AIN SHAMS UNIVERSITY
FACULTY OF COMPUTER & INFORMATION SCIENCES
Department of Information Systems

HIGH PERFORMANCE DATA MINING IN DISTRIBUTED DATABASES

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science

In

Computer and Information Sciences

By

Mahmoud Fouad Anwar Darwish
B.Sc. in Computer and Information Sciences, Information Systems Department, Ain Shams University

Under Supervision of

Prof. Dr Hossam El Deen Faheem
Professor
Computer Systems Department,
Faculty of Computers and Information Sciences,
Ain Shams University, Egypt

Prof. Dr Nagwa Badr
Professor
Information Systems Department,
Faculty of Computers and Information Sciences,
Ain Shams University, Egypt

Prof. Rania El Gohary
Assistant Professor
Information Systems Department,
Faculty of Computers and Information Sciences,
Ain Shams University, Egypt

Abstract

The massive volume of data generated on daily basis decreases the ability of current data mining techniques to generate knowledge in a short time. The constant change in data requires constant updating of the existing patterns. It is computationally intensive to repeat the knowledge discovery process on the whole databases with every update. Therefore, there is a need to enhance the performance of association rules mining methodologies when dealing with incremental updates.

In order to enhance the performance of incremental association rules mining, this thesis focus on the utilization of current hardware and software advances in high-performance computing. This thesis proposes a distributed incremental association rules mining approach based on MPI. In addition, the thesis also proposes a hybrid incremental mining approach based on OpenMP and MPI to work in high performance computing environments. In order to reduce the need to reprocess the entire database, this thesis depends on pre-large and negative borders approaches.

To evaluate the applied approaches, this thesis considered the output accuracy, processing time and the acceleration as our primary evaluation metrics. In fact, experimental results have proved that our distributed method reduces processing time by 40% when compared to

serial existing approach and our hybrid approach reduces processing time by 19% when compared to distributed approach.

Acknowledgments

I would like to Thank God for his blessings and support to finalize this thesis.

I would like to thank Mrs. Fahima Moustafa, my mother, the real hero behind this thesis and any success in my life. I am so lucky to have a wonderful mother like you.

I would like to thank my supervisors for their support and help to accomplish my master. I wouldn't be able to do this without your guidance and support. A Special Thank you goes to Dr. Rania El Gohary, Dr. Nagwa Bader, and Prof. Hossam El-Deen Faheem.

I also thank the teaching assistants at Faculty of Computers and Information Science for their support during performing experiments at high performance computing laboratory. Thank you. I would also like to Thank Mohamed Gawish for his continuous support throughout this thesis.

I would like to thank Ain Shams University, Faculty of Computer and Information Science for providing the appropriate atmosphere and facilities during my master. I always feel so proud for being a graduate from the faculty of Computer & Information Science, Ain Shams University.

Table of Contents

CHAPTER PAGE

List of Tables

Table Page

List of Figures

List of Abbreviations

ARM	Association Rules Mining
DDM	Distributed Decision Miner
Di	Database number i
FDM	Fast Distributed Mining
FP-Growth	Frequent Pattern Growth
FUP	Fast Update
FUP2	Fast Updated2
HPC	High Performance Computing
ICSC	Incremental candidate support count
IDC	New incremental database count
LP	Local Pattern
LS	Large support
MAAP	Maintaining Association Rules with Apriori Property
MPI	Message Passing Interface
NLSC	New large minimum support count
NPSC	New pre-large minimum support count
OCSC	Original candidate support count
ODAM	Optimized Distributed Association Rule Mining
ODC	Original database transactions count
OpenMP	Open Multi-Processing
PLS	Pre large Support

SIMDAR	Secure Incremental Maintenance of Distributed Association Rules
SLP	Suggested Local Pattern
UCSC	updated candidate support count in both original and incremental databases

Chapter 1: Introduction

1.1 Overview

In our present time, data volume is increasing dramatically, sources of data are becoming decentralized, and the rate of data updates is tremendously speeding up. Today's data characteristics make it hard for humans to understand the big picture or get accurate insights about these massive amounts of data. Figure 1.1 shows the main challenging characteristics of today's dataset. The first challenge is the massive volume of data generated every second by various systems. The second one is velocity as the speed of data generation is increasing every second. The third challenge is variety of datatypes as nowadays there are different types of structured, unstructured and semi-structured data sources. These problems makes it necessary to have an automated analytical method that could deal with today's data and generate useful insights in a short time.

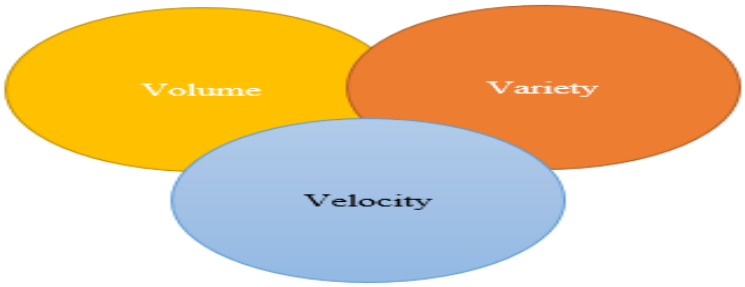

Figure 1.1 Today's dataset characteristics

Data mining provides the ability to discover hidden knowledge from data. It comprises different techniques which reveal the useful patterns that are implicit in the data [1-2]. Through data mining enable us to extract useful patterns and relationships between data using the association rules mining method. It also enables us to group similar data together using the clustering technique. In addition, data mining makes it possible for us to predict the future through analyzing current data [3], [4]. With data mining one can also analyze data in less time especially when compared to the time consumed to examine it manually. Consequently, data mining is revolutionary as it provides hope to develop automated tools to process data to generate useful knowledge [5].

Association rules mining is considered an important method of data mining and a well-studied problem as well. It is used to predict the occurrence of an item depending on the occurrence of another item [6], [7]. The problem originates in supermarket sales systems. For example, the rule {bread, cheese} -> {eggs} predicts that a customer is going to buy eggs in case he buys bread and cheese. Such rules are very important in empowering decision makers to take the right decision in promotional pricing and marketing. Besides market basket analysis, Association rules mining is used in different areas like detecting intrusion [8-10]; mining usage of the web [11-13]; medical diagnosis [14-16]; and others [17-19].

With the continual and constant increase of data, the generated association rules based on old data may become invalid and outdated. To maintain the existing rules, Association rules mining should be applied again on the whole dataset after every change. This process is called incremental association rules mining [20-24]. Researchers proposed different approaches for handling incremental association rules mining, however, performance is still a hot area of research. Some could handle incremental mining but require more than one database scan for original data in some cases like FUP and FUP2. Others methods like FP-Growth could take one scan but can't fit in memory. [25-30]

High performance computing is the use of parallel processing for running applications efficiently and quickly. The latest advances in technology are pushing towards parallel processing. Many computing cores could exist in single microprocessor. Current hardware clusters support hybrid memory nature with a distributed memory across all nodes and non-uniform memory access in each node. All cluster nodes are connected together through a high speed network. Various high performance programming modules that could be used in parallel processing has been introduced lately like MPI and OpenMP. Developing a hybrid data mining applications that could make use of cluster's hybrid memory nature is a promising field that could improve performance dramatically [31].

1.2 Problem Definition

The massive amount of data generated in the real-world applications has a significant effect on the speed of processing. Therefore, the speed of maintaining association rules decreases whenever data increases. There are multiple problems in improving association rules discovery and maintenance processes. The first problem is to reduce any need to reprocess the whole database whenever data changes. The second problem is to reduce the execution time for processing newly added dataset. The third problem is to allow incremental mining approach to process multiple data concurrently.

1.3 Motivation

Rapid development realms of parallel computing paradigms and parallel processing could have been more efficiently utilized to tackle the problem of incremental association rule mining in order to achieve much more advantageous operation.

1.4 Thesis Objectives

The objective of this thesis includes:

A) Providing a high performance distributed incremental association rules mining approach that can deal with today's dataset in an efficient and quick way.

B) Using high-performance computing environments to reduce processing time for incremental association rules mining.

C) Reducing the need to reprocess the whole database to maintain association rules.

1.5 Thesis Outlines

This thesis is structured as follows:

Chapter 1 is the Introduction to the thesis. Chapter 2 presents the background and literature review. Chapter 3 focuses on the proposed approach. Chapter 4 deals with experimental results and findings. Chapter 5 presents the Conclusion and future work.