Community Detection on Social Network – A Survey

Greeshma T S, Subu Surendren

Abstract: ***Social network is an important application in the internet which represent the geographically dispersed users. Social network provides a variety of methods for explaining patterns and entities. Social networks are mostly represented as graphs, which contain nodes and edges. Nodes are used to represent actors such as people and organizations whereas edges show the relationship between these nodes. Several data sources involved in the social network forms communities which work in self-descriptive manner. A collection of nodes which are connected by edges with high similarity is called a community. The community detection in social network, intend to partition the the graph with dense region which correspond to closely related entities. The selection of data sources and determination of community detection approaches can enhance the accuracy, efficiency and scalability of community. In this survey, different community detection approaches are discussed.***

Keywords: social network, community detection, community structure

I. INTRODUCTION

Social networks portray interactions among the interconnected nodes represented as graph [1]. The community detection process (CDP), provide common relations between users and analyze each related part of a network. Communities in social networks can be performed by different methodologies which have high importance for understanding the types, detecting and analyzing useful and hidden patterns in aforesaid network [2]. Community detection can be performed in a series of process such as:

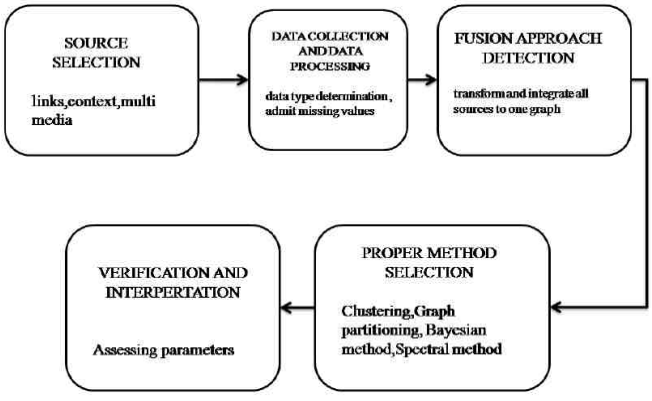

Figure: Community Detection Process

Revised Version Manuscript Received on July 02, 2016.

Greeshma T S, Department of Computer Science, Sree Chitra Thirunal College of Engineering Trivandrum (Kerala). India.

Subu Surendren, Department of Computer Science, Sree Chitra Thirunal College of Engineering Trivandrum (Kerala). India.

A. Source Selection

Data source selection where the sources are contents (opinions, comments and weblogs), links (co-author, friend list or reference to other web pages, links are user's explicit connections), attributes and features (profile information) and other media sources such as tagged photos or shared videos. The accuracy and reality level of the social network depends on inclusion of data source.

B. Data Collection and Data Processing

Raw data which are collected will decrease the result quality, because there are incomplete, inconsistent and noisy data in the dataset. So pre-processing is to done for clean and consistent data, which is on the basis of the accepted quality and the availability of background knowledge. The important techniques are:

Fill with null values: Missing relations can be detected using link prediction [6]. If nodes have properties and there are some null fields, filling them manually by background knowledge or by a default value based on their properties or replace nulls by the most probable value [7].

Select distinctive features. Feature selection is to reduce the quality of data and data dimension since the data source is large. Feature selection is applicable on the properties of users, if no new information is there to add features then it can be eliminated. Feature selection will decrease the runtime of the algorithm and facilitates results interpretation [8].

C. Fusion Approach Detection

There are different methods to use information of multiple sources which provide scalability and interpretability. By using the multiple sources will produce better results in quality and inclusiveness. If the well-known link and content methods are reviewed and the effect of each source is analyzed then finally, a mixture model can be presented. After declaring what is significant to the aim of CDP, it is important to determine the right multi-source fusion approach, in order to facilitate the selection of the right algorithm [9, 10].

D. Proper Method Selection

Based on the fusion approach and data types of the network the best algorithm or a framework has to be selected. For different kinds of graphs such as (un)weighted graphs or (un)directed graphs have useful and effective methods. Existence of membership determination or hierarchical relations which are hidden in communities are important [11, 12].

E. Verification and Interpretation of Detected Community

The accuracy and reality of results (communities) need to be evaluated. Using multisource fusion approach, traditional parameters produce low value which does not mean the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

low accuracy of the results, because traditional parameters are based on one single source and one single concept. But now the informative network is divided into communities based on multiple sources which have their own effects on CDP. Cliques, k-cliques, k-clubs, quasi-cliques verify whether extracted communities satisfy the definition. Modularity is a parameter that most widely used and effective one to get optimized communities. The good partitioned communities have high value of modularity.

II. DETECTION OF COMMUNITIES

Social networks contain communities based on common interests, location, occupation, etc. The social networks properties can be characterized as static properties and dynamic properties. The static properties are referred by the structure of sub-graphs; while dynamic properties are described by the network structure itself over time [11]. The major advantage of community detection is the gathering of information from multiple clusters and sources. Community detection provides exchanging and offering of information due to the similarities in the members within one group and also provide the structure of network [12].

A. Approaches of Community Detection

There are different approaches for community detection such as:

1) Integrated Internet of Things and Social Network Architecture

The people linked to internet can be viewed as number of things. This method considers only the function of community which has occurrence of two nodes of atmost one hop and at least two mutual friends. The friend suggestion can be made by mutual friends. The result of these approach in an integrated environment are more significant on intra community methods than inter-community methods [13].

2) Community and Sub-Community Detection methods in Social Networks

Newman-Girvan algorithm [14][20] helps to detect community and sub community within a social network. In this method the edges are removed involving an iterative method using split of communities that is *'betweenness'* measure is calculated during each edge removal. This algorithm provides strength of each common structure and can be represented using objective metric.

3) Weighted network

The communities in a social network can be identified through clustering and its main objective is to maximize total weight of all selected clusters that minimize the similarity between the selected clusters. In this approach every data object is assigned to exactly one cluster that is it assures that every cluster has atleast one object and also ensures that only certain numbers of clusters are selected[15].

4) Distributed environment, in Web-Scale Networks

This method helps to find partition in community from billions of edges using learning schema in community detection which provides an high quality partition. In this approach data are stored in main memory so execution require less time and easily scalable. The network with large number of edges can be scaled by distributed core detection algorithm and vertices form small groups belongs to similar community. The pre-processing method of community detection are done through core group detection method [16].

5) Edge Content in Social Media Networks

In this method each edge provides richer characteristics of each community in social network and this structure is called linkage structure. In this structure a pairwise interaction determines specifies information denoting nature and relationship among individuals. The effectiveness in CDP can be leveraged using edge constant [17].

6) Influence Ranking in Social Networks

This method is based on connectivity and determined proximity with in topology of network. An influence vector generated by every node by using influence cascade model that capture node information throughout the network. The closeness of each pair is based on similarity measure that defines a meaningful and measure of connectivity. A latest influence diffusion model which is embedded to a node is parsed around within the network [16].

7) DBSCAN Algorithm

Community Detection consist of three members namely core, border and outlier based on influence. The detection of outlier are done by changing the radius of each cluster and thus they are eliminated. The method points to core having high influence among the node and clustering accuracy is achieved by eliminating outliers [18].

8) Bayesian network and Expectation Maximization technique

In this method communication between nodes are determined and social network are familiarized using statistical model, so Basiyan network helps to show relation among the variables. The approach is based on expectation and maximization which helps to obtain the estimation for model parameter [19].

9) Graph mining technique

Here the subgraph is generated based on the adjacency among the nodes. The reachability between nodes are determined if there is path between them. The nodal degree is computed based on actors adjacency within the group members or not[12].

10) Spectral clustering

The effectiveness of spectral clustering is based on the feature of detection on the basis of their role and correctness. The number of communities are determined based on eigen value distribution of Laplacian matrix and clustering is done using k-means. This approach divides network accurately into two halves and common communities are not detected [13,22].

11) Overlapping communities

In this method overlapping communities which is an attribute on community detection are detected based on when a person belongs to more than one social graph. The overlapping communities are divided into two such as node based overlapping and link based overlapping. Node based overlapping categories nodes in network. Link based overlapping categories the edge within the network [11].

Table 1 shows the advantages and disadvantages of different community detection approaches. Detection using graph mining and detection method for distributed environment in web scale network can be implemented in

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

large networks. All the detection methods ignore outliers for improving accuracy.

Table 1: Advantages and disadvantages of community detection method

Algorithms	Advantages	Disadvantages
DBSCAN algorithm	Notion of noise Robust to outlier	Not deterministic
Integrated Internet of Things and Social Network Architecture	Similar to real life situation Used to suggest friends	Not generalized Fails in directed graph
Edge Content in Social Media Networks	Better Supervision	Applicable on email and flicker images based on clustering only
Weighted network	Total weight of cluster is calculated Similarity in between the cluster	Not clear for type of network
Newmann Girvan Algorithm	Can detect even sub communities	Computation cost is high
Distributed environment in web scale network	Applicable to large graph	
Bayesian network and Expected Maximization techniques	Work both in (un)directed and (un) weighted network	Have to specify number of communities
Graph mining techniques	Useful for large number of nodes	
Spectral clustering	Find solution with underlying hierarchical structures and fuzzy nodes	
Overlapping communities	Detect and analyse overlapping communities	Only modularity method has been used as a fitness function

III. CONCLUSION

The different community detection processes and methods have been reviewed. Several methods have been developed that are flexible enough to apply quite general network structures. These methods will improve the speed and sensitivity of community structure algorithms. Many interesting networked systems awaiting analysis using these methods.

REFERENCES

1. Michel Plantic, Michel Crampes ,"Survey on Social Community Detection", Springer Publishers, 25 March 2013.
2. James P Bagro, "Evaluating Local Community Method in Network ", Journals of Statistical Mechanism Theory and Experience, 2008.
3. C.C Aggarwal, H.Wang,"Survey of clustering algorithm for graph data managing and Mining Graph Data ",Springer, 2010.
4. A.Pothen,"Graph Partitioning Algorithm with Application to Scientific Computing ", Springer, 1997.
5. M.Girvan and M.E.Newman ,"Community Structure in Social and Biological Networks", Proceeding of National Academy of Science, June 11,2002.
6. A.Hasian and M.J Zahi,"A Survey of Link Prediction in Social Network", Springer , March 11,2011.
7. J.Han,M.Konnber and J.Pei,"Datamining Concepts and Techniques", Morgen Kaufmann,2006.
8. R.Xu and D.Wunsch,"Survey of Clustering Algorithm ",Neural Network ,IEEE Transaction, May 2005.
9. N.F.Chikki,B.Rothenburger and N.Aussenac Gilles," Combining Link and Content for Community detection : a discriminative approach", Proceedings of 15th ACM SIGMM workshop on Social Media, June 28,2009.
10. F.Moser, R.Ge and M. Ester,"Joint Cluster Analysis of Attribute and Relationship Data without a -prior-specification of number of clusters" Proceedings of 13th ACM SIGDD International Conference on Knowledge Discovery andDatamining, August 07, 2007.
11. S.Fortunate,"Community Detection in Graph", Physics report ,2009.
12. S.Papadopacelos, Y.Kompatsiaris,A.Vakali, and P.Spyridones, "Community Detection in Social Media ", Datamining and Knowledge Discovery, June 14,2011.
13. Yangyang Li,Ruachen Liu and Jiamhe Wu, "A Spectral Clustering Based Adapive Hybrid Multionjective Harmony Search Algorithm for Community Detection ", WCC12012 IEEE World Congress on Computational Intelligance, June 15,2012.
14. Deepjyoti Chaudhery, Saprativa Bhattachayie , Anirban Das,"An Empirical Study of Community and Sub community Detection in Social Network Applying Newmann Girvan Algorithm," Emerging Trends and Application in Computer Science ,Sep 14,2013.
15. Ganjaliyev.F," New Method for Community Detection in Social Network Extracted from the Web" , Problems of Cyberneties and Information ,Sep 14, 2012.
16. Michael Ovelganne ,"Distributed Community Detection in Website Network",Advance in Social Network Analysis and Mining ,IEEE, Aug 28, 2013.
17. Guo-Jun Qil,Charu C,Aggarwal and Thomas Huangl ," Community Detection with Edge Content in Social Media Network ", Data Engineering ,IEEE , April 1,2012.
18. Yomna M.ElBarawy, Ramedan F Mohammad and Naveen I Ghali," Improving Social Network Community Detection Using DBSCAN Algorithm" , Computing Application and Research ,Jan 20,2014.
19. Ahmed Ibrahem Hafez, Abaul Ella Hassanien , Aly A. Fahm and M.F. Talba, "Community Detection in Social Network by Using Bayesian Network and Expectation Maximization Technique", IEEE , Dec 16,2013.
20. M .E.J .Newmann ,"Community Structure in Social and Biological Network " IEEE, April 6, 2002.
21. Hastic , T.R. Tibshirani and J.H Friedmann,"The elements of Statistical Learning," IEEE ,Auguest 2008
22. A.Y.Ng, M.I.Jordan, Weiss,"On Spectral Clustering Analysis and Algorithm ", Stanford Alhab, 2001.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.