Volume 2, No. 2, February 2011
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

Community Discovery in Mobile Blog

Dr. M. Mohamed Sathik[*1] and A. Abdul Rasheed[2]

[*1]Associate Professor,
Department of Computer Science,
Sadakathullah Appa College,
Tirunelveli, Tamilnadu, India
mmdsadiq@gmail.com[1]

[2]Assistant Professor,
Department of Computer Applications,
SRM Valliammai Engineering College,
Chennai, Tamilnadu, India
profaar@gmail.com[2]

Abstract: Social network gains popularity due to its ease of use, as an application of Web 2.0. The study of networks is an active area of research due to its capability of modelling many real world complex systems. One such interesting property to investigate in any typical network is the community structure which is the division of networks into groups. Discovering communities in a social network environment is graph partitioning problem, which subdivides the entire graph into smaller partitions. Graph partitioning is believed as NP – hard problem, due to its complexity to split the number of vertices. We introduced the method of mutual accessibility to find communities in social networking environments. Existing work presents community discovery from blog posts. In this research, we discovered community structures from blogs which are posted by mobile devices such as mobile phones, specialized devices like personal digital assistants (PDA). We also applied the method called "mutual accessibility" for discovering communities. We presented the results obtained through our method. It is feasible that our method produced accurate identification of community structure in large complex networks.

Keywords: Social Network; Moblog; Graph Clustering; Community; Mutual Accessibility

INTRODUCTION

Social networks gained popularity recently with the advent of sites such as MySpace, Friendster, Orkut, Twitter, Facebook, etc. 133 million blog records indexed by Technorati since 2002 and 900000 blog posts in 24 hours. By June 2008, Technorati tracked blogs in 81 languages and there are 77.7 million unique visitors in the US by August 2008. The number of users participating in these networks is large, e.g., a hundred million in these and growing. Social network represented a graphical representation of people who are connected by relationships, groups connected by any relations, and organizations connected by relations. Social Network Analysis (SNA) provides a spectrum of tools and theoretical approaches for holistic exploration of the interaction patterns among individuals, groups and even organizations. SNA is a field of research that provides a set of tools and theoretical approaches for holistic exploration of the communication and interaction patterns of social systems. Mobile Blog (Moblog) are specialized blog posts hosted by the bloggers who have mobile devices like mobile phones, PDAs etc., These blog posts have uniqueness like the bloggers are registered users. They can host their blogs "on-the-move". Moreover, these blog posts do have limited number of characters, as it is being supported by the mobile devices or PDAs.

A fundamental problem related to these networks is the discovery of clusters or communities. One of the most important research and review questions in social networks is the "identification of communities". A community is a set of real-world entities that form a closely knit group. Communities provide a natural division of graph nodes into densely connected subgroups. There are a lot of methods proposed in the past decade to discover communities and those methods are discussed in the literature survey section of this paper. It is generally considered that the community discovery is a graph clustering problem. Graph clustering is the task of grouping the vertices of the graph into clusters taking into consideration the edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters. That means nodes in intra – cluster vertices are denser than inter – cluster. Community detection in complex networks has attracted a lot of attention in recent years. Communities can be defined as collections of individuals who interact unusually frequently. A community is a densely connected subset of nodes that is only sparsely linked to the remaining network. The identification of communities often reveals the properties, such as related topics or common view points, shared by the members like occupations, social functions, or some other common hobbies like dating. Given a graph G = (V, E), where V is the set of vertices and E the set of edges that determines the connectivity between the nodes. The graph partitioning problem consists on dividing G into k disjoint partitions. The goal is to minimize the number of cuts in the edges of the partition.

Consider the figure shown in Figure 1, which contains three groups of communities. This also shows the interaction level

among the members of intra – community and also the interaction with inter – community members.

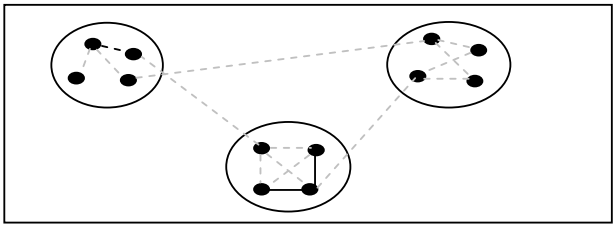

Figure 1. A group of three communities and the interaction among the members

Graph clustering is the process of grouping graph nodes, such that most edges are inside individual clusters, and inter-cluster edges are comparatively few. Though there are some popular classes of clustering algorithms like geometric, hierarchical and partitioning methods, such methods cannot be used for graph clustering. A cluster in a graph is called a community. For a given data set, the goal of clustering is to divide the data set into clusters such that the elements assigned to a particular cluster are similar or connected in some predefined sense. The goal in graph partitioning is to minimize the number of edges that cross from one subgroup of vertices to another, usually posing limits on the number of groups as well as to the relative size of the groups.

We already introduced a method called "mutual accessibility" for discovering communities. In this article, we used our method for a mobile blog dataset.

LITERATURE SURVEY

Community detection in complex networks has attracted a lot of attention in recent years. The researchers are putting their effort by applying different methodologies to discover such communities. In this section, we provide some of the existing methods which are reviewed in the past decades. Through the existing literature, we came to know that no such existing method talks about how one person (vertex) knows the other (vertex). That means there should be a strong tie between the two vertices in the entire graph. This purpose can be solved by using Strongly Connected Components (SCC), as it identifies the paths between any two vertices involved. The communities are formed in such a way that when there is a path from a vertex u to v, then there should also be a path from v to u. Hence, the intermediate vertices can also have the similar kind of relationship, equivalence relationship, to form strong components, and hence communities.

An improved spectral clustering method for discovering communities in social network is presented in [1]. To make full use of the network feature, the core members are used in this method for mining communities. The authors utilized Page Rank method for discovering communities. In this work, the authors proved that their method is better in terms of time and accuracy.

A good survey on various community detection algorithms can be found in [2]. This gives an elaborate description about different algorithms along with the results that are obtained by those algorithms. In this paper, the authors tested several methods against a recently introduced class of benchmark graphs, with heterogeneous distributions of degree and community size and the results produced in the form of charts.

Biologically inspired algorithms are applied for wide variety of problems. Community discovery is no way exempted from this phenomenon. Hence, a genetic algorithmic approach is applied by [3]. The algorithm uses a fitness function able to identify groups of nodes in the network having dense intra – connections, and sparse inter – connections.

A random graph is a graph that is generated by some random process. A random graph is a graph in which properties such as the number of graph vertices, graph edges, and connections between them are determined in some random way. The random graph is defined by the join distribution of the presence or absence of vertices. The inclusion of vertices can be combined to form communities. This method is introduced by [4], as a method of discovering communities in networks. In this paper, the authors used block structures model for the purpose in the context of social sciences, using a Bayesian approach.

Communities are emerging in various types both in good and bad groups. One such ideal way to identify hate group through blogs are done by [5]. The authors proposed a semi-automated approach to analyze virtual communities and to monitor for activities that are potentially harmful to society. The authors used blogs as their data source for this work.

Community discovery is basically a clustering problem, in data mining perception. As inter – cluster members may either be included in one or more clusters, which is so called overlapping of communities. Identifying overlapping of communities is done by [6]. The authors devised a novel algorithm to identify overlapping communities in complex networks by fuzzy c – means clustering approach.

A simple label propagation algorithm for community discovery is done by [7]. The authors used the network structure alone as its guide for the work. This work didn't require any pre-defined objective function or prior information about the communities.

The concept of modularity matrix for community detection is introduced by [8]. In this paper, the authors defined the maximization process that can be written in terms of the eigenspectrum of a matrix, called the modularity matrix, which plays a role in community detection. The algorithms and measures proposed are illustrated with applications to a variety of real-world complex networks.

[9] Showed how community detection can be interpreted as finding the ground state of an infinite range spin glass. In this paper, the community structure of the network is interpreted as the spin configuration that minimizes the energy of the spin glass with the spin states being the community indices.

Random walks has several important advantages like it captures well the community structure in a network, it can be computed efficiently, and it can be used in an agglomerative algorithm to compute efficiently the community structure of a network. This approach for community discovery is used by [10]. The authors proposed a measure of similarities between vertices based on random walks for community discovery.

An extremal optimization method for community discovery was proposed by [11] which is a divisive algorithm for graph partitioning. It optimizes the modularity using a heuristic search based on the extremal optimization EO algorithm. The authors produced the results by taking computer-simulated and real networks and compare them with other approaches.

Community detection using modularity was proposed by [12]. It is an agglomerative hierarchical clustering method. The basic idea of the algorithm was modularity. The author

produced the results by taking various applications to prove the efficiency of the proposed method, as it is faster than other previous algorithms.

Problem decomposition to discover communities was applied by [13]. As per this approach, the network is decomposed into manageable sub networks using a multilevel graph partitioning procedure.

We introduced the method called *mutual accessibility* by using strongly connected components. The description of the method was given in [14]. In this method, the members know each other in the network. The cluster (community) can be constructed only if the members of the cluster known each other by themselves. Our method provides the stability and enhancement of the members of the community, when compared with all other existing methods.

MATERIAL AND METHOD

In this section, first we explain our methodology and then we present the details of the dataset used for community discovery.

Methodology

Suppose a graph G has V vertices and E edges, mathematically represented as G = (V, E). A strongly connected component of a directed graph G is a maximal set of vertices C ⊆ V such that for every pair of vertices u and v, there is a directed path from u to v and a directed path from v to u. A directed graph is called strongly connected if there is a path from each vertex in the graph to every other vertex that is both the nodes are mutually reachable. The strongly connected components (SCC) of a directed graph G = (V, E) are its maximal strongly connected sub graphs. Strong connectedness is an equivalence relation on vertices, and the resulting equivalence classes are called the strongly connected components of the graph. Within a strongly connected component, any vertex can be reached from any other vertex. This strong connectivity provides "mutual accessibility" among the nodes of the graph G. The sub graphs generated through the strongly connected components are the partitioned graphs, also called clusters.

Strong connectedness is an equivalence relation on vertices, and the resulting equivalence classes are called the strongly connected components of the graph. Within a strongly connected component, any vertex can be reached from any other. We can more formally generalize the strongly connected components as follows: Given a graph G = (V, E), where V is a set of vertices (say size n) and E is a set of edges (say size m), the connected components of G are the sets of vertices such that all vertices in each set are mutually connected (reachable by some path), and no two vertices in different sets are connected. Given a strongly connected digraph G, we may form the component digraph G^{SCC} by the following two properties:

- The vertices of G^{SCC} are the strongly connect components of the digraph G.
- There is an edge from v to w in G^{SCC}, if there is an edge from some vertex of component v to some vertex of component w in digraph G.

Algorithms for finding strongly connected components may be used to solve 2 – satisfiability problems. A 2-satisfiability is the problem of determining whether a collection of two - valued variables with constraints on pairs of variables can be assigned values satisfying all the constraints. A 2 – satisfiability instance is unsatisfiable if and only if there is a variable v such that v and its complement are both contained in the same strongly connected component of the implication graph of the instance.

There are two properties of Strongly Connected Components of a directed graph:

1. There should be at least a path from each vertex in the graph to every other vertex
2. There should not be a cycle or loop in the resultant SCC

Tarjan has devised an O(n) algorithm for determining strongly connected components[15]. The algorithm's running time is therefore linear in the number of edges in G (i.e) O(|V| + |E|). The basic idea of the algorithm is to apply a depth-first search (DFS) begins from a start node. The strongly connected components form the subtrees of the search tree, the roots of which are the roots of the strongly connected components. The nodes are placed on a stack in the order in which they are visited. When the search returns from a subtree, the nodes are taken from the stack and it is determined whether each node is the root of a strongly connected component. If a node is the root of a strongly connected component, then it and all of the nodes taken off before it form that strongly connected component.

Fig. 2 is used to explain a digraph and the number of components in it. The vertices of the digraph are numbered 1 through 12. There are four different communities of variable in size for the given digraph. There is a single member community indexed as A, two member community mentioned by C, and three member communities is specified as B and six members community is represented as D. The outline boundaries are used to draw the number of components as communities. The final digraph is also satisfying the properties and 2-satisfiability of SCC.

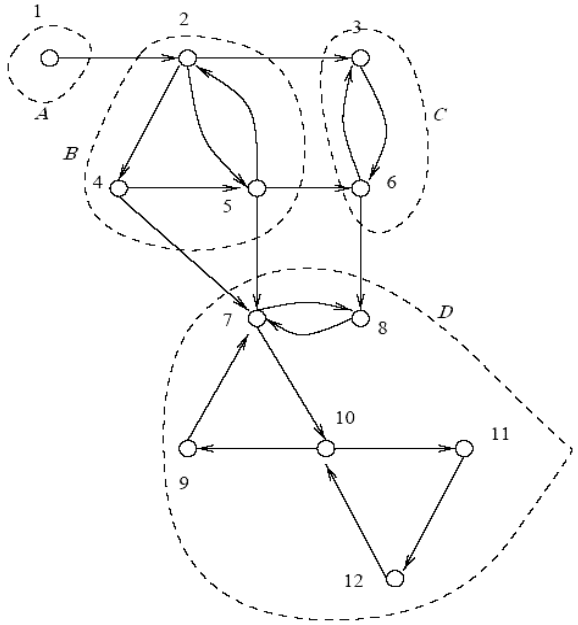

Fig 2: A Sample digraph and its subcomponents as communities

Material

We crawled a moblog website using a web crawler provided by [16]. We extracted the user (blogger) information from the XML content. To preserve the privacy of the bloggers, we used numbering system of bloggers. We considered only the

blog posts which have at least one response for the corresponding blog post. When a blog post has no response, we just removed such blogger from the list. We considered this as a preprocessing step. Suppose for a blog post i there was a response of blog post say j, then we created an edge represented as (i, j). In this way, we created 63592 vertices and 813098 vertices. This is the size of the dataset that we considered further for discovering communities using our method, called "mutual accessibility". We discovered 142 communities (clusters) for our graph dataset. As we already explained, our method provided the strength in such a way that members of the individual clusters are known among themselves. Social Networks are established based on relations and strong connectedness is an equivalence relation. Hence, our method has its uniqueness.

We also computed the average clustering coefficient. Clustering coefficient can be applied to a single node; where as the average clustering coefficient can be applied to an entire network. The communities are formed in such a way that when there is a path from a vertex u to v, then there should also be a path from v to u. If every node in the neighborhood of u is connected to every other node in the neighborhood of u, then the neighborhood of u is complete and will have a clustering coefficient of 1. If no nodes in the neighborhood of u are connected, then the clustering coefficient will be 0.

A clustering coefficient (C) for the whole graph is the average,

$$C = \frac{1}{n} \sum_{v \in G} c_v$$

where n is the number of nodes in the entire graph G and c_v is the number of vertices in a cluster. In this way, we computed the average clustering coefficient of our graph as: C = 0.2168.

Summary of the dataset is provided as in Table I:

Table I. Summary of dataset used for our study

Number of Vertices (Size of the graph)	*63592*
Number of Edges	813098
Number of Discovered Communities	142
Average Clustering Coefficient of the graph	0.2168

Result and discussion

In this article, we discovered communities from mobile blogs (moblogs). We used the method called mutual accessibility introduced by us in our earlier work. We created a moblog dataset from a moblog website, for our study. We also came to know that our method discovered communities so that the members of the clusters are known among themselves. The size of the dataset and the result obtained are provided as a summary in Table I.

Conclusion

Social network is represented as relations that exist among the members or even groups. They gain popularity due to its ease of use as an application of Web 2.0. Blog posts are organized as a tool in different subjects. Blog based websites have tremendous growth in today's scenario. Bloggers too have common platform to express their own idea and freedom to respond for an existing blog post. Community discovery in a complex network is a graph partitioning problem. There were several methods introduced in the past decade. We already introduced a method of knowing the cluster members among each other. We called this method as mutual accessibility. The underlying mathematical concept of our method is through strong connectivity of the graph. In this article, we applied a moblog dataset and the results are also explained in the previous subsection.

REFERENCES

[1] Shuzi Niu, Daling Wang, Shi Feng, Ge yu, 2009, An improved spectral clustering algorithm for community discovery, Ninth Intl. Conf. on Hybrid Intelligent Systmes, Vol. 3, 262-267.

[2] Andrea Lancichinetti, Santo Fortunato, 2009, Community detection algorithms: a comparative analysis, arXiv: 0908.1062vl physics soc-ph.

[3] Clara Pizzuti, 2008, Community detection in social networks with Genetic Algorithms, Proceedings of the 10th annual conference on genetic and evolutionary computation, 1137-1138.

[4] Daudin J. J , Pichard F and Robin S, 2008, A mixture model for random graphs, statistical computing 18, 173-183.

[5] Michael Chaua, Jennifer Xu, 2007, Mining communities and their relationships in blogs: a study of online hate groups, Int. J. human – computer studies 65, 57-70.

[6] Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, 2007, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A374, 483-490.

[7] Raghavan U N, Albert R and Kumara S, 2007, Near linear time algorithm to detect community structures in large – scale networks, Physical Review E76, 036106.

[8] Newman M E J, 2006, Finding community structure using the eigenvectors of matrices, Physical Review E74, 036104.

[9] Richardt J and Bornholdt S, 2006, Statistical mechanics of community detection, Physical Review E74, 016110.

[10] Pascal Pons, Matthieu Latapy, 2005, Computing communities in large networks using random walks, LNCS 3733, 284-293.

[11] Jordi Duch an dAlex Arenas, 2005, Community detection in complex networks using extremal optimization, Physical review E72, 027104.

[12] Newman M E J, 2004, Fast algorithm for detecting community structure in networks, Physical Review E69, 066133.

[13] Narasimhamurthy A, D. Greene, N. hurley and P. Cunningham, 2008, Community finding in large social networks through problem decomposition, 19th Irish conference on Artificial Intelligence and cognitive science (AICS'08).

[14] Dr. M. Mohamed Sathik, A. Abdul Rasheed, 2010, discovering communities in social networks through mutual accessibility, Intl. Jnl on computer science and engineering, vol. 02, no. 04, 1423-1428.

[15] Robert Tarjan, 1972, Depth – first search and linear graph algorithms, SIAM J. Computing, Vol. 1 , No.2, 146-160.

[16] Robert C Miller, Krishna Bharat, "SPHINX: A framework for creating personal, site-specific web crawlers", Proceedings of the seventh international World WideWeb Conference. Printed in Comuter Network and ISDN Systems v.30, pp119-130, 1998.

SHORT BIODATA OF ALL THE AUTHOR

Dr. M. Mohamed Sathik received his Ph.D., in Computer Science from Manonmaniam Sundaranar University, Tirunelveli, INDIA in 2006. He also received M. Phil., in Computer Science, MBA., M. Tech., in Compute Science and Information Technology. He has more than 25 years experience in teaching. He is a recognized supervisor for M. Phil., and Ph.D., in various universities. He published several papers in international journals. He is also a review member in several journals of international repute. He chaired international conferences. His research interest includes virtual reality, data mining and image processing.

Mr. A. Abdul Rasheed is graduated in MCA and M. E., with specialization in Computer Science and Engineering. He has around 15 years experience in teaching. He is pursuing his research in Computer Science and Engineering in Manonmaniam Sundaranar University, Tirunelveli under the guidance of Dr. M. Mohamed Sathik.