# CHAPTER-14

# Mining Association Rules in Large Databases

# 14.Mining Association Rules in Large Databases

## 14.1Introduction

Association rule mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuosly being collected and stored , many industries are becoming interested in mining association huge amounts of business transaction records can help in many business decision making processes, such as catalog design, cross-marketing, and loss-leader analysis.

A typical example of association rule mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread(and what kind of bread)on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space. For example, placing milk And bread within close proximity may further encourage the sale of these items together within single visits to the store.

How can we find association rules from large amounts of data, where the data are either transactional or relational? Which association rules are the most interesting ?How

can we help or guide the mining procedure to discover interesting associations? What language constructs are useful in defining a data mining query language for association rule mining.

**14.2 Association Rule mining**

Association rule mining searches for interesting relationships among items in a given data set. This section provides an introduction to association rule mining. We begin by presenting an example of market basket analysis, the earliest form of association rule mining. The basic concepts of mining associations are given and we present a road map to the different kinds of association rules that can mined.

**14.3 Market Basket Analysis: A Motivating example for Association Rule Mining**

Suppose, as manager of an ABCompany branch, you would like to learn more About the buying habits of your customers. Specifically, you wonder, “Which groups or Sets of items are customers likely to purchase on a given trip to the store ?”To answer Your question, market basket analysis may be performed on the retail data of customer Transactions at your store. The results may be used to plan marketing or advertising Strategies, as well as catalog design. For instance, market basket analysis may help Managers design different store layouts. In one strategy, items that are frequently Purchased together can be placed in close proximity in order to further encourage the Sale of such items together. If customers who purchase computers also tend to buy

Financial management software at the same time, then placing the hardware display Close to the software display may help to increase the sales of both these items. In an Alternative strategy, placing hardware and software at opposite ends of the store may Entice customers who purchase such items to pick up other items along the way. For Instance, after deciding on an expensive computer, a customer may observe security Systems for sale while heading towards the software display to purchase financial Management software and may decide to purchase a home security system as well. Market basket analysis can also help retailers to plan which items to put on sale at Reduced prices. If customers tend to purchase computers and printers together, then Having a sale on printers may encourage the sale of printers as well as computers.

If we think of the universe as the set of items available at the store, then each item Has a Boolean variable representing the presence or absence of that item. A Boolean Vector of values assigned to these variables can then represent each basket. 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 financial management software at the same time is represented in Associated Rule below:

Computer=> financial_management_software

[support = 2%,confidence == 60%]

Rule support and confidence are two measures of rule interesting that were described earlier. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule means that 2% of all the transaction under analysis show that computer and financial management software are purchased a computer also bought the software. Typically ,association rules are considered interesting if they satisfy both a minimum support thresholds and a minimum confidence threshold. Users or domain experts can set such thersholds.

**14.4 Basic Concepts**

Let $\tau = \{i_1, i_2 \ldots\ldots i_m\}$ be a set of items. Let D, the task-relevant data, be a set of databse transactions wher each transaction T is a set of items such that $T \subset \tau$ each transaction is association with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$ . An association rule is an implication of the form A=>B, where $A \subset t$, $B \subset t$ and $A \cap B = f$. The rule A=> B holds in the transaction set D with support s, where s is the percentage of transaction in D that contions $A \cup B$ (i.e both Aand B ). This is taken to be the probability, $P(A \cup B)$. The rule A=>B has confidence c in the transaction in the transaction set D if 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

Support (A=>B)=$P(A \cup B)$

Confidence (A=>B)=p( B/A)

Rules that satisfy both a minimum support threshold (min_sup) and a minimun confidence threshold (min_conf) are called strong. By convention, we write support and confidence value so as to occure between 0% and 100% rather than 0 to 1.0.

A set of items is referred to as an itemset. Anitemset that contains k item is a k-itemset. The set{computer, financial_management_software} is a 2-itemset. The occurrence frequency of an itemset is the number of transaction that contain the itemset. This is also know, simply,as the frequency, support count or count of the itemset. An itemset satisfies minimum support if the occurrence frequency of the itemset is greater than or equal to the product of min_sup and the total number of transactions in D. The number of transaction required for the itemset to satisfy minimum support is

therefore referred to as the minimum support count.If an itemset satisfies minimum support,then it is a frequent itemset .The set of frequent K-itemsets is commonly denoted by LK.

"How are association rules mined from large databases?" Association rule mining is a two-step process:

1**.Find all frequent itemsets:** By definition,each of these itemsets will occur at least as frequently as a pre-determined minimum support count.

**2.Generate strong association rules from the frequent itemsets**:By definition ,these rules must satisfy minimum support and minimum support and minimum confidence.Additional interestingness measures can be applied,if desired.The second step is the easiest,of the two.The overall performance of mining association rules is determined by the first step.

**14.5 Association Rule Mining:A Road map**

Market basket analysis is just one form of association rule mining,in fact,there are many kinds of association rules.Association rules can be classified in various ways,based on the following criteria:

Based on the types of values handled in the rule; if a rule concerns associations between the presence or absence of items,it is a Boolean association rule.For example,the rule above is a Boolen association rule obtained from market basket analysis.

If a rule describes associations between quantitative items or attributes,then it is a quantitative association rule.In these rules ,quantitative values for items or attributes are partitioned into intetervals.The following rule is an example of a quantitative association rule,where X is a variable representing a customer:

age(X,"30.....39")^income(X,"42K.....48")implies buys(X,high resolutionTV)

Note that the quantitative attributes,age and income,have been discretized,

**Based on dimensions in the data:** If the items or attributes in an association rule reference only one dimension,then it is a single-dimensional association rule,Note that above rule could be rewritten as buys(X,"computer") implies buys(X,"financial_management_software")The first rule above is a sinle-dimensional association rule since it refers to only one dimension,buys.If a rule references two or more dimensions,such as the dimensions buys,time_of_transaction,and customer_category,then it is a multidimensional association rule.

**BASED on the levels of abstraction in the rule set:**Some method for association rule mining can find rules at differing levels of abstaction. For example, suppose that a set of association rule mined includes the following rules:

age(x,"30...39") buys(x,"laptop computer")

age(x,"30...39") buys{x,"computer")

In above rules the items bought are referenced at different levels of abstraction. (e.g.,"computer" is a higer-level abstraction of "laptop computer"). We refer to the rule set mined as consisting of multilevel association rules. If, instead,the rules within a given set do not reference items or attributesat different levels of abstraction, then then the set contains single-level association rules.

- **Based on various extensions to association mining:**Association mining can be extended to correlation analysis, where the absence or presenc of correlated items can be identified. It can alos be extended to mining maxpatters (i.e.,maximal freqent patterns) and frequent closed itemssets. A **maxpattern** is a frequent pattern,p,such that any proper sub pattern of p is not frequent. A frequent closed itemeset is a frequent closed itemset where an itemset c is closed if there exists no proper superset of c,c ' such that every transaction containing c also contains c'. Maxpatterns and frequent closed itemset can be used to substantially reduce the number of frequent itemsets generated in mining.

**14.6 Mining single-dimensional Association Rules from Transactional Databases**:

In this section, you will learn methods for mining the simplest from of assocaition rules-single-dimensional, single-level. Boolean association rules, such as those discussed for market basket analysis later. We begin by presenting a priori, a basic algorithm for improved efficiency and scalability are presented.Then methods for mining association rules that,unlike a priori,do not involve the generation of "candidate" frequent itemsets.The last section describes how principles from priori can be applied to improve the efficiency of ansewering iceberg queries,which are common in market basket analysia.

The a priori Algorithm:Finding Frequent Itemsets Using Candidate Generation

A priori is an influential algorithm 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,as we shall see below.A priori employs a iterative approach known as a level-wise search,where f-itemsets are4 used to explore (k+1)-itemsets.First ,the set of frequent 1-itemsets is found.This finding of each Lie requires one full scan of the database.

To improve the efficiency of the level-wise generation of frequent itemsets,an important property called the a priori property,presented below,is used to reduce the search space.We will first describe this property,and then show an example illustrating its use.

In order to use the a priori proprety,all nonempty subsets of a frequent itemset must also be frequent.This property is based on the following observation.By definition,if an itemset(i.e.,I U A)cannot occur more frequently than I.Therefore,I U U is not frequent either,that is ,P(I U A)<min_sup.

This property belongs to a special category of properties called anti-monotone in the sense that if a set cannot pass a test,all of its supersets will fail the same test as well.It is called anti-monotone because the property is monotonic in the context of failing a test.

To understand how the a priori property is used ,let us look at how LK-1 is used to find LK.A two-step process is followed ,consisting **of join** and **prune** actions.

**1.The join step:**To find LK,a set of candidate k-itemsets is generated by Joining Lk-1 with itself.This set of candidates is denotedCk.Let I1 and I2 be itemsets in Lk-1;The notation L1[j] refers to the jth item in li;(e.g.,l1[k-2]} refers to the second to the last item in li) By convention,a priori method assumes that items within a transaction or itemset are sorted in kexicograpic order.The join ,Lk-1 C Lk-1, is performed,where members of Lk-1 are joinable if their first (k-2) items are in comon.That is ,members l1 and l2 of Lk-1 are joined if (l1[1]=L2[1]^(l1[k-2]=L2[k-2]^(l1[k-1]=l2[k-2]),The conditional l1[k-1]<l2[k-2].,simply ensures that no duplicates are generated.THe resulting itemset formed by joining l1and l2 is l1[l1]l1[2]...l1[k-1]l2[k-1].

2.The prune step: q is a superset of Lk that is,its members may or may not be frequent,but all of the frequent k-itemsets are included in ck.A scan of the database to determine the count of each candidate in Ckwould result in the determination of Lk (i.e., all candidates having a count no less than the maximum suport count are frequent by definition,and therefore belong to Lk).ck,however,can be huge,and so this could involve heavy computation.To reduce the size of Ck,the a priori property is used as follows.Any(k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset.Hence,if any (k-1)-subset of a candidate k-itemset is not in Lk-1 then the candidate cannot be freuent either and so can be removed from Ck.This subset testing can be done quickly by maintaining a hash tree of all frequent itemsets.

10.7 Generating Association Rules from Frequent Itemsets

Once the frequent itemsets from transactions in a database D have been found it is straightforward to generate strong association rules from them(where strong association rules satisfy both minimum support and minimum confidence).This can be done using the following equation for confidence,where the conditional probablity is expressed in terms ofitemset support count;

confidece(A implies B)=P(B\B)= support count{A u B}/supprot_count(a)

where support_count (Au B) is the number of transactions containing the itemsets AuB ,and support_count(A) is the number of transactions containing the itemset A.Based on this equation,association rules can be generated as follows:

- For each frequent itemset 1,generate all nonempty subsets of 1.
- For every nonempty subset 5 of 1, output the rule,s implies(1-s).

Since the rules are generated from frequent itemsets,meach one automatically satisfies minimum support.Frequent itemsets can be stored ahead of time in hash tables along with their counts so that they can be accessed quickly.

## 10.8 Iceberg Queries

The Apriori algorithm can be used to improve the efficiency of answering ice-berg queries.Iceberg queries are commonly used in data mining,particularly for market basket analysis.An iceberg query computes an aggregate function over an attribute or set of attributes in order to find aggregate values above some apecified threshold:

Given a relation R with attributes a_1,a-2,....,a-n and b, and an aggregate function,agg-f,an iceberg query is of the form

Given the large quantity of input data tuples,the number of tuples that will satisfy the threshold in the having clause is relatively small.The output result is seen as the "tip of the iceberg," where the "iceberg"is the set of input data.

## 14.9 Review Questions

1 Expalin about Association Rule mining

2 Expalin about Association Rule Mining:A Road map

3 Discuss Mining single-dimensional Association Rules from Transactional Databases

4 How can we Generating Association Rules from Frequent Itemsets

5 Expalin about Iceberg Queries

**14.10 References**

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition