DOI 10.29042/2018-4077-4081

Helix Vol. 8(5): 4077- 4081

Network Based Community Detection By Using Bisecting Hierarchical Clustering

[1]Miss Snehal Prakash Mahajan, [2]Prof Abhijeet R. Raipurkar
Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur-440013, Phone: (91)-(712) - 2580011
Email: mahajansp@rknec.edu, raipurkarar@rknec.edu

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract

Social media is the platform where human interacts with each other, it is becoming part of our life. One of the task of detecting communities in social networks are real time networks that can be classified by using clustering techniques to extract and find hidden communities from the network. For example Facebook, the different communities can be extracted like people belonging to different interests and communities and this information can be used for marketing purposes. Nowadays, studies involving this clustering processes are basically composed of modularity maximization. In this paper the author proposes a bisecting hierarchical clustering based on a measure known as inter and inter cluster similarities to detect communities within the networks. The experiments indicates a successful performance.

Keywords: Social Networks, Semantic Measures, Clustering.

Introduction

Detection of communities in social sites or networks aims at the finding partitions of a graph with clusters with "similar" data points. It is the method that involves detecting communities in social networks in order to explain similar behaviors among groups of individuals [10]. Hence, the representation of social networks by using graphs that enables their analysis by using graph clustering methods. Apart from the sociology, there are also some other areas where the community detection problem might be suitable, for example, biology and many various areas [9].

To analyse the structure of the original systems, often the corresponding network structure is studied using groups of nodes having more intra group and less inter group edges. Such groups exist in most real world networks and influence the behaviour of the underlying system. The community detection is an important problem and it has the potential to solve many real world problems. Most of the existing community detection algorithms lack the ability to detect accurate community boundaries if the difference between the internal and the external node degree does not exceed a detect ability threshold. Community detection is the method of discovering groups in a network where different groups possesses different properties.

Need of community detection in social media:-

• Human beings are social.

•Easy-to-use social media allows people to extend their social life in unprecedented ways.

•Find difficult to meet or contact friends in the physical world, but much easier to find friend online with similar interests.

•Interactions between nodes can help determine communities.

Literature Review

In the complex social networks, community detection is performed by looking out for the nodes which are similar to each other and keeping those nodes in same community. When the nodes of a network, belongs to the same community, can be arranged to form a group, then that network is said to have a community structure. Community Structure is quite common in real-time networks. The community detection problem has many wide-spread applications and hence proven to be very important. The main advantage of community detection is accessing the information from diverse sources and clusters. A community structure consists of members with similar interests. Detection of communities makes exchanging or offering information easier because members of same community often have similar tastes.

Louvain Modularity:

This method is a greedy method that extracts communities from large networks & attempts to optimize the "modularity" of a partition of the network. It performs two steps. First, it finds "small" communities by optimizing modularity in a locally. Second, it performs aggregation of nodes of the similar community and builds a new network whose nodes are the communities.

Girvan and Newman Algorithm:

This method[7] is popular because it marks the beginning of a new era in community detection field. It identifies edges present in a network that lie between communities and then by removing them, leaving behind just the communities themselves.

Infomap Algorithm:

In this method[3] the neighboring nodes are joined into modules, which are joined into super modules and so on. First, every single node is assigned to its own module. After this, in random sequential order,

each node is moved to the neighboring module this results in the largest decrease in the map equation. Suppose if no move results in a decrease of the map equation, then the node remains in its original module. This procedure is repeated each time in a new random sequential order, until no move generates a decrease in the map equation.

Clique guided community detection:

This is a new approach that is developed for fast and efficient community detection. Clique guided community detection consists of two phases. In the first phase, the framework finds the disjoint cliques. For the second phase, the cliques from the first phase are used to guide the process of merging of individual vertices until a good quality result is reached.

Graph Partition method:

This method[2] is based on min-max clustering principle which was proposed by Ding and Zha et al. Its principle states that the similairity between two sub graphs is minimized, while the similarity within each sub graph is maximized. Luo and Wang et al. propose a framework that identifies modules within a biological network. Networks are firstly divided into sub networks and the identification of these modules is based on their topology.

DBSCAN Algorithm:

It is one of the most efficient community detection method is using this algorithm [6] which is most effective unsupervised clustering algorithm. The Database SCAN algorithm successfully identifies and classifies clusters in the very large spatial data (large range) sets by checking at the local density of database elements, using only one input parameter. Although, the user also gets a suggestion on which parameter value that it would be suitable.

MCL(a cluster algorithm for graphs) :

The algorithm[8] is responsible for simulating flow using two simple algebraic operations on matrices. Presently, there are no high-level procedural instructions for that are used for the process of assembling of groups, joining the groups, or splitting of the groups - the structure of the cluster is bootstrapped via a flow process that is inherently affected by any cluster structure present.

Proposed Methodology

The proposed method includes five steps, they are as follows:

I. Collection of data sets from Social sites.

Collection of dataset from Social sites may include online communities. For example chat applications, apps that provides online services like zomato, amazon etc.

II. Application of hierarchical clustering.

Here we have used divisive clustering where the whole data set is considered as one cluster.

Application of Bisecting hierarchical clustering technique.

- In this module, the data is divided into the n number of clusters one by one.
- One by one clusters are formed to get more intra clusters.
- More numbers of intra cluster will be used to find Communities.

III. Evaluating Statistical measures on clustered data. Here, the similarity between two clusters is measured by using the count of tf (term frequency) for each term in document.

IV. Processing statistical measures to divide in sub communities.

V. Evaluation of results and optimization based on results requirements.

As shown in above steps, we will perform step by step approach.

Previous Clustering methods used k means clustering approach to divide clusters and then after dividing the different communities were detected. We observe that such approach may fail to resolve communities of smaller sizes in networks with communities of significant size variations. We will use k_means clustering to generate clusters, but here the initial value of k, is k = 2. That is a new approach which we will perform.

In this paper, the new approach is introduced i.e. Bisecting Hierarchical Clustering. We will use this approach because it gives us a better understanding of the text and help us to gain details of the clusters. The text based documents are used to extract communities from the social networks.

Algorithm is given below:

ALGORITHM : BISECTING HIERARCHICAL CLUSTERING

INPUT: Data Set (Text files).

OUTPUT : K Clusters

STEP 1 : Select Data Set

STEP 2 : Divide data into two clusters. STEP 3 : REPEAT

STEP 4 : FOR i=1 to n

STEP 5 : IF C1 > C2 & S>T

Divide C1 into two clusters. STEP 6 : OR IF C2 > C1 & S>T

Divide C2 into two clusters.

STEP 7 : END IF STEP 8 : END FOR

STEP 9 : Add two clusters from bisection. STEP 10 : Until K cluster is formed.

Nomenclature: C1- Cluster one, C2- Cluster two, n- number of iterations, S - Similarity metrics, T- Threshold value.

This Algorithm is applied on the dataset directly to gain the insight of clusters.

Bisecting Hierarchical Algorithm

STEP 1: First step is to select relevant database (for this project text file is used). The data set should contain number of text files.

STEP 2: By using Bisecting Hierarchical Clustering the dataset is divided into two clusters, C1 and C2. K_means Algorithm that calculates centroid and then grouping nearest neighbours to their nearest centroid. (in this algorithm k = 2).

STEP 3: After getting two clusters C1 and C2, it will check

IF C1 > C2 & S>T

Divide C1 into two clusters. ELSE IF C2 > C1 & S>T Divide C2 into two clusters.

STEP 4: Process is repeated until desired number of clusters is reached.

C1 and C2 are clusters that we get after dividing dataset into two clusters. If the cluster C1 is larger than C2 and the similarity metrics exceeds threshold value T, then the cluster C1 is further divided into two sub clusters to get C3 and C4. Similarly for, If C2 > C1 & S > T.

Threshold, **T** is the Threshold value that defines the maximum similarity allowed (given by user). Similarity, **S** is the cluster similarity measure

IF S > T, Divide into two clusters.

Flowchart

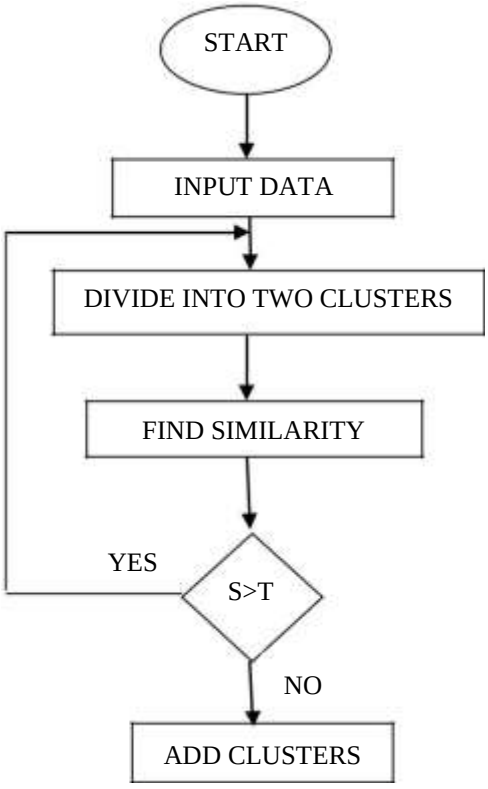

Fig 1. Detecting Communities by using Bisecting Hierarchical Algorithm

The goal of this technique is to detect hidden communities from the social network and classify them for future use.

This uses text files for clustering by applying clustering on text. For this, the tools that are used are

1. Eclipse 3.4
2. Programming Language - java 1.7

Dataset

Dataset used in this project is extracted from Reddit Pizza Requests. This dataset contains a collection of 5671 textual requests for pizza from Reddit community "Random Acts of Pizza" together with their outcome (successful/unsuccessful) and meta-data. All requests ask for same thing, a free pizza. The outcome of each request whether its author received a pizza or not is known. Meta data includes information such as time of the request, activity of the requester, community - age of the requester etc.

Experiments and Results

To perform the task of Social network clustering and classifying into communities. We have used online social network that is "Reddit".

Let us study the experiment performed and its outputs.

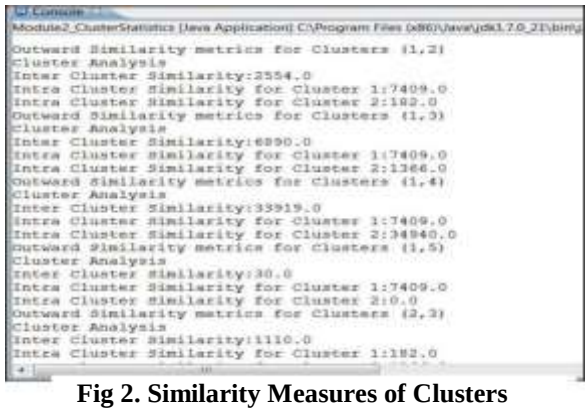

Fig 2. Similarity Measures of Clusters

In Fig 2, similarity measures of cluster is calculated.

for example,

No. of clusters formed = 5

then each cluster is compared to other cluster i.e (1,1), (1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5), (4,5).

After comparison, inter cluster and intra cluster similarities are calculates for each comparison.

Here , in this case, let us suppose the given value for threshold T, is T = 12000, that means the maximum inter cluster similarity is 12000, if any cluster C, exceeds this threshold value , then it will bisect that cluster C in two sub clusters naming C1 and C2.

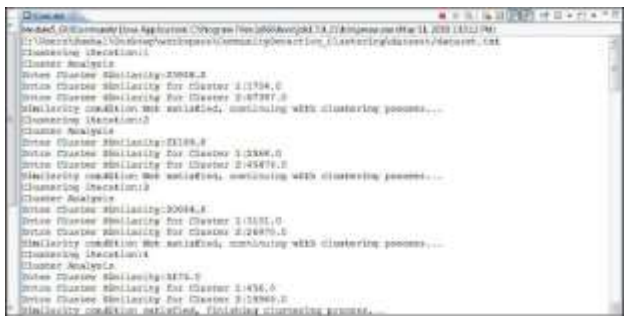

Fig 3: Bisecting Hierarchical Clustering

Fig 3, Bisecting hierarchical algorithm is applied on the selected data set (dataset can be selected manually).

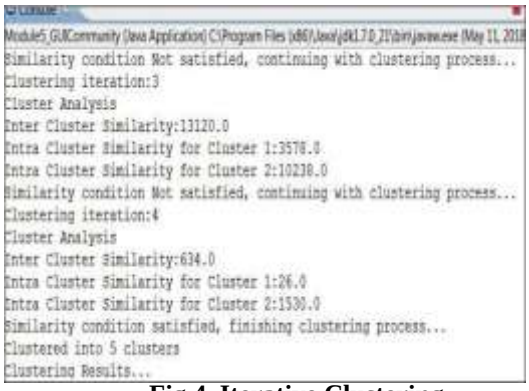

Fig 4. Iterative Clustering

Fig 4, It checks whether similarity S, exceeds threshold T

i.e. **S > T**

if yes then it will perform iterative clustering till the desired result is reached. Here , in this case we have selected the maximum numbers of clusters to be formed is n = 6.

Hence, the value of n is manually given by the user according to requirements.

Here, maximum n = 6, clusters can be created,

& the loop is terminated.

As it is iterative clustering, we need to give where should it stop.

(Otherwise, if there is no specific value for n then it will create infinite loop)

Fig 5. Final Clusters and Intra Cluster Similarities

Fig 5 shows the final number of cluster generated with their intra cluster similarity.

It also shows time needed for clustering is 13029 ms(milliseconds).

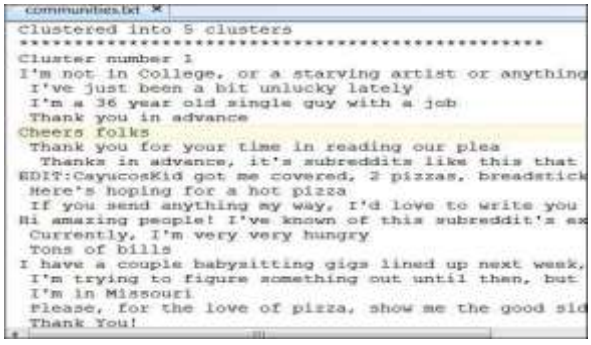

Fig 6. Community.txt Output file

Fig 6 shows the output file "communities.txt", a Text file with clustering results is created automatically.

The summary of communities shows us uniquely present data inside the generated clusters. It will remove similar texts from the clusters, this file is updated automatically.

Fig 7. Community_summary.txt Output file

Fig 7 shows that the summary of texts is stored automatically in the file named "Community_summary.txt " file.

S. no	Thresh-old (given)	Max no of cluster	No. of Cluster formed	Time (ms)
1	12000	6	5	13029
2	120	8	6	12050
3	150	8	7	13293
4	1500	6	4	12285
5	120	6	6	9789

TABLE I: RESULTS

As we can observe the above results, it can be said that the previous clustering techniques used for cluster classification and community detection failed to gain inter and intra cluster knowledge. To deal with cluster details that is, inter and intra cluster similarities this Bisecting Hierarchical Algorithm successfully finds the inter and intra cluster similarities and gained details of clusters that are formed. These measures can be used in many ways according to the trends.

Conclusion

This algorithm successfully classifies clusters into different community groups. Previous algorithms failed to gain knowledge intra cluster details, to overcome this problem this technique can be used to gain detailed intra cluster knowledge. For this project, we have taken data from Social Network Site Reddit pizza services which gives online food services and we successfully classified the different communities based on their texts, comments, orders etc.

And as a result, we got different communities of the users.

Acknowledgement
We would like to thank "Reddit pizza requests" for permits to use and revise the dataset. Any recommendations, results, conclusions or views presented in this paper are those of author(s) and do not express the views of the Reddit online service.

References

[1] Arif Mahmood and Michael Small, Senior Member, IEEE "Subspace Based Network Community Detection Using Sparse Linear Coding", March 2016.
[2] Apeksha P. Naik & SachinBojewar, "A Survey paper on Techniques used for Community Detection in Social Networks," International Conference On Emanations in Modern Technology and Engineering (ICEMTE-2017).
[3] "Infomap-community-detection" http://www.mapequation.org/code. html.
[4] Mehjabin Khatoon, W. Aisha Banu, "A Survey of Community Detection Methods in Social Networks",I.J.Education and Management Engineering, 2015, 1, 8-18.
[5] LOuvian method "http://arxiv.org/abs/0803.0476".
[6] Yomna M. El Barawy, Ramadan F. Mohamedt and Neveen I. Ghali, "Improving Social Network Community Detection Using DBSCAN Algorithm", Computer Applications & Research (WSCAR), 2014 World Symposium, 2014 IEEE.
[7] "Girvan-Newman Method" https://arxiv.org/pdf/0906.0612v2.pdf
[8] https://micans.org/mcl/index.html ? sec_characteristics. "MCL Algorthm".
[9] Community detection in networks via a spectral heuristic based on the clustering coefficient Mariá C.V. Nascimento↤ Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, Rua Talim, 330 - Vila Nair - São José dos Campos/SP CEP: 12231-280, Brazil.
[10] M. Girvan, M. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences 99 (2002) 7821–7826.
[11] L. Adamic and N. Glance, "The political blogosphere and the 2004 US election: Divided they blog," in Proc. 3rd Int. Workshop Link Discovery Link Discovery, 2005, vol. 411, pp. 36– 43.
[12] Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," Nature, vol. 466, no. 7307, pp. 761–764, 2010.
[13] A. L. Barabasi. Network Science (online available). 2012.
[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J. Statist. Mech.: Theory. Exp., vol. 2008, no. 10, p. P10008, 2008.
[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Found. Trends Mach. Learn.

