Thesis Abstract

Efficient Techniques for Real-Time Frequent Pattern Mining in Business Applications

Rajanish Dass

Indian Institute of Management Calcutta, 2005
Supervisor & TAC Chairman: Professor Ambuj Mahanti
TAC Members: Prof. Sougata Ray, Prof. P.S. DasGupta

Frequent Pattern mining in real-time decision making is of increasing thrust in numerous business applications. Applications such as e-commerce, recommender systems, supply-chain management and group decision support systems are to name a few. Finding frequent patterns from databases has been the pre-requisite and the most time consuming process of the association rule mining. Till date, a large number of algorithms have been proposed in the area of frequent pattern generation. However, all of these algorithms produce output only at the completion and are not amenable to the real-time need. The need for real-time frequent pattern mining for online tasks and real-time decision making is increasingly being felt. Moreover, with dense datasets, where there are many long frequent patterns, the performances of the existing algorithms significantly degrade. A couple of recent developments use Diff-Set techniques for improving the performance of the vertical mining algorithms in dense datasets, but the performance of these techniques degrade in sparse datasets and have to calculate the density of the whole dataset before the user can decide on which process to use. The objective of this thesis is to address the problem of frequent pattern mining in real-time. In doing this, our main focus has been on the design of efficient search algorithms and powerful heuristics. Thus, in this thesis, we present BDFS(b), an vertical mining algorithm to perform vi real-time frequent pattern mining with limited available computer memory and user defined completion time. We have developed a few versions of this algorithm. One such version is named as BDFS(b)-diff sets in which we have implemented BDFS(b) with Diff-Sets for performing real-time frequent pattern mining in dense datasets. We have also incorporated two domain independent heuristics, h1 and h2 that improve the performance of these algorithms. The technique for using these heuristics has been implemented in extensions of BDFS(b) and we have named them as BDFS(b)-h1 and BDFS(b)-h2, using the heuristics h1 and h2 respectively, for finding the set of all frequent patterns for a given database and a given user-defined support threshold. Empirical evaluations show that these algorithms can make a fair estimation of the probable frequent patterns and reaches the possible longest length frequent pattern much faster than the existing algorithms and can estimate the final set of frequent patterns even in a smaller percentage time slice of the full execution time. More than that, scalability tests show that our algorithms are highly scalable with the number of items and number of transactions in the database. Use of BDFS(b)-h1 and BDFS(b)-h2 can complete the search at much lesser time of completion, checks lesser number of patterns and can approximate the actual set of all frequent patterns with very high accuracy. Comparisons with present state-of-art algorithms show that BDFS(b) and its variants (i.e. BDFS(b)-diff-sets, BDFS(b)-h1 and BDFS(b)-h2) can perform much better than the currently existing efficient algorithms like Apriori, FP-Growth, Eclat, dEclat etc.both in cases of complete execution and in real-time execution.