ABSTRACT

Graphs are very appropriate for modelling complicated structural data, such as circuits, images, communication network, molecular structures, biological networks, protein interactions and so on. More specifically, labelled graphs have been a promising abstraction to capture the characteristics of data sets rising in these fields. Hence, efficient graph mining algorithms are necessary for increasing our understanding of the information represented by these large datasets of graphs.

Identifying similarities of graphs is a challenging and essential process in various domains, such as pattern recognition, information retrieval and bioinformatics. There are various similarity measures in the domain of graph mining, out of which the most familiar measures are maximum common subgraph, frequent subgraph search and query graph search to visualize the common and frequent structures of the graph database.

This dissertation focuses on two types of graph databases to find the similarity patterns. The first one is the graph database of communication network where the state of the communication network is captured as a time series of graphs which has periodic snapshots of logical communications within the network. The second one is the chemical graph database which consists of chemical compounds which are easily visualized as undirected labelled graphs where atoms are the nodes with labels representing the name of the atoms and bonds representing the edges.

The devised algorithms for chemical graph database can handle at most two input chemical graphs at a time, whereas the algorithms of communication network can handle k input graphs simultaneously. In this dissertation, various issues on the domain of graph similarity search in communication and chemical databases such as common subgraph prediction, frequent subgraph mining using

the concept of edge relaxation and induced subgraph mining in large graph databases have been discussed.

There are six works associated with this thesis to perform similarity searches in large graph databases. They are explained briefly below.

The vertex cover-based binary tree algorithm detects all Maximum Common vertex Induced Subgraphs (MCIS) in large communication networks. The process of determining the similarity of a communication network graph database is possible by finding all MCIS of the network database and the identified subgraphs help to understand the common substructures which are always active and retain the links between any pair of nodes exactly as in all graphs of the database. The concept of vertex cover is used with a newly defined data structure which is a caterpillar-based binary tree to reduce the search space of the problem.

The centrality measures-based maximal common induced subgraph algorithm is used to find a maximal Common vertex Induced Subgraph (CIS) of the network graph database, which gives promising effect in the resulting graph in terms of a large number of vertices. It is possible to develop an algorithm which finds one maximal CIS with polynomial time complexity. Hence, the problem of finding an active pattern in a communication network is modeled as a detection of a maximal CIS. Centrality measures are used to assess the variation in successive graphs and to identify the behavior of each node in the time series graph.

An algorithm to search edge relaxed query graph with a minimum support threshold in large communication networks is devised, which decides whether the query graph is frequent as expected times in the graph database, if it is not so, then the algorithm relax infrequent edges from the query graph to detect the subgraph of the query graph with the given minimum support threshold. It is often required to determine the frequency of a query graph in the graph database

to realize how much query graph has occurred in the communication network. When a query graph does not have an exact match with the graphs in the graph database, the idea of finding approximate matches of the query graph is coined to determine the frequency of the query graph by relaxing few edges of it.

A new proposed algorithm is designed to detect edge relaxed induced subgraphs of the query graph in a large communication network. By searching induced subgraph, it is possible to give the importance of the connections of the query graph. The process of analyzing the performance of a particular interested group of systems and knowing how well the group maintains the relationship over a long period of time in a computer network are achieved by forming a query graph with those interested systems as nodes and possible connections as edges and determining their occurrences in the graph database. It may be possible to get non-positive result when searching a query graph directly in a graph database, so that the concept of edged relaxed gets significant in such situations.

An algorithm is defined to detect a large sized maximal Common Connected Edge induced Subgraph (CCES) of two given chemical graphs using a new technique incorporating centrality measures. The idea is to use a DFS search tree whose root node is chosen as the one with the highest average of closeness and reach centrality measures from the tensor product graph of the input graphs. This measure of average narrows down the search space of the problem. This algorithm ensures at least one maximal CCES and more than one may also be generated.

A new algorithm is formulated to detect a maximal Common Connected vertex induced Subgraph (CCS) by checking the induced property of the vertices which are collected by performing a DFS search on the tensor product graph of two input graphs. CR-maximal CCES algorithm discussed in chapter 6 analyzes the similarity of two chemical graphs by finding one maximal common connected edge induced subgraph. Another familiar approach to capture the

structural similarity between two chemical compounds is to detect a maximal CCS in their molecular graphs. The preprocessing algorithm constructs a tensor product graph of input graphs G and H, and performs a DFS search on any one of the nodes of the constructed tensor product graph to construct a DFS search tree. The proposed algorithm checks the induced property of any one of the branches of the DFS search tree to extract a maximal CCS.