1

**Chapter One**

Introduction

**1.1 Introduction**

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques [DANI05].

The amount of data stored in database continues to grow fast. Intuitively, this large amount of stored data contains valuable hidden knowledge which could be used to improve decision making process of an organization. For instance, data about previous sales might contain interesting relationship between products and customers. The discovery of such relationships can be very useful to increase the sales of a company. However, the number of human data analyst growth at much smaller rate than the amount of stored data. Thus, there is a clear need for automatic (or semi-automatic) methods for extracting knowledge from data. This need has led to the emergence of a field called data mining and knowledge discovery [WEIS98].

Modern humans find themselves living in an expanding universe of data in which there is too much data and too little information. The development of new techniques to find the required information from huge amount of data is one of the main challenges for software developers today.

The politicians, the managers, marketers, and any decision makers require the computer to deeply understand the data universe to extract the hidden and unintentionally constructed knowledge and information depending on new tools, especially after failure of conventional tools to mine and analyze such knowledge and information. Data mining is an approach to overcome these problems. It is the process of Knowledge Discovery from Databases (KDD).

The process of knowledge discovery (where data is transformed into knowledge for decision making) is data mining. The term is misnomer; mining gold from sands is referred to as gold mining and not sand mining. Thus, data mining should have been more appropriately named “Knowledge Mining”, but such a misnomer that carries both “data” and “mining” become a popular choice in information systems literature [MICH04].

Over the last four decades, Database Management System’s processed data by using the database technology available to users that supports query languages. The problem with query languages they are structural languages which assume that the user is aware of database schema. For example the query language “SQL” supports operation of selection from table or join related information from tables based on common fields. Today’s database users need functionality of consolidation, aggregation, and summarization of data, which require viewing the information along multiple dimensions that is what SQL is unable to do. The automatic Knowledge discovery tools have emerged in order to overcome this difficulty, and have taken the attention of the researchers of database literature. Knowledge Discovery in Databases includes all pre-processing steps in stored data, discovering interesting patterns on the data are referred to as Data Mining; which refers to the discovery of interesting patterns from data in knowledge discovery process. These interesting patterns may be in the form of associations, deviations, regularities, etc [HUSS02].

**1.2 Literature Survey**

The term Knowledge Discovery in Database “KDD” and Data Mining was first formally put forward by Usama Fayaad, (who began working in the field in 1989 at NASA’s Jet propulsion Lab, compiling data on astronomical phenomena.) at the first International conference on Knowledge Discovery and Data Mining, held in Montern in 1995[PAOL03].

The problem of discovering co-occurrence of an item in small data is a very simple task. However, the large volume of data makes this problem massively difficult and efficient algorithms are needed [RAKE94].

The problem of discovering association rule is decomposed into two parts: i) discovering all frequent patterns (represented by large itemsets or frequent itemsets), in the database, and ii) generating the association rules from those frequent itemsets. The second part is a straightforward problem and can be managed in polynomial time. On the other hand, the first task is very difficult especially for large databases. The typical example of association rule is the basket data analysis. In the basket database, all the records consist of two fields: Transaction ID (TID) and the item the customer bought in the transaction. Usually the transaction consists more than one item. An itemset is a set of items. It may be frequent, (large), or infrequent, (small). It is called frequent if the number of the occurrences of its items together in a database is greater then or equal to a use-defined threshold known as minimum support (minsup); otherwise it is called small or infrequent [RAKE94].

The current association rule mining algorithms [RAKE94, RAKE96a, OGIH97a, SEGE97, ROBE99, CHAR98, OGIH00, BUND01, HIPP01, and GOUD01] are iterative and use repeated scans on the database causing massive I/O traffic. Most of them use complex data structures such as hashing trees. Also, most of them work to mine frequent itemsets, except the algorithms presented in [GOUD01 AND BUND01] which work to mine maximal itemsets. Maximal itemsets is a frequent itemset and not subsets of other frequent itemsets. Usually the number of frequent itemsets is very large and increases with the decreasing of minsup value. There for, the mining of them is computational intensive process and has proved as NP-complete problem [OGIH98]. The number of maximal itemset is very small relative to the cardinality of frequent itemsets. There for, the algorithm that mine maximal itemsets are fast algorithms. But these algorithms are doomed to fail in the second phase of rule mining because it requires the maximal itemsets to be degenerated to there subsets [HIPP01, MICH02].

Another important defect with the existent algorithms for association rules mining is their insensitivity to the knowledge changes that happen when the database is updated. These algorithms must be re-run to discover the withered and emerging itemsets of the updated database. In addition to these drawbacks mentioned the available association rules algorithms consist of steps that require programming languages statement and facilitates. Such statements are not usually supported by DBMSs, Therefore the nature of algorithms is the cause of the bad integration of association rules mining and DBMSs[ABRA97, HIPP01, and RAME01].All the algorithms but DICT [SEGE97], can not be implemented by using DBMS tools such as SQL. DICT undergoes large and many problems with large and dense databases.

Recently data mining has been the province of high level specialist academic researchers and has featured complex custom programming and/or very high end software products, high cost and prolonged delivery schedules. Now, some front end tool venders are attempting to provide packaged commercial software products that combine simple capabilities with relatively easy to use GUI interfaces. These tools provide simpler capabilities instead of complex, esoteric ones. The idea is to get limited, but may be useful, data mining to a larger audience; the customer gets some of the benefits of data mining in a compressed time frame and at reduced cost [ODWH01]. Although many efforts have been put in the association area, and some commercial products, such as MineSet, intelligent miner, and Zhang tool have been developed. There are several limitations in these commercial products. Some of them are listed below:

* The existing miners manipulate data file in specific format, (usually from flat file) and they can not connect to multiple database management systems. For example the Intelligent Rule Miner can use the text file or the data of DB2 database as the data source [HUSS02].
* The available rule miners can only mine rules from one file or table. It is often required to combine data from more than one data source; therefore; it is important to develop miner having the ability to connect many files or tables to generate a suitable data set for mining [HUSS02].
* The available rule miners do not split the two phases: the frequent item generation and the rule generation and make one run for the mining. The separation of these two phases can reduce work time significantly by working at dead time in many cases.
* Because of the pruning process during the rule generation, there is no chance to find weak association rules instead of strong ones.

**1.3 Aim of the Study**

The aim of this study is to overcome some of the limitations and restrictions, which listed in the previous section, through proposing a new approach for mining association rules.

The proposed technique divides the running time of the association rule mining process into two parts first part runs at any time that there is no need for the server this part finalizes all process expensive tasks and other part runs at user query that not require many process time as a result users not wait too much for computer to answer queries. This is done by adding a layer between two processes which is output of the first process and second process uses this ready and easy data for finding association rules.

**1.4 Thesis organization**

This thesis is organized as follows: Chapter 1 of this thesis has been devoted to give a general definition to data mining and knowledge discovery with a brief historical survey. The rest of the thesis is organized as follows: Chapter 2 presents the fundamentals of data mining and association rules in databases. The main activities of data mining are explained and the principles and algorithms of association rules mining are presented. Chapter 3 describes the design of the Databot, which consists of the proposed algorithms and efficient methods to store and retrieve itemsets and their frequencies from database. Chapter 4 involves technical implementation by using java/MySql and algorithm verifications. Finally, conclusions and suggestions for future researches are given in chapter 5.