ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Survey of Community Detection in Online Social Network

B. Padma Priya[*1], K. Sathiyakumari[2]
[*1]Research Scholar, PSGR Krishnammal College for Women, India
[2]Assistant Professor, PSGR Krishnammal College for Women, India
priyaa.19885@gmail.com

Abstract

In this paper we present a large Scale Community detection and analysis of Facebook, which gathers more than one billion active users in 2012. Characteristics of this online social network have been widely researched over these years. Facebook has affected the social life and activity of people in various ways. One major fact in today's technical world, people are very active users of Online Social Networks. They share every details of their day to day life and are in touch with their loved ones no matter in which part of the world they live. The impact is considerably taken into account as this online Social Network play a very important role in people lives. We study the structural properties of these samples in order to discover their community Structure. Here two Clustering algorithms are used to discover the communities in Complex networks and is compared.

Keywords: Data mining, Complex Networks, Community Mining, Community Detection.

Introduction

The problem of studying the structure of complex networks has been a latest topic in research field in several fields like Social Sciences, Physics and Computer Science. During the recent years the Study of Online Social Network has developed to a higher rate. The role of online social network is to help the people to enhance the connection among them in the context of Internet. The first phenomena are relationship among people in some areas of social network is very strong.

They have a close connection between family, colleagues, friends and so on. The Second Phenomena is Out going connections with other individuals not belonging to any of these categories are happening a lot these days. This effect reflects in the society, which is called Community structure. The Community is defined as a sub structure defined in a network that represents connection among users .For the structure perspective community is represented by a graph corresponding to the communities. From a Scientific point of view the interesting properties or hidden information from a network tends to explore a lot of commercial and scientific applications in the real world.

Two problems can be specified which discovering communities in a network. The first one is when partitioning the vertices into disjoint subset because entity may belong to several different communities, which is called the problem of overlapping. The second one is represented by network in which the individual does not belong to any of the community.

In order to investigate the community structure of real and online social networks there can be done by two possibilities. They are Partitioning Algorithms and overlapping node community detection algorithms.

A Survey of Community Detection

A. *Partitioning Algorithms*

The discovery of Community Structure in a network has been approached in many different ways. Let us consider a network represented by graph G= (V, E) high value for l_s for each discovered community is detected as dense within their structure and weakly coupled among themselves. As the task of maximizing the function Q is NP-Hard several appropriate techniques have been used which has been portioned into m communities the value of network modularity is

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{|E|} - \left(\frac{d_s}{2|E|} \right)^2 \right]$$

Where l_s is the number of vertices belonging to the $s\text{-th}$ of the community d_s the sum of degrees of the vertices in the $s\text{-th}$ community. Thus high value of Q implies. The most popular partition algorithm used to detect the community structure in a structure is network

ISSN: 2277-9655
Impact Factor: 1.852

modularity proposed by (Girvan and Newman 2002, Newman and Girvan 2004) called the Girvan Newman Algorithm.

The hierarchal clustering method is based on assigning a weight for every edge and placing these edges into an initially empty network starting from edges with strong weights and processing towards the weakest ones. The edges with greatest weight are the central ones. The Girvan Newman Algorithm works the opposite way. Instead of trying to construct a measure that tells us which edges are the most central to communities, it focuses on these edges that are least central.

The Girvan–Newman algorithm extends this definition to the case of edges, defining the centrality of an edge as the number of shortest paths between pairs of nodes that run along it. If there is more than one shortest path between a pair of nodes, each path is assigned equal weight such that the total weight of all of the paths is equal to unity. By removing these edges, the groups are separated from one another and so the underlying community structure of the network is revealed.

The algorithm's steps for community detection are summarized below

1. The betweenness of all existing edges in the network is calculated first.
2. The edge with the highest betweenness is removed.
3. The centrality of all edges affected by the removal is recalculated.
4. Steps 2 and 3 are repeated until no edges remain. First it calculates the edge betweenness $B(e)$ of the given edge e in a network S is defined as

$$B(e) = \sum_{n_i \in S} \sum_{n_l \in S} \frac{np_e(n_i, n_l)}{np(n_i, n_l)}$$

Where n_i and n_l are the vertices of S $np(n_i, n_l)$ is the number of shortest path between n_i and n_l and $np_e(n_i, n_l)$ is the number of shortest path between n_i and n_l containing e.In Girvan Newman algorithm it is possible to maximize the value of Q deleting edges with the high value of betweenness. From this the algorithm ranks all the edges with respect to their betweenness and calculates the Q value and iterates the process until there is an increase in Q value. At each iteration the component S identifies a Community. Its cost is $O(n^3)$ being n the number of vertices in S. So it is suitable for large-scale networks. For instance, if two communities are connected by more than one edge, then there is no guarantee that all of these edges will have high centrality. According to the method, we know that at least one of them will have, but nothing more than that is known. By recalculating the centrality after the removal of each edge it is ensured that at least one of the remaining edges between two communities will always have a high value.

In this method of detecting the community structure the problem of finding the partition of the network that maximizes the network modularity value is not computationally affordable because its NP is hard. Number of improved versions for this approach has been provided.

B. Overlapping Node detection Algorithm

The discovery of Community Structure in a network has been approached in many different ways. Let us consider a network represented by graph G= (V, E) which has been portioned into m communities the value of network modularity is

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{|E|} - \left(\frac{d_s}{2|E|} \right)^2 \right]$$

Where l_s is the number of vertices belonging to the $s\text{-th}$ of the community d_s the sum of degrees of the vertices in the $s\text{-th}$ community. Thus high value of Q implies high value for l_s for each discovered community is detected as dense within their structure and weakly coupled among themselves. As the task of maximizing the function Q is NP-Hard several appropriate techniques have been used. The most popular partition algorithm used to detect the community structure in a structure is network modularity proposed by (Girvan and Newman 2002, Newman and Girvan 2004) called the Girvan Newman Algorithm. The hierarchal clustering method is based on assigning a weight for every edge and placing these edges into an initially empty network starting from edges with strong weights and processing towards the weakest ones. The edges with greatest weight are the central ones.

The Girvan Newman Algorithm works the opposite way. Instead of finding a measure that tells us which edges are the most central to communities, it finds the edges that are least central. The Girvan–Newman algorithm extends this definition to the case of edges, defining the centrality of an edge as the number of shortest paths between pairs of nodes that run along it. If there is more than one shortest path between a pair of nodes, each path is assigned equal weight such that the total weight of all of the paths is equal to unity. By removing these edges, the groups are separated from one another and so the underlying community structure of the network is revealed.

ISSN: 2277-9655
Impact Factor: 1.852

The algorithm's steps for community detection are summarized below

1. The betweenness of all existing edges in the network is calculated first.
2. The edge with the highest betweenness is removed.
3. The centrality of all edges affected by the removal is recalculated.
4. Steps 2 and 3 are repeated until no edges remain.

First it calculates the edge betweenness $B(e)$ of the given edge e in a network S is defined as

$$B(e) = \sum_{n_i \in S} \sum_{n_l \in S} \frac{np_e(n_i, n_l)}{np(n_i, n_l)}$$

Where n_i and n_l are the vertices of S $np(n_i, n_l)$ is the number of shortest path between n_i and n_l and $np_e(n_i, n_l)$ is the number of shortest path between n_i and n_l containing e.In Girvan Newman algorithm it is possible to maximize the value of Q deleting edges with the high value of betweenness. From this the algorithm ranks all the edges with respect to their betweenness and calculates the Q value and iterates the process until there is an increase in Q value. At each iteration the component S identifies a Community. Its cost is $O(n^3)$ being n the number of vertices in S. So it is suitable for large scale networks. For instance, if two communities are connected by more than one edge, then there is no guarantee that all of these edges will have high centrality. According to the method, we know that at least one of them will have, but nothing more than that is known. By recalculating the centrality after the removal of each edge it is ensured that at least one of the remaining edges between two communities will always have a high value.

In this method of detecting the community structure the problem of finding the partition of the network that maximizes the network modularity value is not computationally affordable because its NP is hard. Number of improved versions for this approach has been provided.

The second problem we face in discovering the community structure is finding overlapping nodes belonging to different community at the same time. The first approach has been provided by (Palla et., 2005) .An interesting approach was proposed by (Gregory 2007).Some novel techniques has been proposed lately (McDaid and Hurley, 2012; Lee et el., 2010).One of the popular method for detecting the overlapping community structure is the clique percolation method. The clique percolation method (CPM) is based on the assumption that a community consists of overlapping sets of fully connected sub graphs and detects communities by searching for adjacent cliques. It begins by identifying all cliques of size k in a network. Once these have been identified, a new graph is constructed such that each vertex represents one of these k-cliques.

Two nodes are connected if the k cliques that represent them share k−1 members. Connected components in the new graph identify which cliques compose the communities. Since a vertex can be in multiple k-cliques simultaneously, overlap between communities is possible. CPM is suitable for networks with dense connected parts. The small values of k have been shown to give good results [Palla et al. 2005; Lancichinetti and Fortunato 2009; Gregory 2010]. CFinder1 is the implementation of CPM, whose time complexity is polynomial in many applications [Palla et al.2005]. However, it also fails to terminate in many large social networks. CPMw [Farkas et al. 2007] introduces a sub graph intensity threshold for weighted networks. Only k-cliques with intensity larger than a fixed threshold are included into a community. Despite their simple concept , one may argue that CPM-like algorithms are more like pattern matching rather than finding communities since they aim to find specific, localized structure in a network.

Contribution to the state of art

Emily Ferrera has done a analysis on a large social network using various different techniques. First data has been collected from face book social network. Once data is analyzed community is detected using quantitative and qualitative perspective.

Data Collection

Emily Ferrara has used a special architecture for collecting a face book details. The architecture of the designed sampling platform can be schematized as figure 1.He devised a java cross platform data mining agent which implements the logic of a crawler based on Apache HTTO Library as interface for transferring data through the web.

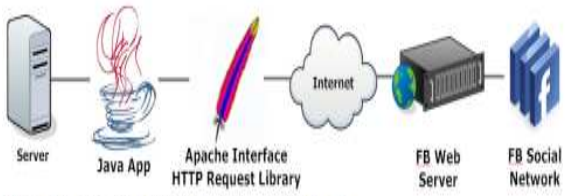

Figure 1 The data collection platform architecture

ISSN: 2277-9655
Impact Factor: 1.852

Figure 2 The logic of the Facebook crawler

The above figure describes the logic of the mining agent. For mining Emilio has used two types of methodology is used .They are Breath first search method and Uniform Sampling Methodology. The first sampling methodology has been implemented as a breadth-first-search (BFS), an uninformed traversal algorithm with the goal of visiting a graph. Starting from a seed node, it explores its neighborhood then, for each neighbor, it visits its unexplored neighbors, and so on, until the whole graph is visited This sampling technique has several advantages, such as the ease of implementation and the efficiency. For these reasons it has been adopted in a variety of OSNs mining studies. The second sampling methodology that has been chosen is a rejection-based sampling technique, namely Uniform sampling. The main advantage of this technique is that it is unbiased for construction, at least in its formulation for the Face book network. Details about its definition are provided by (Gjoka et al., 2010).The process consists of the generation of an arbitrary number of user-IDs, randomly distributed in the domain of assignment of the Face book user-ID system.

Community Structure discovery

The detection of community structure inside a large structure is a complex work. Community detection algorithms such those originally presented by (Girvan and Newman, 2002; Newman and Girvan, 2004) are not viable solutions, respectively because too expensive for the large-scale of the Facebook sample we gathered, or because they require a priori knowledge. Emilo has adopted two fast and efficient cient optimized algorithms, whose performance are the best to date (Label Propagation Algorithm), presented by (Raghavan et al., 2007), and FNCA (Fast Network Community Algorithm), more recently described by (Jin et al.,2009), have been adopted to detect communities from the collected samples of the network.

C. Label Propagation Algorithm

LPA (Label Propagation Algorithm) (Raghavan et al., 2007) is a linear time algorithm for community detection. LPA uses only the network structure as its guide, is used for large-scale networks. Its functioning is reported as described in (Raghavan et al., 2007)

The first step is to initialize each vertex is given a unique label. The second step is that each vertex updates its label with the onethat is used by the greatest number of neighbors. If more than one label is used by the same maximum number of neighbors, one is chosen randomly. After many iterations, the same label is associated with all the members of a community. The third step is the Vertices labeled alike are added to one community.

After further the author has found that this result is too optimistic between quality of results and amount of time required for computation.

D. Fast Network Community Algorithm

The Second efficient algorithm used is Fast Network Community Algorithm (Jin et al., 2009). FNCA is an optimization algorithm which aims to maximize the value of the network modularity function, in order to detect the community structure of a given network. The network modularity function has been introduced by (Newman and Girvan, 2004) and has been largely adopted in the last few years by the scientist. Given the undirected un weighted network G = (V,E) let $i \in V$ be a vertex belonging to the community r(i) denoted by $c_r(i)$ the network modularity is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left[\left(A_{ij} - \frac{k_i k_j}{2m}\right) \times \delta\left(r(i), r(j)\right)\right]$$

Where A_{ij} is the element of the elementary matrix $A = (A_{ij})_{n\times n}$ representing the network $A_{ij} = 1$if i and j are tied by an edge $A_{ij} = 0$.The function $\delta(u, v)$ namely kronecker delta .The value k_i represents represents the the degree of the vertex I defined as $k_i = \sum A_{ij}$ while m is the maximum no of edgesin the network defined as

.

$$m = \frac{1}{2}\sum_{ij} A_{ij}$$

while the above equation can be represented as

$$Q = \frac{1}{2m}\sum_{i} f_i, \qquad f_i = \sum_{j\in c_{r(i)}}\left(A_{ij} - \frac{k_i k_j}{2m}\right)$$

where function f represents the difference between actual and expected no of edges which fall within communities thus indicating how strong the community structure is. Thus, each node needs to calculate its f function only for the labels of its neighbors, instead of for all the nodes of the network. Moreover, authors put into evidence that, if the labels of neighbors of one node do not change at last iteration, the label of that node is less likely to change in the current iteration. Like LPA the accuracy of the result is similar. As for the community detection algorithm, we found that the LPA method has been proved to be a good choice among the heuristic methods based on local information in order to discover the underlying

ISSN: 2277-9655
Impact Factor: 1.852

community structure of a large network. Results compared against another well-known similar Community Structure Discovery method, called FNCA, seems to be slightly better.

Conclusion

In this paper the data collection methods for facebook and their community detection techniques are discussed.Thus by the evidence put together by the authors we conclude that Fast network community algorithm better than any other algorithm as for now. The performance of this technique in the context of the community structure discovery on large scale. Online Social Networks such as Facebook deserves further investigation.

References

[1] Clauset, A., 2005. Finding local community structure in networks. Physical Review E72 (2), 026132.

[2] Clauset, A., Newman, M., Moore, C., 2004. Finding community structure in very large networks. Physical Review E 70 (6), 066111.

[3] Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., Provetti, A., 2011. Crawling facebook for social network analysis purposes. In: Proceedings of the International Conference on Web Intelligence, Mining And Semantics. pp. 52:1-52:8

[4] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Crawling facebook for social network analysis purposes. In Proc. International Conference On Web Intelligence, Mining And Semantics, pages 52:1-52:8, 25-27 May 2011.

[5] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Generalized louvain method for community detection in large networks. In Proc. 11th International Conference On Intelligent Systems Design And Applications, pages 88{93. IEEE, 2011.

[6] Duch, J., Arenas, A., 2005. Community detection in complex networks using extremal optimization. Physical Review E 72 (2), 027104.

[7] Girvan, M., Newman, M., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99 (12), 7821.

[8] Gjoka, M., Kurant, M., Butts, C., Markopoulou, A., 2010. Walking in facebook: a case study of unbiased sampling of osns. In: Proceedings of the 29th Conference on Information Communications. IEEE, pp. 2498-2506.

[9] Gregory, S., 2007. An algorithm to _nd overlapping community structure in networks. Knowledge Discovery in Databases: PKDD 2007, 91-102.

[10] Lancichinetti, A., Fortunato, S., Kert_esz, J., 2009. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11, 033015.

[11] McDaid, A., Hurley, N., 2010. Detecting highly overlapping communities with model- based overlapping seed expansion. In: 2010 International Conference on Advances in Social Networks Analysis and Mining. IEEE, pp. 112-119.

[12] Newman, M., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical review E 69 (2), 026113.

[13] Raghavan, U., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76 (3), 036106.

[14] Shah, D., Zaman, T., 2010. Community detection in networks: The leader-follower algorithm. In: Proceedings of the Workshop on Networks Across Disciplines: Theory and Applications. pp. 1-8