Graph mining techniques

* 1. D-walks

The D-walks (discriminative random walks) is a technique introduced by Callut et al in [1] that can tackle semi-supervised classification problems in large graphs. The input graph is interpreted as a Markov chain (MC) in which random walks (D-walks) are performed. This technique is based on betweenness measures in which the betweenness of a node *‘x’* with respect to class ‘*y*’ is measured as the average number of times ‘*x*’ is visited during D-walks. The D-walks is able to classify the unlabeled nodes of both directed or undirected graphs with a linear time complexity Ɵ(|*y*|.m.L) and memory requirement of Ɵ(m + L.n) where |*y*| is the number of classes, ‘m’ is the number of edges in the input graph and ‘L’ is the maximum walk length considered. With such low complexities this technique is able to deal with very large graphs containing several millions of nodes and edges. Different experiment performed with the implementation of this technique in CORA database shows that this technique can efficiently and accurately classify the unlabeled nodes of the graphs.

1.2 Multi-level kernel k-means

This is a general algorithm for graph clustering presented by Dhillon et al. in [2] that is based on multilevel methods using weighted kernel K-means objective function as refinement algorithms. This technique can handle graph having large number of nodes and very large number of edges. The technique is divided into three phases: Coarsening phase, Initial clustering phase and refinement phase. In the coarsening phase, a graph G0 is passed through a transformation function which transforms G0 into a smaller set G1,G2,…,Gm. In the initial clustering phase, a parameter which indicates how small the coarsest graph should be is specified using the spectral algorithm of Yu and Shi [3]. the refinement phase is an improvement. The technique has been implemented on the IMDB movie dataset. The dataset has 1.2 million nodes and 7.6 million edges. The proposed techniques compute 5000 cluster and 5000 eigenvectors.

1.3 Bi-partite graph co-clustering

Chen et al, in [4] proposed a graph model used to handle the many-to-many correspondences problem among concepts in ontologies. Their proposed technique made use of weighted bi-partite graph to model ontologies. Weights of graph edges are calculated through the use of similarity measure techniques; the similarity degree are then assigned as weights of the edges in the graph. Graph partition techniques are applied to co-cluster the vertex sets of the bipartite graph and mappings is established between two ontologies based on the resulting cluster pairs. The proposed technique used a threshold in which the edges in the graph having weight greater than the threshold are maintained while others are purged out. The main success of this technique is that many-to-many mapping can be established among ontologies.

[1] Callut J., Françoisse K., Saerens M., Dupont P. (2008) Semi-supervised Classification from Discriminative Random Walks. In: Daelemans W., Goethals B., Morik K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science, vol 5211. Springer, Berlin, Heidelberg.

[2] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. 2005. A fast kernel-based multilevel algorithm for graph clustering. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD). Association for Computing Machinery, New York, NY, USA, 629–634.

[3] Yu and Shi. "Proceedings Ninth IEEE International Conference on Computer Vision," Proceedings Ninth IEEE International Conference on Computer Vision, 2003, pp. i-, doi: 10.1109/ICCV.2003.1238306.

[4] Fonseca, Yiling Chen Frederico. "A bipartite graph co-clustering approach to ontology mapping." *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data. Colocated with the Second International Semantic Web Conference (ISWC-03), CEUR-WS. org*. 2003.