Association Rule Mining Using *k*-Map Model in Data Mining

Daban Abdulsalam Abdullah

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the Degree of

Master of Science
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
July 2015
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Serhan Çiftçioğlu
Acting Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Prof. Dr. Nazım Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Asst. Prof. Dr. Ersin Kuset Bodur
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev

2. Asst. Prof. Dr. Ersin Kuset Bodur

3. Asst. Prof. Dr. Yücel Tandoğdu

ABSTRACT

In data mining, many algorithms were suggested to define the frequent rules within the data set. One of the problems is to choose a correct algorithm for the problem and the determination of the efficiency of the algorithm has important role during the investigation of hidden knowledge.

The thesis describes how to handle data set with Association Rules Analysis/ Market Basket Analysis with the popular Apriori algorithm and k-Map algorithm of data mining. The goal of this thesis is to find the most frequent patterns within the data set and then using different measurements to do further investigation on the obtained frequent patterns.

Keywords: Data Mining, Association Rules Analysis, Market-Basket Analysis

ÖZ

Veri madenciliğinde, anlamlı kurallar tanımlamak için bir çok algoritma önerilmiştir. Doğru algoritmayı seçmek ve algoritmanın kullanırlığının kararı verinin içindeki gizli bilginin bulunması için önemli problemlerdir.

Bu tez verinin Birliktelik Kuralları Analizinde sıklıkla kullanılan Apriori algoritması ve *k*-Harita (Karnaugh Haritası) algoritmasının nasıl kullanılacağını tanımlar. Bu tezin amacı verinin içindeki anlamlı kuralları bulmak ve sonrasında ise farklı ölçüler kullanıp anlamlı kurallar için ileri analizler yapmaktır.

Anahtar kelimeler: Veri Madenciliği, Birliktelik Kuralları Analizi, Sepet Analizi

To my family, and to my brother…

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor Asst. Prof. Dr. Ersin Kuset Bodur. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly.

I would also like to thank all the members of staff at Eastern Mediterranean University who helped me in my supervisor's absence.

I also like to thank to Professor Dr. Rashad Aliev for his support during my study in numerous occasions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Information technology plays an important role in all fields in human life. Collecting the data from different resources the storage stage may become the important part of database to turn the raw data to information by the way the knowledge can be used for different purposes. But sometimes we have very huge data from different registers, at this point, data mining is essential to discover the information from data.

During the knowledge discovery process, different tools of data mining and their improvements are discussed by the author in [1]. The most used algorithms are announced by IEEE, International Conference on Data Mining in 2006 which are C4.5, k-Means, SVM, Apriori, EM, Page Rank, Naïve Bayes and CART in [2].

Association rule generates the connection between different components to analyze the exhibit scenario, relation or relationship with those patterns (also called as itemsets) with specific goal to find out knowledge or to analyze data. Information is the pattern which can be utilized to build or improve data or knowledge. Association rule mining is the scientific method to discovery interesting frequent association rules within data to reach the information, [3].

Mining process focuses on finding important, useful relationships within large data that provides assistance in decision making. Different algorithms were designed to

identify frequent patterns in effective ways which are Apriori, FP-growth and Eclat. Different researchers did various surveys about association rules to obtain useful information of data by proposing different algorithms, [4].

Different types of algorithms are developed to investigate frequent itemsets in data, but one of the problems is the number of the plurality of required process of the used algorithms. Partition algorithm can be used to mine the frequent itemsets but lot of steps required in the algorithms that makes the analyzing process very slow. Recently, k-Partition is developed by means of k-Map to identify frequent itemsets in order to decrease the number of steps, [5].

Many algorithms are developed for Association rule, and this rule is known as the most the effective method in data mining. In association rule, commonly used algorithm is Apriori algorithm; most of the algorithms are based on Apriori algorithm to scan candidate itemsets. For example, in [6], the authors used Fp-tree algorithm but the process of generation candidate itemsets is based on Apriori algorithm.

Different methods of data mining such as classification, clustering, prediction, association rule or regression can be used to analyze the large data to investigate the useful relationship among the attributes of data. In [7], different algorithms have been discussed to measure the performance factors of association rule.

A number of techniques and algorithms in data mining are utilized for knowledge discovery from database and also numbers of establishments are designed using data mining techniques to increase the business performance or to increase the income of

the companies. The authors have been discussed the importance and the significant role in technology in his research in [8], [9]. There are many different interesting research problems that need further investigation in data mining. Before discovery the frequent pattern within the data set in data mining, the author stated a fundamental method in data mining application such as mining data stream, web mining and multimedia data mining. Overview of data mining methods, the extensions of the methods and their applications are all discussed in the study, [10].

In 1994, Apriori algorithm is presented in the VLDB conference by Agrawal and Srikant to find out association rules within the large data. And, also Apriori algorithm is compared to the other known algorithms of data mining, [11]. Apriori algorithm is one of the common and helpful algorithms of Association rule mining of data mining. The aim of Apriori algorithm is to discover frequent items in order to expose hidden data, and it is known as a traditional algorithm of association rule mining, [12], [13].

A variety of data mining methods can be utilized in protection data and affair control system to improve ability of the system. Data mining technique is applied in security information and event management system to improve the capability of the system in [14]. Todays the increasing numbers of information technique are developed to protect an unauthorized user or to share the secure data on web. Secure data sharing on web mining is analyzed in [15].

This study has five chapters, as well as these five chapters of this study are ordered as follows. Chapter 1 is the review part of data mining. Basic definitions and related concepts to Association rules are presented in Chapter 2. Apriori algorithm and k-

Map algorithms are discussed in Chapter 3. Finally, the experiments solved by Apriori algorithm and k-Map algorithm are given in Chapter 4. Chapter 5 which is the Conclusion part includes the study results.

Chapter 2

ASSOCIATION RULE MINING

2.1 Review of Fundamental Definitions

Association rule is the most commonly used technique for market basket analysis in data mining to dig up the interesting, unknown relationships within the data set. Association rule mining reveals the unexpected relationships, frequent patterns between the items in transaction data. There are some special definitions that we use in association rule like support number and confidence. Usually the minimum number of support number is predefined by the data miner to reduce the number of transactions or to delete unnecessary or uninteresting knowledge.

Let $I = \{i_1, i_2, i_3, i_4, \cdots, i_m\}$ the set of m items or attributes such that m is the positive integer. Let $T = \{T_1, T_2, T_3, ..., T_n\}$ be a set of transactions where n is the positive integer. *Let* A_1, A_2 be itemsets such that $A_1, A_2 \subseteq I$. The association rule between A_1 and A_2 can be represented by $A_1 \Rightarrow A_2$, where $A_1 \in I$ and $A_2 \in I$ hence A_1 and A_2 are independent, i.e. $A_1 \cap A_2 = \varnothing$.

In order to measure the interestingness of association rules in data set, there are some probabilistic measures like support and confidence. In addition, lift, leverage and conviction are also alternative measures.

Definition 2.1: The *support count* of an itemset A is the amount of transactions that consist of A in data set and it is denoted by $supp(A)$.

Support is an important and useful measure for example if a rule has a low support the used algorithm finds it to eliminate, sometimes no need further investigation in order to eliminate the rule when two different frequent rules have different supports. The concept of support is introduced by R. Agrawal and R. Srikant in 1994, [11].

Definition 2.2: The *support* of the itemset A in data set is equal to the ratio of support count of A to the total number of transaction in the data set denoted by $s(A)$, and it is defined by

$$s(A) = \frac{supp(A)}{n} \qquad [1]$$

where n is the number of transactions in data set, and the support count of A is the number of transactions that consists of A.

In addition, a technique for mining traditional association rule is designed, in 1993 by R. Agrawal and R, Srikant**,** [11].

A common strategy for Association rule mining algorithm is to split up the problem into sub problems or subsets. Then each sub problem is discussed by the algorithm then finally the solutions are merged. We are doing the following for investigating frequent itemsets, [16]:

The main steps of frequent itemsets generation are

- All itemsets that satisfy the minimum support threshold are obtained.
- Rule Generation
- To reduce the number of itemsets by using any algorithm from each frequent itemsets high confidence rules are produced.

Most researchers have focused on taking out frequent item set, [17]. The most favorable, important association rule is generated by evaluating the association rule in order to compare different rules such as $A \Rightarrow B$ or $B \Rightarrow A$. Just comparing the confidence results we may delete one of the association rules. In our calculations we will use the support count of the itemsets instead of the support.

Mining of frequent item sets is an essential issue for association rule mining. Many algorithms are used for mining maximal frequent item set, for example Apriori algorithm and frequent pattern, FP-growth, algorithm, [18].

Definition 2.3: The *confidence* value of a frequent rule, $A \Rightarrow B$, is equal to the ratio of support of $A \cup B$ to the support of A, i.e.

$$Conf\,(A \Rightarrow B) = \frac{supp(A \Rightarrow B)}{supp(A)} = \frac{supp(A \cup B)}{supp(A)} \qquad [2]$$

Association rule must satisfy a minimum confidence threshold, that means $conf\,(A \Rightarrow B) \geq \text{minimum threshold}$. Support eliminates uninteresting rules (rule with low support). Confidence measures the reliability of the association rule. Moreover, high value of confidence indicates strong association rule, and low confidence value indicates weak association rule.

Definition 2.4: *Lift* is the probabilistic measuring. Lift measures the goodness of the rule and it is calculated by

$$Lift(A \Rightarrow B) = \frac{s(A \cup B)}{s(A) \times s(\mathrm{B})} \qquad [3]$$

The *Lift* value is between 0 and infinity, if Lift is greater than 1, the rule is strong or the performance of rule is good but if Lift is less than 1, the negating the rule gives a better result, [19]. The rule with high lift is more significant.

The rule is not interesting when it has a low support and at that time, the support value has an important role to eliminate the unnecessary rule. The other measures are used by the researcher for this purpose, the common used measurement is known as the confidence, and it measures the stability of the rule.

The target of Association rules is to specify all frequent rules among the data set. If the item-set is frequent that means its support number is greater than or equal to threshold. And, also the frequent rule has support count and confidence that are greater than or equal to minimum support and minimum confidence, respectively, [11].

Definition 2.5: Coverage is also called antecedent support; coverage measures how a rule is appropriate in a data set. It is given by

$$coverage(A \Rightarrow B) = s(A) \qquad [4]$$

Definition 2.6: Conviction is known as an alternative measure to confidence and is defined by

$$conviction(A \Rightarrow B) = \frac{1 - s(\mathrm{B})}{1 - conf\,(A \Rightarrow B)} \qquad [5]$$

Definition 2.7: Leverage measures the difference of A and B appearing together in the data set and it is defined by

$$leverage(A \Rightarrow B) = s(A \cap B) - s(A) \times s(B) \qquad [6]$$

In order to find out Association rule, two substances are specified:

i. Frequent *itemset* mining: all frequent $k-itemsets$ are scanned for association rules. The support count of $k-itemsets$ will be bigger than or equal to minimum support.

ii. Rule mining: the rule among the frequent itemsets is produced by comparing the confidence between the frequent $k-itemsets$.

For example, let $I = \{ i_1 = \mathrm{camera}, i_2 = \mathrm{watch}, i_3 = \mathrm{mobile}\}, m = 3$, be the set of items. The set of transactions is $T = \{T_1, T_2, T_3, T_4, T_5, T_6\}$ where $T_1 = \{i_1, i_2\}$, $T_2 = \{i_1\}$, $T_3 = \{i_1, i_2\}, T_4 = \{i_2, i_3\}, T_5 = \{i_1, i_3\}, T_6 = \{i_1, i_3\}$ with minimum *support* %30 and minimum *confidence* is %50.

The items i_1, i_2 and i_3 are frequent $1-itemset$ such that i_1, i_2, i_3 since $supp(i_1) = 5$, $supp(i_2) = 3$ and $supp(i_3) = 3$ and $n = 6$.

Only, the rules $\{i_1 i_2\}$ and $\{i_1 i_3\}$ are frequent $2-itemsets$ such that $i_1 i_2 \in L2$, $i_1 i_3 \in L2$ because $supp(i_1 \cup i_2) = 2$, $supp(i_1 \cup i_3) = 2$ and $supp\{i_2 \cup i_3\} = 1$ and

$$s(i_1 \Rightarrow i_2) = \frac{supp(i_1 \cup i_2)}{n} = \frac{2}{6} = \%33.3 > \%30,$$

$$s(i_1 \Rightarrow i_3) = \frac{supp(i_1 \cup i_3)}{n} = \frac{2}{6} = \%33.3 > \%30.$$

$$s(i_2 \Rightarrow i_3) = \frac{supp(i_2 \cup i_3)}{n} = \frac{1}{6} = \%16.6 < \%30$$

so the rule $\{i_2, i_3\}$ is not frequent. In addition, there is no $3-itemsets$. For further investigation below measurements are evaluated for the frequent itemsets: $\{i_1, i_2, i_3, i_1 i_2, i_1 i_3\}$.

We calculate the confidence for those rules since their support counts are the same.

$$Conf(i_1 \Rightarrow i_2) = \frac{supp(i_1 \cup i_2)}{supp(i_1)} = \frac{2}{5} = \%40$$

$$Conf(i_2 \Rightarrow i_1) = \frac{supp(i_1 \cup i_2)}{supp(i_2)} = \frac{2}{3} = \%66.6$$

$$Conf(i_1 \Rightarrow i_3) = \frac{supp(i_1 \cup i_3)}{supp(i_1)} = \frac{2}{5} = \%40$$

$$Conf(i_3 \Rightarrow i_1) = \frac{supp(i_1 \cup i_3)}{supp(i_3)} = \frac{2}{3} = \%66.6$$

And, the confidence results for the rules $i_2 \Rightarrow i_1$ and $i_3 \Rightarrow i_1$ are equal, then Lift measurements are evaluated for finding the best rule.

$$Lift(i_2 \Rightarrow i_1) = \frac{s(i_1 \cup i_2)}{s(i_1) \times s(i_2)} = \frac{2 \times 6}{5 \times 3} = 0.8,$$

$$Lift(i_3 \Rightarrow i_1) = \frac{s(i_1 \cup i_3)}{s(i_1) \times s(i_3)} = \frac{2 \times 6}{5 \times 3} = 0.80$$

The lift is symmetric, and the lift results for both frequent rules are less than 1 and equal.

By comparing the confidence for the frequent rules, we say that the rules $i_3 \Rightarrow i_1$ and $i_2 \Rightarrow i_1$ are the most frequent rules, in fact we do not reach to the same result just considering the support results. Usually the confidence is strong measure even the support numbers of the rules are the same.

2.2 Lattice Structure

Lattice structure is a representative tool and stage for analyzing data and discovering information in classifying or associating the data set, the lattice algorithm is an important part in the application of the lattice concept. Moreover, more than ten algorithms were published for the concept of lattices.

As real data sets for mining increases, concept lattice structure suffers from its complication issues on such data. The concepts of algorithms, efficiency and performance are dissimilar from one another. In addition, it is necessary to build concept lattices to the increase the efficiency of lattice-based algorithms in data mining.

A comparison between the algorithms is required for the lattice algorithms to develop more efficient algorithm, of course this point is in our hands. Lattice algorithm performance is very important because of the capacity of real data sets in Data Mining (DM). These types of data sets are very large for example the customer data of companies.

Lattice node generation increases exponentially in worst case. Concept lattice algorithms efficiency is different from one to another. The existing lattice algorithms need to be compared with large data by satisfying the mining or learning task through the usage of an efficient algorithm. So that the lattice based algorithm which can be used in real applications increases the capacity of the algorithm, [20].

Concept Lattice – as mathematical abstraction of concept system can support human to discover information and then to create knowledge.

- The frequency of an item-set makes all of its subsets to be frequent, $X \rightarrow Y$
- The less frequent of an item-set causes all of its supersets to be less frequent too, $Y \rightarrow X$.

2.3 Compact Representation of Frequent Itemsets

The number of the resulting production from a transaction data of a frequent item-set can be too wide. It is therefore necessary to divide into smaller set of frequent items can be selected among the large group of frequent items to represent the whole data set.

In this section, we present two of such representations in the form of closed and maximum frequent itemsets. The lattice structure of the data set in the example with four items can be seen in the following Figure 1.

Table 1: Data set for example

Tid	ItemSet
1	x_1
2	x_3
3	x_4
4	x_1, x_4
5	x_2, x_4
6	x_3, x_4
7	x_1, x_2, x_4
8	x_2, x_3, x_4
9	x_1, x_2, x_3, x_4

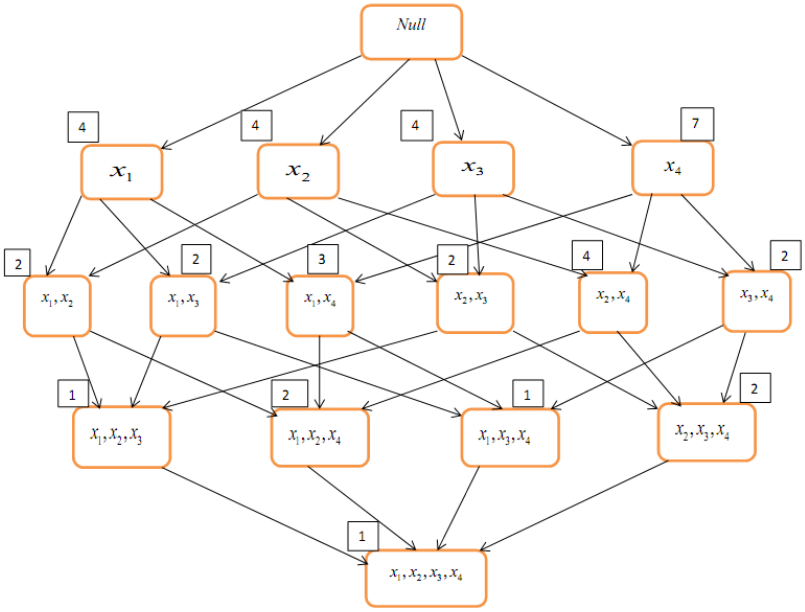


Figure 1: Lattice structure

In the figure, the itemsets fall into two categories: frequent and infrequent.

In the lattice structure, Figure 1, $\{x_3\}, \{x_1, x_4\}$ and $\{x_2, x_4\}$ are maximal frequent itemsets due to their infrequent immediate supersets.

An item-set like $\{x_1, x_4\}$ can be considered as largest frequent because of their superset for example in the example $\{x_1, x_2, x_4\}$ and $\{x_1, x_3, x_4\}$ are infrequent.

Conversely, $\{ x_1 \}$ cannot be considered as maximal since it has a frequent immediate superset $\{ x_1, x_4 \}$.

Closed Itemset: Let X be $k-itemset$. X is called closed if its immediate supersets do not have exactly the same or bigger support count as X, [21].

Closed Frequent Itemsets: a closed itemset is used in some algorithms like Close algorithm. Let X be frequent k-itemset. X is called a closed k-itemset if the support of a frequent $(k+1)$-itemset that consists of X is less than the support of X, [21].

Maximal Frequent Item-set: a maximal item-set refers to a frequent item-set in which no immediate supersets can be found as frequent, [21]. Let X be frequent k-itemset. X is called a maximal k-itemset if none frequent $k+1$-itemset consists of X.

For example, in above example i_3 is closed because $supp(i_3) = 3 > supp(\mathrm{i_1 i_3}) = 2$ but it is not maximal since $\mathrm{i_1 i_3}$ is also frequent. And, 2-itemsets, $\mathrm{i_1 i_3}$, is maximal because there is no any 3-itemset

Relationships between Frequent Item-sets: to conclude the relationship between frequent itemsets it should be pointed out maximal frequent itemsets and closed frequent itemsets. As earlier mentioned maximal and closed frequent itemsets are subsets of frequent itemsets but more compact representation are maximal frequent itemsets because they are subsets of closed frequent itemsets.

Three types of itemsets relationships are shown in the diagram below. Maximal frequent itemsets are less frequent used than closed frequent itemsets because they provide us with the support of the subsets when efficiency is more important that space, so no additional pass is needed to find this information, [22].

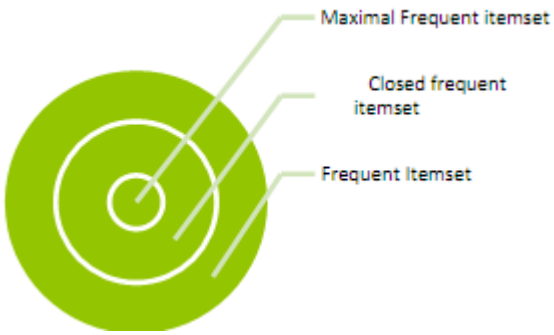


Figure 2: Maximal-Closed itemsets

2.4. Market Basket Analysis

Market Basket Analysis is a technique formed on the fact that you will more likely by certain type of item if you buy an item that is in one way or another has a relation to the item that you already bought and other way around as well for instance if a watch is bought and not the mobile phone it may be or more likely to buy TV than a person who didn't purchase the watch.

An itemsets is the set of items a customer buys, and market basket analysis tries to discover interaction between the items that are bought. The relationship will be typically in a rule form:

IF {watch, no mobile} THEN {TV}.

Support for the rule is shown when there is a probability that a buyer will purchase watch without buying a mobile (i.e. that the antecedent is true). Confidence is the conditional probability that TV will be purchased by a customer.

In applying the market basket analysis the algorithms are clear. The problems will rise when preventing combining explosion (one thousands or more line items may be stocked by supermarket), exploiting taxonomies and dealing with the availability of a large amount of transaction data.

Anyone familiar with the business may find trivial major difficulty with a large number of rules. Although there is reduction in the volume of data, user are still we are still being asked to find a needle in a haystack. When rules are required to have a high confidence level risks missing and a high minimum support level any exploitable result might be found, [23].

Most purchases are bought on impulse in retailing. Clues as to what a customer might have bought are given by market basket analysis if the idea had occurred to them.

At first the place and promotion of the items inside the market by the help of market basket analysis, for instance the buyer didn't count buying candy while purchasing Barbie dolls if it wasn't located beside or near the Barbie doll that can be seen by the costumer.

But this analysis is only the first level. Interesting results can be found in differential market basket analysis and can also potentially high volume of trivial results problem can be eliminated, [24].

The results in different analysis are compared between different stores, between different days of the week, between customers in different demographic groups, different seasons of the year, etc.

There are other alternative ways to increase the improvement of selling rate of items in the store other than the market basket analysis which they can be more efficient in some areas or some countries in way that market basket analysis cannot be used in order to improve or increase rate purchasing the frequent items in the market.

Market Basket Analysis can be used in the following areas:

- Bank transactions
- Purchase of Telecommunication services
- Drug Industry
- Online shopping purchase

Note that however the terminology, to be purchased at the same time there is no requirement for all the items. The bought goods that have been purchased over period of time can be looked at as a sequence and it can be fitted to the algorithms groups of bought items (or events) that generally occur in sequence may be identified *by a* predictive market basket analysis, [25].

Requirements of Market Basket Analysis: over the past fifteen years, eBay developed from a basic web site for online auctions to full-scale E-commerce companies that methods petabytes of web data to build a greater purchasing experience.

Data mining is very important in developing a great knowledge at eBay. Data Mining is a systematic method of extracting knowledge from data. Methods consist of pattern mining, trend discovery, and association. For eBay, knowledge discover plays a significant role in the right here spot. Whenever the consumer queries an item, how can we find the beneficial results for the consumer? Usually, a consumer

search of a number of keywords could match a lot of items. As an example, “Verizon Cell phones” is a common search at eBay; also it matches a lot more than 13,040 listed items.

We are able to generate graphs between searches and products, and between several items. As an example, the consumer who searches for “Verizon cell phones” may select the Samsung SCH U940 Glyde item, and the LG VX10000 Voyager, right now understand the query is related to these two items, as well as the two items have an association to each other because a consumer seen (or perhaps considered ordering) both. Recommending similar items is an important part of eBay. A great item recommendation can help to save hours of search valuable time and pleasure our users.

The eBay try to increase confidence of the marketing and advertising product sales and to grow a marketing technique, so marketing functions perform an important role to be offered for sale to buyers based on the marketing survey.

Chapter 3

ASSOCIATION RULES ALGORITHMS

3.1 Partition Algorithm

Studies have been conducted in order to increase the capability of Apriori algorithm to get better results; this birthed the existence of Partition algorithm. Study work has been recently done discovering the basis of partition algorism. Partition algorithm as the name implies logically splits the data-base into partitions after that for every partition P_i where $i = 1, 2, ..., n$, those frequent itemsets Li will be identified in a single data-base scan. Global candidated itemsets created by those local frequent item-set.

The support count after that determined through the use of much more data-base scan so as to get last group of frequent itemsets. At this approach Apriori performance is being improved but still some scope is there for further improvement. Mining frequent patterns is definitely an excellent subject of study for investigators. Researchers have been developed various algorithms for discovering of frequent patterns efficiently in other words there are lots of algorithms to save time, to save money or to decrease the number of the scans in the algorithms. The increase amount of the numbers database can has been the drawback and fall out of a lot of researches.

Partition algorithm is one of the kind of the methods looking for the useful information within the data set, however a lot of number of database checking is required inside of the algorithm and this makes mining operation very slow. Not many developments have been able to decrease the number of database scans as a little succeeded cutting the number to two.

An effort has been made in this study to add to a k-Partition algorithm obliging just single database examines. There is the pressure of an entire database as Karnaugh Map, having little size i.e. a small amount of the entire database. The partition algorithm took a gander at can consequently be utilized to check the successive examples utilizing k-Map model. The method shows effectiveness regarding the time utilized by processor for mining regular patterns, [33].

k-Partition algorithm based on Karnaugh Map uses the k-Map of the whole data set which requires just one step of calculation of the support numbers of each partition in the data set. At that point Partition algorithm can be developed for realizing continuous examples. The name k-Partition is exploited for joining the idea of Partition algorithm and k-Map, [5].

3.2 Apriori Algorithm

Apriori algorithm is the most common used and important algorithm for mining frequent item sets in data set, and is generated by R. Agarwal in 1994. The Priority algorithm performs an expansiveness first pursuit in the inquiry space by creating frequent $(k+1)$- item sets from successive k-item sets.

The frequency of a thing set is figured by including its event every transaction. Apriori is a compelling algorithm for digging continuous item sets for Boolean affiliation rules. The algorithm which is known as Apriori utilizes earlier information of frequent item set. It is an iterative level examination algorithm, where k - item sets are operated to investigate $(k+1)$ -itemsets.

Firstly, the set of frequent $1-$ *itemset* is established. Generally, this set is denoted by $L1$. The set $L1$ is used to determine the set $L2$, when the set of frequent $2-itemsets$ is created, it is used to determine $L3$ and so on, until no more frequent $k-itemsets$ can be found. At every stage, for every Lk requires one full scan of database. This process takes a lot of time.

Many Algorithms have been proposed to mine association rule that uses support and confidence as constraint. We are proposing a method that can be combined with Apriori algorithm and reduces storage required to store candidate and the execution time by reducing CPU time, [26].

There are two steps for understanding that how $L(k-1)$ is used to find Lk:

The join step: a set of candidate $k-itemsets$ is produced by joining $L(k-1)$ with $L(k-1)$ to realize candidate set Ck. The set of candidates is denoted by Ck.

The prune step: the members of Ck may be or may not be frequent, but every frequent k-item sets are in Ck. In the prune step any $(k-1)$- itemsets that is frequent can be a subset of a frequent k-itemsets, [27].

Ck : candidate k -item sets of size k .

Lk : frequent k -itemset of size k .

Let D, the assignment significant data, be a set of transactions in data set where every transaction T consists of the purchased items of customers and it is called TID. The set of items is denoted by I where $I = \{i_1, i_2, i_3, i_4, \ldots..., i_m\}$. And, every item set contains k items is known as $k-itemset$. All $k-itemsets$ are in the set Ck. If $k-itemset$ justifies the minimum support then it is called a frequent $k-itemset$, is represented by Lk.

At first, a set of candidates are generates by Apriori algorithm, that could be candidate $k-itemsets$, represented by Ck. When the candidate itemsets satisfies the minimum support and then it is called frequent itemsets.

Description of the algorithm:

1. Threshold for support and confidence are decided according to the problem or aim.
2. The first search is done on the dataset to obtain the set of candidate $1-itemset$, $C1$, considering the number of occurrences of each item is determined. The set of frequent $1-itemset$, $L1$, is then determined among candidate $1-itemset$ in $C1$. The algorithm uses $L1 \infty L1$ in order to list candidate $2-itemsets$, $C2$,
3. In this step, the dataset is scanned to list the candidate $2-itemsets$, $C2$ and the set of frequent $2-itemsets$, $L2$, is formed according to the minimum

threshold among the $candidate\ 3-itemsets$, and then $C3$ is generated by $L2 \infty L2$.

4. Then, the dataset is scanned to discuss the support count of each candidate itemsets, $C(k-1)$, with minimum-support. And this set is merged by itself, $L(k-1) \infty L(k-1)$ to generate the set Ck when the set of frequent elements in $L(k-1)$ is produced. Consequently, this process lasts when there is no more candidate itemsets, [28].

3.3 *k*-Map Algorithm

An important component of association rule mining is the mining frequent patterns. Recently, investigators have attempted in order to develop models and algorithms for association rule mining but there were major obstacles experienced by these algorithms of number of database scans. Because Apriori algorithm required large number of database scans makes this mining process slow. Success in decreasing the number of database scans to two has been made through improvements by investigators. A model which requires less than two database scans was attempted on here to develop. This brought about the development of Karnaugh Map model to compress the whole database in terms of frequency of item sets. Thus the whole size of Karnaugh Map matrices will be lesser than that of whole database for the mining process carried on the Karnaugh Map matrix, [29]

The Karnaugh Map matrices will have size equivalent to very small fraction of whole database as the whole database will be scanned only once Thus efficiency in association rule mining is enhanced by this approach, [30].

Karnaugh Map (*k*-Map) principle to find out ensemble association rules by experiment transaction data [31]. The algorithm used in this study *k* -Map model was proposed by N. Sharma and A. Singh in 2012, [5].

A pictorial method of grouping together expressions sharing common factors is provided by Karnaugh map, thus eliminating unrelated variables. Extensive calculation need is reduced by Karnaugh map for taking advantage of the human´s pattern recognition capability. Thus the rapid identification and elimination of potential race conditions are therefore permitted, [32].

Chapter 4

EXPERIMENTAL EVALUATIONS

4.1 Problem Statement

It has been submitted that there are many algorithms which are used in association rule for facilitate the daily life, particularly for reducing the amount of consuming for solution of place and reducing the loss of time.

Association rule includes different types of algorithms in Data Mining. The famous of these algorithms is Apriori algorithm. Recently, Partition algorithm based on k-Map is proposed and it is used by different researches because this algorithm reduces the number of scans. In our work we will use both Apriori and Karnaugh Map (k-Map) algorithms.

In our study, we focus on k-map algorithm to get a useful amount of information from the data set. Our data is obtained from the Digital SDL cameras customer's sales in order to find the useful information in the sales of digital cameras through internet. For example if someone decides to go to a camera shopping center in Cyprus to buy Digital SL camera, and then when that person visits the store at a moment he/she will see lots of camera with different features and the other equipment's in front of the store. Strategically placed cameras will identify shoppers as they enter a store. Because of strategically placed cameras, this strategy will attract the customers to go inside of the store, and when they enter to the store they

will find the related items arranged in front of them such as Macro or Micro lenses, 16 GB Scandisk or 64 GB Scandisk, sunscreen, 8 or 10 ND filters, UV protector, spare batteries, recharger, camera case, tripod etc.

It is pointed that purchasing DSLR camera we face on some problems such as:

i. How to select and buy Digital SDL camera.
ii. At the beginning, which accessories might be chosen when purchasing a camera?
iii. Which types of the equipment's must be chosen?

The total expenses of buying DSLR camera may be higher because of different varieties of other extras. Therefore, choosing or buying camera can be extraordinary experience. Especially, to select more features are incredible. Actually, even they are not interested in buying the other items but they will buy them anyway because the item is related to the other item that they want to buy. How these algorithms are very effective association rules to attract customers?

In this study, the mentioned algorithms will be used to find frequent items; the aim is to find the related items that have the bestselling and to examine the most popular/sold quantities in the store. And, after that finding the most association between those items and arranging them in a pattern so the costumer do not waste time during the shopping, and the algorithms can be used as a tool to raise the rate of selling of some items that are less desired by the costumers, because the demand on these items will increase if it was put near the desirable items. At this point, our aim would be to compare the interesting association between the listed products for camera store.

In our data set, we have 40 transactions and 6 attributes A, B, C, D, E and E where

A: Nikon DSLR D7000 Camera Body.

B: Spare battery and charger kit for Nikon DSLR D5200.

C: 10 stop ND filter for 52 mm.

D: Screen Protector for 52 mm.

E: UV Filter.

F: Silk Pro 700 DX Tripod.

The whole data set is presented in Table 2, we have listed all transactions with purchases, and the items are represented by letters A, B, C, D, E and F, respectively. Using Karnaugh Map algorithm we get frequent items, the degree of interesting and connection between the cameras and accessories.

The used algorithms will help us to indicate those materials which are more sold or more efficient or more sold together, and again cameras. The hidden information will be helpful to buy suitable cameras with suitable extra parts when the customers will buy more suitable equipment he or she will save money or time but may be the customer will spend more but in this case the owner of shop will earn more money.

Before running the algorithms each transaction is converted to binary number, for example the first transaction bought the items C, E, F and its binary code is 001011. Similarly, the Binary code has gotten for all data set. Then, the second column is added to Table 2 for the each of the transaction in data set.

Table 2: Data set with Binary code

Transactions	Items	Binary Code
1.	C, E, F	001011
2.	B ,E, F	010011
3.	A B D E F	110111
4.	B D	010100
5.	D E	000110
6.	A B C D E F	111111
7.	C D E F	001111
8.	C	001000
9.	A B C E F	111011
10.	A C F	101001
11.	A D E	100110
12.	D	000100
13.	A C D F	101101
14.	A B F	110001
15.	B C F	011001
16.	A C E	101010
17.	A D	100100
18.	D E F	000111
19.	A D F	100101
20.	A D E F	100111
21.	A C D E F	101111
22.	C D F	001101
23.	B C D E	011110
24.	D F	000101
25.	B D F	010101
26.	B F	010001
27.	A B C D	111100
28.	A B C E	111010
29.	A B E F	110011
30.	A B C D E	111110
31.	B C D	011100
32.	A B	110000
33.	A B D E	110110
34.	B C D F	011101
35.	A B C	111000
36.	A B E	110010
37.	A D E	100110
38.	B C F	011001
39.	A D F	100101
40.	B D F	010101

4.2 Problem Based on Apriori Algorithm

Of Apriori is to count up the frequencies, called the supports, of each member item separately would be to count up the frequencies, known as the supports, of every members item individually.

We assume that the minimum support is 30% that means the minimum support count is 12, so the corresponding itemset will be frequent if it is contained in at least 30%

in data set. In the following algorithm, we obtain the candidate itemsets, denoted by Ci, and we obtain frequent itemsets, denoted by Li discussing the number of occurrences of each candidate item sets.

Step 1: Generating frequent 1-itemset Pattern

The set of frequent 1-itemset, L1, consists of the candidate 1-itemset satisfying minimum support count. In the first iteration of the algorithm, each item is a member of the set of candidate. All the item sets of sizing 1 include a support for a minimum of 3, so that they are frequent. The next stage is to produce a listing of all pair of the frequent items, demonstrate in Table 3. In the following iteration by using only the frequent 1-itemset the candidate 2-itemsets are generated since according to Apriori principle all supersets of the infrequent 1-itemsets must be infrequent.

Table 3: Candidate and Frequent 1-itemset

*C*1			*L*1	
1-itemset	**Support count**		***1-itemset***	**Support count**
A	22		*A*	22
B	22	⇒	*B*	22
C	18		*C*	18
D	25		*D*	25
E	19		*E*	19
F	23		*F*	23

Step 2: Generating frequent 2-itemsets: The number of candidate 2-itemsets generated by the algorithm is the combination rule ${}_6C_2 = \binom{6}{2} = 15$, and to find the group of frequent two itemsets, L_2, the algorithm utilizes L_1 Join L_1 to produce a candidate set of 2-itemsets, C_2.

Then, the transactions in database are scanned once more and the support count for every candidate item set ,C_2, is stored. The list of 2 -itemsets is shown in the Table 4.

Table 4: Candidate 2-itemsets

C2		⇒	*C2*	
2-itemsets	**Support count**		***2-itemsets***	**Support count**
AB	12		*BF*	12
AC	10		*CD*	10
AD	13		*CE*	9
AE	12		*CF*	11
AF	11		*DE*	12
BC	11		*DF*	12
BD	11		*EF*	10
BE	10			

The set of frequent 2-itemsets, L2, after that decided, which includes those candidate 2-itemsets in C2 including minimum support, and 6 of these 15 candidates, {AC}, {AF},{BC},{BD},{BE},{CD},{CE},{CF},{EF} But computing their supporting values illustrated that they were infrequent. The remaining six candidates are frequent, see Table 5.

Table 5: Frequent 2-itemsets

L2	
2-itemsets	**Support count**
AB	12
AD	13
AE	12
BF	12
DE	12
DF	12

The algorithm will terminate here since the set {A,D,F}produced at the next stage does not have the desired support.

The candidate 3-itemsets can be seen in the Table 6. But all the support counts are less than the minimum support count, 12, none of 3-itemstes are frequent.

Table 6: Candidate 3-itemsets

C3	
3-itemsets	**Support count**
ABD	5
ABE	8
ABF	5
ADE	8
ADF	7
BDF	5
DEF	6

And finally the frequent patterns represented by L_D is

$L_D = \{A,B,C,D,E,F,AB,AD,AE,BF,DE,DF\}$, L_D consists of only 1-itemset and 2-itemsets.

4.3 Problem Based on k -Map Algorithm

Let T_i be the transactions, $i = 1,2,\ldots,40$ and let P_j represent the partitions, $j = 1,2,3,4$ and let A_i represent the attributes in each transaction. In the dataset there are 40 transactions with 6 items. In order to run the algorithm, the data set is divided into four partitions and every partition consists of ten transactions. And, also the minimum support is assumed as %30.

For Partition 1: The k-Map is created for partition 1 which is given in Table 7 and this table is used to evaluate the support numbers of candidate k-itemsets in every stage for partition 1 with support count 3.

Table 7: k-Map for partition 1

		$\bar{F}$			
		$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
$\bar{E}$	$\bar{A}\bar{B}$	0	0	1	0
	$\bar{A}B$	0	1	0	0
	$A\bar{B}$	0	0	0	0
	AB	0	0	0	0

		F			
		$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
$\bar{E}$	$\bar{A}\bar{B}$	0	0	0	0
	$\bar{A}B$	0	0	0	0
	$A\bar{B}$	0	0	1	0
	AB	0	0	0	0

		$\bar{F}$			
		$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
E	$\bar{A}\bar{B}$	0	1	0	0
	$\bar{A}B$	0	0	0	0
	$A\bar{B}$	0	0	0	0
	AB	0	0	0	0

		F			
		$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
E	$\bar{A}\bar{B}$	0	0	1	1
	$\bar{A}B$	1	0	0	0
	$A\bar{B}$	0	0	0	0
	AB	0	1	1	1

i. Using Table 7, the support numbers of each transaction in partition one is evaluated. The set of frequent 1-itemset is obtained considering candidate 1-itemset from Table 7. Here, we present some calculations, and then Table 8 is drawn using similar calculations.

$$supp(A) = (0+0+0+0)+(0+0+0+0)+(0+0+1+0)+(0+0+0+0)+ (0+0+0+0)+(0+0+0+0)+ (0+0+0+0)+ (0+1+1+1) = 4.$$

$$supp(B) = (0+0+0+0)+(0+0+0+0)+(0+0+1+0)+(0+0+0+0)+ (0+0+0+0)+(0+0+0+0)+ (0+0+0+0)+ (0+1+1+1) = 4.$$

All of $1-itemset$ are frequent since their support counts are greater than the minimum support count, that means $supp(A)$, $supp(B)$, $supp(C)$, $supp(D)$ and $supp(E) \geq 3$. The candidate and frequent 1-itemset are given in Table 8 for partition 1.

Table 8: 1-itemset in partition 1

$1-itemset$	$C1$	$L1$
A	$supp(A)=4$	A
B	$supp(B)=4$	B
C	$supp(C)=6$	C
D	$supp(D)=5$	D
E	$supp(E)=7$	E

From Table 8, all possible 2-itemsets are listed in the candidate 2-itemsets set,

$C2=\{AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, CF, DE, DF, EF\}$

ii. Then the support numbers of candidate 2-itemsets are calculated and discussing their values the frequent 2-itemsets are obtained, all results are in Table 9.

$$supp(AB)=(0+0+0+0)+(0+0+0+0)+(0+0+0+0)+(0+1+1+1)=3$$
$$supp(AC)=(0+0+0+0)+(1+0+0+0)+(0+0+0+0)+(0+0+1+1)=3$$
$$supp(AD)=(0+0+0+0)+(0+0+0+0)+(0+0+0+0)+(0+0+1+1)=2$$

Table 9: 2-itemsets in partition 1

$2-itemsets$	$C2$	$L2$	$2-itemsets$	$C2$	$L2$
AB	$supp(AB)=3$	AB	BF	$supp(BF)=4$	BF
AC	$supp(AC)=3$	-	CD	$supp(CD)=2$	-
AD	$supp(AD)=2$	-	CE	$supp(CE)=4$	CE
AE	$supp(AE)=3$	AE	CF	$supp(CF)=5$	CF
AF	$supp(AF)=4$	AF	DE	$supp(DE)=4$	DE
BC	$supp(BC)=2$	-	DF	$supp(DF)=3$	DF
BD	$supp(BD)=3$	BD	EF	$supp(EF)=6$	EF
BE	$supp(BE)=4$	BE			

The set $L2=\{AB,AC,AE,AF,BD,BE,BF,CE.CF,DE,DF,EF\}$ consists of the frequent 2-itemsets.

iii. In this step candidate 3-itemstes are obtained, then the support numbers of candidate 3-itemsets are calculated to find frequent 3-itemsets

$$C3=\{ABC, ABD, ABE, ABF, ACE, ACF, ADE, ADF,$$
$$AEF, BCE, BCF, BDE, BDF, BEF, CDE, CDF, CEF, DEF\}$$

$$supp(ABC)=\{(0+0)+(0+0)+(0+0)+(1+1)\}=2.$$

$$supp(ACE)=\{(0+0)+(0+0)+(0+0)+(1+1)\}=2.$$

Similarly, other support counts also are calculated and the following table, Table 10, shows the candidate and frequent 3-itemsets for partition 1.

Table 10: 3-itemsets in partition 1

3−itemsets	C3	L3	3−itemsets	C3	L3
ABC	$supp(ABC)=2$	-	BCF	$supp(BCF)=2$	-
ABD	$supp(ABD)=2$	-	BDE	$supp(BDE)=2$	-
ABE	$supp(ABE)=3$	ABE	BDF	$supp(BDF)=2$	-
ABF	$supp(ABF)=3$	ABF	BEF	$supp(BEF)=4$	BEF
ACE	$supp(ACE)=2$	-	CDE	$supp(CDE)=2$	-
ACF	$supp(ACF)=3$	ACF	CDF	$supp(CDF)=2$	-
ADE	$supp(ADE)=2$	-	CEF	$supp(CEF)=5$	CEF
ADF	$supp(ADF)=2$	-	DEF	$supp(DEF)=3$	DEF
BCE	$supp(BCE)=2$	BE			

$L3=\{ABE,ABF,ACF,AEF,BEF,CEF,DEF\}$ is the set of frequent 3-itemsets. Then the next stage creates four itemsets $C4$ by using frequent itemsets $C3$. The list of $3-itemsets$ for this partition can be seen in Table 11.

iv. From $L3$ we create $C4$ of 4-itemsets will be recognized as follows:

$C4=\{ABEF,ABCF,ACEF,ADEF,BCEF,BDEF,CDEF\}$

Here some calculations are given:

$supp(ABCF)=(0+0+1+1)=2,$

$supp(ACEF)=(0+0+1+1)=2.$

Table 11: 4-itemsets in partition 1

$4-itemsets$	$C4$	$L4$
$ABEF$	$supp(ABEF)=3$	$ABEF$
$ABCF$	$supp(ABCF)=2$	-
$ACEF$	$supp(ACEF)=2$	-
$ADEF$	$supp(ADEF)=2$	-
$BCEF$	$supp(BCEF)=2$	-
$BDEF$	$supp(BDEF)=2$	-
$CDEF$	$supp(CDEF)=2$	-

There is no candidate 5-itemsets since there exists a single element which is 4-itemsets, $L4=\{ABEF\}$. The frequent candidate set of Partition 1 is

$$L_{P_1}=L1\cup L2\cup L3\cup L4=$$

$$\{A,B,C,D,E,F\}\cup\{AB,AC,AE,AF,BD,BE,BF,CE.CF,DE,DF,EF\}\cup$$

$$\{ABE,ABF,ACF,AEF,BEF,CEF,DEF\}\cup\{ABEF\}$$

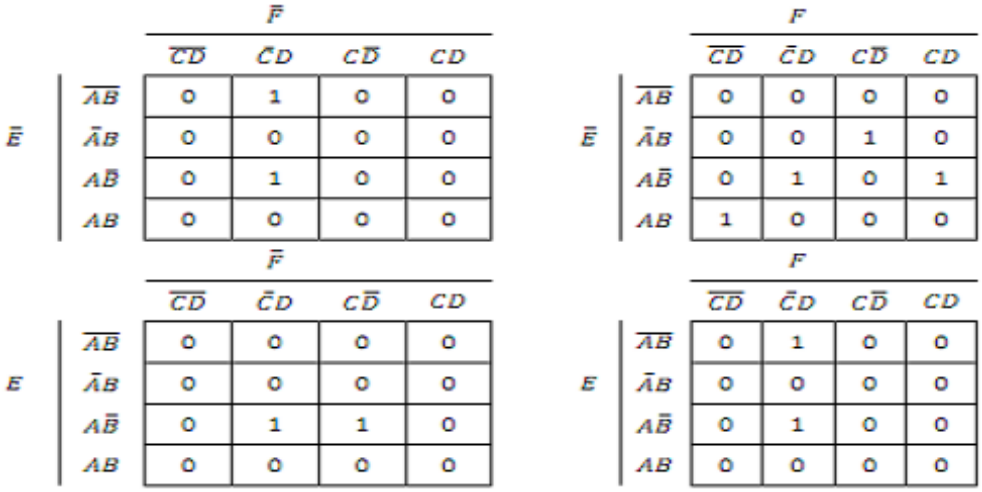
$$=\{A,B,C,D,AB,AC,AE,AF,BD,BE,BF,CE.CF,DE,DF,EF,ABE,ABF,ACF,$$

$$AEF,BEF,CEF,DEF,ABEF\}$$

The same process is applied on the other partitions.

Partition 2: The k-Map is created for partition 2 which is given in Table 12 and the table is used to evaluate the support numbers of candidate k-itemsets in every stage for partition 2 with minimum support count 3.

Table 12: k-Map for partition 2



The process in partition 1 is done for this partition 2 by the way the following table is obtained for every k-itemsets. The k-itemsets results are summarized in Table 13. In Table 13, the first column consists of the candidate itemsets and the second column consists of frequent itemsets for Partition 2.

Table 13: Frequent k-itemsets for partition 2

$k-itemsets$	*Candidate* $k-itemsets$	*Frequent* $k-itemsets$
$1-itemset$	$\{A,B,C,D,E,F\}$	$\{A,C,D,E,F\}$
$2-itemsets$	$\{AC,AD,AE,AF,CD,$ $CE,CF,DE,DF,EF\}$	$\{AD,AE,AF,DE,DF\}$
$3-itemsets$	$\{ADE,ADF,AEF,DEF\}$	$\{ADF\}$

At the end, the frequent set is found as

$$L_2 = L1 \cup L2 \cup L3 =$$

$$= \{A,C,D,E,F\} \cup \{AD,AE,AF,DE,DF\} \cup \{ADF\}$$

$$\{A,C,D,E,F,AD,AE,AF,DE,DF,ADF\}$$

Partition 3: The k-Map is created for partition 3 which is given in Table 14 and this table is used to evaluate the support numbers of candidate k-itemsets in every stage for partition 3 with minimum support count 3.

Table 14: k-Map for partition 3

$\bar{E}$	$\bar{F}$					$\bar{E}$	F				
	$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD			$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
$\bar{A}\bar{B}$	0	0	0	0		$\bar{A}\bar{B}$	0	1	0	1
$\bar{A}B$	0	0	0	0		$\bar{A}B$	1	1	0	0
$A\bar{B}$	0	0	0	0		$A\bar{B}$	0	0	0	0
AB	0	0	0	1		AB	0	0	0	0

E	$\bar{F}$					E	F				
	$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD			$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
$\bar{A}\bar{B}$	0	0	0	0		$\bar{A}\bar{B}$	0	0	0	0
$\bar{A}B$	0	0	0	1		$\bar{A}B$	0	0	0	0
$A\bar{B}$	0	0	0	0		$A\bar{B}$	0	0	0	1
AB	0	0	1	1		AB	1	0	0	0

The process in partition 1 is done for this partition 3 by the way the following table is obtained for every k itemsets. The candidate and frequent k-itemsets are summarized in Table 15.

Table 15: Candidate and frequent k-itemsets for partition 3

$k-itemsets$	*Candidate* $k-itemsets$	*Frequent* $k-itemsets$
$1-itemset$	$\{A, B, C, D, E, F\}$	$\{A, B, C, D, E, F\}$
$2-itemsets$	{AB, $AC, AD, AE, AF, BC,$ $BD, BE, BF, CD, CE, CF,$ $DE, DF, EF\}$	{AB, $AC, AD, AE,$ $BC, BD, BE, BF,$ $CD, CE, DE, DF\}$
$3-itemsets$	$\{ABC, ABD, ABE, ABF, ACD,$ $ACE, ADE, ADF, BCD, BCE,$ $BCF, BDE, BDF, BEF, CDE,$ $CDF, DEF\}$	$\{ABC, ABE, ACD, ACE,$ $BCD, BCE, CDE\}$

The frequent set is found as

$$L_3 = L1 \cup L2 \cup L3 =$$

$$= \{A, B, C, D, E, F\} \cup \{AB, AC, AD, AE, BC,$$

$$BD, BE, BF, CD, CE, DE, DF\} \cup \{ABC, ABE, ACD, ACE,$$

$$BCD, BCE, CDE\}$$

$$L_3 = \{A, B, C, D, E, F, AB, AC, AD, AE, BC,$$

$$BD, BE, BF, CD, CE, DE, DF, ABC, ABE, ACD, ACE,$$

$$BCD, BCE, CDE\}$$

Partition 4: The k-Map is created for partition 4 which is given in Table 16 and this table is used to evaluate the support counts of candidate k-itemsets in every stage for partition 4 with minimum support count 3. At the end, we form the candidate and frequent k-itemsets, they can be seen in Table 17 for every stage, and the set of frequent itemsets for partition 4 is obtained.

Table 16: k-Map for partition 4

$\bar{E}$	$\bar{F}$					F				
	$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD		$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
$\bar{A}\bar{B}$	0	0	0	0	$\bar{A}\bar{B}$	0	0	0	0
$\bar{A}B$	0	0	0	1	$\bar{A}B$	0	1	1	1
$A\bar{B}$	0	0	0	0	$A\bar{B}$	0	1	0	0
AB	1	0	1	0	AB	0	0	0	0

E	$\bar{F}$					F				
	$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD		$\bar{C}\bar{D}$	$\bar{C}D$	$C\bar{D}$	CD
$\bar{A}\bar{B}$	0	0	0	0	$\bar{A}\bar{B}$	0	0	0	0
$\bar{A}B$	0	0	0	0	$\bar{A}B$	0	0	0	0
$A\bar{B}$	0	1	0	0	$A\bar{B}$	0	0	0	0
AB	1	1	0	0	AB	0	0	0	0

Table 17: Candidate and frequent k-itemsets for partition 4

$k-itemsets$	*Candidate* $k-itemsets$	*Frequent* $k-itemsets$
$1-itemset$	$\{A,B,C,D,E,F\}$	$\{A,B,C,D,E,F\}$
$2-itemsets$	$\{AB,AC,AD,AE,AF,BC,$ $BD,BE,BF,CD,CE,CF,$ $DE,DF,EF\}$	$\{AB,AD,AE,$ $BC,BD,BF,DF\}$
$3-itemsets$	$\{ABC,ABD,ABE,ABF,ADE,$ $ADF,BCD,BCF,BDF\}$	-

The frequent item sets are

$$L_4 = L1 \cup L2$$
$$= \{A,B,C,D,E,F\} \cup \{AB,AD,AE,BC,BD,BF,DF\}$$

and

$$L_4 = \{A,B,C,D,E,FAB,AD,AE,BC,BD,BF,DF\}$$

Last Step: Finally, the following Table 18 is obtained for all data and support counts are discussed with minimum support count for the candidate set which is the union of the frequent itemsets for every partition such that $C = L_1 \cup L_2 \cup L_3 \cup L_4$ using that Table 18.

The set C_{data} is the candidate itemsets obtained by the previous steps which is

$$C_{data} = L_1 \cup L_2 \cup L_3 \cup L_4 = \{A,B,C,D,E,F,AB,AC,AD,AE,AF,BC,BD,BE,BF,$$
$$DF,CD,CE,CF,DE,EF,ABC,ABE,ABF,ACD,ACE,ACF,ADF,AEF,BCD,$$
$$BCE,BEF,CDE,CEF,DEF,ABEF\}$$

These results are presented in Table 19.

Table 18: k -Map for data set

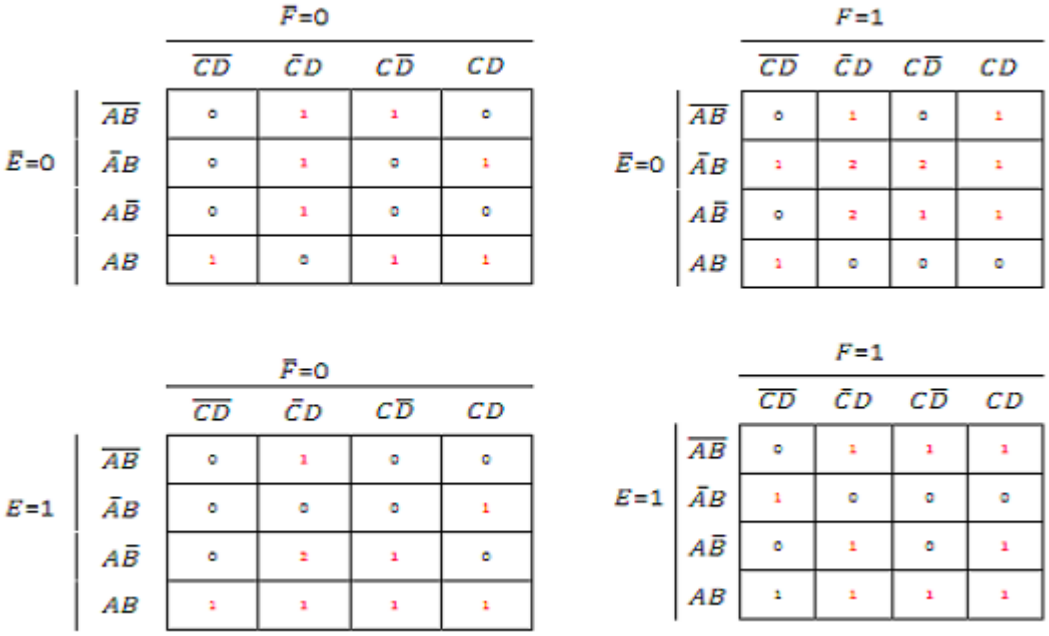


Table 19: Candidate k -itemsets for data set

C_{data}			C_{data}	
itemsets	Support count		*itemsets*	Support count
A	22		*DE*	12
B	22		*DF*	14
C	19		*EF*	10
D	25		*ABC*	6
E	19		*ABE*	8
F	23		*ACD*	5
AB	12		*ACE*	6
AC	10		*ABF*	5
AD	13		*ACF*	5
AE	12		*ADF*	7
AF	11		*AEF*	6
BC	11		*BCD*	6
BD	11		*BCE*	5
BE	10		*BEF*	5
BF	12		*CDE*	5
CD	10		*CEF*	5
CE	9		*DEF*	6
CF	11		*ABEF*	4

Table 20: Frequent k-itemsets for data set

L_{data}	
itemsets	Support count
A	22
B	22
C	19
D	25
E	19
F	23
AB	12
AD	13
AE	12
BF	12
DE	12
DF	14

Now minimum support to all transactions is 30% of 40, and we calculate new minimum support counts for all itemsets.

Then, $\frac{30\times 40}{100}=12$ will be the minimum support count for all data since threshold is %30.

Only, the itemsets in the following set are satisfying the minimum support count, and the frequent set for all data is formed as

$L_D=\{A,B,C,D,E,F,AB,AD,AE,BF,DE,DF\}$ from Table 20. The frequent $1-itemset$ and $2-itemsets$ are in Table 20. The lattice structure is given with support numbers for the frequent itemsets in Figure 3.

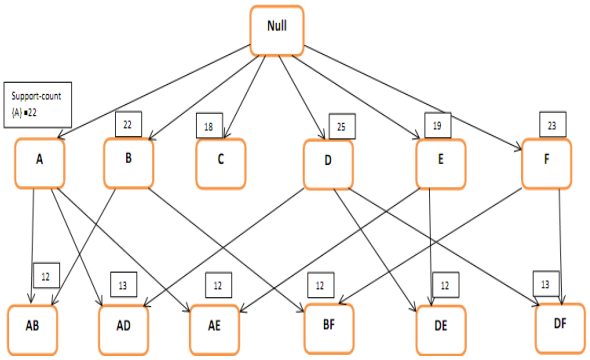


Figure 3: Lattice structure of frequent items

The frequent $1-itemset\ A,B,C,D,E,F$ are closed frequent itemsets since the support counts of their immediate supersets are less than the support counts of these frequent items.

The item C is $1-itemset$ and it is maximal, because there is no any $2-itemsets$ containing C. The other $1-itemsets\ A,B,D,E$ and F are not maximal since all of their immediate supersets are frequent. And, $2-itemsets\ \ AB, AD, AE, BF, DE, DF$ are also maximal frequent itemsets since there is no any frequent $3-itemsets$.

4.4 Probabilistic Comparison for Association Rules

Applying both algorithms, it has been found the same set of frequent itemsets. But we wanted to do further investigations on the set of frequent itemsets. Therefore, we applied different measures on the set of frequent itemsets to find out the most useful information within these frequent itemsets. In this section, we would like to present these calculations.

The following calculations show the confidence and lift measures of the frequent itemsets.

$$Conf\,(\mathrm{A} \Rightarrow B) = \frac{supp(\mathrm{A} \cup B)}{supp(\mathrm{A})} = \frac{12}{22} = 0.545$$

$$Conf\,(B \Rightarrow A) = \frac{supp(B \cup A)}{supp(B)} = \frac{12}{22} = 0.545$$

$$Conf\,(A \Rightarrow D) = \frac{supp(A \cup D)}{supp(A)} = \frac{13}{22} = 0.591$$

$$Conf\,(D \Rightarrow A) = \frac{supp(D \cup A)}{supp(D)} = \frac{13}{25} = 0.52$$

$$Conf\,(A \Rightarrow E) = \frac{supp(A \cup E)}{supp(A)} = \frac{12}{22} = 0.545$$

$$Conf\,(E \Rightarrow A) = \frac{supp(E \cup A)}{supp(E)} = \frac{12}{19} = 0.632$$

$$Conf\,(B \Rightarrow F) = \frac{supp(B \cup F)}{supp(B)} = \frac{12}{22} = 0.545$$

$$Conf\,(F \Rightarrow B) = \frac{supp(F \cup B)}{supp(F)} = \frac{12}{23} = 0.522$$

$$Conf\,(D \Rightarrow E) = \frac{supp(D \cup E)}{supp(D)} = \frac{12}{25} = 0.48$$

$$Conf\,(E \Rightarrow D) = \frac{supp(E \cup D)}{supp(E)} = \frac{12}{19} = 0.632$$

$$Conf\,(D \Rightarrow F) = \frac{supp(D \cup F)}{supp(D)} = \frac{14}{25} = 0.56$$

$$Conf\,(F \Rightarrow D) = \frac{supp(F \cup D)}{supp(F)} = \frac{14}{23} = 0.609$$

$$Lift(A \Rightarrow B) = \frac{s(A \cup B)}{s(A) \times s(\mathrm{B})} = \frac{12 \times 40}{22 \times 22} = 0.992$$

$$Lift(A \Rightarrow D) = \frac{s(A \cup D)}{s(A) \times s(D)} = \frac{13 \times 40}{22 \times 25} = 0.945$$

$$Lift(A \Rightarrow E) = \frac{s(A \cup E)}{s(A) \times s(E)} = \frac{12 \times 40}{22 \times 19} = 1.148$$

$$Lift(B \Rightarrow F) = \frac{s(B \cup F)}{s(B) \times s(F)} = \frac{12 \times 40}{22 \times 23} = 0.949$$

$$Lift(D \Rightarrow E) = \frac{s(D \cup E)}{s(D) \times s(E)} = \frac{12 \times 40}{25 \times 19} = 1.011$$

$$Lift(D \Rightarrow F) = \frac{s(\mathrm{D} \cup F)}{s(\mathrm{D}) \times s(\mathrm{F})} = \frac{14 \times 40}{25 \times 23} = 0.974 .$$

In our experiments, the results of both Apriori and Karnaugh Map (*k*-Map) algorithms are the same, but we couldn't find appropriate place for these frequent items and so we need more information about the frequent patterns. In this section, the definitions of the confidence and lift measurements have been used. In Table 21, all the results of measurements are proposed.

Table 21: Confidence and Lift results

Frequent itemsets	$Conf(X \Rightarrow Y)$	$Conf(Y \Rightarrow X)$	$Lift(X \Rightarrow Y)$
$\{A, B\}$	0.545	0.545	0.992
$\{A, D\}$	0.591	0.52	0.945
$\{A, E\}$	0.545	0.632	1.148
$\{B, F\}$	0.545	0.522	0.949
$\{D, E\}$	0.48	0.632	1.011
$\{D, F\}$	0.56	0.609	0.974

The confidence results of the frequent itemsets $E \Rightarrow D$ and $E \Rightarrow A$ are high and the same, so we need further investigations in order to find out the most frequent pattern. That's why we discuss the lift measure between the frequent rules.

The $Lift(E \Rightarrow A) = 1.148$ is greater than one, and it is the highest value, so it indicates the strong relationship between these two items. Then we may say that the most frequent pattern is $E \Rightarrow A$ among the other frequent patterns in data set.

Chapter 5

CONCLUSION

We would like to emphasis that how it can be cooperative with new science to implement to our daily life and to save our time during our works. Then we have mentioned a kind of science that it has participated that in many new science and field of technology, it is very useful and cooperative it is science of data mining that it has participated in many fields of nowadays of science. In this thesis, we have mentioned a field that it is participating in these fields association rule. It is cooperative with some aspects in data mining. We have illustrated in data mining we have written a summary on it. How can we use it and how does it appropriate with nowadays science? How can it work with science field? One field of the fields that we have pointed it is data mining it is pointes some algorithms in this association rule that they are used.

In this study, our data is investigated using two different algorithms: Apriori and k-Partition based on Karnaugh map. In general we have worked on algorithms' association rule that it can be facility in a work it is called saving time for sellers and customers in the big markets in the world and universal web sites. We have worked on some simple careers. So we have used some measuring instruments like lift and confidence support.

REFERENCES

[1] Begum, S. H. (2013). Data Mining Tools and Trends. *International Journal of Emerging Research in Management & Technology*. ISSN: 2278-9359. 6-12.

[2] Wu, X., Kumar, V. & et al. (2007). Top 10 Algorithms in Data Mining. *Springer-Verlag London Limited*. 1-37.

[3] Kajal, A. & Kajal, I. (2012). Multilevel Association Rules in Data Mining. *Indian Journal of Computer Science and Engineering (IJCSE).* 3, June, 518-521.

[4] Saxena, A. & Gadhiya, S. (2014). A Survey on Frequent Pattern Mining Methods Apriori, Eclat, FP growth. *International Journal of Engineering Development and Research, IJEDR*. 2, 92-96.

[5] Sharma, N. & Singh, A. (2012). *K* -Partition Model for Mining Frequent Patterns in Large Databases. *International Journal on Computer Science and Engineering (IJCSE)*. 1505-1512.

[6] Singh, S. & Singh, J. (2012). Association Rules and Mining Frequent Itemsets using Algorithms. *International Journal of Computer Science and Engineering Technology (IJCSET)*. ISSN: 2229-3345, 3, 8 August, 370-373.

[7] Trupti, A., Kumbhare, S. & Chobe, V. (2014). An Overview of Association Rule Mining Algorithms. *International Journal of Computer Science and Information Technologies, (IJCSIT)*. 5, 927-930.

[8] Ramageri, B. M. (2010). Data Mining Techniques and Applications. *Indian Journal of Computer Science and Engineering*. ISSN: 0976-5166, 1, 4. 301-305.

[9] Goele, S. & Chanana, N. (2012). Data Mining Trend in Past, Current and Future. *International Journal of Computing and Business Research*, In Proceeding I-Society 2012. http://www.researchmanuscripts.com/isociety2012/15. ISSN: 2229-6166.

[10] Han, J., Cheng, H. & Xin, D. (2007). Frequent Pattern Mining: Current Status and Future Directions. *Springer Science +Business Media, LLC 2007*. DOI 10.1007/s10618-006-0059-1, published online 27 January 2007. 55-86.

[11] Agrawal, R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94). Santiago, Chile, ISBN: 1-55860-153-8, 487–499.

[12] Bansal, D. & Bhambhu, L. (2013). Execution of Apriori Algorithm of Data Mining Directed Towards. *International Journal of Advanced Research in Computer Science and Software Engineering*. 3, 9. ISSN: 227 128X, 54-62.

[13] Singh, J., Ram, H., & Sodhi, J. S. (2013). Improving Efficiency of Apriori Algorithm Using Transaction Reduction. International Journal of Scientific and Research Publications. 3, 1, January. 1-4.

[14] Zope, A. R., Vidhate, A. & Harale, N. (2013). Data Mining Approach in Security Information and Event Management. *International Journal of Future Computer and Communication*. 2, 2, April, 80-84

[15] Dekate, S. K., Adhikari, J. & Parate, S. (2014). Enhancing Data Mining Techniques for Secured data Sharing and Privacy Preserving on Web Mining. *International Journal of Scientific and Research Publications*. 4, 12, 1-5.

[16] Jiawei, H. & Kamber, M. (2006). Data Mining Concepts and Techniques. *Morgan Kaufman Publishers, Elsevier*, 2nd Ed.

[17] Chandraveer, S. D., Arora, S. & Makani, Z. (2013). Comparsion of Interestingness Measure: Support-Confidence Framework versus Lift-Irule Framework. *IJERA*. ISSN: 2248-9622, 408-412.

[18] Nageswara, R. G. & Suman, K. G. (2011). Mining frequent item sets without candidate generation using FP-Trees. *Journal of Computer Science and Information* Technologies, ISSN: 0975-9646, 2677-2685.

[19] Berry, M. J. A. & Linoff, G. S. (2004). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. Wiley Publishing, Inc. 287-319

[20] Aditi, S. & Muralidhar A. (2014). Hierarchial Approach for Frequent Closed Itemset Generation in Distributed Environment. *International Journal of Computer Applications*. DOI: 10.5120/19001-0476, 8-10.

[21] Luccheses, C., Orlando, S. & Perego, R. (2006). Fast and Memory Efficient Mining of Frequent Closed Itemsets. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 18, 21-36.

[22] Arun, K. P. (2003). Data Mining Technique. *Universities Press (India) Private Limited Publisher*, ISBN: 81-7371-380-4, 69-114.

[23] Boztug, Y. & Hildebrandt, L. (2006). A Market Basket Analysis Conducted with a Multivariate Logit Model. *Springer Berlin Heidelberg Publisher*. ISSN: 1431-8814, 558-565.

[24] Phani, P. J. & Murlidher, M. (2008). A Study on Market Basket Analysis Using a Data Mining Algorithm. *International Journal of Emerging Technology and Advance Engineering*, (*IJETAE).* ISSN: 2250-2459. 361-363.

[25] Trnka, A. (2010). Market Basket Analysis with Data Mining Methods. *Networking and Information Technology (ICNIT).* ISBN: 978-1-4244-7578-0, 446–450.

[26] Priyanka & Sharma, Er. V. K. (2014). Apriori Algorithm for Mining Frequent Itemset –A Review. *International Journal of Computer Application and Engineering Technology*. ISSN: 2277-7962, 232-236.

[27] Mishra, R. & Choubey, A. (2012). Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data. *International Journal of Computer Science and Information Technologies, JCSIT*, ISSN: 0975-9646, 4662-4665.

[28] Patel, B., Vijay, K., Chaudhari, K., Rajneesh, K. & Rana, Y. K. (2011). *International Journal of Soft Computing and Engineering (IJSCE).* ISSN: 2231-2307, 24-26.

[29] Setiabudi, D. H, Budhi, G. S. , Purnama, I. W. J. & Noertjahyana, A. (2011). Data mining Market Basket Analysis' Using Hybrid-dimension Association Rules, Case Study in Minimarket X. *Uncertainty Reasoning and Knowledge Engineering (URKE),* ISBN: 978-1-4244-9984-7, 196-199.

[30] Rajput, D. S., Thakur, R. S. & Thakur, G. S. (2014). Karnaugh Map Approach for Mining Frequent Termset from Uncertain Textual Data. *British Journal of Mathematics & Computer Science*. ISSN: 2231-0851, 333-346.

[31] Gautam, P. & Pardasani, K. R. (2010). A Fast Algorithm for Mining Multilevel Association Rule Based on Boolean Matrix. *International Journal on Computer Science and Engineering (IJCSE).* ISSN: 0975-3397 746, 746-752.

[32] Singh, J. & Singh, R. (2014). Karnaugh Map. *International Journal of Research (IJR)*, ISSN: 2348-6848, 456-461.

[33] Schulz, C. (2013). High Quality Graph Partitioning. *Karlsruher Instituts für Technologie*. PhD Thesis, 117-124.