Using Complex Network Features for Fast Clustering in the Web

Jintao Tang[1,2], Ting Wang[1], Ji Wang[3], Qin Lu[2], Wenjie Li[2]

[1]School of Computer, National University of Defense Technology, Changsha, P.R. China

[2]Department of Computing, Hong Kong Polytechnic University, Hong Kong

[3]National Laboratory for Parallel and Distributed Processing, Changsha, P.R. China

{tangjintao, tingwang, wj}@nudt.edu.cn, {csluqin,cswjli}@comp.polyu.edu.hk

ABSTRACT

Applying graph clustering algorithms in real world networks needs to overcome two main challenges: the lack of prior knowledge and the scalability issue. This paper proposes a novel method based on the topological features of complex networks to optimize the clustering algorithms in real-world networks. More specifically, the features are used for parameter estimation and performance optimization. The proposed method is evaluated on real-world networks extracted from the web. Experimental results show improvement both in terms of Adjusted Rand index values as well as runtime efficiency.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval - *Clustering*;

General Terms

Algorithms, Performance, Experimentation

Keywords

Graph clustering, complex networks, parameter estimation.

1. INTRODUCTION

Graph clustering is an important technology in social network analysis. With the rapid development of the Internet, network applications must deal with very large networks, often comprising millions of nodes. How to find a good tradeoff between speed and accuracy becomes an open challenge. Furthermore, how to discover a priori knowledge of real-world networks to determine parameters is crucial to the performance of clustering algorithms.

Most real-world networks are complex networks which have some topological features that are not apparent in simple networks. These topological features reveal additional information of the communities in complex networks. For instance, the small world [1] hypothesis suggests that the diameter of the complex networks is small; the high clustering coefficient [2] suggests that most of the nodes are only connected to some neighbors in the same cluster; the low diameter mainly depends on a few "weak long ties" among the clusters. Scale-free feature [3] indicates that the distribution of degrees follow a power-law to suggests that a few active nodes consume a lot of edges, and other common nodes only consume very few edges. From these works in complex networks we can deduce two hypothesis: (1) most of the common nodes are connected to its neighbors to form some clusters towards the so called **active nodes**, and (2) a few **weak ties** among clusters make greater contribution to network connections. In this paper, we employ the hypothesis to optimize the graph clustering algorithms for large-scale complex networks such as link-based data in the web. The first hypothesis is also used to determine clustering parameters. Two well-known algorithms are used to evaluate the performance of the proposed methods. The selected algorithms include the k-medoids algorithm [4] and the Girvan-Newman algorithm [5], which are widely studied in social networks analysis. Experiments show that the performance of clustering algorithms is improved by approximating the shortest paths algorithm based on the hypothesis.

The research is supported by the National Natural Science Foundation of China (60873097, 60933005), the National Grand Fundamental Research Program of China (2011CB302603), and the HLT Lab in the department of Computing, Hong Kong Polytechnic University.

Copyright is held by the author/owner(s).

WWW 2011, March 28-April 1, 2011, Hyderabad, India.

ACM 978-1-4503-0637-9/11/03

2. COMPLEX NETWORK CLUSTERING

Many graph-clustering algorithms are time-consuming because of the bottleneck of all-pairs shortest paths problem (APSP). Since the aforementioned hypothesis says that the weak ties between clusters make greater contribution to network connections, a novel shortest paths approximation algorithm using center distance to zone [6] (CDZ) is proposed to identify active nodes for complex networks analysis. CDZ identifies a path which goes through the active nodes and the weak ties to replace the actual shortest path, thus it can be used in the clustering algorithms to reduce the complexity of the algorithms with good approximation.

The incorporation of the CDZ algorithm into the *k*-medoids algorithm is straightforward. The bottleneck in the *k*-medoids algorithm is in comparing the closeness centrality, which is defined as the average distance from the given node to every other node. We use the distances reported by CDZ to replace the actual distances when computing the closeness centrality values. The Girvan-Newman (G-N) algorithm is a divisive clustering technique based on the edge betweenness centrality, which is the proportion of all shortest paths in the network that run through a given edge. Again, CDZ uses the identified paths to approximate the shortest paths to obtain the betweenness value. The approximation of edge betweenness takes expected $O(m)$ time(m is the number of edges) in complex networks, which make it feasible to apply the G-N algorithm even in large-scale networks.

When applying clustering algorithms in real-world networks analysis, it is difficult to determine the algorithm parameters because of the lack of a priori knowledge. For instance, the *k*-medoids algorithm requires the information of the number of clusters, which is unknown for most real-world networks. As analyzed above, the topological features in complex networks can provide some information of the clusters. That is, most common nodes tend to be connected to the neighbors in the same cluster towards the active nodes. Consequently, CDZ divides the graph

into zones based on active nodes. These zones are considered as the clusters in complex networks. So for the *k*-medoids algorithm, the size of the active nodes discovered by CDZ is used for the input parameter *k*. And the active nodes are used as the initial medoids instead of randomized initialization to make the *k*-medoids algorithm deterministic. The G-N algorithm requires the number of removed edges, which is hard to determine without a priori knowledge. We change its terminal condition to overcome this limitation by repeating the main steps of the G-N algorithm until the number of clusters reaches the estimated number *k*.

3. EXPERIMENTS

Two well-known datasets extracted from the web are used to evaluate the proposed algorithm: the American College Football network [7] and the Cora citation network [8], whose clusters are known for evaluation purposes. Adjusted Rand index (ARI) is used as the effectiveness measure of clustering algorithms.

Table 1 ARI of clustering algorithm and the optimization

	Football	Cora
K-medoids	0.322	0.247
+CDZ	0.329	0.252
+Parameter Estimation	0.452	0.241
Girvan-Newman	0.884	0.633
+CDZ	0.671	0.592
+ Parameter Estimation	0.658	0.508

The two original clustering algorithms are used as the baseline. The actual cluster number is used as the parameters for the baseline. If we run *k*-medoids more than once, the clustering results will be quite different because of its randomized initialization. In the experiments, its fluctuation range of ARI can reached 10%. So, we take the average of 10 runs rather than taking a one run data. On the other hand, Applying CDZ makes the *k*-medoids algorithm deterministic, thus CDZ based *k*-medoids only needs to run once.

As shown in Table 1, the results of CDZ based *k*-medoids keeps at the same level as the original *k*-medoids. CDZ has high accuracy on distance approximation, which makes the correctness of medoids generation is not heavily affected. The *k*-medoids using CDZ with parameter estimation has a high clustering accuracy on the Football network, and a lower but acceptable accuracy on the Cora network. Further investigation shows that the estimated number of conference (13) in the Football network is very close to the actual conference number (12), whereas the estimated number of clusters (124) in Cora is much higher than the actual clusters' number (71). However, the acceptable reduction on accuracy indicates that using the complex network features is a practicable solution to guide the parameter estimation in real-world networks without any priori knowledge.

Different from *k*-medoids algorithm, the accuracy of G-N algorithm using CDZ slightly reduces the performance of G-N. Since CDZ uses the paths through the active nodes to approximate the actual shortest paths, it makes the betweenness value of some edges overestimated. So the approximation G-N algorithm may remove these edges instead of removing the real important edges. However, the G-N algorithm takes $O(m^2n)$ time, whereas the CDZ based G-N can reduce the algorithm complexity to $O(m^2)$, where m is the number of edges and n is the number of nodes. In consideration of the dramatic reduction in time complexity, the tradeoff in accuracy is quite acceptable.

To evaluate the computational complexity of the proposed method in real-world applications, we extracted several sub-networks of Cora with the number of nodes ranging from 1,000 to 30,000. The practical runtime of different algorithms in these sub-networks are investigated and the results are plotted as shown in Figure 1.The approximation algorithms based on CDZ have much lower bound on the runtime. When mining the clusters, the runtime of CDZ based clustering algorithms are nearly one ninth of that of the original algorithms. This result demonstrates that using the complex networks features can obviously improve the performance of clustering algorithms.

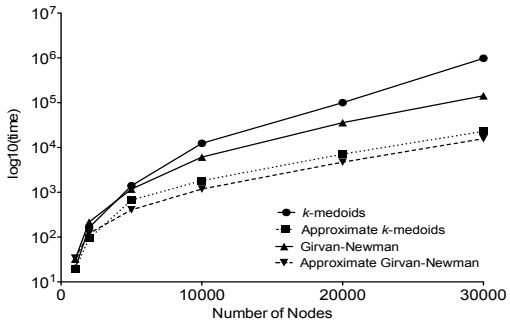

Figure 1. The $\log_{10}$ of run time of Girvan-Newman algorithm and the approximate algorithm on Cora network

4. CONCLUSION

In conclusion, using the topological features of complex networks to guide clustering algorithms can bring some obvious benefits. The main contributions of the proposed methods include: (1) CDZ based clustering algorithm is efficient on a run time and scalable to network sizes; (2) the selected approximation are effective for real-world networks; (3) The parameter estimation provides a way to assign the parameters, with no request for a priori knowledge; and (4) The incorporation of CDZ with *k*-medoids clustering algorithm make the latter to be a deterministic algorithm with predictable and improved performance.

5. REFERENCES

[1] S. Milgram. The small world problem. 1967. *Psychol Today* 2, 60-67.

[2] D. J. Watts and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 440–442.

[3] A-L. Barabási, R. Albert. 1999.Emergence of scaling in random networks. *Science* 286, 509-512.

[4] Leonard Kaufman and Peter J. Rousseeuw. 1990. Finding groups in data: an introduction to cluster analysis. *WILEY SER PROB STAT*. Wiley.

[5] M. Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks.*PNAS 99* (12), 7821-7826.

[6] Jintao Tang,Ting Wang, Ji Wang, and Dengping Wei. 2009. Efficient social network approximate analysis on blogosphere based on network structure characteristics. In *Proc. SNA-KDD 09*.

[7] M. E. J. Newmanand M. Girvan. 2004. Finding and evaluating community structure in networks. *PHYS REV E* 69 (2), 026113.

[8] A. McCallum, K. Nigam, J. Rennie, and K.Seymore. 1999. A Machine Learning Approach to Building Domain-Specific Search Engines. In *Proc.IJCAI '99*. 662-667.