Welcome students.

Today we'll be talking about the course titled Data Mining.

This course is for semester 5 of B.Sc. Computer Science.

In this session we are dealing with Unit 5: Association Analysis.

Module name: Mining Multi-Level association rules. Mining Multidimensional association rules, Other applications of association rule mining.

The outline for the today's session is

We will be learning about mining multi-level association rules, mining multidimensional association rules and other applications of association rule mining.

At the end of this session the learning outcomes are:

The student will be able to explain the mining of multi-level association rules.

Understand the different approaches for mining multi-level association rules.

Explain mining of multidimensional association rules and understand the different applications of association rule mining.

Let's start with the first topic that is Mining multi-level association rules.

Now, multilevel associations involve concepts at different abstraction levels.

It is interesting for us to examine how to develop the effective methods for mining patterns at multiple abstraction levels, with sufficient flexibility for easy traversal among different abstractions spaces.

For many applications it is difficult to find strong associations among data items at low or primitive levels of abstraction due to sparsity of data at those levels.

Association rules.
Mining Multi Level Association rules generated from mining data at multiple levels of abstraction are called as multiple level or multi-level association rules.

Multi-Level Association rules can be mined efficiently using concept hierarchies under a support confidence framework.

A concept hierarchy defines a sequence of mappings from a set of low level concepts to higher level more general concepts.

Data can be generalized by replacing low level concepts within the data by their higher level concepts or ancestors from the concept hierarchy.

Here we will be using the top down strategy which is employed, wherein the counts are accumulated for the calculation of frequent itemset at each concept level starting at the concept level one and working downwards in the hierarchy towards the most specific concept levels under no more frequent item sets can be found.

For each level, any algorithm for discovering frequent itemsets may be used, such as Apriori or its variations.

Let's consider an example.

For Mining Multilevel association rules, suppose we are given the task relevant set of transactional data for sales in a store showing the items purchased for each transaction.

So let's consider this database that task relevant data which is represented by D.

We have the transaction ID's and we have the items purchased.

That is the customer with transaction ID.

This has bought this product so all these are represented with respect to transaction ID's and items purchased.

This is the concept hierarchy for the database D for all computer items.

We have different levels.

This is level 1, level 2, Level 3, level 4 and so on.

So each of this level has been represented or mapped from lower level concepts to higher level concepts.

Similarly, from here it is mapped to a higher concept.

Similarly, from here it is mapped to a higher concept, so we have a concept hierarchy over here representing all computer items.

Now we'll see the different approaches for mining the multi-level association rules.

A number of variations to the primary approach are described where each variation involves playing around with the support threshold in a slightly different way.

The approaches are as follows.

The first one is using Uniform minimum support for all levels, which is referred to as. Uniform Support.

Here the same minimum support threshold is used when mining at each level of abstraction.

When a uniform minimum support threshold is used, the search procedure is simplified.

The method is simple in that the users are just required to specify only one minimum support threshold.

So here the uniform minimum support approach has some difficulties.

That is, it is unlikely that items at lower level of abstraction will occur as frequently as those at higher levels of abstraction.

If the minimum support threshold is set to very high, it could miss some meaningful associations occurring at low abstraction levels, whereas if the threshold is set to very low, it may generate many uninteresting associations occurring at high abstraction levels.

Figure below shows the representation of uniform minimum support which is set to 5% at all levels.

We have level 1, level 2, so we have computer with support value as 10% and we have laptop computer with support as 6% and this desktop computer with support as 4%.

So at each level we have set the minimum support value same that is 5% at level 1 and level 2.

So this is an example or the representation wherein you are representing or the multi-level mining with uniform support value that is 5% as minimum support value.

The second approach is using Reduced minimum support at lower levels of abstraction which is referred to as Reduced Support.

Now in this each level of abstraction has its own minimum support threshold.

The deeper the level of abstraction, the smaller the corresponding threshold is.

For example, in the figure below, this is the figure which represents the multi-level mining with reduced support over here.

We have level 1 and level 2 wherein we have at level 1 we have the minimum support value as 5% and at level 2 we have a lesser minimum support of 3%.

So here the minimum support threshold for level 1 and level 2 are 5% and 3% respectively.

So in this way computer, laptop computer and desktop computer are all considered as frequent that is computer, laptop computer and desktop computer are considered as frequent and here we have used the reduced support as in when the level is decreasing.

The third approach is using Item or group based minimum support, which is referred to as Group based support.

Here the users or the experts often have an insight or information as to which groups are more important than others.

It is sometimes more desirable to set up user specific item or group based on minimum support thresholds when mining multi-level rules.

For example, if we consider a user could set up the minimum support thresholds based on the product price or on items of interest, such as by setting particularly low support thresholds for laptop computers and flash drives in order to pay particular attention to the association patterns containing items in this categories.

Next is Mining Multidimensional association rules.

Let's consider an instance wherein in mining AllElectronics database.

This is the name of the database.

We may discover the Boolean association rule as buys X is a random customer, a customer who buys digital camera will also buy HP printer.

So X is a random customer.

So this is an association rule which represents X implies Y.

So if a customer buys digital camera, the customer will buy HP printer also.

So in this, multidimensional database is referring to each distinct predicate in the rule as dimension.

So buys is my predicate which will be represented as the dimension.

Here we can refer to rule above as Single dimension or Intra dimensional.

This rule is called as Single dimension or Intra dimensional association rules because it contains a single distinct predicate that is buys with multiple occurrences, that is, predicate occurs more than once within the rule.

Such rules are commonly mined from the transactional data.

Now, instead of considering the transactional data only, sales and related information are often linked with relational data or integrated into a data warehouse data.

Such data stores are multidimensional in nature.

For instance, other than in addition to keeping track of all the items purchased in sales transaction, a relational database may record attributes associated with the items and transactions such as the item description or the branch location of the sale.

Additional relational information regarding the customers who purchased the items, for example, customers age, occupation, credit rating, income, and address may also be stored.

Now let's consider each transactional or database attribute or warehouse dimension as a predicate, wherein we can therefore mine the association rule containing multiple predicates such as age.

Let's consider a person X whose age is between 20 to 29 and occupation is student.

Then he will buy the laptop.

So this is my association rule, wherein the predicate we have age, occupation, and buys so single dimension dimensional or intra dimensional association rule will contain single distinct predicate.

For example, buys with multiple occurrences over here that is, the predicate will occur more than once in the rule, so this is an example of single dimensional or intra dimensional association rule.

Now association rule that involves 2 or more dimensions or predicates can be referred to as multidimensional association rules.

So this example wherein a person with age between 20 to 29 and occupation as a student, then he will buy he or she will buy the laptop.

So this is my association rule and here I have two or more dimensions or predicates that is, age, occupation and buys.

So it's a multi-dimensional association rule.

Above rule contains 3 predicates.

That is age, occupation and buys, each of which occurs only once in the rule.

Hence we say that it has no repeated predicates over here.

So multidimensional association rules with no repeated predicates are called as Interdimensional association rules.

Here we can also mine multidimensional association rules with repeated predicates, which will contain multiple occurrences of some predicates, so these rules will be called as Hybrid dimensional association rules.

Let's see what are the applications of association rule mining?

The first one is Medical Diagnosis: Association rules in medical diagnosis can be useful for assisting physicians for curing patients.

Diagnosis is not only an easy process and has a scope of errors which may result in unreliable end results.

Using relational association rule mining we can identify the probability of the occurrences of an illness concerning various factors and symptoms.

Using learning techniques, this interface can be extended by adding new symptoms and defining the relationships between the new signs and the corresponding diseases.

The second application of association rule mining is Census data.

Every government has tonnes of census data.

This data can be used to plan efficient public services, that is, education, health, transport as well as help the public businesses for setting up new factories, shopping malls and even marketing particular products.

This application of association rule mining and data mining has an immense potential in supporting sound public policy and bringing forth an efficient functioning of a democratic society.

The third application of association rule mining is Protein sequence.

Proteins are sequences which are made up of twenty types of amino acids.

Each protein bears a unique 3D structure, which depends on the sequence of this amino acids.

A slight change in the sequence can cause a change in structure which might change the functioning of the protein.

This dependency of the protein functioning on its amino acid sequence have been a subject of great research.

The nature of associations between different amino acids that are present in the protein can be identified by generating association rules.

Here are my references.

Thank you.