An Abstraction Based Communication Efficient Distributed Association Rule Mining

P. Santhi Thilagam[1] and V.S. Ananthanarayana[2]

[1] Sr. Lecturer, Dept. of Computer Engineering, NITK-Surathkal, India-575 025
santhi_soci@yahoo.co.in
[2] Professor, Dept. of Information Technology, NITK-Surathkal, India-575 025
anvs@nitk.ac.in

Abstract. Association rule mining is one of the most researched areas because of its applicability in various fields. We propose a novel data structure called Sequence Pattern Count, *SPC*, tree which stores the database compactly and completely and requires only one scan of the database for its construction. The completeness property of the SPC tree with respect to the database makes it more suitable for mining association rules in the context of changing data and changing supports without rebuilding the tree. A performance study shows that SPC tree is efficient and scalable. We also propose a Doubly Logarithmic-depth Tree, *DLT*, algorithm which uses SPC tree to efficiently mine the huge amounts of geographically distributed datasets in order to minimize the communication and computation costs. DLT requires only $O(n)$ messages for support count exchange and it takes only $O(log\ log\ n)$ time for exchange of messages, which increases its efficiency.

Keywords: Association rule mining, Distributed databases, Sequence Pattern Count Tree, Incremental mining, Doubly Logarithmic-depth Tree.

1 Introduction

Due to the explosive growth in the number, size and complexity of databases, many geographically distributed organizations, the ever-growing number of applications and the high scalability of distributed systems, there is a need for mining distributed databases [4]. Association Rule Mining (ARM), the mining of frequent patterns in large transaction databases and many other types of databases has been studied popularly in data mining research. The most important component affecting the performance of any ARM algorithm is the number of disk accesses required [1]. Recently alternative data structures were employed in order to improve the efficiency of existing and new algorithms [5]. This motivated us to propose an approach that employs an abstraction called Sequence Pattern Count, SPC, tree which is more compact and complete, and suitable for incremental mining and changing support [10]. A SPC tree is constructed using a single database scan and can be updated dynamically. Algorithm based on this structure does not require any more database scans to generate frequent itemsets [6].

S. Rao et al. (Eds.): ICDCN 2008, LNCS 4904, pp. 251–256, 2008.

All proposed algorithms for mining association rules in distributed databases focus on reduction of communication [3], efficient usage of memory, processing power, ability to scale up the number of processors and associated data, ability to increase the size of the database, decrease in response time with addition of processors[4][8]. However, the majority of the parallel mining algorithms suffer from high communication and synchronization overhead [2][7]. In this work, a new distributed association rule mining approach is proposed that decreases communication costs by introducing a new message exchange procedure and a new computational efficient technique that reduces computation time. Our main contributions reported in this paper are:

1. Communication optimization: Doubly Logarithmic-depth Tree algorithm is used to minimize the communication costs in terms of number of messages for support count exchange and time taken for exchange of these messages.
2. Minimization of the number of database scans: SPC tree structure is used to make the mining process more efficient in terms of database scans in the distributed environment, which requires only one scan of the database for mining frequent itemsets.

2 Distributed Association Rule Mining

Let 'DB' be a database and 'n' be the processors of nodes namely $P_1, P_2, \ldots, P_n$ which are connected over a computer network. Each processor has a local memory and a local disk. The processor can communicate only by passing messages and we assume that there is no loss of message during communication and network is completely reliable. Let the database DB be partitioned into n non-overlapping blocks $DB^1, DB^2, \ldots, DB^n$ where n is the number or processors available and each partition DB^i has the same schema. Let the size of DB and the partitions DB^i be D and D^i respectively. For a given itemset X, let $X.sup$ and $X.sup^i$ be the respective support counts of X in DB and DB^i. The problem is to find all frequent itemsets in DB[9].

2.1 Solution Approach

Distributed Association Rule Mining is explained in Algorithm 1 which consists of the following phases:

SPC tree Construction – This preprocessing step allows us to store the DB^i compactly in main memory. The frequent-1 itemsets can be generated during the construction of the SPC tree which will be used in mining process later. SPC tree construction is explained in Algorithm 2. *SPC tree Growth* – This algorithm adopts pattern-growth approach to mine all frequent itemsets from the SPC tree which requires no more database scans. *Communication between processors* – It uses DLT algorithm to exchange the local counts between the processors in order to calculate the global count of each itemsets which is explained in Algorithm 3.