International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 2, Mar-2017, ISSN (Online): 2395-566X

Community Detection on Social Media: A Review

Sweta Rai*
Department of Computer Science & Engineering
Radharaman Institute of Technology and Science, Bhopal
*Email: swetarai76@gmail.com

Shubha Chaturvedi
Department of Computer Science & Engineering
Radharaman Institute of Technology and Science, Bhopal

Chetan Agrawal
Department of Computer Science & Engineering
Radharaman Institute of Technology and Science, Bhopal

Abstract – **Social media mining is a process of visualizing, evaluating and extracting applicable patterns over the social network. Numbers of methods and algorithms are introduced for massive investigation of social media data. Recently Community detection attracts attention of researchers. Community detection identifies groups of nodes that are more densely interconnected relatively to the rest of the network. In this paper influence, homophily, confounding in community detection and measures; Modularity, Normalized mutual information over and discuss major issues in this area such as "link prediction".**

Keywords – **Social Media Mining, Community Detection, Link Prediction, Homophily, 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 types of groups. To know the behavior of the user it is need to understand the background of user. It is not that easy in social network to identify the behavior of the single use, 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

Social media mining is a process of representing extracting and analyzing actionable patterns from social media data. Social media shatters the boundaries between the real world and the virtual world. We can now integrate social theories with computational methods to study how individuals interact and from communities. The uniqueness of social media data is for novel data mining techniques that can effectively handle user generated content with rich social relation. There are many emerging research area in social media mining. The most known research area of social media mining is community detection.

Community detection is a process of detecting communities form in social media on the basis of ground truth given from social media data. Here we are doing community detection based on influence. In community detection data points are defined as actors in social media and similarity between actors are defined based on the interest these user shares. In social networking sites only fraction of user gets influence by other users. Community detection has received attention in all kinds of networks, such as social network, biological network and World Wide Web. Now we discuss social forces through which users or nodes are connect in social communities. Three common forces are as influence, homophily, and confounding. Influence is the process by which an individual (the influential) affects another individual such as influenced individual becomes more similar to the influential figure.

II. SOCIAL MEDIA MINING

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]

III. COMMUNITY DETECTION

It is a process of detecting communities in the social network. Community detection is essential in social media, due to many reasons. First, users create group on the basis of their interest. There are two type of communities; real-world communities and virtual communities. Real-world communities are a body of user with common economic, social, or political interest/characteristics. Virtual communities are those communities which are created on the social sites. There are varieties of community detection algorithm. When we detect communities, we are interested in two type of detection .Specific member's Specific form of communities. We also denote the former as:-Member-based community detection, Group-based community detection.

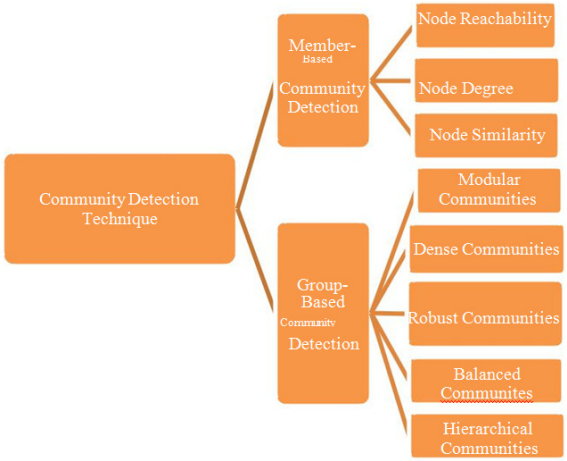

Fig.1. Hierarchical Classification of community detection Technique [3].

Member-based community detection is done on the basis of characteristics of the member, because it is said that similar member's will we 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 connect. 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.

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.

IV. LITERATURE REVIEW

R. Hosseini et.al [8] proposes an improved label propagation algorithm called memory-based label propagation algorithm (MLPA) for finding community structure in social networks. In the proposed algorithm, a simple memory element is designed for each node of graph and this element store the most frequent common adoption of labels iteratively.

W. Wang et.al [1] approach is not differentiates the influence ranking but also effectively find communities in both directed and undirected network. These two task is incorporates into one integrated framework. They demonstrate its superior performance with extensive tests on asset of real-world networks and synthetic benchmarks. They provide a new perspective on the influence based connectivity of network graph topology, and define a novel influence centrality and shared-influence-neighbor (SIN) similarity in an integrated framework. The SIN similarity is well-suited as a refined proximity metric for community detection. Experiments on both real-world and simulated network show the effectiveness and superior performance of the algorithm (IGSK) in both directed and undirected network.

N. Barbieri et.al [2] proposed a stochastic framework which assumes that item adoptions are governed by an underlying diffusion process over the unobserved social network, and that such diffusion model is based on community-level influence. When parameters of model get fitted to user activity log. We learn the community membership and the level of influence of each user in each community. In this paper author define two models: the extension to the community level of the classic (discrete time), Independent cascade model, and model that focuses on the time delay between adoptions. This the first work studying community detection without the network. The experiments show that both the models are robust and effective, and be can profitably employed to discover communities and region of influence in situations where the social connections are not visible.

F. Jiang et.al [3] present as efficient and effective framework based on local influence to detect both overlapping and hierarchical communities. Also they try to illuminate two fundamental questions.1) whether local influence regarded as a new property can affect the formation of communities; 2) how to quantify node‟s local influence and utilize it to detect communities. This paper show a new quantity of local influence is presented to measure influence of nodes. Due to fully use of local topological structure, local influence also measure the mutual relations between two nodes and quantify the importance of nodes precisely. They proposed a uniform framework for both community detection and influence maximization, heuristic selecting nodes with largest total influence applied, which gives excellent result efficiently. It is tested on both real network and artificial network.

E. Kafeza et.al [4] present an approach that extend the notion of influence from users to networks and consider

personality as a key characteristic for detecting influential networks. We describe the twitter personality based influential communities extraction (T-PICE) system that creates the best influential communities in a twitter network graph. The approach of author is demonstrated by sampling the twitter graph and comparing the influence of the created communities with and without considering the personality factor. They proposed the influential communities extraction methodology (T-PICE), a unified framework that extracts users personality based on several aspects of user behavior and colors the network graph.

V. Sathanur et.al [5] introduces PHYSENSE, a scalable framework for influence computation and activity prediction on large online social networks (OSNs). PHYSENSE estimates and set up sociological influence model to compute the diffusion of activity potential in the neighborhood of each of the nodes. PHYSENSE then scales these to significant parts of the entire OSN by propagating these activity potentials through an equivalent Helmholtz Green‟s function. PHYSENSE formulates the challenge of influence detection on large online social network as a driven problem.

Influence of the different members in a community is not the same. Every community have some core members their influence is far greater than the others. In this view, a community discovery algorithm is proposed to find core members of the community. Selecting initial members from these core members will have the greatest influence. The contribution of this paper is mainly reflected in three aspects. 1.) Clearly pointed out that the influence of the vertices in the network is different, the more close to core member of community have greater influence. 2.) It pointed out that community detection algorithm can be used to find the greatest influence vertices in the social network. 3.)For a given network, calculate the appropriate value k which let influential throughout the network and without getting wasted [6].

S. Maiti et.al [7] observe that a communication of an influential user is likely to reach many more users than the same made by a user having lesser influence in the network. Based on this observation, they have formulated a method using the spread of communications. (i.e., the number of users the communication reaches). We have verified the method on three datasets downloaded from „Twitter‟ and results are found to be the best among existing methods on the said datasets. In this paper they discussed about determining most influenced nodes in a social networking site. Spread of the communication has been introduced here to determine the influence users. They have considered only “Twitter” in describing the proposed method. As the social networking sites differ in structure and goal.

V. RESEARCH GAP

1) Extraction of appropriate features from social media data. Which represent complex relationships which stem from different data origin and subsequently use them to identify influential communities

2) Second, research gap is ambitious problem of inferring and the community structure when the social graph is not available.

3) Most of the community detection approaches do not consider about the information of formation of communities in social network.

4) In real-world communities, community members have individual social roles: leaders, elite‟ members. Traisons to other communities, some are more influential than others, and some are more susceptible to influence individual roles.

5) The identification of influential users in a social network is a problem that has received significant attention in recent research.

VI. PROBLEM DEFINITION

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.

There is a problem of influence maximization in the social network. In this work the main focus is on detecting influence flow in the community with influence-user of the community. As it is known that most influential user increase the flow influence in the community with this one more issues of community detection is taken i.e. scalability in large network.

VII. DATASETS

Dataset is a collection of data in the form of table. It represent as a single database table in that column represent the variable and row represent the member with respect to that variable. There are two types of dataset is used i.e. Zachary's karate club dataset and Twitter dataset.
Zachary's karate club dataset -It is karate club in U.S University created by Wayne Zachary's with their respective members, but due to some disputes between the ideas it is split into 2 sub groups. This club dataset is published online by Network Repository of University of California.

Description of dataset

- It contain 34 nodes
- It has 78 edges
- It is undirected type of graph
- It is static in nature
- It is unwieghted

Table 1 shows

Attributes	Description
ID	It is one of the attribute of dataset it shows uniqueness of the node.
Vertex	It is the column in dataset workbook which shows the number nodes in the graph.
Label	It shows the name of the nodes in the graph.
Edges	Connecting line between two nodes or path is called edge.
Weight	Weight of the edges it is the number of nodes connected to that edges.
In-degree	It is the number of edges coming to the nodes.
Out-degree	It is the number of edges going out of the nodes.
Degree Centrality	Degree centrality of the node is identifying by the in-degree and out-degree of the node. First we calculate the in-degree and in-Degree represent that node is prominence/ prestige. Secondly we calculate the out-degree of the node that represent node is gregariousness. Total degree centrality is the summation of in-degree and out-degree.
Eigenvector Centrality	This centrality is calculated on the basis Of neighbor. Eigenvector centrality normalized the centrality of degree. Neighbor of the node is more important in calculating the eigenvector centrality.
Katz Centrality	In Katz Centrality we added a bias β to all the nodes that are present in the node and after that the centrality that calculate is called Katz centrality.

Twitter dataset- Twitter data is available in various twitter corpus and also at twitter API.

VIII. EVALUATION MEASURES

Evaluation measures are the parameter through which communities detected by the algorithm is as per the ground truth or not. There three type of evaluation parameters are as follow.

Modularity:- Modularity is one measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is often used in optimization methods for detecting community structure in networks.

Normalized mutual information:- Mutual information is used in determining the similarity of two different clustering's of a dataset. Commonly used computer-generated benchmarks start with a network of well-defined communities.

Omega/Rand index- The Rand index or Rand measure (named after William M. Rand) in statistics, and in particular in data clustering, is a measure of the similarity between two data clustering's. A form of the Rand index may be defined that is adjusted for the chance grouping of elements; this is the adjusted Rand index. From a mathematical standpoint, Rand index is related to the accuracy

IX. CONCLUSION

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. Apart from the recent work there exist possibilities for further research with some improvement or optimization. In existing work, different type of clustering algorithm is combining with the logic for community detection. It need further improvement as per perspective of space and time complexity in the large-complex network.

REFERENCES

[1] W. Wang and W. N. 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, Beijing, 2014, pp. 555-560

[2] N. Barbieri, F. Bonchi and G. Manco, "Influence-Based Network-Oblivious Community Detection," 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, 2013, pp. 955-960. doi: 10.1109/ICDM.2013.164

[3] F. Jiang, S. Jin, Y. Wu and J. Xu, "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, Beijing, 2014, pp. 27-32.

[4] E. Kafeza, A. Kanavos, C. Makris and P. Vikatos, "T-PICE: Twitter Personality Based Influential Communities

Extraction System," 2014 IEEE International Congress on Big Data, Anchorage, AK, 2014, pp. 212-219.

[5] V. Sathanur, V. Jandhyala and C. Xing, "PHYSENSE: Scalable sociological interaction models for influence estimation on online social networks," Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on, Seattle, WA, 2013, pp. 358-363.

[6] J. Li and Y. Yu, "Scalable Influence Maximization in Social Networks Using the Community Discovery Algorithm," Genetic and Evolutionary Computing. (ICGEC), 2012 Sixth International Conference on, Kitakushu, 2012, pp. 284-287

[7] S. Maiti, D. P. Mandal and P. Mitra, "Detecting influential users using spread of communications," Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in, Trivandrum, 2013, pp. 288-292.

[8] R. Hosseini and R. Azmi, "Memory-based label propagation algorithm for community detection in social networks," 2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP), Mashhad, 2015, pp. 256-260.