A Survey of Message Filtering & Community Detection in OSN

Rahul Nema, Dr. Anjana Pandey

Abstract— Today a large number of users access social networking websites. During access of online social network a large number of users who are un-authorized also access data and sometimes posts unwanted messages on the social network. These networks may sometimes contain hidden structure to link between numbers of users. For the detection of communities various algorithms are proposed in which WEBA and Greedy are the techniques used for the community detection. The concept of community kernels means the shared properties of various community groups so that the groups can be distinguished based on their entities. Here in this paper a complete survey of all the techniques that are used in online social network for the message filtering and community kernel detection techniques, hence by analyzing various advantages and issues in the techniques a new and efficient technique is implemented in future.

Index Terms— Social Network, Message Filtering, Communities, WEBA, Greedy.

1 INTRODUCTION

Defining a social network it is a social structure which consists of individuals or organizations) which are in some manner related to each other. The social network viewpoint provides a set of methods which analyzes the structure of social entities and the theories that explains the patterns observed in these structures. The study of these structures uses social network analysis for identification of local and global patterns, locating influential entities and examining network dynamics. **S**ocial network approaches to understand social interaction which should be first visualized and investigated through the properties of relations between units and not unit properties itself. In a social network different types of relations that may be singular or combination form the network configurations and network analytics. Whereas a social networking service provides platform for building social networks or social relations among users who share interests, activities, backgrounds etc. With the help of social network service each user maintains its profile containing his or her social links and other additional services.

Therefore with the help of Social networks users can create a public profile and can maintain a list of users to share connections and view and cross the connections inside the system. Social network services are web based facilitating the user to interact over the Internet, in the form of e-mail and instant messaging. Social network allow multiple information and communication tools like mobile connectivity, photo, video, sharing, blogging etc.

Social networks provide community services which are group centered and sometimes social networks are also considered as individual centered service. Social networks can be divided into communities consisting of users that may have same like, features, dislikes etc. Considering the main types of social networking services are those which are categorized in groups or communities of users like schoolmates, politicians, celebrities etc. and a recommendation system linked to trust of its users. Some of the widely used Social networks are Facebook, Google+, YouTube, LinkedIn, Integra, Pinterest, Tumblr and Twitter.

Each user has its own social network that may be online or offline consisting of friends, families and people they are acquainted with. Basic fundamental aim of online social networking is to make users social networks visible to others who are not in his/her immediate network. In a social network people are held together by kinship, friendship, classmates, colleagues, business partners, etc. which is a pre-established interpersonal relationships. The connections between the users are built one at a time. The primary reason behind people joining a social networking site is to maintain old relationships with others and form new ones for expansion of their network. Social networks are extremely unique in their own way as users collectively identify others if they are fake and also users generally do not compartmentalize their life i.e. they don't have only one social network. Communities in the social network are held together by common interest. The users may have

Common hobby for which the community members are passionate, a common goal, project, similar lifestyle, geographical location, profession etc. Thereby in social networks, there are two types of users exhibiting different influence and different behavior [1].

As the internet consists of information of various type and kind. To exploit and gain knowledge from this data there are users who reside on internet and continuously use it. The users share, disseminate and communicate multiple type of information among them. The information being in texts, audio, video, images etc. [2]. The users belong to various communities that consist of similar type of users in behavior which influence each other to share various type of data. The medium for this type of data exchange can be is in the form of mails, messages and social networks.

Social Networks can be visualized as an internet service which helps the user to build a social networks over the internet and relations with other users for sharing of interests, backgrounds and even establish real life connections and participate in multiple activities with the users that can be characterized by communities.

Social Network services are web based group centered services in which users share almost every type of data. One of the most important tasks while studying the networks is of identifying network communities. Communities discover

groups of interacting objects (i.e. nodes) and formalize the relations between them. For example, in social networks, communities can be defined on the basis of groups of friends attending the same school or people coming from the same home town etc [3]. The communities can correspond to scientific disciplines, family, friends, similar interests etc.

With the help of network communities functionally related objects study can be analyzed for studying interactions between modules, infer missing attribute values and predict unobserved connections. Social networks sometimes contain the community structure property i.e. groups of vertices with denser connections inside each group are divided and fewer connections crossing groups division in which vertices and connections represent network users and social interactions, respectively. Members of communities of a social network share things in common like interests in photography, movies, and music or discussion topics thus interacting more frequently with each other than with members outside of their community. Community detection in a network can be explained as gathering of network vertices into groups in such a way that nodes in each group connect inside densely and sparsely outside. [4, 5]

The problem associated with community kernel detection has some of the practical applications in form of representative user finding, friend recommendation, network visualization, and marketing. The problem being non-trivial in nature poses a set of challenges like true influential user's identification is difficult. The number of followers can be used as indicator but the follower count poses no information about who follows them. Influential users interact with each other is a bit non clear process and how does it take place explaining this with example as the questions arises that whether an actress will follow another actress or a sports person?

Since real world social networks with thousands of millions of vertices is increasing fast an algorithm with high scalability to solve the problem of community kernel detection is required with subtasks involving identifying influential (kernel) members and detecting the structure of community kernels.

Social Networks are the complex systems in which the nodes (or vertices) represent entities that have some relationships. Examples of such systems other than social networks are web graphs, telecommunication networks, biological networks, trade networks etc. The community detection problem or clustering symbolizes the identification of groups of nodes within which the connections (or edges) are numerous and between which they are scarce.

Spectral clustering methods are used for clustering which are based on the eigen decomposition of a Laplacian matrix that is derived from the data. This interpretation has an advantage of being extension of the clustering model to out of sample nodes. The clustering model thus can be trained and visualized on a small subset of the whole graph and thereby can be applied to the rest of the network in a learning framework platform while dealing with large and complex networks. The out-of-sample extension in the community detection field algorithm easily solves the problem of online clustering of large and increasing networks. This process when applied on every new node arriving in a data stream every node does not have to run on a new graph [6].

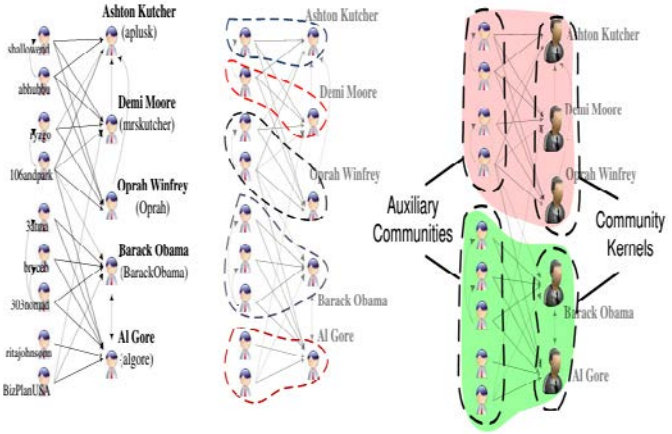

Figure 1. An illustration of community kernel detection on: The left figure shows the original Twitter network, the middle figure shows the five communities detected by Newman's algorithm and the right figure shows two community kernels and their corresponding auxiliary communities detected by WEBA

Social networks often rely upon Social influence. It governs the dynamics of all social networks. Users are influenced by other users while sharing information, exchanging data etc. thereby forming communities that may have users influenced by similarities like hobby, likes and dislikes etc. Social network analysis basically is focused on macro level models like degree distributions, clustering coefficient, communities, small world effect etc. Social influences define that a user can have higher influence over field than other user while the other may higher influence with completely different field. Thus these users are needed to be analyzed on the basis of influences for community formation based on common or similar influences. Social influence does not define global measure of importance of nodes or users; it defines the measure on links between nodes or the users.

Social network

Social networks have become very popular in this days and age due to the increase in proliferation and affordability of internet devices like personal computers, mobile and other more recent t hardware innovations like internet tablets. This is proved by the burgeoning popularity of many online social networks like Twitter Facebook .generally social network can be defined as a network where node (individual or organization) are related to each other by various interdependencies like friends. Where the nodes consist of actors and the dyadic ties (edges) denotes the relation or interactions between these actors. A generalization of the idea of social networks is that of information networks, in which the nodes could comprise the actors. Clearly, the concept of social networks no restrictions found for the specific case of an internet-based solution for example as facebook ,the problem of social networking has been studied in terms of generic interactions between any group of actors in the field of sociology. Such interaction should be done either in any conventional or non conventional form, whether they be, telecommunication interactions, email interactions, postal mail interactions or may be face-to-face

interactions.

Community detection

Community detection is relevant in many disciplines of science and modularity optimization is the widely accepted method for this purpose. It has recently been shown that this approach presents a resolution limit by which it is not possible to detect communities with sizes smaller than a threshold which depends on the network size. The problem of community detection in complex networks has recently attracted the attention of researchers in different areas of scientific knowledge.

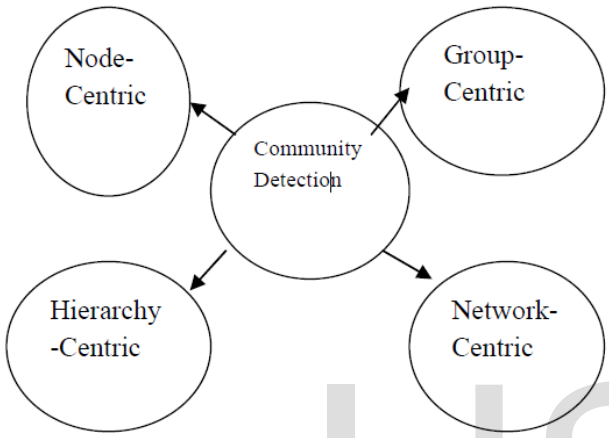

Figure 2. Architecture of Community Detection

The figure shown below is the parallelization performance of the exising WEBA Algorithm. Here WEBA is used as weighted algorihtm which is used to detect community kernel in online social networks.

The algorithm successfully predicts the link establishment between various nodes in thee network and finds the auxiliary nodes that are used to find the community kernels.

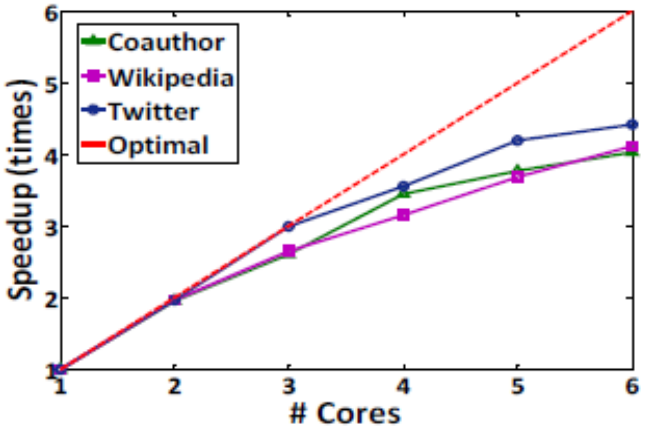

Figure 3. Performance of WEBA Algorithm

The figure shown below is the analysis of three existing techniques used for the efficient detection of community kernels in huge datasets such as co-authors and twitter and facebook. The analysis is done on the basis of precision and recall in which WEBA has high rate of precision and recall.

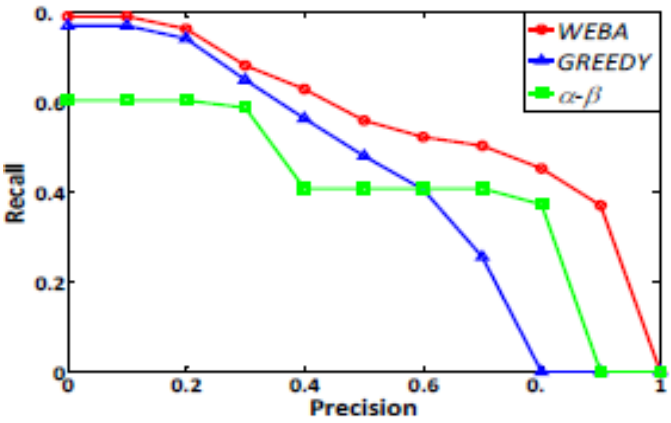

Figure 4. Comparison of existing community kernel methods

2 Literature Survey

L. Wang [1] et.al. explained different type of users exhibiting different behavior and influence. Considering the facts related with twitter they remarked that less than 1% of the users have behavior of producing their own content and that is also just 50% while other users behave differently on social network and poses less influence. They explained and defined the problem of community kernel detection and explained community structure in social networks. They analyzed the fact of users influenced by other users who are similar or common in some manner generating natural partition in various community kernels. Proposing GREEDY and WEBA algorithms they found community kernels in social network explaining GREEDY as maximum cardinality search and WEBA specifies the problem in an optimization framework. A community kernel distinguishes different groups of social entities and captures the common property shared by each group. Their suggested algorithm show that WEBA improved the performance over traditional cut based and conductance based algorithms. With the help of WEBA meaningful community kernels can be found, which reveals the common profession, interest, or popularity of groups of influential individuals [1].

J. Yang [7] et.al. remarked that with the help of Community detection algorithms organizational principles in networks can be visualized. Communities can be detected from the information of the network structure, features and attributes of nodes. They developed Communities from Edge Structure and Node Attributes (CESNA) detecting overlapping communities in networks with node attributes with accuracy and scalability and does not focus only upon network structure and node attributes as done by other community detection algorithms. CESNA proposed by them formalizes the interactions between the network structure and the node attributes thereby accurate community detection and provides robustness in network structure when noise is present. CESNA helped in interpretation of detected communities by determining relevant node attributes for each community. It has a linear runtime in the network size and can process networks an order of magnitude larger than other comparable approaches. CESNA can be made to handle more types of general attributes and it can be

IJSER

incorporated with information diffusion and edge attributes [7].

J. Leskovec [8] et.al. remarked that due to problem of detection of clusters and communities in social network user with the help of objective function determines the perception of clusters in network as nodes and apply algorithms or heuristics extraction of set of nodes that are similar or looks similar to good communities and are related to objective function. They compared various network detection methods to formalize the performance and to identify systematic biasing in clusters and studied algorithms for optimization of objective functions giving size resolved version of optimization problem to fix objective and the cluster. They proposed that determination of clustering structure for large networks is complex and the algorithms can easily optimize community score function on the basis of size scales and obtained cluster score. With the complex notion of quality cluster community score optimization fails. They also explained regularization concepts in machine learning and data analysis which causes effects due to extreme sparsity of real networks and generating results about the formalization of a notion of regularization by approximate computation [8].

Kevin S. Xu [9] et.al. studied the communities in social networks and focused upon temporal dynamics related with the networks and communities rather focusing on static networks as done by others. They tracked the communities in dynamic social networks over time by adaptive evolutionary clustering for tracking the communities. They obtained the temporal evolution of communities based upon real data sets generating statistical value to identify change points in network. With the help of adaptive evolutionary clustering variations in the communities become stable by temporal smoothing. Their proposed strategy is capable to track illegal activities in the network by communities as in case of multiple identities or changing members instantly. They highlighted challenges associated with tracking of communities with the experiments like validation of communities, selection of number of communities at each time which if wrong leads to unusual appearance of merging or splitting communities etc [9].

N. P. Nguyen [10] et.al. defined the community structure property of social networks. The community structure helps in developing social aware strategies for the problems related to social network and provides applications enabled by mobile networking like routings in Mobile Ad Hoc Networks (MANETs) and worm containments in cellular networks. They presented Quick Community Adaptation (QCA). QCA is an adaptive modularity based method which identifies and traces the community structure of online social networks updating network communities quickly and efficiently and tracing the evolution of community structure over time. Testing QCA in real world social networks and by realistic application on routing strategies in MANETs outperformed all the other methods presently being deployed. The algorithms adopted by them effectively updated and identified high quality network community structure while having fast running time. Via practical social aware routing strategy in MANETs QCA algorithm gave realistic applications in mobile computing as it is possible to combine or integrate as a community detection core [10].

A. Lancichinetti [11] et.al. suggested that with the help of community structure complex systems can be understood without depending upon the local organization of their constituents. They proposed several methods against introduced class of benchmark graphs, with heterogeneous distributions of degree and community size to overcome the drawbacks of other algorithms and test that are subjected to small networks with known community structure and artificial graphs with a simplified structure. They proposed that the Infomap method by Rosvall and Bergstrom is best on the criteria's of GN (Girvan and Newman) and LFR (Lancichinetti Fortunato Radicchi) benchmarks and random graphs for community detection. The benchmarks GN and LFR they used have low clustering coefficient and LFR benchmark is the generalizes the GN benchmark by power law distributions of degree and community size [11].

Nina Mishra [12] et.al. analyzed that due to ubiquitous behavior of social networks close knit-clusters are being found in the network. They defined clusters as collection of entities with dense relation patterns internally and sparse relation externally. With the help of tightly knit communities in networks target marketing schemes can be formed based on these clusters. While clusters follow some criteria's while forming like all vertices are clustered, external sparsity is ignored and clusters should not overlap. These limitations are overcome in their proposed scheme by combination of internal density and external sparsity and thereby monitoring internally dense and externally sparse clusters combinatorial properties. They proposed ρ-champion algorithm for determining (α, β)-clusters and explained various criteria's on which clusters depend giving results related to non overlapping of the clusters in various conditions and on multiple values of the clusters [12].

Jie Tang [13] et.al. explained about the influences of users by other users in social network. The influencing may be in form of colleague towards work or friends towards daily life etc. they proposed Topical Affinity Propagation (TAP) which models the topic level social influence on large networks. TAP performs topic level influence propagation through topic modeling and the network structure. They designed TAP with efficient distributed learning algorithms implemented and tested under Map Reduce framework. TAP describes the problem through graphical probabilistic model. They also implemented distributed learning algorithm under Map reduce programming model providing scalability. They suggested that discovered topic based influences by expert finding can improve its performance [13].

P. huetz [14] et.al. identified strongly connected substructures in social networks for reviewing the coarse grained organization. They gave multistep extension of greedy algorithm (MSG) with the help of which in each iteration step more than one pair of communities can be merged preventing premature condensation into large communities. They also proposed a refinement procedure for MSG as vertex motor (VM) which reassigns vertices to neighboring communities for improving modularity value thus the combined MSG-VM algorithm can find solutions of higher modularity without scaling computational cost of greedy algorithm. The MSG-VM algorithm run required similar computer time as required by greedy algorithm. With the help of MSG-VM algorithm network partitions with high modularity can be found [14].

V. D. Blondel [15] et.al. proposed heuristic method based on modularity optimization for extraction of community structure of networks. The proposed scheme outperformed every other scheme on the basis of computation time verifying the algorithms accuracy on ad hoc modular networks. With the help of algorithm networks of unmatched size can be studied. The speed of their proposed algorithm can be increased or improved heuristics like in the first phase of the algorithm when the gain of modularity is below a given threshold it is stopped or by deleting the nodes of degree 1 from the original network and then after the community computation they are added. They formalized a hierarchical community structure completely for the network in which hierarchical level is given through intermediate pass. They also gave the accuracy of the intermediate partitions which are local maxima of modularity in which modularity cannot be increased by movement of entity between the communities as these entities are at first nodes and then converts into large sets of nodes in the following passes. The algorithm then forms the final partition which has high modularity value [15].

David Crandall [16] et.al. explained the interaction between similarity and socialites. They remarked their study as people are similar to their neighbors because they resemble their current friends by social influence and form new links with others who are similar or are like them. This leads to uniformity of behavior and fragmentation respectively. They developed technique to identify and model the interactions between social influence and selection by data collected from online communities in which social interaction and changes in behavior over time is measurable. They monitored social interaction as an effect and a cause of selection and formalized that people become aware of others by shared, recent activity etc [16].

M. Rosvall [17] et.al. suggested that for easier understanding of structure of huge networks that may be social or technological the network can be subdivided into modules or clusters or set of nodes or subunits. They developed information theoretic foundation which explains modularity concept in networks, identifying the modules which are network composed through optimal compression of network topology. They found node communities which are clustered on the basis of their links. hub versus periphery distinction can also be extracted from network structure with the help of appropriate encoder. They remarked that while abstracting problem of finding pattern in networks and data compression, their proposed information theoretic scheme provides information out of a network structure easily and in large amount [17].

Jiyang Chen [18] et.al. presented datasets in graphical format or in networks. The nodes in the graph are entities and edges are relationships between pairs of entities. The community structure is a cluster of densely connected groups of vertices and has sparser connections between the groups. They presented community mining technique as Max-Min Modularity considering connected pairs and criteria. These are defined by domain experts to find communities and detect communities in networks through hierarchical clustering algorithm. Their scheme gave robustness against noise. Their approach takes domain knowledge into consideration thereby improving the community detection accuracy and maximizes connected node pairs and minimizes unrelated pairs in the same community [18].

REFERENCES

[1]Liaoruo Wang, Tiancheng Lou, Jie Tang and John E. Hopcroft "Detecting Community Kernels in Large Social Networks", 2011

[2] N. Sanyog Choudhary, Himanshu Yadav and Anurag Jain "Message Filtering Techniques in Social Networks over Web Environment- A Survey", IJETAE, 2014

[3] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multi-scale complexity in networks. Nature, 2010

[4] G. Palla, P. Pollner, A. Barabasi, and T. Vicsek. Social group dynamics in networks. Adaptive Networks, 2009.

[5]M. Girvan and M. E. J. Newman. Community structure in social and biological networks. PNAS, 99, 2002.

[6] Rocco Langone, Carlos Alzate and Johan A. K. Suykens "Kernel spectral clustering for community detection in complex Networks", 2011

[7] Jaewon Yang, Julian McAuley and Jure Leskovec "Community Detection in Networks with Node Attributes" ,2013

[8]Jure Leskovec, Kevin J. Lang and Michael W. Mahoney "Empirical Comparison of Algorithms for Network Community Detection", ACM, 2010

[9]Kevin S. Xu, Mark Kliger, and Alfred O. Hero III "Tracking Communities in Dynamic Social Networks", 2010

[10]Nam P. Nguyen, Thang N. Dinh, Ying Xuan and My T. Thai "Adaptive Algorithms for Detecting Community Structure in Dynamic Social Networks", 2009

[11]Andrea Lancichinetti and Santo Fortunato "Community detection algorithms: A comparative analysis", 2009

[12]Nina Mishra, Robert Schreiber, Isabelle Stanton and Robert E. Tarjan "Finding Strongly-Knit Clusters in Social Networks", Internet Mathematics, 2009

[13] Jie Tang, Jimeng Sun, Chi Wang and Zi Yang "Social Influence Analysis in Large-scale Networks", ACM 2009

[14]PhilippS huetz and Amedeo Caflisch " Efficient Modularity Optimization by Multistep Greedy Algorithm and Vertex Mover Refinement", 2008

[15]Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment, An IOP and SISSA, 2008.

[16] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg and Siddharth Suri "Feedback Effects between Similarity and Social Influence in Online Communities", 2008

[17] Martin Rosvall and Carl T. Bergstrom "An information-theoretic framework for resolving community structure in complex networks", PNAS, 2007

[18] Jiyang Chen, Osmar R. Zaiane and Randy Goebel "Detecting Communities in Social Networks using Max-Min Modularity", 2007.