Social Theories Based Unsupervised Community Detection Over Social Media: A Review

Rishank Rathore, Ravi Singh Pippal

Department of Computer Science & Engineering

RKDF University, Bhopal

Abstract Social media mining is an emerging field with lot off research areas such as, sentiment analysis, link prediction, spammer detection, and community detection. In today's scenario researchers are working in the area of community detection and sentiment analysis because the main component of social media is user. Users create different types of communities in social world. The ideas and discussions in the community may be n0egative or positive. To detect the communities and their behavior researchers have done a lot of work, but still there are two major issues present as per survey that are scalability and quality of the community. These issues of community detection motivate to work in this area of social media mining.

Keywords: Social Media mining, Social media, Community detection, big data, influence, homophile, confounding

I. INTRODUCTION

In today's scenario social media is an emerging field for many researchers. In social media the data generated through user side is huge. To maintain the user-generated data there are many mining tasks are present in social media mining. There are many social networking sites where user makes their own community on the basis of their interest. As it is known that social media is a big virtual world in that many users have their profile and they are connected to different type of groups. To know the behavior of the user it is necessary to understand the background of user. It is not that easy in social network to identify the behavior of the single user, therefore it is needed to perform community detection in social network. Many researchers had done lot of work in this field of the social network.

There are most popular sites in the web world for social media where users can make different social relationships and groups of different people of different thinking and views. This type of sites is known as social networking sites. Here users can share contents and resources. Examples of these sites are (https:// www.facebook.com), (https:// www.twitter.com), (https://www.friendsnet.com), and so on

Jack and Scott coined something about social media in 2011 to describe online social network. Social media is the collection of web-based broadcast technologies, where user gets an ability to emerge from consumers of content to publishers. In 2011 oxford university defines that social media is a web-based application for social networking **[1].**

Social networking is defined as the use of dedicated social sites for communication with other users, or to find people with similar interests. After further research Kalpan and Haenlien in 2010 said according to organization for economic cooperation and development content given by the user must meet three basic requirements to qualify as UGC (user-generated content).

1.) Content must be published to all web users or to a selected group (excluding mails and instant messages).

2.) It should be creative and original not the replica of another's content of the other user.

3.) It should be created apart from professional routines and not used for commercial purpose.

Kalpana and Haenlien give an argument that the systematic classification of social media can be difficult, because new sites develop every day. This argument brings social media as an emerging field for researcher in research area. **[1]**

Social media mining is a process of visualizing, evaluating and extracting applicable patterns over the social network [3]. Through Social media mining they have integrated social theories with the computational methods.

Social media mining define basic principle and concepts for investigating huge amount of social media data. In this mining they have discuss different disciplines such as, computer science, data mining, social media, machine learning etc.

For social media mining they have encompasses the tools to formally represent, model, measure and

extract meaningful pattern for large social media networks. Social media sites generate user data which is different from traditional attribute-values of data for Hellenic data mining. The data which is generated from social sites is noisy, distributed, not in proper structure and frequent. All the characteristics of social media data pose challenges for data mining task and for that new techniques and algorithm have to be developed. [3]

II. COMMUNITY DETECTION

Community detection is a process of finding communities in social network. Communities are known as groups, clusters or cohesive subgroups. Human nature is to form groups and community on the basis of their interest and characteristics. To understand the complex social network better, for that we have to figure out the number of communities through that we can divide complete network into small groups or in the form of clusters.

There are certain question arises regarding community detection. Some of the questions are as follow; first, is about the detection of communities, second evolution of communities, third, about the evaluation measures of detected communities. Whenever we detect communities it is based on either specific member or specific form of communities. **[3]**

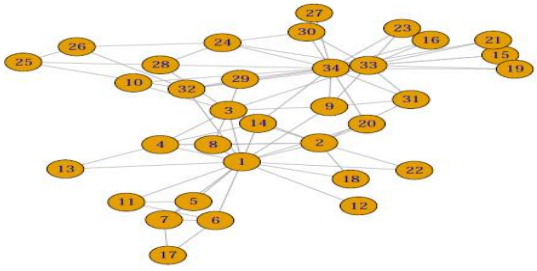

Fig 1 Blueprint of communities

III. COMMUNITY DETECTION ALGORITHM

Community detection algorithm is divided into two categories; Member-based and group-based community detection. Member-based community detection is based on characteristics of the member and Group-based community detection based on the interest of the user. [3]

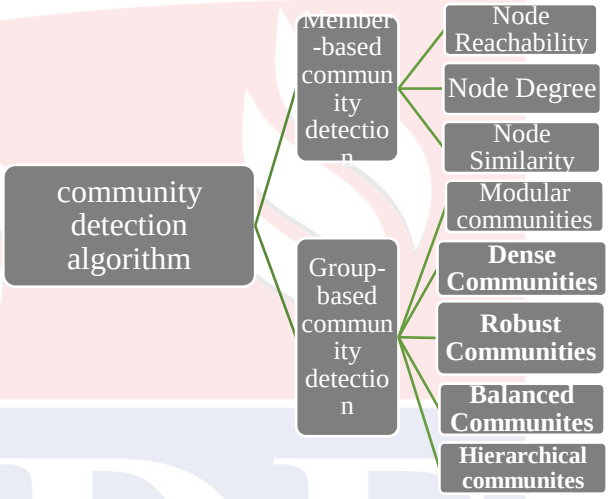

Fig 2 Hierarchical Classification of community detection algorithm [3]

Community detection algorithm is divided into two categories; Member-based and Group-based. These two categories are further divided into sub categories.

IV. MEMBER-BASED COMMUNITY DETECTION

Member-based community detection is done on the basis of characteristics of the member, because it is said that similar member's will be in same communities. If we consider a graph or network , then nodes that form cycle are consider to be a community, because they are closely connected. Sub graph of a graph is considered as a community on the basis of some characteristics that are node degree, node similarity and node reachability.

a) Node degree

Subgraph is searched in the network based on node degree and node degree is also known as clique. Clique is complete subgraph, in which all the nodes and pair of nodes inside the subgraph are connected. Suppose clique size of a subgraph is k and it consist k nodes where all nodes induced node degree K-1. There are two algorithms for finding out the cliques from the network or graph are as follow; Brute-force clique identification and clique-percolation Method.

b) Node reachability

Reachability means two pairs of nodes are in reach to each other. Whenever we deal with node reachability, we look for the subgraph in which nodes are

reachable from other nodes via a path. Two extreme points of reachability is achieved when nodes are supposed to be in the same community. There are some guidelines of node reachability; 1) there should be a path between them (without concerning about distance); 2) they should be close enough to be an immediate neighbor. For justifying the first property we can use any traversal algorithm to identify connected components. There are some predefined subgraph are as follow.

c) Node similarity

It is process of calculating similarity between the two nodes. If two nodes are similar enough they are assumed to be in the same community. Once the similarities between the nodes are determined then by applying classical algorithm, we can find the communities. In structural equivalence similarity is taken or determine on the basis of neighbor.

$$O^{'}(v_i, v_j) = |N(v_i) \cap N(v_j)|.$$

For large network the value of the equation increases rapidly, because number of neighbors is more in the large network. There are different normalization procedure are; Jaccard similarity and cosine similarity. Overlapping of the communities can change the similarity value of the node.

V. GROUP-BASED COMMUNITY DETECTION

In group-based community detection we consider the characteristics of the group. This category of community detection consists of some communities; balanced communities, modular communities, robust communities, dense communities, and hierarchal communities.

a) Balanced communities

They have used graph-based clustering [3], because it is proven most useful in identifying the communities, they cut the graph into several partitions and these partitions are represented as communities. Here they have used minimum-cut method for finding the balance communities. It is one of the oldest algorithms for dividing the graph into several partitions. It is mostly used for load balancing in order to minimize the communication between processor nodes. In this method network is divided into predefined number of parts, approximately in same size. When minimum-cut method is applied on network for community detection sometimes it result such communities which have only single node.

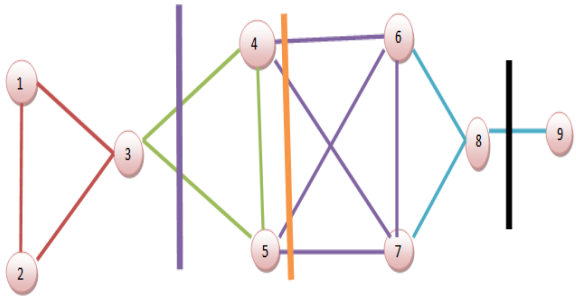

Fig 3 Minimum-cut method

b) Robust communities

In robust communities, they seek to find out such subgraph in which if one edge or node is removed then it does not disconnect the subgraph. Suppose there is an m-vertex graph in that m is the minimum number of nodes that must be removed to disconnect the graph. It means there are at least m- independent paths between pair of nodes. There is a similar subgraph is the m-edge graph, in that at least m-edges must be removed to disconnect the graph.

c) Modular communities

Modular communities are those communities that are identifying on the basis of the modularity. The structure of the community should be random. They have consider an undirected graph G (V, E), |E| = m. where degree of the node is known but edges are not known. Consider two nodes V_a and V_b, with degree d_a and d_b, respectively. They have to find out the expected number of edges between these two nodes. Now the probability of edge going out from node V_a and connected to V_b is (di / 2m).

d) Dense communities

Dense communities are those communities in which interaction is frequent. This type of communities is formed on the basis of same interest. For identifying the density of the community, some properties are there. Cliques, clans, and clubs are the dense community. Complete connected or interconnected subgraph/clique is a dense community.

e) Hierarchical communities

A hierarchical community means every community has its own sub/super communities. They have used hierarchical clustering algorithm, in which n nodes are consider to be one or an communities. In this clustering algorithm first they divide the graph/network equal to number of nodes. After that they start merging the adjacent communities into one

new community and then after that final number of communities is displayed.

VI. Background And Literature Review

Kuang Zhou [14] present a Median variant of Evidential C-means (MECM) prototype based algorithm for overlapped community detection. MECM relaxes the restriction of a metric space embedding for the objects and investigated the obtained creedal partitions of graphs for better understanding of the graph structures. MECM work over single center community network and ignore "multi-center" to avoid the troubles brought by the need for an initial seed using ESC and the definition of a threshold to control the distance between prototypes. Yu Xin [16] present Link-Block-Topic modelforoverlapped community detection based on the link-field-topic (LFT) model independently with context sampling, the number of communities. This clustering algorithm establish the semantic link weight (SLW) depending on the analysis of LFT and separate the SSN into clustering units independently with context sampling, the number of communities.Alexander G. Nikolaev [15] proposed an entropy centrality-based clustering algorithm. Defines a variation of entropy based on a discrete, random markovian transfer process and present their utility over the originally introduced path-based network. Clustering algorithm calculate weighted dataset for character co appearances in the text of "Les Miserables" and discovered communities with previously reported literature. W. Fan [17] discover people's underlying community structure based on people interaction and profile information over different social networking to analyze people's behaviors and the relationships among them. Yafang Li [18] present a parameter-free community detection method (K-rank-D). This algorithm select initial seeds and the number of communities from the decision graph draw after evaluate the importance of a node through PageRank centrality algorithm. But it's very hard for K-rank-D to extract optimal number of communities. Samira Malek [19] present a Fuzzy Duo centric Community Detection Model to detect overlapping duo centric communities in the complex networks. Duocentric community is fabricated nearby two central nodes that connected enough to each other to shape the center of the community.The network's nodes membership values are defined as type-2 fuzzy numbers that indicate the degree of belonging to both central nodes as upper and lower membership values. If communities do not have sharp boundaries, interval type-2 fuzzy membership values are able to describe how nodes are shared between the communities and formed overlapping communities. YunfengXu [22] present a community forest model for disjoint community detection based on social and biological properties to characterize the structure of real-world large-scale networks. Community forest model use backbone degree to measure the strength and similarity of edge and vertices and developed an algorithm that based on backbone degree and expansion to discover communities from real social networks.Fengjiao Chen [23] work over hierarchical structure of community members and present a structure to dig finer information by partitioning the members into several levels according to their belonging coefficients.

VII. Conclusion

Community detection is the most important feature of social media. It is similar to the clustering feature of data mining. In member-based community detection a lot of work has been done, but identifying community through influence is a different way of detection in social media mining. The main objective of the dissertation is to identify the community using Influence. Social network is large and complex, due that most of researchers do not consider the knowledge of community formation i.e. ground truth in their approaches and as it is known that every individual has its important role in the formation of community and social group. This is the researcher gap where work can be carried out for better community detection through two parameters; influence and user attributes.The user-generate content is used for many research work in the field of social media. Community detection is one of the emerging fields of the social media mining. Researcher has done lot of work in community detection. Major issues of community detection are scalability and quality of the community.

VIII. Reference

[1] Malliaros, Fragkiskos D., and Michalis Vazirgiannis. "Clustering and community detection in directed networks: A survey." *Physics Reports* 533.4 (2013): 95-142.

[2] Gong, Maoguo, et al. "Community detection in networks by using multiobjective evolutionary algorithm with decomposition." *Physica A: Statistical Mechanics and its Applications* 391.15 (2012): 4050-4060.

[3] Social media mining by Reza Zafarani and Mohammad Ali Abbasi (http://dmml.asu.edu/smm,)

[4] Authors: A. *Lancichinetti and S. Fortunato*. Presented by: Ravi Tiwari (https://www.cise.ufl.edu/research/OptimaNetSci/**slides**/22Apr'10.**ppt**)

[5] Kafeza, Eleanna, et al. "T-PICE: Twitter personality based influential communities' extraction system." *Big Data (BigData Congress), 2014 IEEE International Congress on*. IEEE, 2014.

[6] Jiang, Fei, et al. "A uniform framework for community detection via influence maximization in social networks." *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 2014.

[7] Barbieri, Nicola, Francesco Bonchi, and Giuseppe Manco. "Influence-based network-oblivious community detection." *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013.

[8] Wang, Wenjun, and W. Nick Street. "A novel algorithm for community detection and influence ranking in social networks." *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 2014.

[9] Sathanur, A.V.; Jandhyala, V.; Chuanjia Xing, "PHYSENSE: Scalable sociological interaction models for influence estimation on online social networks," in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, vol., no., pp.358-363, 4-7 June 2013 doi: 10.1109/ISI.2013.6578858

[10] Li, Jinshuang, and Yangyang Yu. "Scalable influence maximization in social networks using the community discovery algorithm." *Genetic and Evolutionary Computing (ICGEC), 2012 Sixth International Conference on*. IEEE, 2012.

[11] Maiti, Saptaditya, Deba P. Mandal, and Pabitra Mitra. "Detecting influential users using spread of communications." *Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in*. IEEE, 2013.

[12] N. Shrivastava, A. Majumder and R. Rastogi, "Mining (Social) Network Graphs to Detect Random Link Attacks," 2008 IEEE 24th International Conference on Data Engineering, Cancun, 2008, pp. 486-495. doi: 10.1109/ICDE.2008.4497457.

[13] Catanese, Salvatore & De Meo, Pasquale & Ferrara, Emilio &Fiumara, Giacomo&Provetti, Alessandro. (2012). Extraction and Analysis of Facebook Friendship Relations. Computational Social Networks: Mining and Visualization. . 10.1007/978-1-4471-4054-2_12.

[14] Kuang Zhou, Arnaud Martin, Quan Pan, Zhun-ga Liu, Median evidential c-means algorithm and its application to community detection, Knowledge-Based Systems, Volume 74, January 2015, Pages 69-88, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2014.11.010.

[15] Alexander G. Nikolaev, RaihanRazib, AshwinKucheriya, On efficient use of entropy centrality for social network analysis and community detection, Social Networks, Volume 40, January 2015, Pages 154-162, ISSN 0378-8733, https://doi.org/10.1016/j.socnet.2014.10.002.

[16] In Yu, Jing Yang, Zhi-QiangXie, A semantic overlapping community detection algorithm based on field sampling, In Expert Systems with Applications, Volume 42, Issue 1, 2015, Pages 366-375, ISSN 0957 4174, https://doi.org/10.1016/j.eswa.2014.07.009.

[17] W. Fan, K.H. Yeung, Similarity between community structures of different online social networks and its impact on underlying community detection, In Communications in Nonlinear Science and Numerical Simulation, Volume 20, Issue 3, 2015, Pages 1015-1025, ISSN 1007-5704, https://doi.org/10.1016/j.cnsns.2014.07.002.

[18] Yafang Li, CaiyanJia, Jian Yu, A parameter-free community detection method based on centrality and dispersion of nodes in complex networks, In Physica A: Statistical Mechanics and its Applications, Volume 438, 2015, Pages 321-334, ISSN 0378-4371, https://doi.org/10.1016/j.physa.2015.06.043.

[19] Samira MalekMohamadiGolsefid, Mohammad HosseinFazelZarandi, Susan Bastani, Fuzzy duocentric community detection model in social networks, In Social Networks, Volume 43, 2015, Pages 177-189, ISSN 0378-8733, https://doi.org/10.1016/j.socnet.2015.04.009.

[20] ArieCroitoru, N. Wayant, A. Crooks, J. Radzikowski, A. Stefanidis, Linking cyber and physical spaces through community detection and clustering in social media feeds, In Computers, Environment and Urban Systems, Volume 53, 2015, Pages 47-64, ISSN 0198-9715, https://doi.org/10.1016/j.compenvurbsys.2014.11.002.

[21] PooyaMoradianZadeh, ZiadKobti, A Multi-Population Cultural Algorithm for Community Detection in Social Networks, In Procedia Computer Science, Volume 52, 2015, Pages 342-349, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.05.105.

[22] YunfengXu, Hua Xu, Dongwen Zhang, A novel disjoint community detection algorithm for social networks based on backbone degree and expansion, In Expert Systems with Applications, Volume 42, Issue 21, 2015, Pages 8349-8360,

ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2015.06.042.

[23] Fengjiao Chen, Kan Li, Detecting hierarchical structure of community members in social networks, In Knowledge-Based Systems, Volume 87, 2015, Pages 3-15, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2015.05.026.