Name: Syed Zubair Ahmad Shah
Supervisor: Dr. Mohammad Amjad
Department: Computer Engineering
Title of Thesis: Preceding Clustering by Pattern Preservation

Abstract

Our age is the age of information. Large amounts of data are being produced every day. From the past two decades, tremendous work has been done to deduce knowledge from data. Data mining is a field of computer science mainly concerned with this job. Data mining includes techniques like frequent pattern mining, clustering, classification, outlier detection etc. This research is related to the first two techniques. I propose a methodology for performing document clustering. The way I perform clustering is summarized in these steps – first frequent termset mining is performed by a novel bipartite graph based algorithm, then the dataset is modified in accordance with the result of first step, and then finally clustering is performed on the modified dataset. I claim that the proposed approach produces better quality clusters and also reduces the dimensions of input dataset by a huge margin. It also assists in interpreting the resultant clusters with ease. I support my claim by implementing the proposed approach on example as well as real datasets. The results have been extensively shown in this thesis. Example dataset has been considered so that every step of execution of the proposed approach could be clearly understood and represented. Step-by-step execution and the changes that the bipartite graph undergoes in every step have been demonstrated. Real datasets have also been considered to show the practicality of my approach.

The greatest hurdle while implementing a frequent pattern mining algorithm is its execution time. To overcome this hurdle, in this thesis I propose two algorithms for distributed mining of frequent patterns. On example datasets, it has been shown that these algorithms will run smoothly and will speed up the execution many folds.

I also show that my proposed algorithm has great applicability. It is useful in reducing the dimensions of datasets by huge margins; it is helpful in precise detection of cancers; it is applicable in scripture analysis, in market basket analysis, in document categorization and in many other fields. In this thesis, I have shown through implementation that how my proposed algorithm can be used for analyzing the Quran. In addition to that, I have also proposed another algorithm for this sort of analysis. The analysis is that of the frequency of presence of words in different chapters of the Quran and also of the repetition of verses in the Quran. My usage of frequent pattern mining for scripture analysis is based on a fundamental concept of frequent pattern mining: that, with regard to some collection of documents, the lexical frequency profiles of individual documents are a good indicator of their conceptual content. The results obtained here show that the proposed approach produces results that are useful in getting onto the fundamental concepts of the Quran and can be of much help to the people of theology and to those who are interested in objective study of religious scriptures. The proposed algorithms are not limited to the analysis of Quran only. They can be used for analyzing any religious or non-religious scripture.

A method for diagnosis of breast cancer using frequent pattern mining has also been proposed in this thesis. This method suggests using the output of a frequent pattern miner as input to an artificial neural network. The main concern of the proposed method is to reduce the dimensions

of the input medical dataset. This method on one hand reduces the number of inputs to the artificial neural network and on the other produces better classification results.