A DATA EXTRACTION METHOD FOR COMMUNITY DETECTION ON SOCIAL MEDIA USING GRAPH THEORY

Pradeep Bhawsar[1], Sumit Sharma[2]
Dept. of CSE, Vaishnavi Institute of Technology & Science, Bhopal, India[1, 2]

Abstract: **The current situation is causing some observers to start concentrating their attention on web-based media. The client side of web-based media systems can generate enormous amounts of data. Online media mining provides a wide range of mining tasks in order to keep up with the information that the clients produce. There are numerous long-distance interpersonal communication locations where the client can designate their own local area in accordance with what is most practical for them. As has come to be known, a sizable chunk of the entire virtual world is made up of web-based media. This is as a result of the large number of users who manage their own profiles and take part in various online communities. Online media's main trait is its emphasis on community discovery. The clustering phase of the information mining process is analogous. On the other hand, identifying local area by effect is another identification method that may be applied to online media mining. Although this has served as a foundation for much work on local area awareness, it is still crucial. Honoring people in the neighborhood who are effectively using Leverage is the exhibition's main goal. Community detection also needs to address issues with scalability and community quality. According to those findings, some algorithms are scalable in vast networks and produce better outcomes than other algorithms. Based on the social data from Twitter, we have examined the variations and parallels among the algorithms. As a result, it has been proven that the algorithms, when applied to the vast network, are scalable in accordance with the evaluation criteria. This thesis differs from others in part because we thoroughly tested every algorithmic component on a big social network.**

Keywords: **- Community Detection method, Graph Theory, Social Networks, Twitter.**

1. INTRODUCTION

For some commentators, the sector of web-based media is now emerging. The amount of information generated by the client side in web-based media is enormous. Numerous mining tasks are available in internet media mining to keep up with the client-produced information. There are several long-distance communication locations where the client creates their own local region according to their advantage. As is well known, online media is a sizable virtual world where many users have profiles and affiliations with various kinds of groups. Understanding the client's base is crucial to understanding their behavior. Since it is not always easy in informal organizations to understand a single client's behavior, interpersonal organizations are expected to carry out local area discovery. Many experts had completed a significant amount of work in this area of the informal organization.

In the internet media realm, there are certain well-known locations where users can connect with people from all walks of life and join groups with similar interests. Informal communication destinations are places of this nature. Clients can contribute resources and content here. The websites (https://www.facebook.com), (https://www.twitter.com), (https://www.friendsnet.com, etc.) are examples of these locations.

In order to illustrate online informal organization, Jack and Scott wrote something in 2011 about web-based media. Online media is a collection of electronic transmission innovations that allow users to transition from being content buyers to content distributors. Online media are described as an electronic application for interpersonal contact by Oxford College in 2011 [1].

After more research, Kalpan and Haenlien in 2010 stated that by association for monetary participation and improvement content provided by the client should meet three fundamental requirements to qualify as UGC (User Generated Content).

1. Content should be distributed to all web clients or to a chose bunch (barring sends and texts).
2. It ought to be imaginative and unique not the reproduction of one more's substance of the other client.

3. It ought to be made separated from proficient schedules and not utilized for business reason.

Kalpan and Haenlien [1] give a contention that the orderly characterization of online media can be troublesome, in light of the fact that new locales foster each day. This contention brings web-based media as an arising field for analyst in research region. [1]

2. RELATED WORK

In this paper creator's goal is to acquire proper components from online media information that address complex relationship which is utilized to recognize powerful networks. Their commitment is in a few perspectives: first and foremost they broaden existing methodologies for character dependent on client's conduct extricated from online media information; then, at that point, they distinguish the mining calculations that best arranged for every character characteristic. They proposed the compelling networks extraction philosophy (T-PICE), a bound together structure that extricates clients character dependent on a few parts of clients conduct [5].

In this paper creator introduced a successful and effective structure dependent on nearby impact to identify both covering and various leveled networks. There proposed structure is appropriate to handle the issue of heuristic impact amplification. Their calculation initially evaluated the neighborhood impact and afterward discovers the local area [6].

In this paper creator proposed a structure which is straightforwardly distinguishing networks in an organization absent setting. They have utilized social diagram and log movement information for discovering persuasive hub and local area. They suggested that system to manage the aspiring issue of inducing the local area structures [7].

In this paper creators approach is to discover networks in both organization (coordinated and undirected) successfully. Here creator utilizes the term shared-impact Neighbor similitude. It implies two hubs have same arrangement of neighbor hubs [8].

In interpersonal organizations, the impact of the various individuals locally isn't something similar. Each people group has some center individuals their impact is far more prominent than the others. In this view, a local area disclosure calculation is proposed to discover center individuals from the local area. Select introductory individuals from these center individuals that will have the best impact [10].

In this paper creator see that a correspondence of a powerful client is probably going to arrive at a lot a larger number of clients than the equivalent made by a client having lesser impact in the organization. In light of this perception, they have detailed a strategy utilizing the spread of correspondences. We have confirmed the technique on three datasets downloaded from 'Twitter' and results are observed to be awesome among existing strategies on the said datasets. In this paper they examined about deciding most affected hubs in an interpersonal interaction site. Spread of the correspondence has been acquainted here with decide the impact clients. They have thought about as it were "Twitter" in depicting the proposed technique. As the long range interpersonal communication are locales contrast in construction and objective [11].

In this paper creator have overview the distinctive local area identification calculation and applied a portion of the calculation on the genuine organization and manufactured organization. The fundamental focal point of this paper is to actually take a look at the exhibition of the calculation in genuine organization. Aftereffect of the review is that there are two significant issues; 1) nature of the identified networks and 2) adaptability of the calculation [12].

In this paper creator is attempt to utilize hereditary calculation for recognizing the networks. As contrast with other local area discovery calculation, hereditary calculation is more versatile in huge organization and it needn't bother with any earlier information about number of networks or any edge. Creator have tried the exactness of the calculation on notable datasets; Zachary karate club, school football. Select email dataset is utilized to actually look at the versatility of the calculation. Objective of the creator is to streamline the organization particularity utilizing hereditary qualities [13].

In this paper author use the approach of detecting community on the basis of node attributes. It is said that one node can be a part of two or more communities on the basis of their attributes. So, for performing clustering on the basis of node attribute two sources of data is require; first is the data/known properties of the object (nodes) and second, source of data come from the network, a set of connection between objects. Through this paper author developed CESNA method for identifying overlapping community detection in network [14].

Today's research and science has provided significant advances to understand the complex network. The most relevant feature of graph is to represent real system as community structure or clustering. Main focus of the author is to develop the technique through which complex network is divided into small communities through which it better to understand the flow the network [15].

The author presented a survey on the fundamental concepts and methodological principles of clustering algorithm. Small modules of graph are known as cluster or community. Firstly author is concerned regarding the principles of the clustering algorithm and secondly it is concerned regarding the properties of the good cluster. Author has represent methods and metrics for evaluating graph clustering results [16].

3. METHODOLOGY

Data Extraction: It is a process of retrieving data from the data source and the data which is extracted from the source system is unstructured in nature. Extracted data contain noise and irrelevant attributes to remove this, data need to be preprocessed. It is procedure to retrieve the data which is generated by the users through internet or by some applications.

Framework for Data Extraction through Social Sites

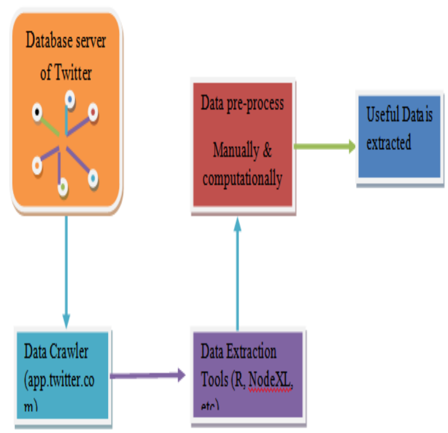

Fig: 1 Framework of data extraction

It is the framework of data extraction through social site, in the process of extraction there are five steps procedure. In first step request is sent to the database server by data crawler for the access token. After receiving the request from the data crawler, database server reply-back with access token. As the data crawler receives the access that token is used by the tool and data is extracted. The data which is extracted is in the form of unstructured data. Fourth step is to preprocess the data by manually as per the need. In last step useful data is received and used for the work.

Flow Chart

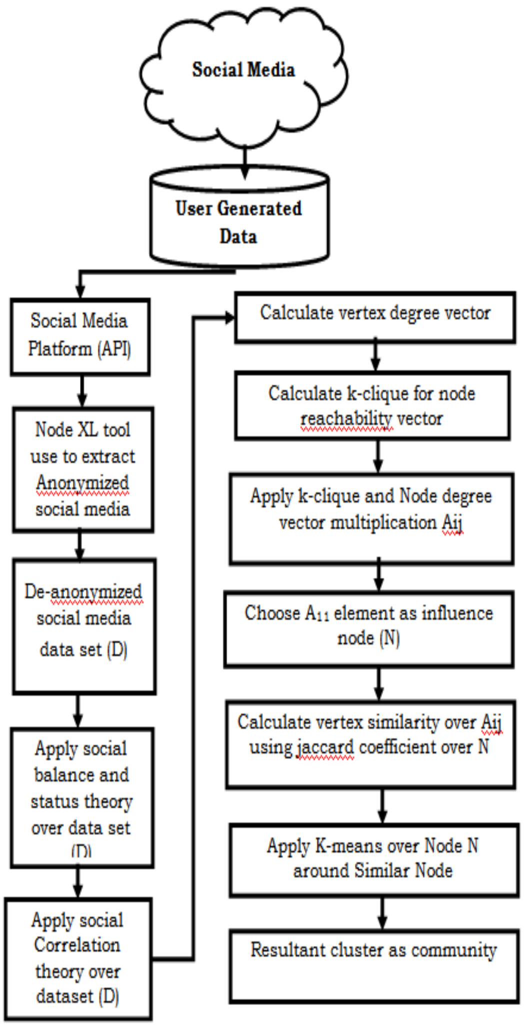

Fig: 2. Flow chart

4. RESULT ANALYSIS

Evaluation measures are the parameter through which communities detected by the algorithm is as per the ground truth or not.

We have used modularity as our evaluation parameter. Modularity is one measure of the structure of networks or a graph which is designed to measure the strength of division of a network into modules called groups, clusters or communities. Networks having dense connections between the nodes within modules have high modularity than the sparse connections between nodes in different modules. Modularity is also used in optimization methods to detect community structure in the networks. Modularity is one such measure, which when maximized, leads to the appearance of communities in a given network. The value of the modularity lies within the range [−1/2, 1). If the number of edges within groups exceeds the expected number on the basis of chance it is positive. Example of modularity by using tweets of the twitter.

The experimental setup for community detection on different dataset by two algorithms is; Pentium 4 machine, 1800MHz, 1GB RAM, Windows 2007 with graphics, R-studio and Dataset should be preprocessed. hedre, we report the results of experiments on the Social network dataset described in Table 1:

Table 1 Modularity of Community Detection algorithms on Social Network Datasets

DATASETS	Walktrap	Fast greedy	Edge betweenness
KARATE	0.353222	0.380671	0.4012985
DOLPHINS	0.488845	0.495491	0.5193821
FOOTBALL	0.602914	0.549741	0.599629

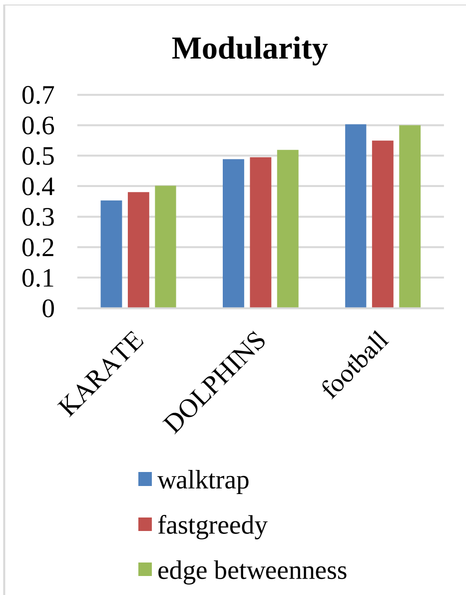

Fig.3 Results of Modularity for comparison algorithms

Table 2: Results of Execution Time

DATASETS	Walktrap	Fast greedy	Edge betweenness
KARATE	0.00399	0.00298	0.004987955
DOLPHINS	0.00398	0.002	0.01478505
FOOTBALL	0.00399	0.00499	0.2853062

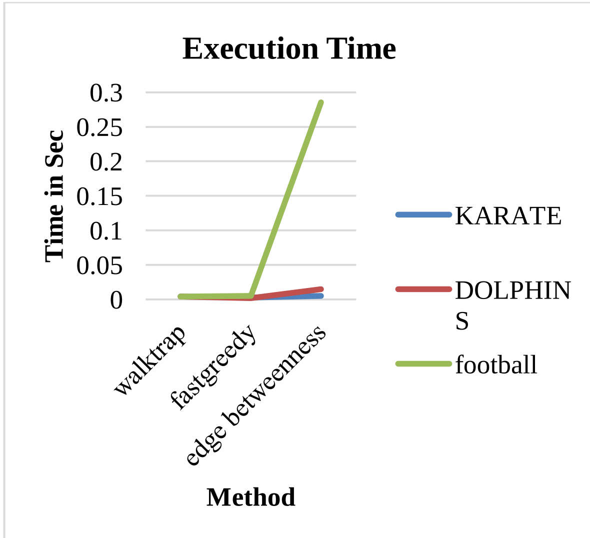

Fig: 4 Results of Execution time for Comparison algorithms

Table 3: Results of Modularity with Twitter Data Set

Dataset	Walktrap	Fast greedy	Edge Betweenness	Proposed Method
Twitter	0.6129143	0.559731	0.609128	0.629128

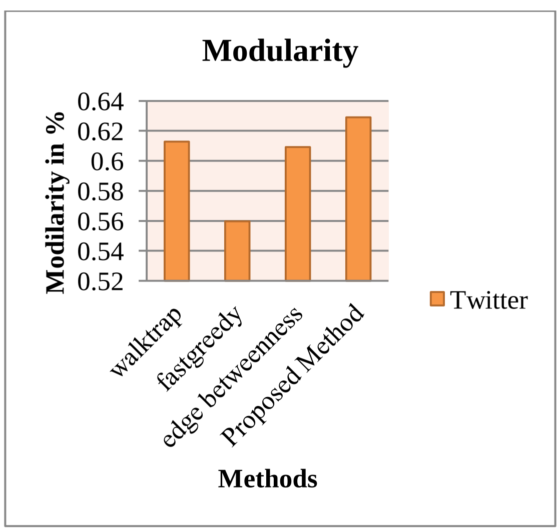

Fig.5.Modularity for with Proposed Approach

5. CONCLUSION

Social media plays a significant role in people's social lives. Positive and negative traits are part of human nature. The fact that some people utilize technology for harmful purposes as well as beneficial ones shows that many people use it. The user enters their personal information when creating their social media profile. This data is vulnerable to third-party hacking and can be exploited to create bogus profiles. Some people build their profiles with fraudulent information and then start communities or groups to spread harmful information. Researchers begin examining the social media data to find these communities and groups. Social media data is used in a variety of academic domains. Massive amounts of user content were produced through social media. The content created by users is frequently utilized in social media research projects. One of the newer subfields of social media mining is community detection. A lot of research has been done on community detection. Scalability and community quality are two major problems with community detection. Some algorithms are scalable in big networks and offer superior outcomes to other algorithms. On the basis of Twitter's social data, we compared the algorithms. As a result, it is demonstrated that algorithms can scale in a big network according to the assessment criterion. This thesis' distinctive quality is that we assessed every algorithmic feature on a sizable social network.

REFERENCES

[1] Qiuling Yan , Shaosong Guo, Dongqing Yang Advanced Data Mining and Applications Volume 7121 of the series Lecture Notes in Computer Science pp 82-95

[2] Tang, Lei, and Huan Liu. "Community detection and mining in social media." Synthesis lectures on data mining and knowledge discovery 2.1 (2010): 1-137.

[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] Chintalapudi, S. Rao, and MHM Krishna Prasad. "A survey on community detection algorithms in large scale real world networks." Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on. IEEE, 2015.

[13] Mursel Tasgin, Amac Herdagdelen, Haluk Bingol (Submitted on 4 Nov 2007)

[14] J. Yang, J. McAuley and J. Leskovec, "Community Detection in Networks with Node Attributes," 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, 2013, pp. 1151-1156. doi: 10.1109/ICDM.2013.167

[15] Fortunato, Santo. "Community detection in graphs." Physics reports 486.3-5 (2010): 75-174.

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

[17] 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.

[18] Shen, Huawei, et al. "Detect overlapping and hierarchical community structure in networks." Physica A: Statistical Mechanics and its Applications 388.8 (2009): 1706-1712.

[19] Xing, Yan, et al. "A node influence based label propagation algorithm for community detection in networks." The Scientific World Journal 2014 (2014).

[20] Bonchi, Francesco. "Influence Propagation in Social Networks: A Data Mining Perspective." IEEE Intelligent Informatics Bulletin 12.1 (2011): 8-16.

[21] Yoonseop Kang, Seungjin Choi Neural Information Processing Volume 5863 of the series Lecture Notes in Computer Science pp 175-184

[22] Eleanna Kafeza, Andreas Kanavos, Christos Makris , Dickson Chiu Volume 8697 of the series Lecture Notes in Computer Science pp 7-13

[23] http://www.defence.gov.au/pathwaytochange/docs/socialmedia/1.%20Social%20media%20and%20its%20origins%20SM.pdf