Volume 2, Issue 2 (2014) 459-464 ISSN 2347 - 3258
International Journal of Advance Research and Innovation

Algorithm and Techniques for Overlapping Community Detection

Sakshi Singh
Department of Computer Science & Engineering, Teerthanker Mahaveer University, Moradabad, India

Article Info

Article history:
Received 5 March 2014
Received in revised form
20 April 2014
Accepted 28 May 2014
Available online 15 June 2014

Keywords

Modularity
MAX-CUT
Chromosome

Abstract

A lot of phenomenon, real world and otherwise can be conveniently represented as graphs, with the nodes corresponding to the entities and the edges representing the interaction be- tween those entities. Communities or modules, which are groups of nodes densely connected to each other within the community but sparsely linked to other communities and the rest of the graph, often having similar structural and functional properties. A lot of algorithms have been proposed to partition the set of vertices into communities; such a partition exclusively puts a node into one community or the other. But in real life a node can belong to multiple communities simultaneously, i.e. the communities can overlap. Different metrics have been proposed. We reduce the modularity maximization problem for splitting the graph into two communities to the MAX-CUT problem with both positive and negative weights. We introduce and analyze three approximation algorithms to maximize modularity for the two community case; recursive bi-partitioning can be carried out as long as modularity increases to split into more than two communities.

1. Introduction

Many real world phenomenon can be represented as graphs, directed or undirected, where the nodes represent the entities and the edges represent the interaction between those entities. Given an undirected unweighted graph $G = (V, E)$, a community is an induced subgraph on a subset C of vertices satisfying some properties.

A rough guideline of a community is by Newman and Girvan [20]: "a community is a subgraph containing nodes which are more densely linked to each other than to the rest of the graph or, equivalently a graph has a community structure if the number of links into any subgraph is higher than the number of links between those subgraphs".But only rarely do real world graphs separate into non overlapping communities or 'hard- partitions', most real networks have well defined overlapping and nested communities. There will be a set of nodes that can be put in more than one community, where a certain strict classification into one community or another is inaccurate, if not totally wrong. We reduce the modularity optimization for the two community 'hard-partition' case to the MAX- CUT problem with both negative and positive weights. Further we propose and analyze three approximation algorithms for modularity maximization for the two community 'hard-partition' case.We extend the Modularity metricso as to allow overlapping communities,we further discuss several properties of the extension.Reduces the extended modularity maximization problem into a Genetic Optimization problem, the algorithm is presented to maximize modularity. the results of application of our algorithm on real world graphs and computer generated overlap models and are presented and analyzed.

Corresponding Author,
E-mail address: sakshisingh0009@gmail.com

1.1 Modularity:A Goodness Measure

It is a measure of how good a particular division of a graph G into a set of communities C is larger values of Modularity indicate a better partition, or a more modular graph.

1.2 Conductance

For a Graph, G = (V, E) and a partition of the vertex set V into non-empty subsets S, S' Conductance of the cut (S, S') is defined as follows: $\emptyset(s) = \frac{cut(s,s')}{\min(vol(s),vol(s'))}$

Here vol(S) = vol(S, V) $= \sum_{i \in s, j \in s} a_{ij}$ and cut(s, s') $= \sum_{i \in s, j \in s} a_{ij}$

2. Approximation Algorithm For Modularity Maximization

We formulate the problem of partitioning the graph into two communities so as to maximize modularity as a Strict Quadratic Program, we then relax it into a Vector Program which can be solved upto any degree of accuracy, and we finally round the solution to get back to the solution of the original problem. We use the rounding techniques based on Rietz' method of averaging with the Gaussian measure to get a randomized approximation algorithm with an approximation guarantee of $\frac{4}{\pi} \approx 0.27$.We reduce the modularity maximization problem into the MAX-CUT problem and then use the techniques described in Alon et.al [49] to im prove the approximation guarantee to $\frac{2 \ln(1+\sqrt{2})}{\pi} \approx 0.56$