, is defined as follows:

$$Q_q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \tag{1}$$

A is the adjacency matrix of the network, k_v is the degree of node v, and m is the number of edges in network A. c_v represents the community to which the node v belongs. If $i = j, \delta(i, j) = 1$, otherwise $\delta(i, j) = 0$. In our approach, we adopt improved community modularity performance metrics taking user's interests, locations, and social relations into account.

$$Q = Q_q + Q_\alpha + Q_\beta \tag{2}$$

$$Q_\alpha = \frac{1}{2m_e} \sum_{vw} \alpha(v, w)\delta(c_v, c_w), \tag{3}$$

$$Q_\beta = \frac{1}{2m_e} \sum_{vw} \beta(v, w)\delta(c_v, c_w) \tag{4}$$

$\alpha(v, w)$ means that if different users v, w interest are similar, then $\alpha(v, w) = 1$, otherwise $\alpha(v, w) = 0$. Since the user self comparison is not meaningful, set $\alpha(v, v) = 0$ so that when the number of communities equals the number of users, the Q_α value must be zero. $\beta(v, w)$ means that if different users v, w are similar in orientation, $\beta(v, w) = 1$, otherwise $\beta(v, w) = 0$. Similarly, $\beta(v, v) = 0$ is set so that when the number of communities is equal to the number of users, the Q_β value must be zero. Whether the different users' v, w interests are similar or not, is determined by whether the user v, w is most interested in the same topic. Whether the user's v, w orientation is similar or not, is determined by whether the user's v, w is most often in the same orientation. m_e represents the number of vw pairs that make $\delta(c_v, c_w)$ not equal zero. This makes the value of $2m_e$ increase dramatically when the number of clustering is too small, which can penalize the modularity. Intuitively, two users with similar interests and similar locations will lead to higher probability of being assigned to same community. If two users are assigned to same community, the more similar they are, the higher the module metric. Therefore, considering the social similarity degree of the user's location, the higher modularity indicates that the community clustering performance is better. Q_q value range $[-1/2, 1]$, Q_α value range $[0, 1/2)$, Q_β value range $[0, 1/2)$, so Q generally takes the range of $[-1/2, 2)$.

4.4 Experimental Result

We compare the performance of our algorithm with the benchmark algorithms. According to Table 1, the performance of the proposed algorithm is better than

Table 1. Algorithm performance comparison

Algorithms	Q	Q_q	Q_α	Q_β
FastNewman	0.662	**0.466**	0.018	0.176
LPA	0.619	0.433	0.018	0.168
Interest spectral clustering	0.659	0.094	**0.330**	0.234
Geographic location	0.575	0.052	0.023	**0.499**
Our algorithm	**0.720**	0.058	0.289	0.373

the benchmark algorithms. The Fast Newman algorithm achieved the highest Q_q value of 0.466334, but it is far behind our proposed algorithm in terms of community interest score and community geographic location score. The LPA graph model algorithm also achieved high Q_q values, but it is far behind our proposed algorithm in terms of Q_α and Q_β. Based on the geographic location algorithm, each user's most frequently visited geographic location is used as a classification basis, it get the highest Q_β score of 0.499253. However, the algorithm does not take into account the network modularity and interest similarity, so it does not perform well on other scores. The interest spectral clustering method uses the topic model algorithm LDA to extract the user's interest vector, and then clusters the users with spectral clustering. So get the highest Q_α score of 0.330194. However, the algorithm does not take into account the location module, so it does not perform well on other scores. The algorithm we proposed takes into account the interest, location, and friend relationship. Due to dimension reduction to reduce complexity, not only the computational efficiency is high, but also the overall performance is best in terms of Q. Overall, our approach is better than traditional benchmark algorithms by combining location interests, location, and friendships.

4.5 Parameter Sensitivity Analysis

Next, we perform parameter sensitivity analysis by comparing system performance under different parameters, including K-means clustering parameter K, fused data type.

K-means Clustering Parameter K. K-means clustering parameter K is an important system parameter of the model. The K-means clustering algorithm depends on the clustering group number K. In the experiment, we increased K from 60 to 480. The result is shown in Table 2. The K-means parameter K has an impact on performance. The performance is best when the K-means parameter $K = 120$. This means that bigger or smaller community number cannot improve clustering performance. Small community number means that users with different interests are assigned to the same community, making the interests of the communities different. Big community number shows that the network module degree Q_q value not perform well.

Table 2. The influence of K

K	Q	Q_q	Q_α	Q_β
60	0.613	0.067	0.196	0.349
120	**0.720**	0.058	0.289	0.373
240	0.716	0.041	0.308	0.367
480	0.684	0.028	0.302	0.353

Table 3. Fused types of data

Data	Q	Q_q	Q_α	Q_β
Loc.	0.500	0.012	0.041	**0.446**
Int.	0.703	0.050	**0.422**	0.229
Int. & Loc.	0.704	0.058	0.275	0.370
ALL	**0.720**	**0.058**	0.289	0.373

Fused Types of Data. Our model incorporates user interests, location, and social information. The types of data fused in the model is an important system parameter of the model. In the experiment, we increased this parameter from only the location information to all kinds of data participated in the operation. The result is shown in Table 3. The more types of data is fused, the better our approach performs: based on the interest and location, the method is slightly better than only the interest-based method. After adding the social vector, the overall Q performance is the best. Overall, our approach is better than traditional benchmark algorithms by combining location interests, location, and friendships. The types of data has an impact on performance. When the number of data types is too small, it means that the amount of information considered in clustering is few, so that the clustering performance is not high under comprehensive consideration. The model performs best when all types of data is included. This means more data can bring more information and improve performance.

5 Conclusion

In this paper, we propose a location-interest-aware community detection approach for mobile social networks. We develop a spatial-temporal topic model to describe users' location interest, and introduce an auto encoder mechanism to represent users' location features and social network features as low-dimensional vectors, based on which a community detection algorithm is applied to divide users into sub-graphs. We conduct extensive experiments based on a real-world mobile social network dataset, which demonstrate that the proposed community detection approach outperforms the baseline algorithms in a variety of performance metrics.

Acknowledgment. This work was partially supported by the National Key R&D Program of China (Grant No. 2018YFB1004704), the National Natural Science Foundation of China (Grant Nos. 61672278, 61832008, 61832005), the Key R&D Program of Jiangsu Province, China (Grant No. BE2018116), the science and technology project from State Grid Corporation of China (Contract No. SGSNXT00YJJS1800031), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Sino-German Institutes of Social Computing.

References

1. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761 (2010)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**(10), P10008 (2008)
3. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1082–1090. ACM (2011)
4. Darling, W.M.: A theoretical and practical implementation tutorial on topic modeling and GIBBs sampling. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 642–647 (2011)
5. Duan, L., Xu, L., Guo, F., Lee, J., Yan, B.: A local-density based spatial clustering algorithm with noise. Inf. Syst. **32**(7), 978–986 (2007)
6. He, D., Yang, X., Feng, Z., Chen, S., Fogelman-Soulié, F.: A network embedding-enhanced approach for generalized community detection. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) KSEM 2018. LNCS (LNAI), vol. 11062, pp. 383–395. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99247-1_34
7. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
8. Jin, D., Liu, Z., He, D., Gabrys, B., Musial, K.: Robust detection of communities with multi-semantics in large attributed networks. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) KSEM 2018. LNCS (LNAI), vol. 11061, pp. 362–376. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99365-2_32
9. Ma, H., Xie, M., Wei, J., He, T.: An overlapping microblog community detection method using new partition criterion. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) KSEM 2018. LNCS (LNAI), vol. 11062, pp. 313–323. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99247-1_28
10. Chen, M., Li, W., Qian, L., Lu, S., Chen, D.: Interest-aware next POI recommendation for mobile social networks. In: Chellappan, S., Cheng, W., Li, W. (eds.) WASA 2018. LNCS, vol. 10874, pp. 27–39. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94268-1_3
11. Newman, M.E.: Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**(6), 066133 (2004)
12. Newman, M.E.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
13. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856 (2002)
14. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**(3), 036106 (2007)