CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Summary of Contributions

This thesis has investigated two issues in ranking of mined subgraphs the efficiency of algorithms and the conciseness of results. The proposed GG, AFSMG and FPGBG algorithms were implemented in Java. All experiments reported in this chapter have been performed on Intel Pentium(R) Dual core E5400 @ 2.70 GHz with 1.99 GB main memory, running Microsoft Windows XP and performance evaluations are also studied. The contribution to this thesis can be summarized in each of two areas as below.

7.2 Efficiency of Algorithms

A variety of novel algorithms have been proposed in the recent past for the efficient mining of subgraphs, each in turn claiming to outperform its predecessors on a set of graph databases and no one gave the sufficient amount of work in ranking of the subgraphs mined. In this thesis, the proposed approach quantifies the algorithmic performance of frequent subgraph mining algorithms with ranking. The Graphgain algorithm utilizes a statistical strategy “lift” to determine the discounted cumulative gains. It uses rules for the determination with the support and confidence measures. This work has shown that these rules constructions were based on the association rule mining in all possible relationships between two quantities which are considered.

The proposed algorithm has been divided into four stages. The first one is the evaluation of lift measure for each rule of a subgraph. The second stage is the determination of MDCG at each point of lift values. This measure was obtained by introducing the measure “lift” to the Discounted Cumulative Gain (DCG) instead of relevance. The third stage is the finding of Graphgain, which is the new
normalization technique adopted for the frequent subgraphs in a graph data set. At each position of frequent subgraphs, the graphgain shows the efficient result over the normalized discounted cumulative gain (nDCG). The fourth and final stage is the ranking of subgraphs. The experimental results validate the quality, performance and practical utility of the presented algorithm.

The AFSMG algorithm proposes the finding of frequently occurring subgraphs and ranking them in large graph databases that can be used to discover recurrent patterns in chemical and synthetic datasets with their appropriate ranking order. Such patterns can play an efficient role for understanding the nature of these datasets and can be used as input to other data-mining tasks such as decision support, priority counting, similarity and quality of the output. Key elements to AFSMG’s computational scalability are the highly efficient canonical labeling algorithm and the ranking of subgraphs. The facts proposed, allow AFSMG to identify the various generated subgraphs uniquely and quickly.

Another algorithm FPGBG have proposed a novel and efficient mining of frequent patterns in large databases. There are several advantages of FPGBG over other approaches: (1) it constructs a highly compact FP-tree, which reduces the database size and scanning process, (2) it applies a pattern growth method which avoids costly candidate generation. In this context, mining is not Apriori-like (restricted) generation-and-test but frequent pattern (fragment) growth only, (3) it applies a partitioning-based divide-and-conquer method which dramatically reduces the size of the subsequent conditional pattern bases and conditional FP-trees.

7.3 Conciseness of Results

This thesis proposes the generalized ranking framework in order to rank the subgraphs as the output of frequent subgraph mining algorithms. This framework provides an order of magnitude improvement in subgraph ranking. This is achieved by rule construction between the frequent subgraphs with the user’s specified support. Instead, this framework accepts that the lift values of two itemsets are equal, and then it doesn’t affect the result due to the fact that, applying of lift measure for the calculation of MDCG measure will not yield the same value. This phenomenon will
improve the performance of the graph as well as the performance of GG measure. Hence the ranking of subgraphs will be a perfect one based on the performance factor. The experimental results of the proposed work have shown that ranking was performed in an efficient manner. It has also shown that the accuracy of ranking is maintained for all types of subgraph mining algorithms.

The detailed experimental evaluation of AFSMG shows that, AFSMG can scale reasonably well to very large graph databases. The performance study of FPGBG shows that the method mines and rank both short and long patterns efficiently in large databases, outperforming the current candidate pattern generation-based algorithms.

One of the advantages is that, if more than one similar case is present, then this algorithm will find the solution for the priority of the similar subgraph. This algorithm also considers another possibility of ranking based on lift. If lift is considered for ranking, then there is a need for a decision to fix the priority of rules because many rules may have the same value of lift. In this context, the graphgain method will play an efficient role in predicting the priority of similar cases. Another advantage of this algorithm is that, no rule will have the same GG priority even though they possess the same lift values. Thus, a hybrid approach that uses graphgain for similar patterns and ranking of all the patterns may result in better performance.

The research work contributes too many exciting information from Graph dataset which helps to improve the performance of the graph and subgraph discovery.
7.4 Future Work

This proposed algorithm can also be extended or simply modified for finding the ranking of subgraphs for many type of data sets such as spatial data set, wave data set and medical data set. This algorithm can be considered as a general ranking method for subgraphs. Once the dataset can be represented by the graph model then this algorithm will provide the rules for ranking the subgraphs from the graph dataset. The algorithms designed and evaluated in this thesis address the extension of integration of the subgraph mining algorithms. By implementing suitable procedure, this algorithm may be extended to distinguish the types of graph to be implemented. Another possible extension to be adopted is various types of graph dataset transformation and graph data conversion.

Furthermore, the performance of AFSMG and FPGBG can be adopted in new algorithms in future to extract a new data structure. There is a scope for suitable modification in this algorithm to maintain sufficient memory space and to reduce the time effectively. The algorithm may be extended by suitable modification so that it can fit for real time applications such as medical data set, ECG flow analysis, domain expert result analysis etc.