CHAPTER 1

INTRODUCTION

1.1 System Overview

Frequent Pattern Mining (FPM) is one of the most well-known techniques to extract frequent patterns from data. It plays an important role in association rule mining, finding correlations and trends etc. Finding Frequent Patterns becomes a very difficult task when they are applied to Big Data. Data storage has increased exponentially in the world over the past few years. Data coming from different sources such as web logs, machine logs, human-generated data, etc. are being stored by companies. This phenomenon is known as "Big Data" and nowadays it is trending everywhere. With the incredibly fast growth of data, comes the need to analyze the huge amount of data.

Due to lack of adequate tools and programs, data remains unused and underutilized because many important knowledge which is useful to the mankind remains hidden. Recent improvements in the field of parallel programming have provided good tools to tackle this problem. However, these tools come with their own technical challenges, e.g. balanced data distribution and inter-communication costs. Yang Qiang (2015) has explained, how Hadoop using MapReduce programming paradigm can be used to mine frequent patterns from Big data. The data in Hadoop Distributed File System is scattered and need lots of time to retrieve. So MapReduce a programming model, can be used for processing and generating large data sets with a parallel and distributed algorithm on a cluster.
Some of the real-life applications to which different Data Mining techniques can be applied are (i) forming groups of people based on their interests of the people or grouping similar constraints based on their properties (clustering), (ii) categorizing new insurers based on records of similar old claimants (classification), and (iii) detecting unusual credit transactions. Besides clustering, classification and anomaly detection, frequent pattern mining and association rule mining are also important because the latter two analyses valuable data (e.g., shopper market basket data) and help shop owners/managers by finding interesting or frequent itemsets that reveal customer purchase behavior.

An algorithm for mining customer transaction database item sets has been proposed by Han, Jiawei, Micheline Kamber, and Pei Jian (2006). Most of the algorithms for FPM can be grouped into two categories: Apriori-like algorithm and FP growth algorithm. Apriori generates frequent patterns by repeatedly scanning the database to prune candidate sets. Whereas FP growth algorithm generates frequent patterns by first constructing FP-Tree. In FP-Tree, transactions are stored in a tree structure in a compressed format. Then, using FP growth algorithm frequent itemsets are extracted from the database.

1.2 Introduction to Data Mining

Data Mining is a powerful new technology to extract hidden predictive information from large databases. It helps companies to focus on the most important information in their data warehouses. Data Mining tools predict future trends and behaviors with which business people can make proactive, knowledge-driven decisions. The automated, prospective analysis offered by Data Mining moved beyond the analysis of the past events provided by retrospective tools
similar to that of decision support systems. Data Mining tools can answer business questions that have traditionally been too time-consuming to resolve.

1.2.1 Foundations of Data Mining

Foundations of Data Mining means a systematic study of various notions that form its inherent hierarchical structure, from the basic concepts like data, objects, attribute/features, knowledge, etc. to the theories, algorithms for deriving knowledge from Data Mining algorithms and Data Mining process and to the evaluate and interpret the results. Data Mining techniques are the result of a long process of research and product development. This evolution of Data Mining began when business people started to store business information on computers, continued with improvements in data access, and more recently, advanced technologies that allowed people to navigate through their data in real time. Data Mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Huge data collection
- Powerful multiprocessor computers
- Algorithms for Mining Data

1.2.2 The Scope of Data Mining

Data Mining have gone through several research and developmental phases for many years. Statistics, Artificial Intelligence and Machine Learning are the three key areas with which Data Mining has attained its maximum growth. Data Mining is built on Statistics which is the foundation of most technologies, e.g. Standard Deviation, Standard Variance, Regression Analysis, Standard Distribution, Cluster Analysis, etc. Artificial Intelligence is also the base for Data
Mining, which tries to simulate human thought process or human intelligence in statistical problems. Another core area for Data Mining is Machine Learning and it is the combination of Statistics and Artificial Intelligence. Data Mining is the collection of historical and recent developments in Statistics, Machine Learning and Artificial Intelligence. These techniques are used to study and find hidden patterns or knowledge available in data. Also, Data Mining is being applied to areas such as information security and intrusion detection.

The name Data Mining has been derived from the similarities between searching for valuable business information in a large database, for example, finding linked products in gigabytes of store scanner data and mining a mountain for finding valuable ore. For both processes, it requires intelligently probing to find exactly where the value resides or either shifting through an immense amount of material. Data Mining technology can generate new business opportunities for a small or a big database. It provides the following capabilities:

**Automated prediction of trends and behaviors.** Data Mining is used to automate the process of finding predictive information from large databases. Traditionally to find an answer for a particular Question requires, extensive hands-on analysis. But now the answers can be found quickly and directly from the data. A typical example of a predictive problem is targeted marketing. Data Mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

**Automated discovery of previously unknown patterns.** Data Mining tools sweeps through the databases and identifies previously hidden
patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify the products which are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent the data entry keying errors.

Some successful application areas include:

- A car manufacturing company can analyze its recent sales force activity and their results to improve targeting of upper class people and determine which marketing activities will have the greatest impact in the next few months. The data needed includes competitor market activity as well as information about the residence of high society people. The results can be distributed to the sales force via a wide-area network that enables the sales branches to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.

- Pattern Discovery is another important application of Data Mining. In this, patterns that occur frequently in a database is found. The most well-studied type of patterns are sets of items that occur frequently together in transaction databases such as market basket logs of retail stores.

- A credit card company can use its customer transaction data to identify which customer can be inserted in the new credit card product. To get the attributes of the customers a small test mailing can be sent to them to identify the affinity with them towards the product. Recent projects have indicated that, more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.

- A diversified transportation company with a large direct sales force can apply Data Mining to identify the best prospects for its services. Data Mining can be used to analyze the customers' experience. A unique
segmentation can be build by identifying the attributes of high-value prospects. Applying this segmentation to a general business database can yield a prioritized list of prospects by region.

- Data Mining can be used in educational systems to bridge the knowledge gap between the students of different Universities. The hidden patterns and associations that are extracted from the mining process can improve the decision making processes in higher educational systems. This improvement can bring advantages like increasing the student’s promotion rate, improving the efficiency of the educational system, reducing the cost of system processes, etc.,

- By applying Data Mining, a large consumer package goods company can improve its sales process to retailers. With data collected from consumer panels, shipments, and competitor activity they can understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

Of all these applications, pattern discovery is very important for association rule mining. For example, association rules can be found for market basket or transaction data analysis, classification models can be mined for prediction, clusters can be identified for customer relation management, and outliers can be found for fraud detection.

### 1.3 Association rule mining

One of the fundamental methods from the prospering field of Data Mining is the generation of association rules that describe relationships between items in data sets. Association rule mining is primarily focused on finding frequent co-occurring associations among a collection of items. It is sometimes
referred to as “Market Basket Analysis”, since that was the original application area of association mining. The goal is to find associations of items that occur together more often than you would expect from a random sampling of all possibilities. Generally speaking an Association Rule is an implication of the form:

\[ X \rightarrow Y \]

Where \( X \) and \( Y \) are disjunct sets of items. The meaning of such rule is quite intuitive: Let \( DB \) be a transaction database, where each transaction \( T \in D \) is a set of items. An association rule \( X \rightarrow Y \) expresses that "Whenever a transaction \( T \) contains \( X \) then this transaction \( T \) also contains \( Y \) with probability \( conf \)." The probability \( conf \) is called the rule confidence and is supplemented by further quality measures like rule support and interest. The support is an indication of how the itemset appears frequently in the database. It is sometimes expressed as a percentage of the total number of records in the database. The confidence is an indication of how often the rule has been found to be true.

An Example for Association Rule Mining is identifying the items that occur frequently from a large transactional database. For this, association rule mining can be used, even if the customers who bought the items are unknown. An Association Rule Mining searches for interesting relationship among those items and displays it in a rule form. An association rule "\( \{bread, jam\} (sup = 2\%; conf = 80\%)\)" states that 2% of all the transactions under analysis show that bread and jam are purchased together and 80% of the customers who bought bread also bought jam. Such rules can be useful for decisions concerning product pricing, promotions, and many things. Association rules are also widely used in various areas such as telecommunication networks, market and risk management, inventory control, etc.
Phases of Association Rule Mining:

It consists of two phases:

- Finding all frequent patterns. By definition, each of these patterns will occur at least as frequently as a pre-defined minimum support threshold. Minimum Support threshold is the minimum support for an itemset to be identified as frequent.
- Generating association rules from frequent patterns. Association rules can be formed only by satisfying the pre-defined minimum support threshold and minimum confidence threshold.

The second phase is straightforward and less expensive. Therefore the first phase of FPM is a crucial step of the two and determines the overall performance of mining association rules. In addition to this, frequent pattern plays an essential role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules.

1.3.1 FPM

FPM means finding patterns (itemset, sequence, structure, etc.) that occurs frequently in a data set. FPM helps us to identify the relationships or correlations between items in the dataset. For example, a set of items, such as paint and brush, which appear frequently together in a transaction data set, is a Frequent Itemset. This information helps the shop keeper to arrange these frequent items together which will induce paint buyer to buy brush. Another example is Frequent Pattern discovery from Web Log data which helps to identify the navigational behaviors of the users. Consider the scenario, such as buying first a PC, then a Data Card, and then a Pen Drive, and if this pattern
occurs frequently in a shopping history database, then that pattern is a frequent sequential pattern. Types of FPM are shown in Figure 1.1.

**Figure 1.1 Types of FPM**

- **Sequential Pattern Mining:** It is concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. The mining process finds a frequent subsequences from a set of sequential data set, where a sequence records an ordering of events.

- **FIM:** Extracting sets of products that are frequently bought together. It aims at finding regularities in the shopping behavior of customers of supermarkets, mail-order companies, on-line shops, etc.
- **Structured Pattern Mining:** The mining process searches for frequent substructures in a structured data set. A structure is defined as a general concept that covers many structural forms, such as graphs, lattices, trees, sequences, sets, single items, or combinations of such structures.

The identification of sets of items, products, symptoms and characteristics, which often occur together in the given database, can be seen as one of the most basic tasks in Data Mining. So FIM is a very important mining of all the Pattern Mining types. That too, Frequent Itemsets from Big Data is a highly researched area.

**Kinds of FIM**

- **Constrained frequent itemset:** An itemset $X$ is a constrained frequent itemset in set $S$ if $X$ satisfy a set of user-defined constraints. A Naïve solution is to find all frequent sets and then test them for constraint satisfaction. However, the mining process can be done more efficiently by pushing the constraint as deeply as possible inside the frequent pattern computation.

- **Closed Frequent Itemset:** An itemset $X$ is closed frequent itemset in set $S$ if $X$ is both closed and frequent in $S$.

- **Maximal Frequent Itemset:** An itemset is maximal frequent if none of its immediate supersets are frequent. An itemset $X$ is a maximal frequent item in set $S$ if $X$ is frequent, and there exist no super-itemset $Z$, such that $XZ$ and $Z$ are frequent in $S$. 
- **Top \( k \)** frequent itemset: An itemset \( X \) is said to be top-\( k \) frequent itemset in set \( S \) if \( X \) is the \( k \) most frequent itemset for a user-specified value, \( k \).

- **Near-Match frequent itemset**: An itemset \( X \) is a near-match frequent itemset if \( X \) equals the support count of the near or almost matching itemsets.

### 1.4 Big Data

Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. The data can be a structured, semi-structured and unstructured data which can be mined for information. Here the definition uses size or volume of data as the only criterion. Another interesting definition of Big Data is that it is a technology designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. There are three main characteristics of Big Data: the data itself, the analytics of the data, and the presentation of the results of the analytics.

This definition is based on the 3Vs model coined by Doug Laney in 2001. He did not use the term Big Data but predicted that data management will get more and more important and difficult. He then identified the 3Vs which is shown in Figure 1.2 (data volume, data velocity and data variety) as the biggest challenges of data management. Data volume is the size of the data, data velocity is the speed at which data arrives and data variety is the data extracted from different sources which can be unstructured or semi-structured.
The size of the data is growing at a rapid speed nowadays. A text file is a few kilobytes, a sound file is a few megabytes while a full-length movie is a few gigabytes. These data comes from many sources. In the olden days all data are generated within the company itself by the employees. But the data are now generated by employees, partners and customers. In addition to this, machines also generate data for a company which has many branches. For example, hundreds of millions of smart phones send a variety of information to the network infrastructure. Handling such huge volume of data is obviously the most widely recognized challenge. Processing huge volume of data is not an issue if the data is loaded in bulk and there is enough processing time. But handling small amount of data gets problematic if the data is unstructured, arriving at high velocity and also need to be processed within seconds. From this it is very clear that it is not
the volume which decides whether the data is big but also other data characteristics like velocity and variety must also be considered.

**Data Velocity**

The speed at which data arrives is known as velocity. There are two folds of data velocity. First, the rate at which new data are flowing in and existing data is getting updated called the ‘acquisition rate challenge’. Second, the time acceptable to analyze the data and process while the data is on move, called ‘timeliness challenge’. The first problem is data acquisition rate. The challenges involved in this is of how to receive, filter, manage and store the continuously arriving data. Traditional relational database systems are not suitable for this task as they process many overhead in the form of locking, logging, buffer pool management and formulating threaded operations. One way to handle this problem is to analyze the flow of data for anomalies and discard unnecessary data by filtering and only store important data. This filtering of data stream without missing any important data not only requires an intelligent tool, but also consumes time and resources. Also it is not always possible to filter data. Another necessary task is to automatically extract and store metadata together with the streaming data. This is useful to track how data is stored and measured.

The second problem is regarding reaction to incoming data streams. In many situations real time analysis becomes necessary, otherwise the information gets useless. As mentioned earlier, it is not only sufficient to analyze the data and extract information in real time but also necessary to react on it and apply. The speed of handling the whole case becomes the decisive issue. For mining of data streams not only speed is important but also the time interval of the actual processing of data. This is very important with respect to data streams with characteristic feature of evolving over time.
Data Variety

Variety refers to the diversity of data sources. Data comes from different sources in various types such as web data, social media, machine generated logs, human-generated data, biometrics, transactional data, etc. This not only implies an increased amount of data sources but also structural differences among those data sources. Furthermore, the structure or schema of different data sources is not necessarily compatible and also the semantics of data can be inconsistent. Therefore, managing and integrating of multi-structured data from a wide variety of sources poses many challenges. First comes the storage and management of this data in a database like systems. Relational database management systems may not be suitable for all types and formats of data. The next challenge is related to the semi and fully unstructured character of data. In the context of integrating different data sources, different data, be it structured, unstructured or semi-structured needs to be transformed to some structure or schema that can be used to relate different data sources.

1.5 Need for the study

FPM has proved to be one of the promising fields in carrying out the research work because of its wide use in all Data Mining tasks such as clustering, classification, prediction and association analysis. Mining frequent itemsets enables humans to take better decisions in a wide range of applications including market basket analysis, traffic signals analysis and in Bioinformatics identify frequently co-occurring protein domains in a set of proteins. Many researches have proposed many algorithms to generate FIM, but the execution time and storage space plays a key difference in all these algorithms. Pruning unimportant patterns becomes another important research area in FIM.
Now it is an era of Big Data. There are some applications where frequent patterns have to be extracted from Big Dataset. One such example is Web Log Mining, which helps us to identify frequent web pages visited by the user. By using this information one can improve their advertising process. To handle Big Dataset parallel mining becomes necessary for which MapReduce concept can be used.

1.6 Problem Statement

Generating all frequent itemsets is typically very large which some applications do not require. The subset that is really needed by these applications usually contains only a small number of itemsets. Thus more time is spent in considering all unwanted frequent itemsets to extract frequent itemsets. In addition to this, memory is also wasted in storing all unimportant frequent itemsets. So constraints can be introduced to remove these unimportant itemsets. All the existing algorithms hold good only when the dataset is small. So there is a need to propose an efficient algorithm to find frequent itemsets from Big Dataset using constraints. In almost all FPM algorithms, Frequent 1-itemsets are generated to find the support count (occurrences) of each item in the entire database. This task is itself a tedious task in generating Frequent itemsets when considering the hugeness of modern databases available. No explicit strategy has been outlined in these algorithms to perform the aforesaid task. So an efficient data structure can be proposed to find the support count of each item.

1.7 Objectives of the Study

The major objectives of this study are as follows:

- To extract Frequent Itemsets from Big Dataset
- To reduce the memory wastage using Constraints
To speed up the execution time using support count tree and cache
To extract cumulative Frequent itemsets from multiple files

1.8 Methodology of the Study

In this research work, the paradigm of Constraint-based Itemset Mining in Big Data is introduced. In order to extract Frequent Itemsets from Big data, MapReduce function in Hadoop is used. Constraints provide focus only on the interesting or required data, thus reducing the number of patterns extracted to those of potential interest. To extract Frequent-1 itemsets an efficient Support count tree algorithm has been embedded in the FP growth algorithm to mine Frequent Itemsets from big data set. To still more increase the efficiency of MapReduce a Modified MapReduce algorithm has been proposed. In this algorithm cache has been included in the Map phase to maintain support count tree for calculating the frequent-1 itemset of each mapper. This reduces the total time of calculating Frequent-1 itemsets since it bypasses the shuffle, sort and the combine task of each Mapper in the original MapReduce tasks. This in-turn reduces the execution time of generating Frequent Itemsets of the entire database.

1.9 Organization of the Thesis

A concise outline of various chapters of the thesis is as follows:

Chapter 1 deals with the introduction to Data Mining, Association rule mining and Big Data. An overview of the system, need, objective and methodology of the study are given. In this chapter organization of the thesis is also presented.
In **Chapter 2**, an in-depth analysis is made on most influential algorithms which have given significant contributions to several efficiency issues of FPM problems in Big Data. In addition to this literature survey on FPM using constraints is also dealt.

**Chapter 3** introduces the concept of FIM, its notations and methodologies. FIM Methods based on layout of data (horizontal as well as vertical) are described in this section. An algorithm of each basic category of FIM methods has been explained here such as Apriori, FP growth, Eclat etc. Description of the proposed improved FP growth algorithm. Experimental Result of the proposed work is shown.

**Chapter 4** describes the need for constrained pattern mining and also describes different types of constraints for FIM. Constrained FIM in Big Data has also been explained.

**Chapter 5** deals with description of uses of cache and how it can be used in Modified MapReduce algorithm for FIM in Big Data. Experimental Results of the proposed work is shown.

In **Chapter 6**, conclusions of the research work are given and scope for future study is indicated.