Implicit Community Discovery Based on Microblog Theme Homogeneity

Zefen Sun, Yajun Du*, Xiaoliang Chen and Chunjia Wang

School of Computer and Software Engineering, Xihua University, Chengdu, Sichuan, China

*Corresponding author

Abstract. With the rapid development of various social platforms, social networks have accumulated more and more data. These huge and complex data contain valuable information in many fields. It is of far-reaching significance to conduct statistical analysis of these data and dig out valuable information and apply it in real life. Based on the homogeneity of social networks, our paper proposes a framework of discovery algorithms based on the homogeneity of microblogs to realize the theme community discovery of microblog social networks. Firstly, the data set is preprocessed, the invalid data is removed and classified by microblog users. Secondly, the LDA model is used to extract the topic of microblog user blogs to describe the characteristics of microblog users. Thirdly, two users are calculated based on the theme interest. The homogeneity measure between them is used to represent the relationship of social networks. Finally, the community of the implicit social network is realized through unsupervised algorithms to build communities with the same interest as the theme. In order to verify the validity of our proposed algorithm framework, this paper uses data sets from microblog network to conduct experiments. The experimental results show that the algorithm effectively combines the homogeneity measure of the network and has achieved very good results in the implicit community discovery.

Keywords: Homogeneity measure; Microblog theme homogeneity; LDA model; Community discovery; Unsupervised clustering.

1. Introduction

With the rapid development of the Internet, various social platforms have emerged. In particular, microblog has become the most widely used social platform because of its full-featured, simple and convenient conditions. Users can post micro-posts on microblog platform, forward comments and search real-time users. The extensive use of microblog has led to the rapid growth of microblog data, resulting in a huge amount of data, but these data is also a large amount of valuable information. The classification or induction of these data can find out the regularity and characteristics of the data. By mastering the rules and characteristics, the complex network can often be simplified to solve the problem. Data is mined by analyzing the common characteristics of similar data and finding homogeneity data representations. Different users in microblog focus on the same theme to form a community. Discovering and analyzing the relationship between the community and the community and the community is conducive to the study of massive data information, and also provides important help for the study of social network structure[4].

Community discovery algorithms have evolved in many versions, the Kernighan-Lin (KL) algorithm is a greedy algorithm based on graph partitioning, which divides the known network into two communities of different sizes. The Girvan and Newman (GN) algorithms [1] use the partitioning ideas in cluster
analysis to filter edges in the network and remove the edges to achieve community partitioning. Q.G et al. [6] proposed a label boundary node algorithm (LBN algorithm) based on boundary nodes and label propagation, but the community network lacks stability. J.C et al. [7] proposed a method for combining links and node content for community discovery using the graph regularization method (CLNCCD), but without considering the heterogeneous distribution of node degrees, the framework still has shortcomings. Xie Y et al. [2] proposed a community discovery method based on deep sparse filtering network, but this method does not study the dynamic evolution mechanism of the community, and there are also defects in the network representation. Throughout the community discovery methods in recent years, the details include not only traditional methods such as graph segmentation method, hierarchical clustering method, other clustering algorithms and heuristic algorithms, but also extended probability graph model methods and non-negative matrix factorization methods, dynamic network discovery model [5] and so on. In the past, most community discovery methods only use the social structure or weight relationship between nodes to discover the community, and ignore the important role of homogeneity in community division.

Homogeneity is one of the important factors in the formation of social network structure, that is, members of the same community tend to have the same characteristics [3]. Homogeneity [8] refers to the principle that content users interact and share content more frequently than other users. In fact, users are more inclined to establish contact with users with similar interests. Homomorphism can well express the similarity between two nodes, and the close association between the homogenous nodes can lay a foundation for the discovery of social network communities composed of similar nodes. However, in the current community discovery research, the important role of homogeneity has not received sufficient attention.

Inspired by this, we analyze the homogeneity of users in the network, and innovatively propose a method based on the homology measure of the user’s theme interest and find the community by matching its theme interest homogeneity. Based on the advantages of previous researchers, we clustered based on the homogeneity measure of nodes to realize the division of implicit communities and discover communities.

2. Method

In this section, we will detail the proposed implicit community discovery algorithm based on the homogeneity of microblog themes. Firstly, we present the proposed model framework, and then describe the proposed homogeneity-based community discovery algorithm model framework, and explain the method formulas that need to be used in detail.

2.1. Model Framework Representation

The overall framework of Implicit community discovery based on microblogs theme homogeneity is shown in figure 1.

2.2. Theme Interest Extraction and Homogeneity Measure Calculation of Homogenous Implicit Community

In order to get the user theme interest for community discovery, The EM algorithm with good convergence speed and accuracy is used to solve the LDA model. The specific topic interest vector extraction algorithm is shown in Algorithm 1, as shown in table 1.

Then, each user can be represented by the vector $\mathbf{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik})$, $\mathbf{\theta}_i$ is a vector of length $k$, $\theta_{ik}$ represents the probability of the user $i$’s blog post theme $T_k$. Let $S(i/j)$ represents the homogeneity measure between the user $i$ and the user $j$, which can be calculated by the following formula (1):
Data Collection  
Data Cleaning  
Data Preparation  
Construction Word Vector  
Theme Extraction  
Computational Homogeneity Measure  
Implicit Network Representation  
Implicit Community Discovery  

Figure 1. Model framework for homogenous-based implicit community discovery.

Table 1. Algorithm 1 (Theme Interest Vector Extraction).

| Input: Users’ microblogs, the number of theme k, and stop words dictionary |
| Output: The users’ theme matric TM |
| 1) Segmentation of Chinese microblog; |
| 2) Remove the stop words; |
| 3) Construct a word vector Word-List; |
| 4) The document is converted into a numerical vector; |
| 5) Calculate the TF-IDF of the document; |
| 6) Using PCA to reduce the dimension of TF-IDF; |
| 7) Initialize the LDA model; |
| 8) Train the LDA model with TF-IDF and get Theme Matric; |

\[
S(i,j) = \frac{\|\theta_i \cdot \theta_j\|_2}{\|\theta_i\|_2 \cdot \|\theta_j\|_2}. \quad (1)
\]

Where \(\theta_i\) represents the theme distribution of the blog set sent by the \(i\)th user, and \(\theta_k\) represents the theme distribution of the blog set sent by the \(k\)th user. When \(S_{ik} = 1\), indicates that the user \(i\) and user \(k\) have the same theme distribution; the closer to 1, the more similar the two user themes are, when \(S_{ik} = 0\), Indicates that the theme distributions of user \(i\) and user \(k\) are completely different.

2.3. Implicit Network and Implicit Community Representation

In social networks, the network is usually represented as a graph containing vertices and edges. Vertices represent entities in the network, and edges represent entity relationships in the network. Similarly, in an implicit network, we also use each vertex to represent a microblog user, and the edge represents the relationship between the two vertices. The difference is that in an implicit network, edges are no longer specific vertex connections, but rather homogeneity measures between users.

Here, we use the microblog blog post theme to characterize each user. Specifically, each vertex can be represented by a vector \(\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik})\). Which \(\theta_i\) represents the probability distribution of the theme in the blog post sent by the user node. Since each edge represents the homogeneity measure between two microblog users, which can be calculated by the formula (1), each edge can be composed of two vertices to represent. In order to better use the clustering algorithm [10] to partition the implicit
network, we choose to represent the implicit network based on the blog post by the vertices representing the user topic interest distribution vector. In theory, there is a homogeneity measure between any two users, then the implicit network of $N$ microblog users will be a full connected graph containing $N$ vertices, totaling $N(N-1)/2$ edges. In order to reduce the connection calculation of the graph and simplifying the calculation of the implicit network, we assume that when $S(\theta|\phi) > \lambda, \lambda \in (0,1)$, there exist an edge between the two vertices (representing a relationship), otherwise it does not exist. In a community, the more accurate the user's theme is found in the community, the better the community is divided.

2.4. Implicit Community Discovery Based on Microblog Homogeneity

An implicit network represented by a homogenous measure constructed by theme interests, users with similar theme interests should be divided into the same theme community. In this paper, the unsupervised algorithm is used for clustering, so that users with homogeneity are divided into a community, a network composed of $N$ communities is formed, and the contour coefficients of the clustering results are calculated to evaluate the clustering results. The implicit community discovery algorithm based on microblog homogeneity is shown as Algorithm 2, as shown in table 2.

### Table 2. Algorithm 2 (Implicit Community Discovery based on Microblog Homogeneity).

1) Data acquisition: obtaining a certain number of microblog blog data sets through web crawlers or other means;
2) Data extraction: extract user ID and blog field from the data set;
3) Data cleaning: sorting the obtained microblog data, removing invalid data, and classifying by user to obtain a blog set of each user;
4) Data loading: splicing each user's blog into a blog file, and all users' documents form an N*1 matrix as the next input;
5) Word2Vector: Classify all blog posts, remove stop words, construct word bag vectors, and convert documents into vector representations;
6) Subject extraction: Based on the variational inference EM algorithm to solve the LDA model, the theme distribution vector and word distribution vector of the user microblog set are obtained;
7) Theme dimension reduction: when the theme distribution is sparse, the interest of the user theme is reduced;
8) Homogeneity Measure: Calculate the homogeneity measure between users based on the theme distribution obtained in the previous step;
9) Network representation: The initial network is constructed from the homogeneity measure obtained in the previous step, and the presence of edges between users with greater homogeneity measures forms an initial network;
10) Community discovery: Based on homogeneity measures to cluster users with the same theme, form $M$ implicit communities.

3. Experiment

In this section, we will use the real data of the microblog social network to experiment and validate the effectiveness of our proposed algorithm for implicit community discovery based on the homology of microblog themes. In order to verify the validity of the algorithm, this article uses the microblog dataset from the user portrait contest in the 5th National Social Media Processing Conference (SMP2016) in 2016.

3.1. Results and Discussion

Silhouette Coefficient is an evaluation method which combines the two factors of cohesion and resolution on the basis of the same original data to evaluate the impact of different algorithms or
different operation modes of the algorithm on the clustering results[9,10]. As can be seen from the figure 2, the cohesion of each theme community is very good, which indicates that users in the community have a consistent theme interest, and there are no samples with contour coefficients around 0 in the figure2, that is, no samples are between the two clusters. This means that each user can be divided into the most similar themes of similar communities. Figure 3 is a visualization of the results of community segmentation using t-SNE. The cohesiveness of each cluster and the degree of separation between clusters can be more clearly reflected from the figure 3. The experimental results fully demonstrate the implicit community discovery based on the homogeneity of microblog and the effectiveness of the algorithm framework.

![Figure 2. Clustering Result Contour Coefficients.](image1)

![Figure 3. Clustering Result Visualization.](image2)

4. Conclusion
In this paper, we fully consider and utilize the role of homogeneity in community discovery. Based on the advantages of previous researchers, this paper uses the homogeneity of microblog users and the homogeneity measure of microblog users' theme interest. This paper proposes an implicit community discovery method based on the homology measure of interest of users on microblog users, bringing together users with similar themes to discover implicit communities with the same topic homogeneity. Experiments show that the proposed method can find a user community with common theme interest, and the homogeneity of users in the community is relatively strong.

In the future, our research work will focus on the evolution mechanism of the research community, and analyze the changes of the network community based on homogeneity, that is, the network evolution process of the community from $T$ to $T + 1$.

Acknowledgement
This research was funded by the National Natural Science Foundation of China (Grant Nos. 61872298, 61532009, 61472329, 61802316, 61602389 and 61902324).

References
[1] Easley, D., & Kleinberg, J. (2010). Networks, crowds, and markets (Vol. 8). Cambridge: Cambridge university press.
[2] Xie Y, Gong M, Wang S, et al. Community discovery in networks with deep sparse filtering[J]. Pattern Recognition, 2018, 81:50-59.
[3] Andrea D S, Barbara G, Laura R, et al. Discovering Homophily in Online Social Networks[J]. Mobile Networks and Applications, 2018.
[4] Mohammadi M, Moradi P, Jalili M. SCE: Subspace-based core expansion method for community detection in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2019.
[5] Newman M E, Girvan M. Finding and evaluating community structure in networks.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 69(2):026113.
[6] Gui Q, Deng R, Xue P, et al. A community discovery algorithm based on boundary nodes and label propagation[J]. Pattern Recognition Letters, 2017:S0167865517304646.
[7] Jinxin C, Jianwu D, Di J, et al. Combination of links and node contents for community discovery using a graph regularization approach[J]. Future Generation Computer Systems, 2018:S0167739X17327711.

[8] Andrea D S, Barbara G, Laura R, et al. Discovering Homophily in Online Social Networks[J]. Mobile Networks and Applications, 2018.

[9] Zhihua Cai, Dianhong Wang, Liangxiao Jiang. K-Distributions: A New Algorithm for Clustering Categorical Data[C]// Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, Third International Conference on Intelligent Computing, ICIC 2007, Qingdao, China, August 21-24, 2007, Proceedings. 2019.

[10] Visser J, Corbetta A, Menkovski V, et al. Stampnet: Unsupervised Multi-Class Object Discovery[C]// 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019.