ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

A SURVEY ON CHALLENGES IN GRAPH CLUSTERING FOR INFORMATION AND SOCIAL NETWORKS

Parimala M [1]
1 School of Information Technology & Engineering, VIT University, Vellore, Tamil Nadu 632014, India.
Email: parimala.m@vit.ac.in

Received on 25-10-2016 *Accepted on 02-11-2016*

Abstract

Recent developments in information network and rapid growth of technologies support the storage and managing huge amount of data by various organizations. The analysis of these data is becoming more and more challenging task as they are expanding quickly with the rapid generation of data from various information networks such as social network, marketing network, sensor and Telecommunication network. Graph Clustering is a challenging research problem that has gained its importance in the field of clustering due to its significant use in various and wide range of domains. The survey gives a clear and concise picture various challenges in developing a graph clustering algorithm and provides an insight for future research directions.

Key words: Clustering, Graph Mining, Challenges, Social Network.

1. Introduction

Data Mining is a powerful tool that can find the hidden patterns and relationships within the data. The fundamental law of geography [1] states that everything is related to everything else but nearby things are more related [2] than distant objects. Data mining or knowledge discovery in database (KDD) refers to the non-trivial process of discovering, interesting, implicit and previously unknown knowledge from large databases [3]. Clustering is one of the fundamental and unsupervised learning technique [4] in the field of data mining. The process of cluster analysis is defined as grouping the similar objects and dissimilar objects in different group. It plays a vital role in almost all the thrust areas in the field of bioinformatics [5],image analysis, market research [6],web search and so on. Based on the mode of the application and information they can be broadly classified as partitioning, density-based, fuzzy and graph clustering. However, the common job among all the algorithms is to group the similar objects using any measure such as Euclidean distance [7], density function [8]. The goal of this survey paper is to study the various challenges associated with clustering in graphs and the taxonomy of graph clustering and it is concluded by summarizing the concepts discussed in the paper.

2. Challenges in Graph Clustering

The clustering process on the graph data is considered to be a more challenging task among the current researchers. The taxonomy of graph clustering algorithm is given in Figure 1. Some of the problems are highlighted in the current section which summarizes the basic challenges in grouping the graph data. In the case of networks, the problem of cluster or community refers to grouping of nodes that are highly similar without any prior knowledge while on the contrary nodes across communities present low similarity. The closeness of the object is based on the similarity measure which is defined based on the datasets used for the community detection. The similarity can consider either topological features or features related to the nodes and additional information can be associated with the edge of the graph. Although there are several definitions for graph clustering problem, the main objective is to group the objects which are denser into a cluster and rest in other cluster.

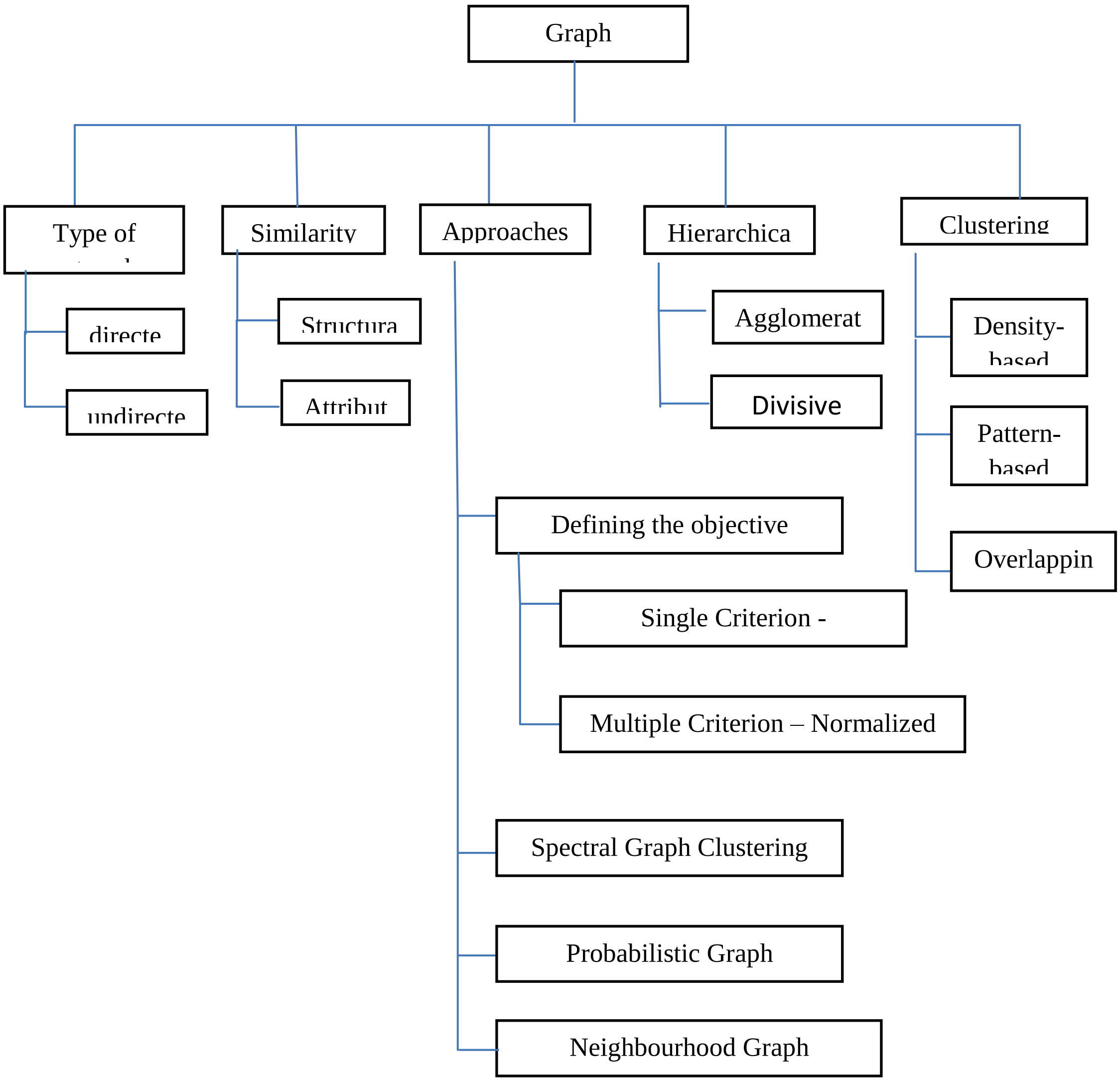

Figure 1. Taxonomy of Clustering in Graph.

It is clearly evident that the application of graph clustering is necessary in various domain such as grouping the authors who have co-authored in citation network, grouping the people with common interest or friendship relationship in social network dataset, donar-recipient relations among genomes in biological data and so on. The major difference between the traditional and graph clustering is that traditional clustering groups the objects based on distance between the objects using any distance measure (Euclidean distance) whereas graph clustering involves in grouping the nodes based on their connectivity(number of possible edge between the nodes) and common neighbours.

2.1 Handling the large dataset

It is always a tedious task to handle the dataset with huge dimension. In that case mapping all these information structures into a graph structure consumes more time [9]. The reason behind the high complexity is due to the time taken to construct of graph network and time taken to group the objects. As stated by Cheng(2011) [11] graph structures is the best way to represent and express the relationship among the objects, we ignore the complexity of graph clustering algorithm. The huge dimensional dataset can be handled by reducing the dimension based on any dimensionality reduction technique or selecting the more dependent attributes by using any statistical methods. This solution can considerably reduce the time complexity. On the other hand the size of the graph can be reduced by partitioning the whole graph into subgraphs which would lead to the modular approach.

2.2 Number of clusters (k)

In general, the clustering algorithms are defined with more number of input parameters. Actually this gives a negative impact on the resultant clusters. When the algorithm varies with respect to the type of information and domain, then the values for the input parameters must also be dynamic and not static. So, the input values provided to the algorithm can be trained from the dataset itself and can be used by the clustering algorithm.In most of the previous related work, the number of clusters that should be formed is given by the user. When the k value given is small, the density of the clusters is high as more number of nodes is grouped in each cluster. Whereas when the k value is high, the density of cluster is less as there will be fewer nodes in each cluster. The k value must be automatically detected based on the distribution of datasets. The idea behind this automatic generation is to find the centroids in the dataset. In graph clustering, the centroids can be detected based on these definitions,

(i) The centroid node should have more neighbours

(ii) The peak node should have high similarity in terms of properties with more number of nodes.

(iii) The capacity or the sum of edge weight should be higher than the other nodes.

(iv) The peak node should have high influence on the other nodes.

Based on the above definition the number of centroids are generated which would indirectly determine the value for number of clusters.

2.3 Overlapping Communities

Mutual clusters are formed if an object can belong to only one clusters whereas if an object can belong to more than one clusters[12][13] then it is known as overlapping clusters. The type of cluster (mutual or overlapping) depends upon the application and the analysis done on the dataset [14]. For example, in citation network the author would have co-authored with authors present with more than one cluster. In social network, the clusters are formed based on the friend relationship. Consider a network with two clusters with A & C as centroids. As the individual B is a friend of both A & C it should be present in both the clusters. For Healthcare dataset the locations are grouped based on their frequent mobility patterns [15] of the infected individuals. The locations can be overlapped in different mobility patterns. Content sharing network such as in Twitter dataset, the messages shared are grouped. We observe that, practically the nature of objects exhibit the overlapping property among them.

2.4 Type of graph network

The edge directionality plays a vital role in detecting the communities or group. The literature survey has defined so many methods for community detection in undirected network. It is difficult and a challenging task to propose a new algorithm or extend the undirected algorithm to the directed network. For the application that exists a symmetric relationship ie., when A is a friend B and vice versa in social network we can use undirected network. Incase of mobility pattern analysis, the number of infected host moving from A to B is not same for B to A. The scenario exhibits a asymmetric relationship for which directed graph can be used.

3. Conclusion: Graph clustering organize, analyse and provides a better understanding of complex structure using graph structure. The survey on this graph clustering proves the importance towards the interesting topics for future research. It clearly visualizes the structure of complex systems and relationship held between them. The various challenges in developing graph clustering algorithm are reviewed and provide an insight for researchers to understand the necessity of graph clustering algorithms which would lead them to continue in the future directions.

References:

1. Tobler, W. R. A computer movie simulating urban growth in the Detroit region. Economic geography.1970:234-240.

2. Parimala, M., Lopez, D., & Senthilkumar, N. C. A survey on density based clustering algorithms for mining large spatial databases. International Journal of Advanced Science and Technology. 2011:31(1): 59-66.
3. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R.Advances in knowledge discovery and data mining: MIT Press; 1996
4. Hu, X., & Pan, Y. (Eds.). Knowledge discovery in bioinformatics: techniques, methods, and applications. 5: 2007.
5. Ng, Raymond T., and Jiawei Han. Clarans: A method for clustering objects for spatial data mining. Knowledge and Data Engineering, IEEE Transactions on 2002: 1003-1016.
6. [6] Punj, G., & Stewart, D. W. Cluster analysis in marketing research: review and suggestions for application. Journal of marketing research. 1983:134-148.
7. Hartigan, J. A., & Wong, M. A.Algorithm AS 136: A k-means clustering algorithm. Applied statistics. 1979:100-108.
8. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd. 1996:96(34): pp. 226-231.
9. Zhou, Y., Cheng, H., & Yu, J. X.Graph clustering based on structural/attribute similarities. Proceedings of the VLDB Endowment. 2009:2(1):718-729.
10. Cheng, H., Zhou, Y., & Yu, J. X. Clustering large attributed graphs: A balance between structural and attribute similarities. ACM Transactions on Knowledge Discovery from Data (TKDD). 2011:5(2)
11. Xie, J., Kelley, S., & Szymanski, B. K. Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Computing Surveys (csur).2013: 45(4): 43.
12. Parimala, M., & Lopez, D. K-Neighbourhood Structural Similarity Approach for Spatial Clustering. Indian Journal of Science and Technology.2015:8(23).
13. Yang, J.,& Leskovec, J. Overlapping community detection at scale: a nonnegative matrix factorization approach. In Proceedings of the sixth ACM international conference on Web search and data mining. 2013:587-596.
14. Yang, J., & Leskovec, J. Structure and overlaps of communities in networks. arXiv preprint arXiv:1205.6228.2012
15. Parimala, M., & Lopez, D. Spatio-temporal graph clustering algorithm based on attribute and structural similarity. International Journal of Knowledge-based and Intelligent Engineering Systems.2016:20(3):149-160.