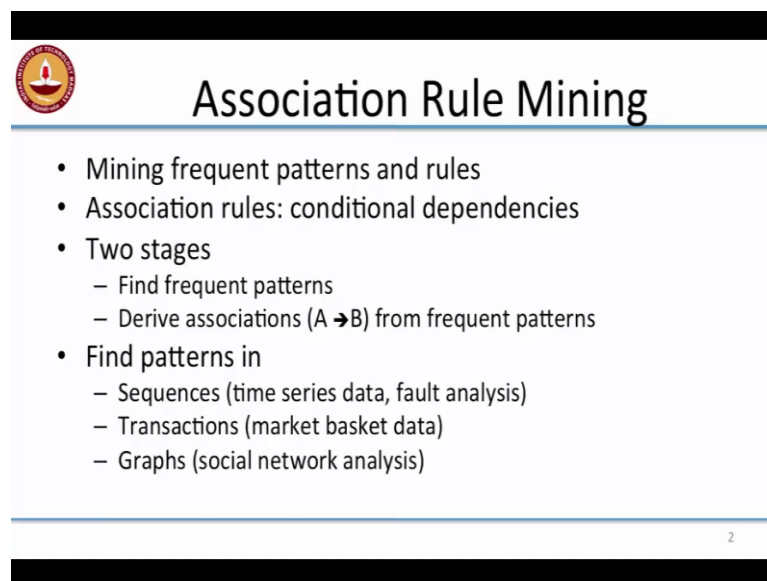**Introduction to Data Analytics**

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

Indian Institute of Technology, Madras

Module - 07

Lecture - 37

Association Rule Mining Frequent Pattern Mining
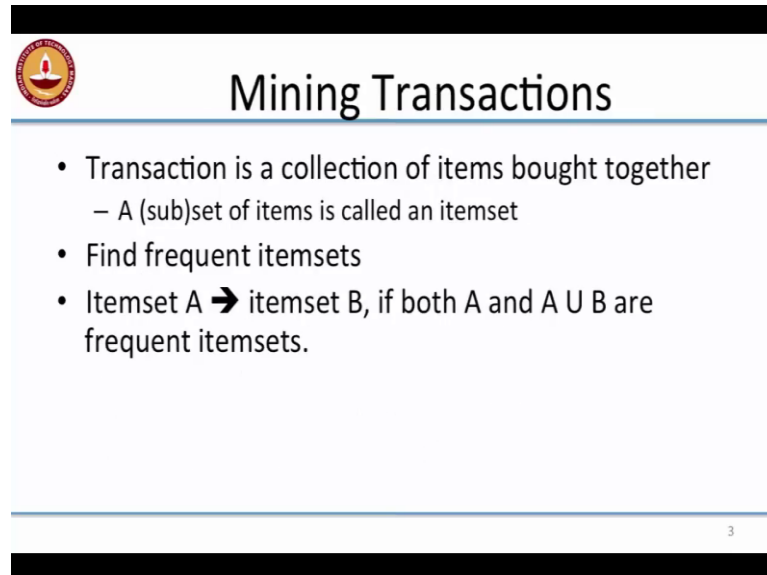
(Refer Slide Time: 00:20)

Hello and welcome to this module on Association Rule Mining or Frequent Pattern Mining, which is essentially the basic problem underline association rule mining. So, we had a very brief look at association rule mining with very beginning of the machine learning section. So, the idea behind association rule mining is to first mined frequent patterns that occur in the data and based on the frequent patterns that you have mined, derive association rules which is to form at if a happens, then b is slightly to happen. So, basically is like a conditional dependence relation that you are mining if A happens, that makes B more like it happened.

You could find such patterns in sequences looking at time series data, like financial data or looking at fault analysis, where one thing causes fault occur; or you can look at in the transactional data column context which is where it was originally proposed and that is what we will look at in more detail in the rest of the module. And more interestingly you

could also look at mining frequent patterns and associations in graphs, which is appropriately used in social network analysis.

(Refer Slide Time: 01:29)

So, let us look at a mining transaction for the rest of the module. So, transaction is a collection of items that were bought together. That is the simple definition that we will use for the purposes of association rule mining. And please note that the set or subset of items, is usually a denoted item set in the association rule mining community. So, the goal here is to find first find frequent item sets, then you would say that an item set A implies item set B; for example, if you could say that somebody buys milk, then they are likely to buy bread, if both A and the event A union B or frequent item sets. That would mean that both A, sorry which is milk in this case and A union B which is bread and milk both should be frequent. In which case, I can say that, if you buy milk then, you buy bread as well.

Let us take a look at exists simple example here. So, here is a set of transactions, so I have 5 transactions and each color here denotes a different kind of item. So, the first 3 transactions are 4 item sets, the 4th one is the 3 item set, and the 5th one is a 2 item sets. And let us assume that we have a frequency threshold of 3.

So, we essentially have the following as frequent 1 item sets. So, we have blue, which occurs in all the 5 transactions and then purple which occurs in 4 transactions. And pink which occurs, big item which occurs in 3 transactions. So, none of the other items occur in 3 or more transactions. So, the frequent one item sets are just these. So, the next thing you have to look at is the frequent 2 item sets, in the frequent 2 item sets you can see or essentially purple and blue which occur in 4 transactions and pink and blue which occur in 3 transactions and so, none of the other combinations are frequent. So, the only other thing which we really have to look at this purple and pink; and purple and pink occur only in 2 of the transactions together therefore, they are not frequent. So, the goal here is to first find such frequent item sets.

(Refer Slide Time: 03:47)

And from these frequent item sets, how do we determine which are interesting association rules. So, the 2 measures of interestingness for association rules or essentially support. So, the support of a rule is the percentage of item sets that contain A union B; right. So, in this case and then the confidence of a rule is the other measure that we are interested in. So, we will go back and look at the data set once we have understood what support and confidences. So, the confidence of a rule is a percentage of item sets containing A, that also contained A union B. So, essentially this tells you how confident you are in making the association.

So, typically we look for rules with both high support and confidence. If you think about it, if both there in the case of support and confidence we really need to find the frequency of the item sets; right. So, once we determine what are the frequent item sets and what their frequencies are, then we can easily determine what are the relevant association rules. So, more effort needs to be focused on counting rather than the association rule itself. So, that is why I said there is frequent pattern mining part of this more interesting than the association rule mining part.
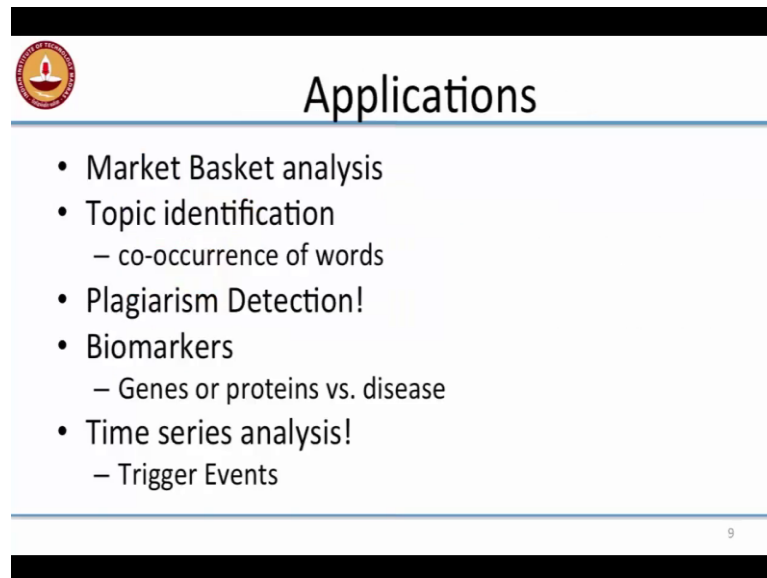
(Refer Slide Time: 05:03)

So, let us go back and look at the pattern set we have mined so far. So, if you remember the one item sets are in really interesting except to establish the frequency part of it. So, let us look at a rule which says that purple implies blue. If you have bought purple then you likely to buy blue. So, our purple occurs in your transaction then blue is likely to occur. So, if you think about it, this is the valid rule to have because both purple as well as purple and blue where frequent. In the earlier slide we saw that both purple and purple and blue are frequency, so it satisfies the A, A union B rule and, what about the support of this rule.

This support is essentially all the transactions the number of transactions in which both A and B occurs. So, that where A union B occurs divided by the total number of transactions. So, A union B occurs in 4 fifth of this data set; and therefore, the support of this rule is 4 by 5. What about the confidence of the rule? The confidence of the rule is, whenever A occurs; how many times A union B is occur. And in this case it turns out that whenever A occurs, A union B occurs; therefore, its confidence is 1.

So, likewise for the pink implies blue rule, you can see that the support is 3 by 5; because that is how many times a pink union blue occurs. And the confidence is 1; because whenever pink occurs, pink union blue also occurs. So, what about this rule? I mean is this the valid rule to look at? Both the blue is frequent and blue and purple is also frequent. So, that is a valid rule to look at, but is that the better rule then the one that we saw earlier. That is not necessarily in the case because the support here is 4 by 5 and the confidence is 4 by 5 as well. So, the earlier rule which is purple implies blue had a higher

confidence than this rule. But both of these are possible rules that you could consider, because both the rules have a high support and a high confidence. So, this is essentially the idea behind the association rule mining.

(Refer Slide Time: 07:14)

So, it has been applied to variety of applications, Market Basket analysis is one. So, is the head mentioned earlier. So, market basket refers to the fact that when you go to super market, you are going to buy a set of items together and put them together in your basket. So, essentially looking at what goes into the basket at the market. So, these baskets are considered as transactions, and you look at frequent pattern mining in these transactions. Could look at co-occurrence of words and that can be used to derive certain kinds of topic relationships. And people are looked at plagiarism detection in terms of frequent pattern mining. Now people have replied this to a biological data looking at bio markers and genes of proteins versus diseases. So, try to find out associations between co-occurrence of a certain protein, abnormalities, and diseases. You also looked at in the contrast of time series, where co-occurrence of events can be used to model trigger events and this identifies trigger events. So, that brings us to the end of the first module on frequent pattern mining and the subsequent module will look at techniques for a efficiently mining frequent patterns.