Comparison of Algorithms for Social Networks using Ontology

Bhaskar Biswas
Department of Computer Engineering
Indian Institute of Technology (BHU)
Varanasi 221005 India

Vijay Nayak
Department of Computer Engineering
Indian Institute of Technology (BHU)
Varanasi 221005 India

Harish Kumar Shakya
Department of Computer Engineering
Indian Institute of Technology (BHU)
Varanasi 221005 India

ABSTRACT

The choice of appropriate algorithm(s) to the large graphs like social networks is always a challenge to the researchers for different applications. In this work, we propose a model to compare algorithms for large graphs such as social networks using ontology. Using ontology, the features and the dimensions of the social networks are summarized properly and through experiments it is shown how to choose any algorithm for a given objectives.

Keywords
Ontology, social networks, algorithms, MCL, RNSC

1. INTRODUCTION

In today's life, Social networks became an important part of our daily routine. It not only connects us from the rest of the world, but also became an important source of news, marketing, shopping, discussions etc. Community detection is one of major tasks in social networks where community or group based on common interest or topic is discovered. The success of community depends on the factors that are taken into consideration while forming the community. These factors, such as common interests, are very diverse in nature, thereby making the task of detecting a successful community a very difficult task. One of the major challenges in the community detection algorithm for social networks is to find the appropriate algorithm. The objective of the community detection algorithms vary from case to case. The objective of the detecting a community depends on the application to application. The success or failure of any community detection algorithm on a particular social network cannot be predicted for its future performance on other application. The repetitions of such experiments on such huge networks with many features and aspects, is a costlier affair. Sometimes minute details of knowledge of the domain are not known to the researcher and sometimes it is not required to know such unnecessary details. Use of statistical methods and/or other methods to sample or reduce the networks may miss some or other features of the network which may lead to results which are wrong or misleading results. Therefore, there is a need for a method to choose the correct or better algorithm for detection of the community. The same can be said for choosing the correct or the better objective function of the algorithm. Such method should be able to give the same result as in the case of actual data and for that the method should have one major property that features of the domain should be same as the actual data. The method is meant for to find the correct algorithm for the said purpose/application, so the cost of repeating the experiments should be low. For this, the volume of data on which the experiments are being conducted should be small. Statistical methods like Principle Component

Analysis (PCA) [15], and spectral methods [16] are used to reduce the dimensions of the data and thereby reducing the size of the data. One of the major drawbacks of such methods, in general, is the degraded performance with reduction in the dimension and sometimes led to wrong results. Correct choice in selecting the dimension can avoid such degradation in the performance. For selection of the dimension, knowledge of the domain is required, which is not always possible for a person to have. Keeping such requirements in consideration, use of Ontology as a base for such said method is proposed. Ontology is a way to represent knowledge of a particular domain. In one of the most sited definitions, ontology is "explicit specification of the conceptualization of a domain" [3]. In this paper, we purpose a model to evaluate the performance of an algorithm of the social networks through the use of ontology.

Ontology is created for a domain by considering all the features/ dimensions. Ontology is conceptualization of the data/domain; therefore the size of data handled by the algorithm will be considerably small. Therefore, the requirement of applying any dimension reduction techniques or size reduction techniques is omitted. Further, the minute detail of knowledge of the domain is not a requirement in forming ontology. Thus, Ontology satisfies all the above discussed requirements for the method. Based on observations, we present in this paper a model using ontology for the social networks so that one can choose efficient algorithm for the given purpose or objective. Rest of the paper is organised as follows: In Section 2, a small literature review for the said domain is presented. Section 3 gives the description of our model, Section 4 presents the description of the experiments we had performed. Section 5 which describes the results and conclusion is followed by references in section 6.

2. LITERATURE REVIEW

Ontology was used to describe the existence of the things [16]. Later, ontology was used to describe different components of semantic web [17]. The language used for describing the semantic web is OWL [17] and that inspired many researchers to use ontology in their own way. In [2], [4] and [7], the authors used ontology for designing web crawler. Pinar Oezden Wennerberg [1] proposed a model where the ontology can be used to discover knowledge in social networks. In 2005, Peter Mika [3] designed a model for social networks which was referred by many other researchers. These models tried to define semantics for the social networks. Similar attempt to define semantics for the social networks was done in [9]. In [8], authors used ontology to visualize social networks. Many attempts were being made to define the architecture for the social networks [13], inter-

connections of the social networks [10] and the applications of the text mining for social networks using ontology [12]. Régine Lecocq et. al. proposed a prototype for the analysis of the social networks based on ontology, motivated by the similar reasons cited in this work [14]. However, in [14] authors proposed a generalized prototype for analysing of the social networks. In this work, we are inspired by the work of [14] and focused only the comparison of the algorithms on the social networks.

3. MODEL

The overview of proposed model is shown in Figure 1. The data extracted from the social networks is done by a web crawler/ apps/ any other tool and is stored in the database. Based on the features of the extracted data, the ontology of the data is created by any ontology creator tool. Different algorithms are then applied to the ontologies thus created. The results are then analysed for the suitability of the objective(s) for which the algorithms are chosen.
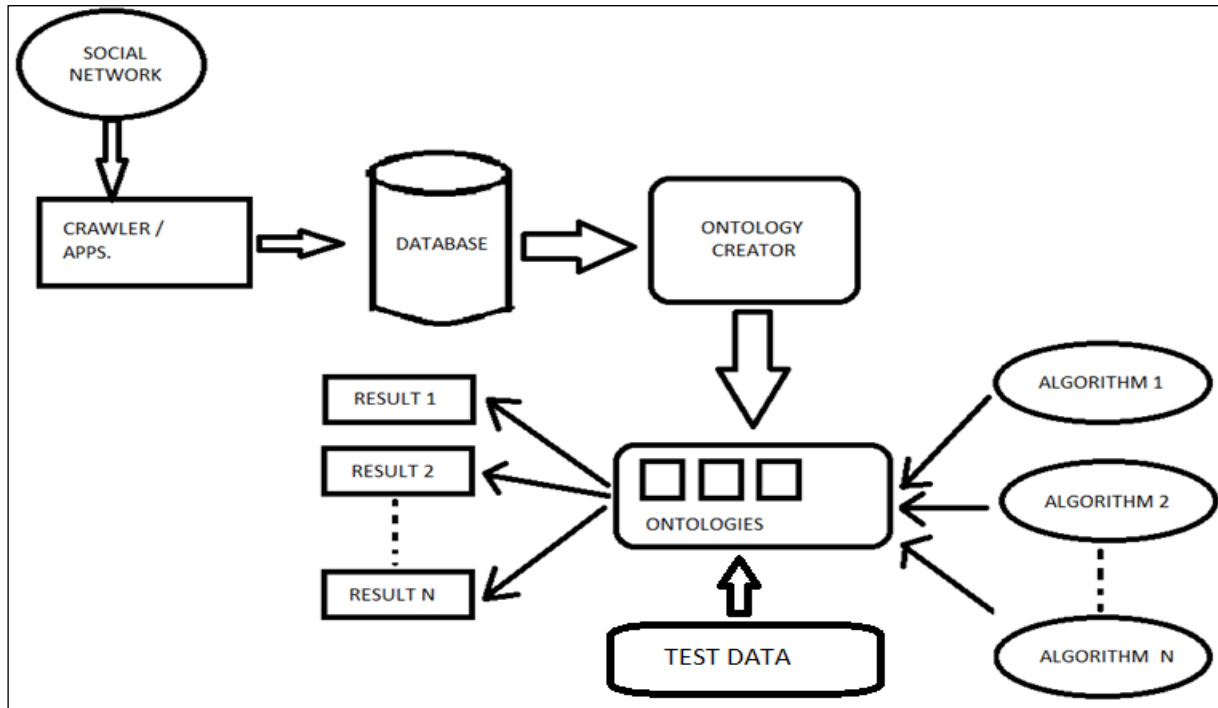

Figure 1: Block diagram of proposed technique

In this work, we have used graph clustering algorithms to test the proposed model. The model is implemented based on the ontology of the clusters formed as a result of the applied algorithms is to be used to compare and determine correctness of the clusters formed. A test data is used on the given ontologies and the results are generated. These results are then compared with each other. The basic aim of this system is to decide which of the algorithms are producing clusters that are strongly connected. This will help in determining which algorithms are producing the perfect clusters that should be assigned to an individual on the basis of its properties. The idea for the system is to test the ontology from each algorithm with same test data sets and find a way to compare the results of these test data sets to find the algorithms which are producing better cluster.

4. EXPERIMENT
4.1. Data preparation

The data used in [6] formed the basis of our experiment. The data contained information about the bloggers from a particular organization. The terms Blogs and Authors are used interchangeably. The algorithms used to test our model are two popular graph clustering algorithms, RNSC and MCL [11]. We had converted the data from [6] into graph by connecting the blogger based the similar tags. For creating ontologies, protégé [19] has been used. The correctness of the algorithms are tested based on a basic property of clusters that

inter-distance between the cluster should be more and the intra-distance should be less. In our case, a single author is classified in less no. of clusters which signifies that the clusters formed are strong as the strongly connected tags are kept in the same and less no. of clusters.

4.2. Ontology Design

The authors (bloggers) in the data [6] have a set of tags that they have mentioned in their blogs. After creating the graph as mentioned above from the data, the authors are clustered (grouped) into different clusters by different algorithms applied. This causes a set of tags for each cluster which become the main property while assigning authors of test data case to a specific cluster. Based on the observations, the ontology was designed and has the following properties:

Classes:
> Authors
> Tags

Object Properties:
> *hasTags:*
>> Domain: *Authors*
>> Range: *Tags*
> *isTaggedBy:*
>> Domain: *Tags*
>> Range: *Authors*

The class *Authors* has subclasses which are the clusters formed by the selected algorithm. The class *Tags* has

subclasses which consists of sets of tags that the authors of each cluster have tagged. *For example,* the class *Tags4* consists of all the tags that all the authors have tagged from the *Cluster4.*

The object properties defined are *hasTags* and *isTaggedBy.* The *hasTags* property has domain as *Authors* and range as *Tags* and this property defines that an individual from class *Authors* 'has tags' as an individual from class *Tags.* The *isTaggedBy* property has domain as *Tags* and range as *Author* and this property defines that an individual from class *Tags* 'is tagged by' as an individual from class *Authors. For example,* authors in *Cluster 4* have *hasTags* property over tags in *Tags4.* The relations between classes are defined as follows:

Object Relations:
1. Cluster[i] *hastags* **some** Tags[i]
2. Tags[i] *isTaggedBy* **some** Cluster[i]

"Cluster[i] *hastags* **some** Tags[i]" states that the class *Cluster[i]* will contain all the authors who have any of the tags present in class *Tags[i]*, where 'i' is the cluster number. "Tags[i] *isTaggedBy* **some** Cluster[i]" states that the class *Tags[i]* will contain all the tags which are tagged by authors present in class *Authors[i]*.

4.3. Individuals

Individuals are the Authors and their Tags. These individuals are classified on the basis of their properties defined earlier to their respective classes. The authors are classified to their respective cluster and their tags to class tags. *For example,* some "author1" is placed in class *Cluster3* as it has *hasTags* relation with tags belonging class *Tags3.*

5. RESULT AND CONCLUSION

MCL and RNSC algorithms were used to create clusters from the same data of 100 nodes of Blogs and their Tags. 13 clusters were created for MCL and 15 were created for RNSC out of the same graph. 20 Blogs and their Tags were used as test data on the ontologies created from the algorithms. Each individual (Blog) was classified into some number of clusters. The following is the number of clusters each individual was classified into and a graph showing us the percentage of clusters each individual was classified into:

Percentage (MCL % or RNSC %) = no. of clusters classified into / total no. of clusters

Table1: Number of clusters of both algorithms

| BLOGS	MCL	RNSC	MCL %	RNSC %
auth_1	3.00	9.00	23.08	60.00
auth_2	4.00	6.00	30.77	40.00
auth_3	3.00	8.00	23.08	53.33
auth_4	1.00	3.00	7.69	20.00
auth_5	2.00	3.00	15.38	20.00
auth_6	5.00	11.00	38.46	73.33
auth_7	3.00	7.00	23.08	46.67
auth_8	5.00	6.00	38.46	40.00
auth_9	5.00	5.00	38.46	33.33
auth_10	4.00	6.00	30.77	40.00
auth_11	1.00	1.00	7.69	6.67
auth_12	1.00	10.00	7.69	66.67
auth_13	2.00	7.00	15.38	46.67
auth_14	1.00	9.00	7.69	60.00
auth_15	5.00	10.00	38.46	66.67
auth_16	3.00	6.00	23.08	40.00
auth_17	4.00	5.00	30.77	33.33
auth_18	2.00	5.00	15.38	33.33
auth_19	4.00	10.00	30.77	66.67
auth_20	2.00	2.00	15.38	13.33
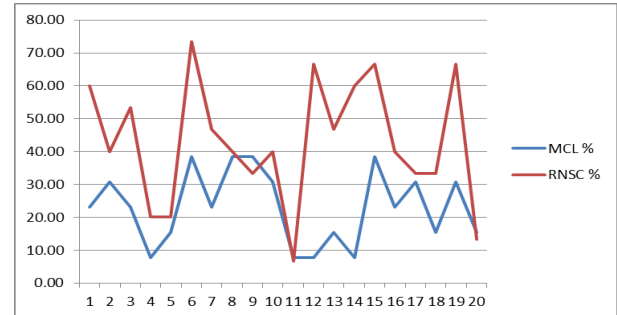
Figure 2: Comparative analysis of both algorithms

6. CONCLUSION

As seen from Figure 2, the percentage of clusters each node was classified into is less in MCL than in RNSC in maximum test cases. This concludes that the tags distributed in the clusters of MCL algorithm are more disjoint than in the clusters of RNSC algorithm. This concludes that the clusters of MCL have the tags more densely and accurately packed such that when tested with the test data provide more accurate results than clusters formed using RNSC algorithm. The data that was used for creating clusters and for testing were real time data, so the correctness depends greatly on this factor.

7. USE AND FUTURE WORK

There are three direct uses, which can be observed: 1. Comparison different algorithms to see which cluster are denser and disjoint 2. Comparison of the clusters formed using same algorithms and different data used to form these clusters. 3. Using different sets of data to form clusters can help in detecting dirty/noisy/inconsistent data which is causing the clusters to not to be perfect/correct/disjoint.

This work may be further extended for bigger graphs for which the efficiency and accuracy is predicted to be better. Application of ontology can be tested for other social network tasks like viral marketing, node ranking, link prediction etc.

8. REFERENCES

[1] Pinar Oezden Wennerberg, "Ontology Based Knowledge Discovery in Social Networks", JRC Joint Research Center, European Commission, Institute for the Protection and Security of the Citizen (IPSC), 2004.

[2] G.Aghila et.al, "Ontology-based Web Crawler", Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), 2004.

[3] Peter Mika, "Ontologies are us: A unified model of social networks and semantics", Lecture Notes in Computer Science (Springer/ The Semantic Web – ISWC 2005) Volume 3729, pp 522-536, 2005.

[4] Debajyoti Mukhopadhyay, Arup Biswas, Sukanta Sinha, "A New Approach to Design Domain Specific Ontology

Based Web Crawler", 10th International Conference on Information Technology (ICIT 2007), 289-291, 2007.

[5] Jason J. Jung, Jérôme Euzenat, "Towards Semantic Social Networks", Lecture Notes in Computer Science (Springer/The Semantic Web: Research and Applications), Volume 4519, pp 267-280, 2007.

[6] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. "Identifying Influential Bloggers in a Community", In Proceedings of the 1st International Conference on Web Search and Data Mining (WSDM08), pp 207-218, February 11-12, 2008.

[7] Wu Chensheng, Hou Wei, Shi Yanqin, Liu Tong, "A Web Search Contextual Crawler Using Ontology Relation Mining", International Conference on Computational Intelligence and Software Engineering (CiSE 2009) Page 1-4, 2009.

[8] Wu Peng, Li SiKun, "Social Network Visualization via Domain Ontology", International Conference on Information Engineering and Computer Science (ICIECS 2009) DOI: 10.1109/ICIECS.2009.5362898, 2009.

[9] Liu Chen, Shan Wei, Zhang Qingpu, "Semantic Description of Social Network Based on Ontology", Proceedings of the 2010 International Conference on E-Business and E-Government (ICEE '10), Pages 1936-1939, 2010.

[10] Morteza Jamalzadeh, Navid Behravan, "Using Semantic Web Ontologies for better inter-operability on social network sites", IEEE International Conference on Control System, Computing and Engineering, Page 103-108, 2011.

[11] Reena Mishra,Shashwat Shukla,Deepak Arora, Mohit Kumar, "An Effective Comparison of Graph Clustering Algorithms via Random Graphs", International Journal of Computer Applications (0975 – 8887) Volume 22–No.1, May 2011.

[12] Sam K. M. and Chatwin C.R., "Ontology-Based Text-Mining Model for Social Network Analysis", Proceeding of the Sixth IEEE International Conference on Management of Innovation and Technology, pp. 226 - 231, 2012.

[13] Moharram Challenger, "The Ontology and Architecture for an Academic Social Network", International Journal of Computer Science Issues (IJCSI), Vol. 9, Issue 2, No 1, PP 22-27, SSN (Online): 1694-0814, 2012.

[14] Régine Lecocq, Étienne Martineau, Maria Fernanda Caropreso, "An Ontology-based Social Network Analysis Prototype", IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), DOI: 10.1109/CogSIMA.2013.6523839, Page(s): 149 – 154, 2013.

[15] http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[16] http://en.wikipedia.org/wiki/Spectral_method.

[17] http://en.wikipedia.org/wiki/ Ontology.

[18] http://en.wikipedia.org/wiki/Web_Ontology_Language.

[19] http://protege.stanford.edu