ABSTRACT

Transactions in the real world is a continuous process. The new transactional records will be continuously/periodically appended to the existing transactional logs. This type of updated data may behave differently from the existing knowledge with respective to the local, temporal aspects and may also have global influence. Henceforth, to the decision maker, the information/knowledge derived based on the existing set of records need to be modified/updated when additional records are added. In general, this knowledge updating process is addressed by the validation of existing knowledge in the additional/appended records. Based on the validation process one need to utilize the influence of the current records for minimizing the complexity of knowledge discovery process. This lead to development of incremental data mining strategies. These techniques are having the advantage of better response time as they avoid processing the previously accumulated data repeatedly. Several attempts were made to adapt some of the partition based data mining algorithms to deal with the process of incremental mining, effectively. This thesis discusses about the recent attempts in the field of incremental mining with emphasis on finding associations among items in transactional database. Mining frequent patterns and deriving association rules is an important task in data mining. In this thesis four different types of data mining strategies based on distributed, incremental, parallel and temporal environments for frequent pattern mining are proposed. Further, the procedures for deriving the association rules are also proposed and investigated. In incremental category two stage incremental CPTree construction methodology is designed in order to build incidental knowledge while building updated knowledge in the form of rules.

Most of the data mining strategies are multipasses (at least two) techniques are employed for mining and discovering the knowledge. In the present scenario, the existing records may not be available because of the privacy. Hence, forces the
researches to focus on single pass data mining constructs. The association rule mining (which is one of the main activities of the data mining) experts start with frequent patterns which in turn, demands item-pattern representation. One of the simplest and feasible item-pattern representation is Transactional Tree. This Transactional Tree attracted the researches and they have come up with single pass construction methods. This Transactional Tree has poor representation from information point of view. Thus, CanTrees are designed and developed to meet the customized activities. Single pass construction process of the CanTree for deriving association rules has been attracting researchers for data mining which includes incremental mining to accommodate the growth of transactional logs. This thesis proposed a five step mechanism for building a CanTree in High Performance Computing (HPC) environment. This representation will not be informative for general purpose as it was designed for specific objective and expected to have more space complexity.

To overcome the space complexity issues, CPTree was discovered which is the optimal representation. As the nature of business is distributed across the continents geographically, the transactions at different locations themselves are sufficient. There is a need to extract location-wise patterns and consolidating those patterns in global level for minimizing data transmission load. This procedure also results implicitly in privacy preserving. This demands distributed computing strategies. This thesis proposed to extract local and global knowledge through compact pattern trees in a hierarchical structure through distributed and parallel computing paradigm. This method also facilitates privacy preserving with a minimal communication load.

It is possible to achieve effective gain in any business due to the adaptive knowledge which demands customized rules for specific conditions. The parallel algorithms are useful to extract frequent patterns from large databases. This thesis developed a set of periodic rules through modified parallel compact pattern tree construction strategy.

An integrated framework has been designed and developed namely, Tree Based Association Rule Miner (TBARM). The functionality and the utility of TBARM has been provided and also demonstrated by using few benchmark datasets. This thesis
experienced different kinds of benchmark datasets such as mushroom dataset, pima Indian diabetes dataset and synthetic data set for proposed mechanisms.

Title .............................................................................................................i
Acknowledgements ......................................................................................ii
Declaration ..................................................................................................iii
Certificate ....................................................................................................iv
Abstract .........................................................................................................v
Table of Contents ..........................................................................................vi
List of Figures ................................................................................................
List of Tables .................................................................................................

Chapter 1:  Introduction

1.1 Data Mining ................................................................................................1

1.2 Frequent Pattern Mining ........................................................................1

1.3 Association Rule Mining .......................................................................2

1.4 Scope of the Thesis .................................................................................4

1.4.1 Tree-Based Structures .....................................................................4