Distributed Frequent Itemset Mining

A dissertation Submitted in partial Fulfilment of the requirements for
MSc degree in computer science

By
Huda Jamal Abdel-hammed Musa

Supervisor
Dr. Mohamed Elhafiz Mustafa Musa
May 2014
DEDICATION

To my family

To my teachers

To my husband

To my best friends
ACKNOWLEDGMENT

First and foremost, I would like to express my sincere thanks to my thesis supervisor Dr. Mohammed Elhafiz Mustafa for providing me their precious advices and suggestions. This Thesis wouldn’t have been a success for me without their comments and suggestions.

Next, I would like to express my family: my father Jamal Abdelhammed Musa, my mother Elham Abdulla Abbas, my brothers and my sisters without their support I would never had dreamt of pursuing higher studies.

Also I would like to express my husband Mustafa Mohammed Ahmed for their unconditional love and support in every part of my life.

Special respect and thanks to my teacher Wafa Fisal for support and help.

Also I would like to thank Sudan University of Science and Technology for providing me such a graceful opportunity to become a part of its family.

Lastly I would like to thanks all the persons those are related to the thesis directly or indirectly.
ABSTRACT

Association rule mining is an important technique to discover hidden relationships among items in the transaction. The problem is that association rules are generated by first mining of frequent itemsets in distributed datasets does not gain the best and most accuracy rules. The goal of the thesis is to experimentally finding the most frequent itemsets from distributed data sources which is first phase of association rules generation. Firstly, the global frequent itemset are generated from global dataset. Secondly, the global dataset are divided into three sites, and then generating the local frequent itemsets from each site. A comprehensive search for the best way to combine the local itemset has been conducted. In this search we find that the union of smallest and biggest of itemsets intersected with the middle always gives result which is equivalent to global itemsets.
المستخلص

(association rules) يعتبر أسلوباً مهماً لإكتشاف العلاقات الخفية بين العناصر. تكمن المشكلة في أن قواعد الربط المستخرجة من أول عملية تكمن في العينات الموزعة لا تعطي القواعد المراد تكوينها بصورة دقيقة. الهدف من هذه الدراسة هو إيجاد مجموعة العناصر الأكثر تكراراً من مصادر البيانات الموزعة، والتي تعتبر المرحلة الأولى من مراحل استخراج قواعد الربط. أولاً يتم توليد مجموعة العناصر الأكثر تكراراً لمجموعة البيانات الشاملة، ثانياً يتم تقسيم مجموعة البيانات الشاملة إلى ثلاثة مجموعات جزئية، ومن ثم يتم توليد مجموعة العناصر المكررة في كل مجموعة. في هذه الدراسة تم إجراء بحث شامل عن أفضل طريقة لدمج مجموعات العناصر المكررة من كل المجموعات، فوجدنا أن إتحاد أصغر مجموعة عناصر (Smallest local frequent itemsets) وتقاطع مجموعة العناصر الوسطى (Middle local frequent itemsets) لعناصر دائماً تعطي عناصر متكررة مكافئة لمجموعة العناصر المكررة في المجموعة الشاملة. إنها تعطي عناصر متكررة مكافئة لمجموعة العناصر المكررة في المجموعة الشاملة. إنها تعطي عناصر متكررة مكافئة لمجموعة العناصر المكررة في المجموعة الشاملة.
# Table of Contents

DEDICATION ..................................................................................................................................... I

ACKNOLEDGMENT ................................................................................................................ III

ABSTRACT ....................................................................................................................................... IV

المستخلص ....................................................................................................................................... V

Table of Contents ........................................................................................................................ VI

List of Figures ................................................................................................................................ VIII

List of Tables ................................................................................................................................ IX

Chapter 1 - Introduction .................................................................................................................. 1

1.1 Introduction .............................................................................................................................. 1

1.2 Problem ..................................................................................................................................... 1

1.3 Objectives .................................................................................................................................. 1

1.4 Methodologies .......................................................................................................................... 2

1.4.1 Local Rules Generating ........................................................................................................ 2

1.4.2 Global Rules Refining ............................................................................................................ 2

1.5 Scope .......................................................................................................................................... 2

1.6 Thesis Structure ....................................................................................................................... 2

Chapter 2 - Association Rule Mining ............................................................................................. 3

2.1 Introduction ............................................................................................................................... 3

2.2 Data Mining .............................................................................................................................. 3

2.2.1 Backgrounds ......................................................................................................................... 3

2.2.2 Data Mining Tasks .................................................................................................................. 4

2.3 Association Rules Mining ......................................................................................................... 4

2.3.1 Parallel ARM ......................................................................................................................... 5

2.3.2 Distributed ARM .................................................................................................................. 5

2.3.2.1 FDM (Fast Distributed Mining of association rules) ....................................................... 6

2.4 Distributed Data Mining (DDM) .............................................................................................. 6

2.5 Literature Review ...................................................................................................................... 8
List of Figures

Figure 2.1: Data Mining Tasks ................................................................. 4
Figure 2.2: A DDM Framework ................................................................. 8
Figure 2.3: Apriori Example ................................................................. 11
Figure 3.1: The Proposed System Structure ........................................... 13
Figure 3.2: Data Set at Site1 ................................................................. 16
Figure 3.3: Visualize Attributes at Site1 .................................................. 17
Figure 3.4: The Local Frequent itemsets and Association Rules at Site1,Site2,Site3... 18
Figure 3.5: The Details of Frequent Itemsets ......................................... 19
List of Tables

Table 3.1: Dataset Statictics.................................................................14
Table 3.2: Dataset Attribute Description..............................................15
Table 3.3: Global CENSUS Dataset.......................................................17
Table 3.4: CENSUS Data Set (3Sites)..................................................18
Table 3.5: The Local Frequent Itemsets in Site1, Site2, Site3.....................20
Table 3.6: Union all and Intersect all local frequent itemsets.....................20
Table 3.7: The Proposed Rule...............................................................21
Table 3.8: Rules applied over datasets ...............................................21
Table 3.9: Differences of Frequent Itemsets .......................................22
CHAPTER 1
INTRODUCTION
Chapter 1

1.1 Introduction

Association rule mining (ARM) is an active data mining research area. Most ARM algorithms focus on sequential or centralized environments where no external communication is required. Although nowadays there is huge data in distributed database and no standard approach to build efficient association rule mining in these data.

1.2 Problem

Modern organizations are geographically distributed. Typically, each site locally stores its ever-increasing amount of day-to-day data. Using centralized data mining to discover useful patterns in such organizations' data isn't always feasible because merging datasets from different sites into a centralized site incurs huge network communication costs. Data from these organizations are not only distributed over various locations but also vertically fragmented, making it difficult if not impossible to combine them in a central location. Most Distributed Association rule mining (DARM) algorithms don't have an efficient message optimization technique, so they exchange numerous messages during the mining process. Distributed data mining has thus emerged as an active sub-area of data mining research.

1.3 Objectives

Distributed ARM system aims to generate rules from different database spread over various geographical sites. Hence, they require external communications throughout the entire process. DARM algorithms must reduce communication costs so that generating global association rules costs less than combining the participating sites' datasets into a centralized site.

1.4 Methodologies

We have two main steps:

1.4.1.1 Local Rules Generating

Each site generates the frequent itemsets. Then it will be used to generate association rules that satisfy minimum confidence.
1.4.2 Global Rules Refining
After generating the local frequent itemsets and the rules at each site, generates
the globally frequent itemsets.

1.5 Scope
We can generate Association Rule from any datasets that distributed among
various sites to discover the most frequent itemsets.

1.6 Thesis Structure
This thesis contains four chapters as follows:
Chapter 2 presents the background of the data mining. It covers in detail about the data
mining, association rule mining and distributed association rule mining. In addition
discusses some related works on distributed association rule mining. Chapter 3 contains proposed system, the experiments and the results. Chapter 4
discusses the conclusion and Future work.
CHAPTER 2
ASSOCIATION RULE MINING
Chapter 2
AssociationRuleMining

2.1 Introduction
In this chapter a background of data mining and association rule mining is discussed. This chapter also covers in detail about the distributed association rule mining algorithms. Also it provides of some related works.

2.2 Data Mining
2.2.1 Backgrounds
There are basically two most important reasons that data mining (DM) has attracted a great deal of attention in the recent years. First, our capability to collect and store the huge amount of data is rapidly increasing day by day. The second reason is the need to turn such data into useful information and knowledge. The knowledge that is acquired through the help of data mining can be applied into various applications like business management, retail and market analysis, engineering design and scientific exploration.[1]

There are many definitions for data mining:

- Data mining (sometimes called data or Knowledge Discovery in Database KDD) is the process of analyzing data from different perspective and summarizing it into useful information. [3]
- Data mining or Knowledge Discovery in Database (KDD) is a collection of exploration techniques based on advanced analytical methods and tools for handling large amount of information. [5]

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among many fields in large databases. Data mining tools and techniques are used to generate information from the data that we have stored in our repositories over the years.
2.2.2 Data Mining Tasks
The process of mining is often controlled by the requirements of the users. The user may be a business analyst or may be a marketing manager. Different users have different need of information. Depending on the requirements we can use different data mining tasks.[2]

![Data Mining Tasks](image)

**Figure 2.1: Data Mining Tasks**

2.3 Association Rules Mining
Association rule mining is an interesting data mining technique. That is used to find out interesting patterns or associations among the data items stored in the database. Support and confidence are two measures of the interestingness for the mined patterns.

Databases or data warehouses may store a huge amount of data to be mined. Mining association rules in such databases may require substantial processing power. A possible solution to this problem can be a distributed system. Moreover, many large databases are distributed in nature which may make it more feasible to use distributed algorithms. Major cost of mining association rules is the computation of the set of large itemsets in the database. Distributed computing of large itemsets encounters some new problems. One may compute locally large itemsets easily, but a locally large itemsets may not be globally large. [2]

Many parallel or distributed ARM algorithms were designed for shared memory parallel environments. Based on the nature and implementation of each algorithm, we can divide the existing algorithms into two groups: parallel ARM and DARM.
2.3.1 Parallel ARM

We can categorize parallel ARM algorithms as data-parallelism or task-parallelism algorithms. In the former, the algorithms partition the datasets among different nodes; each site performs the task independently but must access the entire dataset. [5]

The main challenges associated with parallel data mining include minimizing I/O, minimizing synchronization and communication, effective load balancing, effective data layout, deciding on the best search procedure to use. The parallel algorithms are Count Distribution, Candidate Distribution and Hybrid Count and Candidate Distribution. [6]

2.3.2 Distributed ARM

DARM discovers rules from various geographically distributed datasets. However, the network connection between those datasets isn't as fast as in a parallel environment, so distributed mining usually aims to minimize communication costs.

Distributed ARM algorithms involve distributed association rule learning, collective decision tree learning, distributed hierarchical clustering, other distributed clustering algorithms, collective Bayesian network learning, collective multi-variate regression. [7]

2.3.2.1 FDM (Fast Distributed Mining of association rules):

FDM mine rules from distributed datasets partitioned among different sites. In each site, FDM finds the local support counts and prunes all infrequent local support counts. After completing local pruning, each site broadcasts messages containing all the remaining candidate sets to all other sites to request their support counts. It then decides whether large itemsets are globally frequent or not. Then generates the candidate itemsets from those globally frequent itemsets. [5]

Generally FDM has the following distinct features:
1. Some relationships between locally large sets and globally large ones are explored to generate a smaller set of candidate sets at each iteration and thus reduce the number of messages to be passed.
2. After the candidate sets have been generated, two pruning techniques, local pruning and global pruning, are developed to prune away some candidate sets at each individual site.

3. In order to determine whether a candidate set is large, this algorithm requires $O(n)$ messages for support count exchange, where $n$ is the number of sites in the network.

### 2.4 Distributed Data Mining (DDM)

When data mining is undertaken in an environment where users, data, hardware and the mining software are geographically dispersed, it is called distributed data mining. Thus distributed data mining refers to the mining of distributed datasets. The datasets are stored in local databases hosted by local computers which are connected through a computer network. Data mining takes place at a local level and at a global level where local data mining results are combined to gain global findings. Distributed data mining is often mentioned with parallel data mining in literature. While both attempt to improve the performance of traditional data mining systems they assume different system architectures and take different approaches. In distributed data mining computers are distributed and communicate through message passing. In parallel data mining a parallel computer is assumed with processors sharing memory and or disk. Computers in a distributed data mining system may be viewed as processors sharing nothing. This difference in architecture affected in algorithm design, cost model, and performance measure in distributed and parallel data mining. Typically, such environments are also characterized by heterogeneity of data and multiple users. DDM offers techniques to discover knowledge in distributed data. [3] A typical DDM framework is shown in figure 2.2.
2.5 Literature Review

In Market Basket Analysis If we think of the universe as the set of items available at the store, and then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules.

For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule 2.1 below:

Computer => antivirus software [support = 2%; confidence = 60%](2.1)

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules.
A support of 2% for Association Rule (2.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.

A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. [7]

2.5.1 Frequent Itemsetsand Association Rules

Let I= \{I_1, I_2...I_m\} be a set of items. Let D, the task-relevant data, is a set of database transactions where each transaction T is a set of items such that T is in I. Each transaction is associated with an identifier, called TID.

An association rule is an implication of the form A => B, where A is in I, B is in I, and A and B are disjoint. The rule A => B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A union B. This is taken to be the probability, P(A union B).

The rule A => B has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A). That is,

\[
\text{Support}(A \Rightarrow B) = P(A \cup B) \quad (2.2)
\]
\[
\text{Confidence}(A \Rightarrow B) = P(B|A) \quad (2.3)
\]

Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min conf) are called strong.

From Equation (2.3) we have:

\[
\text{Confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{support}(A \cup B)}{\text{support}(A)} \quad (2.4)
\]
2.5.2 Apriori Algorithm

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Elk requires one full scan of the database.[1]

2.5.2.1 Example

Let Set of items: I= {1, 2, 3, 4, 5}.
Transactions: D = {t100, t200, t300, t400}.
Support of an itemset: Percentage of transactions which contain that itemset.
Large (Frequent) itemset: Itemset whose number of occurrences is above a threshold.
2.6 Related Works

Many algorithms have been proposed to find frequent itemsets from a very large datasets. The number of datasets scans required for the task has been reduced from a number equal to the size of the largest itemsets in Apriori, to typically just a single scan in modern ARM algorithms such as Sampling. When data is saved in a distributed datasets, a distributed data mining algorithm is needed to mine association rules. It has been addressed by some researches and number of distributed algorithms has been proposed. [4]

The partition algorithm is based on apriori algorithm. It consists of two phases. Firstly partitions the data into a number of non-overlapping partitions. For each partition, all frequent itemsets are found. These are referred as local frequent itemsets. A local frequent itemset may or may not be frequent with respect to the entire dataset D. Any itemset that is potentially frequent with respect to D must occur as a frequent itemset in at least one of the partitions. Therefore all local frequent itemsets are candidate itemsets with respect to D. The collection of frequent itemsets from
allpartitions forms the global candidate itemsets with respect to D. Finally the algorithm union all the local frequent itemsets to generate global frequent itemsets. It reduces the number of complete database scans up to two and hence improves the performance of mining algorithm. [10].

Sampling algorithm (mining on a subset of a given data) is also based on apriori algorithm. The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D. In this way, we trade off some degree of accuracy against efficiency. The sample size of S is such that the search for frequent itemsets in S can be done in main memory, and so only one scan of the transactions in S is required overall [10]. Sampling can reduce I/O costs by drastically shrinking the number of transactions to be considered. It can speed up the mining process by more than an order of magnitude. In another hand, because we are searching for frequent itemsets in S rather than in D, it is possible that some of the global frequent itemsets was missed. [15]

E. Ansari, G.H. Dastghaibifard, M. Keshtkaran, H. Kaabi presented a new distributed Trie-based algorithm (DTFIM) to find frequent itemsets. This algorithm is proposed for a multi-computer environment. They added an idea from FDM algorithm for candidate generation step. The point of this algorithm is that every site keeps a copy of Trie locally, and they synchronize their data so that all local Trie copies are the same at the end of each stage. After local support is counted, all sites share their support counts and determine the global support counts, in order to remove infrequent itemsets from their local Trie. These results show Trie data structure can be used for distributed association rule mining not just for sequential algorithms. [12]
CHAPTER 3

PROPOSED SYSTEM
Chapter 3

Proposed System

3.1 Introduction

This chapter describes the proposed system, the dataset, the experiments and the results. Figure 3.1 is an overview of the proposed system for distributed association rules mining. The chapter also reports and discusses the experiments’ results.

![Diagram of the Proposed System Structure]

Figure 3.1: The Proposed System Structure

3.2 Experiments Dataset

The dataset have been downloaded from the university of California site (UCI); this data was extracted from the census income database. It goal is to predict whether income exceeds 50,000$/year. Table 3.1 summarizes details of the dataset and Table 3.2 describes dataset attributes.
Table 3.1: Dataset Statistics

<table>
<thead>
<tr>
<th>Dataset Characteristics:</th>
<th>Multivariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Characteristics:</td>
<td>Categorical, Integer</td>
</tr>
<tr>
<td>Number of Instances:</td>
<td>48843</td>
</tr>
<tr>
<td>Number of Attributes:</td>
<td>13</td>
</tr>
<tr>
<td>Area:</td>
<td>Social</td>
</tr>
<tr>
<td>Date Donated</td>
<td>1996-05-01</td>
</tr>
<tr>
<td>Attribute</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Age</td>
<td>Continuous</td>
</tr>
<tr>
<td>Education</td>
<td>Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool</td>
</tr>
<tr>
<td>Martial-Status</td>
<td>Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse</td>
</tr>
<tr>
<td>Relationship</td>
<td>Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried</td>
</tr>
<tr>
<td>Race</td>
<td>White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black</td>
</tr>
<tr>
<td>Sex</td>
<td>Female, Male</td>
</tr>
<tr>
<td>Gain</td>
<td>Continuous</td>
</tr>
<tr>
<td>Loss</td>
<td>Continuous</td>
</tr>
<tr>
<td>Hours-per-week</td>
<td>Continuous</td>
</tr>
<tr>
<td>Country</td>
<td>United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&amp;Tobago, Peru, Hong, Holland-Netherlands</td>
</tr>
<tr>
<td>Salary</td>
<td>&gt;50K, &lt;=50K</td>
</tr>
</tbody>
</table>
3.3 Preprocessing Stage

We use WEKA tool to generate the association rules from sites. The proposed system is divided into two phases. First generate local frequent itemsets for each site. Second Local frequent itemsets from each site are combined to generate global frequent itemsets.

3.3.1 Generate local frequent itemsets

At each site we apply apriori algorithm. Figure 3.2 shows the datasets uploaded in WEKA and figure 3.3 shows the association rule that collected from the dataset.

![Weka Explorer](image)

**Figure 3.2: Dataset at Site1**

Attributes at Site1 are visualized in Figure 3.3.
3.3.2 Generate Global frequent itemsets

After generates the local frequent itemsets from each site, we combined them to generates the most frequent itemsets. Table 3.3 shows the total number of records, minimum support and frequent itemsets for global CENSUS dataset.

<table>
<thead>
<tr>
<th></th>
<th>Total Rows</th>
<th>MinSup</th>
<th>Frequent Itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENSUS</td>
<td>48,843</td>
<td>0.2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3.4 shows the results after divided the CENSUS dataset into 3 sites.

<table>
<thead>
<tr>
<th></th>
<th>Total Rows</th>
<th>MinSup</th>
<th>Frequent Itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site1</td>
<td>16,280</td>
<td>0.2</td>
<td>3</td>
</tr>
<tr>
<td>Site2</td>
<td>16,280</td>
<td>0.2</td>
<td>4</td>
</tr>
<tr>
<td>Site3</td>
<td>16,280</td>
<td>0.2</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 3.4 and Figure 3.5 shows the result in details.

Figure 3.4: The Local Frequent Itemsets and Association Rules in Site1, Site2, Site3
Figure 3.5: The Details of Frequent Itemsets
3.4 The Results

To generate the most frequent itemset firstly we divide the datasets into 3 sites S1, S2, S3. Then we generate the large itemsets from each site L1, L2, L3. Lastly we combine the large itemsets by using the proposed rule. Table 3.5 shows the total number of records and frequent itemsets at each site in CENSUS, CAR, NURSERY, SAMPLE_MODELING datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>No of Records</th>
<th>Global Frequent Itemset</th>
<th>Local Frequent Itemsets in Site1</th>
<th>Local Frequent Itemsets in Site2</th>
<th>Local Frequent Itemsets in Site3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENSUS</td>
<td>48,000</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>CAR</td>
<td>1728</td>
<td>11</td>
<td>16</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>NURSERY</td>
<td>12960</td>
<td>16</td>
<td>12</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>SAMPLE_MODELING</td>
<td>75,000</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Union and intersection the local frequent itemsets was applied to generate the most frequent itemset, union gives frequent itemset greater than the actual frequent itemset. And intersection gives frequent itemset less than the actual frequent itemset. See table 3.6.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Global frequent itemsets</th>
<th>Union all</th>
<th>Interest all</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENSUS</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>CAR</td>
<td>11</td>
<td>64</td>
<td>0</td>
</tr>
<tr>
<td>NURSERY</td>
<td>16</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>SAMPLE_MODELING</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Because this problem we proposed a rule to generate frequent itemset that equal the actual frequent itemset in global dataset. See table 3.7.

<table>
<thead>
<tr>
<th>No of Distributed Datasets</th>
<th>Large Itemsets</th>
<th>Mining Frequent Itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>L1,L2,L3</td>
<td>(L1∪L3)∩L2</td>
</tr>
</tbody>
</table>

Such that L1 is the large itemset in site1, L2 is the large itemset in site2 and L3 is the large itemset in site3.
We union the maximum number of large itemset (L3) and the minimum number of large itemset (L1), then intersect the result with the third large itemset (L2) to generates the most frequent itemset.

\[(\text{MAX large itemset} \cup \text{MIN large itemset}) \cap \text{Third large itemset} \].

Table 3.8 shows the results of frequent itemsets after apply the proposed rule.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Global frequent itemsets</th>
<th>Union all</th>
<th>Interest all</th>
<th>(L1∪L3) \cap I2</th>
<th>(L1∩L2) \cup I2</th>
<th>(L1∩L2) \cup I3</th>
<th>(L1 \cup I2) \cap I3</th>
<th>(L2 \cup I3) \cap I1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENSUS</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CAR</td>
<td>11</td>
<td>64</td>
<td>0</td>
<td>11</td>
<td>16</td>
<td>30</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>NURSERY</td>
<td>16</td>
<td>20</td>
<td>11</td>
<td>16</td>
<td>29</td>
<td>18</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>SAMPLE</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

In all datasets the proposed rule generate the truth frequent itemsets. But other rules generate number of frequent itemsets greater or less than the actual frequent itemsets. Table 3.9 shows the differences of results.

Table 3.9: Differences of Frequent Itemsets, +I means combined local itemsets is greater than global itemsets by I and –I means less by I. while Ø indicate same numbers

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Union all</th>
<th>Interest all</th>
<th>(L1∪L3) \cap I2</th>
<th>(L1∩L3) \cup I2</th>
<th>(L1∩L2) \cup I3</th>
<th>(L1 \cup I2) \cap I3</th>
<th>(L2 \cup I3) \cap I1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENSUS</td>
<td>+1</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>+1</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>CAR</td>
<td>+48</td>
<td>-16</td>
<td>Ø</td>
<td>Ø</td>
<td>+14</td>
<td>+5</td>
<td>-16</td>
</tr>
<tr>
<td>NURSERY</td>
<td>+9</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>+17</td>
<td>+7</td>
<td>-5</td>
</tr>
<tr>
<td>SAMPLE</td>
<td>Ø</td>
<td>-4</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>-4</td>
</tr>
</tbody>
</table>
CHAPTER 4
CONCLUSION AND FUTURE WORKS
Chapter 4
Conclusion and Future Work

4.1 Conclusion

In this thesis we have discussed a new approach to obtained frequent itemsets from distributed data sources. Firstly, we generate the global frequent itemset from global dataset. Secondly, we divide the global dataset into three sites, and then we generate the local frequent itemsets from each site. A comprehensive search for the best way to combine the local itemset has been conducted. In this search we find that the union of smallest and biggest of itemsets intersected with the middle always gives result which is equivalent to global itemsets. The experiment of this thesis has been conducted on four different datasets. These datasets have different sizes and attribute types.

4.2 Future Work

Some of the future work that could done to find more result on the topic of this thesis could be:

- Doing more experiments for more than 3 sites with different sizes.
- Generating a tool that allows users to obtained frequent itemsets from distributed datasets. And embedding this tool in one of the famous data mining software like weka.
- Suggesting a way for generate the global frequent itemsets from datasets that are not uniformly distributed.
References


[10] Jiawei Han und Micheline Kamber, “Data Mining – Concepts and Techniques,” Chapter 5.2.


Appendix
Appendix

1. CENSUS Dataset

2. Sample_Modeling Dataset
3. Nursery Dataset

|   | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
| 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 3 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 7 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 8 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 9 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 10|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 11|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 12|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 13|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 14|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 15|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 16|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 17|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 18|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 19|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

*Note: The table above contains data from the Nursery Dataset, showing various attributes such as form, child care, housing, finances, and health, along with recommendations for nursery placement.*