Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 2, Number 13; April-June, 2015 pp. 1100-1103

http://www.krishisanskriti.org/jbaer.html

Application of Community Detection Algorithms of Social Networks in E-Governance

Shreekanth Prabhu M[1] and Kavithakala C[2]

[1,2]PES Institute of Technology
E-mail: [1]shreekanthpm@gmail.com, [2]kalakavi25@gmail.com

Abstract—*Social networks are one great example of the networks where entities refer to individuals and relationships among entities refer to friendships, collaborations. In addition, social media such as Facebook, twitter have also become very popular as they facilitate instant communication on a real-time basis. In this work we are looking at using network modeling to solve E-Governance problems. One of the key challenges in Governance is to manage duplication and get relatively distinct set of people and communities, when data is distributed in silos. Accordingly we are exploring application of community detection algorithms such as Girvan Newman and Label propagation algorithms to E-Governance Data. Data related to ration card and voter id which are used for identification of citizens in our country is collected and modeled as networks. The network represented in the form of graph is analyzed using Gephi which is one of the Social Network Analysis Tool.*

Keywords: *Social Networks, E-Governance, Community Detection.*

1. INTRODUCTION

In this era people begin their daily life activity with the use of social media such as facebook for communication, twitter for twittering, Linkedin for professional communication. In order to represent the relations and analyze the activities involved in these social media, social networks came into existence. Social networks is the graph G(V, E) in which vertices V represent nodes and edges E represent relationships, affiliations, collaborations, etc.., between nodes. The important problem in social networks is community detection where the number of edges within the community is more than the number of edges between the communities. Large number of algorithms has been developed to detect communities in the graph [1].

In our work we are trying to apply the community detection algorithms of social networks to E-governance data. Girvan-Newman a divisive clustering based community detection algorithm[2] and Label Propagation algorithm [3] based on labels of the nodes of neighbors are used for the analysis of E-governance data.

In section 2 we explain motivation to carry out this research. In section 3 we explain related work carried out in clustering, E-Governance, Gephi tool and Dataset. In section 4 we describe the algorithms used in our work. In section 5 we give details of implementation of these algorithms in gephi as a plugin. Section 6 explains about future works and section 7 deals with conclusion of our work.

2. MOTIVATION

Most E-Governance solutions currently use prevalent multi-tier architecture with relational model for data. One of the key challenges in Governance is to manage duplication and get relatively distinct set of people and communities, especially when data is distributed in silos. So we have applied community detection algorithms of social networks to detect communities for the relational data by converting the relational data to network data. Because of various features available in plugin we are also able to detect duplicates which ushers the bank agencies in providing loans to the ones who have already taken loan, also helps in identifying the duplicate beneficiaries of any schemes provided by government.

3. RELATED WORK

3.1 Clustering

Network clustering is where the graph is clustered such that number of links between the vertices in the graph is more than the links between the clusters, Initial techniques developed for this type of clustering was Min-max cut[4] and Normalized cut[5] method. The other clustering techniques are agglomerative Hierarchical and divisive clustering

Agglomerative hierarchical clustering is where the similarity for all the vertices is determined, then the vertices with highest similarity are merged to form clusters until there are no more vertices to merge. Alternatively the progression of the algorithm results in the formation of dendogram [1].

Divisive clustering is where the edges with lowest similarity is removed resulting in smaller clusters in the graph[1]. In Girvan Newman clustering instead of using the edges with low similarity we use the edge betweenness to form the clusters. Edge betweenness is the measure that favors more edges

between the nodes within the clusters and less favors the edges between the clusters.

3.2 E-Governance

Services from government to its citizens have been computerized which is called E-governance. It has been described along the following four pillars [6]: -

E-Services: Efficient provision of G2C(Government to Citizen) services through automation and business process re-engineering.

E=Participation: Participation of Citizens in the process of Governance through suggestions, complaints and inputs for policy formulation.

E-Democracy: Using electronic means such as E-voting, online petitions to further the democracy.

E-Management: Management of back-end processes of Governance using automation.

Ours is noval idea where we are trying to use the community detection techniques used for social network analysis to analyze E-Governance data.

3.3 Gephi Tool

Gephi is open source software used for analysis and visualization of social networks which is developed in java on Netbeans platform. So far it was used for the analysis of large networks of twitter, Facebook, biological networks, collaboration networks, affiliation networks. But in our work we are using this tool to analyze E-Governance data. [6,8]

3.4 Dataset

As part of E-Governance dataset we have collected around 3900 citizens ration card details of Karnataka state and we have generated synthetic data of 100 voterid details for karnataka state [8, 9].

4. ALGORITHMS

The two algorithms we have used for analysis purpose are: - Girvan Newman algorithm and Label propagation Algorithm.

4.1 Girvan-Newman algorithm [2]

In this algorithm edge betweenness is determined for all edges then the edge with highest edge betweenness is removed to form groups. In the next step again the edge betweenness is recalculated and process continues to form communities in the graph.

The algorithm described in [2] is as follows:

1. Calculate betweenness scores for all edges in the Network.
2. Find the edge with the highest score and remove it

From the network.

3. Recalculate betweenness for all remaining edges.
4. Repeat from step 2.

4.2 Label propagation algorithm [3]

The main idea behind the label propagation algorithm is the following. Suppose that a node x has neighbors x1, x2, ..., xk and that each neighbor carries a label denoting the community to which they belong to. Then x determines its community based on the labels of its neighbors.

The algorithm described in [3] is as follows:

1. To initialize, every vertex is given a unique label.

2. Then, repeatedly, each vertex x updates its label by replacing it by the label used by the greatest number of neighbors. If more than one label is used by the same maximum number of neighbors, one of them is chosen randomly. After several iterations, the same label tends to become associated with all members of a community.

3. All vertices with the same label are added to one community.

We have analyzed our data using the above discussed two algorithms

5. IMPLEMENTATION AND RESULTS

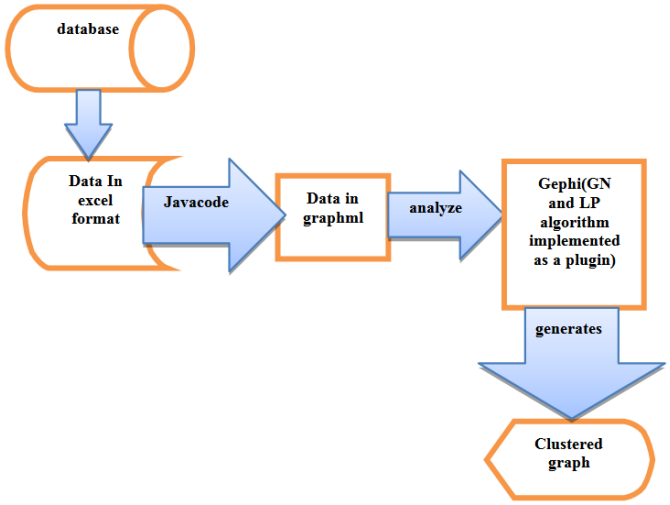

Fig. 1: Data analysis flow diagram

Fig. 1. explains the implementation of our work which works as follows- Dataset collected is stored in excel. It is then converted to graph using java code to generate Graphml file[12]. The graphml file is then loaded to gephi tool to generate the clustered graph using Girvan Newman and Label propagation algorithm in java on Netbeans platform.

We have collected two data sets of proof of Identifications used in India which are ration card dataset having attributes of Ration Card Number, Name, Address, Card type whether APL or BPL, Fair price owner name, taluk, district, state and Voterid datasets having attributes of voterid number, name, part number, AC Number, AC Name, polling station name, age, gender, district, taluk, state.

The two data sets are analyzed as separate graph in which node may be a taluk, state, district, fair price owner and citizen node with attributes of details of ration card for Ration card graph. The graph for voter card has nodes of taluk, district, state, polling station, and citizen node with attributes of details present in voter card.

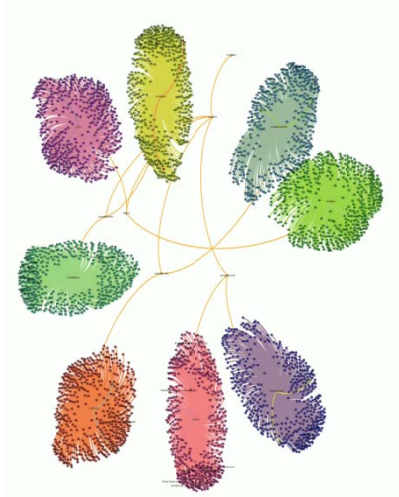

Fig. 2: Clusters of ration card obtained using Gephi tool (GN)

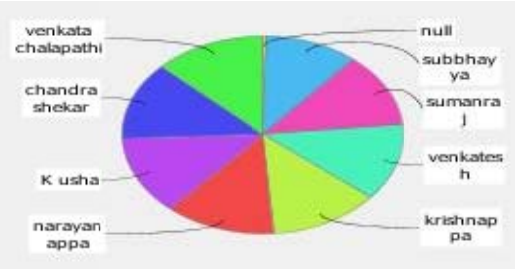

Fig. 2: Colors representing clustering based on fair price owner name.

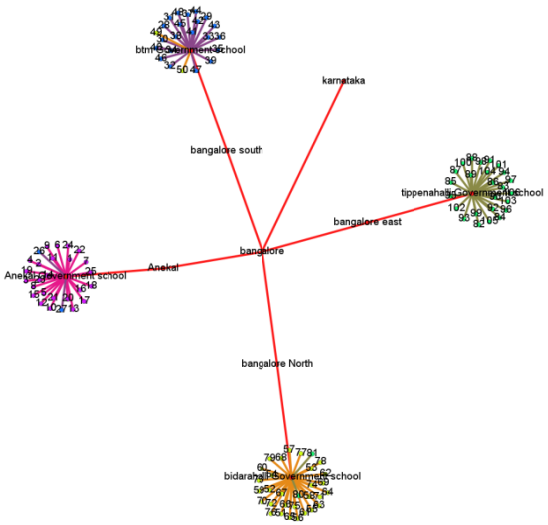

Fig. 4: Voter id dataset clustered using Gephi (LPA)

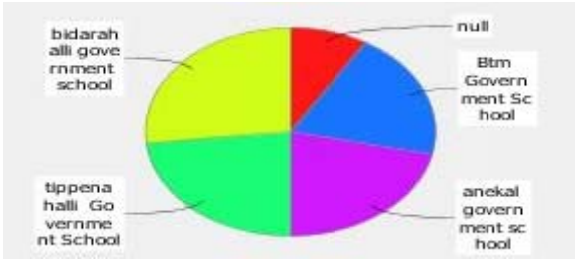

Fig. 5: Colors representing clusters based on their polling station.

The graph can be analyzed based on individual attributes.

We can also merge two graphs using gephi. Edges represent to which state or taluk or district ration card or voter card holder belongs. We can also detect duplicates in the graph and delete the node if duplicates found .The advantage of this type of analysis is that we can maintain a single record for all the identification cards used by a person. It is useful for banks to provide loans because all the details of person will be present will be available in a single storage place. Verification of documents becomes easy. Counting the number of people having ration cards for their identification or any other cards can also be determined.

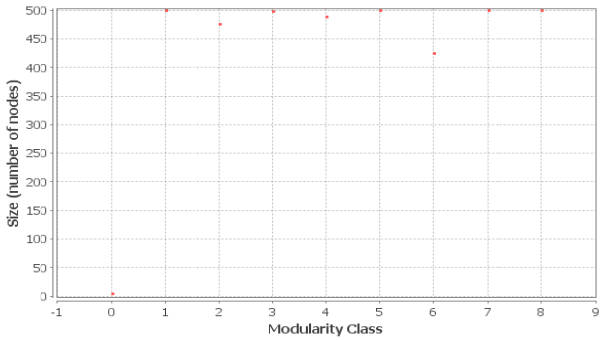

Fig. 6: Modularity determined using Gephi tool for Ration card dataset.

Modularity is one of the measures of Networks to determine the partitions of the graph. In Fig. 6 Modularity class of x-axis represents the number of clusters formed in the network. y-axis represents Number of nodes present in each cluster.

5.1 Comparison of results of two algorithms.

Table 1: Algorithm comparisons with respect to time

Algorithm	No of datasets	Time taken
GN	3900	4 min
LPA	3900	10 sec

From Table 1. Comparisons we can conclude that Girvan-Newman is not the good algorithm for analysis but at the same time the best Quality clustering of data obtained from our analysis was from Girvan-Newman. But it is expensive in

terms of time because of using edge betweenness measure for clustering, Also it is necessary to have advance knowledge about the number of clusters the network must be partitioned. Where as LP algorithm does not require any knowledge about number of clusters in advance but the quality of clustering is not as good as the one obtained from Girvan Newman algorithm.

6. FUTURE WORK

As a future work we can analyze the dataset using different other community detection algorithms such as AHSCAN[9] and develop new algorithms that are based on attributes of nodes. We can also try to show the advantages of using graph database such as Neo4j to store e-governance data rather than using relational database.

7. CONCLUSIONS

We would conclude that with little effort in organizing datasets in graph rather than tables we can analyze the confidential data of citizens using social network analysis algorithms rather than using queries on relational model to obtain data from database or analyse the data and generate the reports of our interest using query languages such as SQL. Much more powerful algorithms can be developed to determine the clusters in the graph based on the attributes of nodes.

REFERENCES

[1] Fortunato S 2010., "*Community detection in graphs*", *Phys. Rep.* **486** 75

[2] Newman M E J., "Fast algorithm for detecting community structure in networks", *Phys. Rev.* E **69** 066133, 2004

[3] Raghavan U N.Albert R and Kumara S., "*Near linear time algorithm to detect community structures in large-scale networks*" *Phys. Rev.* E **76** 036106, 2007

[4] Ding, C., He, X., Zha, H., Gu, M. and Simon, H., "*A min-max cut algorithm for graph partitioning and data clustering*",. Proc. of 2001 IEEE International Conference on Data Mining.

[5] Shi, J. and Malik, J., 2000." *Normalized cuts and image segmentation*". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 22, No. 8, 2000

[6] Bastian, Mathieu; Heymann, Sebastien; Jacomy, Mathieu , "*Gephi : An Open Source Software for Exploring and Manipulating Networks*", *AAAI Publications, Third International AAAI Conference on Weblogs and Social Media*, retrieved 2011-11-22, 2009

[7] Sameer Sachdeva, "*white paper on E-governance strategy in India*", December 2012.

[8] Akhtar, N. “Social Network Analysis Tools”, Communication Systems and Network Technologies (CSNT), IEEE 2014 Fourth International Conference on Date 7-9 April 2014,Bhopal.

[9] Nurcan yuruk, Mutlu Mete, Ziaowei xu, Thomas A.j. Schweiger., "*AHSCAN:Agglomerative Hierarchical structural clustering algorithm for Networks*", Conference of Advances in Social Networks and Mining", 2009

[10] http://ahara.kar.nic.in/

[11] http://www.ceokarnataka.kar.nic.in/

[12] http://graphml.graphdrawing.org/