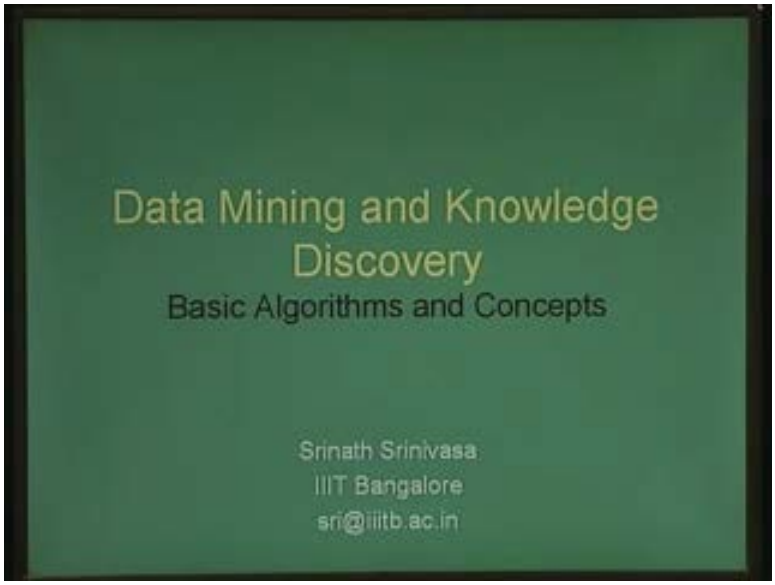
Database Management System
Dr. S. Srinath
Department of Computer Science & Engineering
Indian Institute of Technology, Madras
Lecture No. # 34

Data Mining and Knowledge Discovery

Hello and welcome. In this session today we are going to look at very interesting aspect of or interesting application in which database technologies are used namely the field of data mining and knowledge discovery. In fact in recent years data mining has become an extremely or fields that eliciting an extremely large amount of interest not just from researchers but also from commercial domain. I mean the commercial utility of data mining is probably of more interest than or at least as much interest as the research interest that lies in data mining.

(Refer Slide Time: 01:25)



And in addition to commercial interest, there is also number of public debates that data mining has started which range from topics like legalities and ethics and the rights to certain information and the rights to non-disclosure of information or the rights to privacy and so on and so forth. So data mining actually is in some sense has opened a pan door as box in and only time will tell whether the technology has given, has been on an overall sense completely beneficial or destructive in nature.

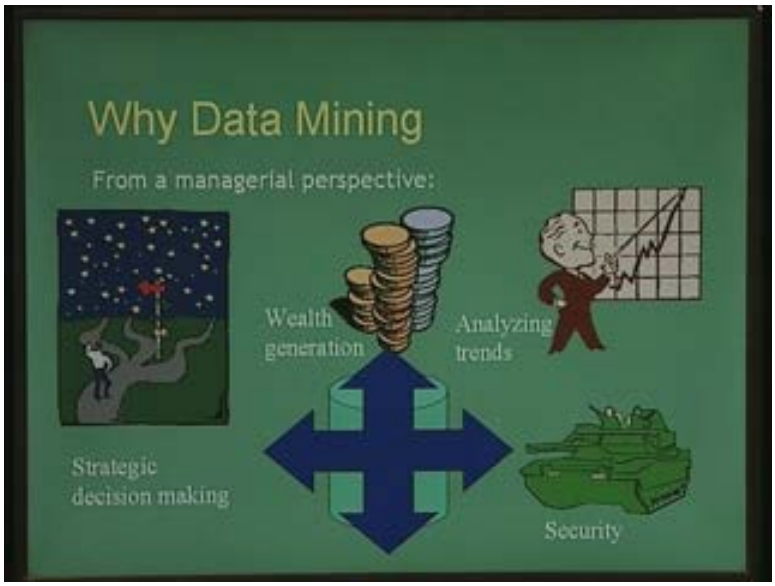
But then there is nothing beneficial or destructive about technology per say it's how we use it, how we use technology which is what matters. So any way in this session, we shall be concentrating mostly on the technical aspects of data mining obviously.

(Refer Slide Time: 03:59)

And we shall look at the basic algorithms and concepts that make up data mining and what exactly is meant by data mining and how does it differ from the traditional operations of databases or traditional way in which databases are used. So the overview of this or this set of two sessions would be as follows. Let us first motivate the need for data mining that is why data mining and what are some of the basic underlying concepts in data mining, what are the building blocks of data mining concepts. Then we look at data mining algorithms and several classes of this data mining algorithms.

We will start with tabular mining as in mining relational tables and we will look at classification and clustering approaches and we will also look at mining of other kinds of data like sequence data mining or mining of streaming data and so on. And data warehousing concepts would be covered as a different session all together. First of all, why data mining from a managerial perspective. Let's first look at what a data mining has for the commercial world first before we go in to looking at the technical aspects of data mining.

(Refer Slide Time: 04:33)



If you were to let us say give an internet search or talk to a manager, let us say about why he or she would invest in data mining, you would encounter a variety of answers. One would say something like strategic decision making that is I look for some kinds of some ways or some patterns in or mind for certain nuggets of knowledge to understand something about strategic decision making or to help in strategic decision making. Somebody would say well it is very useful for something called wealth generation although there is no precise definition of the term wealth generation and you would say that data mining would help me in understanding or making the right decisions that can help me increase my financial portfolio or whatever.

Somebody would say well I would use data mining for analyzing trends, analyzing how my customers behave or analyzing how particular market is behaving and so on and so forth. And more recently data mining has been used extensively for security purposes especially mining network logs or network streaming data in order to look for abnormal behavioral patterns or patterns that might be potentially linked to abnormal activity in the network or in the system and so on. So, security is now relatively recent and very important application area of data mining.

So, what is this data mining all about and why is this so controversial and why is it so interesting from a technical perspective at the same time. Data mining is the generic term used to look for hidden patterns in data or hidden pattern and trends in data that are not immediately apparent by just summarizing the data. So if I want to look for certain patterns, let us say if I have set of all students and their grades if I want to look for certain patterns on how are the students performing over time or what is the is there some kind of relation between subject A and subject B I mean if a student does well in subject A, he or she does badly in subject B or so on and so forth.

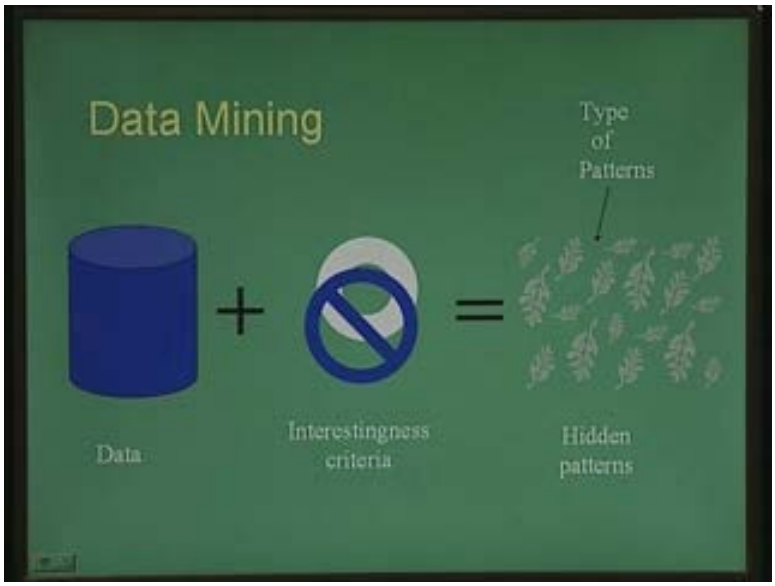
(Refer Slide Time: 06:33)

Such things cannot be discovered by just aggregating the data, by just saying what is the average or what is the summation or whatever. Besides, such things also cannot be discovered by, I mean such things in a sense cannot be within quotes discovered if we have to give queries that finds out these aggregations. That is if we already knew what it is that we are looking for then it's not a hidden pattern any more. We know that such a pattern exists that is students performing in subject A will not perform well in subject B, we know that such a correlation exists and there is nothing hidden in the pattern anyway.

So data mining essentially has no query that is if you are performing a data mining on a on a database, we do not talk of any data mining query. In fact it is the mining algorithm that should give us something which we don't know. Now how do we say something which we don't know, which is putting it in a very broad sense I mean which is making things so vague. So data mining is actually controlled by what are called as interestingness criteria and we just specify to the database that this is what we understand by an interesting pattern.

Let us say correlation between performances in subject A and subject B or some kinds of trends over a period of time. This is what is interesting for us. Now find me something or find me everything which I don't know about or which are interesting according to this criteria.

(Refer Slide Time: 09:24)

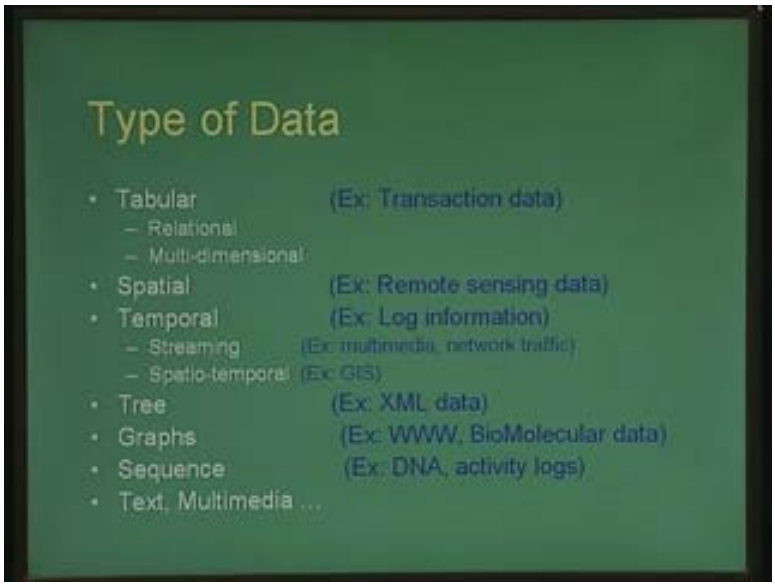


So when we talk about data mining, we have a set of data to begin with that is we have a database and then we give one or more interestingness criteria and the output of which will be one or more hidden patterns which we didn't know exists in the first place. Now given this model, we should say now when we say patterns then the obvious question to ask is what type of patterns, what do you mean by patterns or what do you mean that this is or when do you say that something is a pattern and something is not a pattern.

If we have to answer that we have to ask two further questions that is what is the type of data that we are looking at, what kind of data set is it that we are looking at and what is the type of interestingness criteria that we are looking. What do we mean by interestingness, is it correlation between something, what exactly do we mean by interestingness.

So let us look at the different type of data that we encounter in different situations. The most common kind of data is the tabular data or the relational database which is in the form of set of tables or now slightly different multi-dimensional form of database. And it's very common that any kind of transaction data that is let us say data array coming out from the database from an ATM for example or the data coming out from the transactional database at a railway reservation counter or at a bank or any place like that are all tabular in nature. So it's a most common form of data and which is a rich source of data to be in mine.

(Refer Slide Time: 10:16)

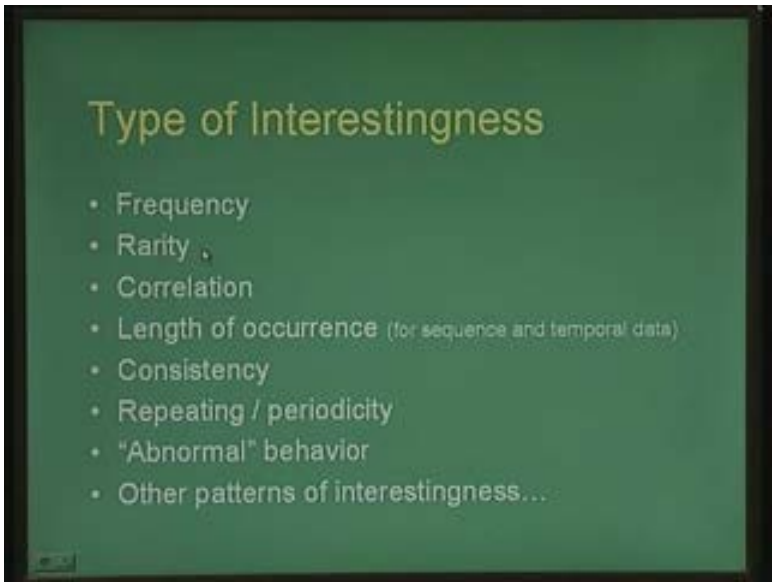


In addition to tabular data, there are spatial data for example where data is represented in the form of either points or regions which have been encoded with certain coordinates X Y Z coordinates. So each point in addition to having certain attributes also has certain coordinates and mining in this context also requires us to know what is the importance of the coordinates system.

In addition to spatial data there are other kinds of data like say temporal data, temporal data in the sense that were each data element has a time tag associated with it. So temporal data could be for example streaming data where network traffic or set of all packets that are flowing through a network forms streaming data which just flows fast and where each packet can be allocated some kind of a time stamp or something like activity logs, your database activity log is a temporal data. There could be also be spatio temporal data that is data that are tagged both by time and coordinates. And other kinds of data like tree data which for example XML databases or graph data where especially bio molecular data or volvoid web is a big graph data and so on.

Then there are sequence data like data about genes and DNAs and so on and again activity, I mean sequence is a kind of temporal data where timestamp need not be explicit in sequence then text data, the arbitrary text or multimedia and so on and so forth. So, the several different kinds of data that can be the source from which we can extract or mine for unknown nuggets of knowledge.

(Refer Slide Time: 13:24)



Similarly when we talk about interestingness criteria, several things could be interesting. If certain pattern of events or certain patterns of data keep occurring frequently then it might be of interest to us, something that happens very frequently. So frequency by itself is an interestingness criteria or interestingness or a criteria on which interestingness can be based.

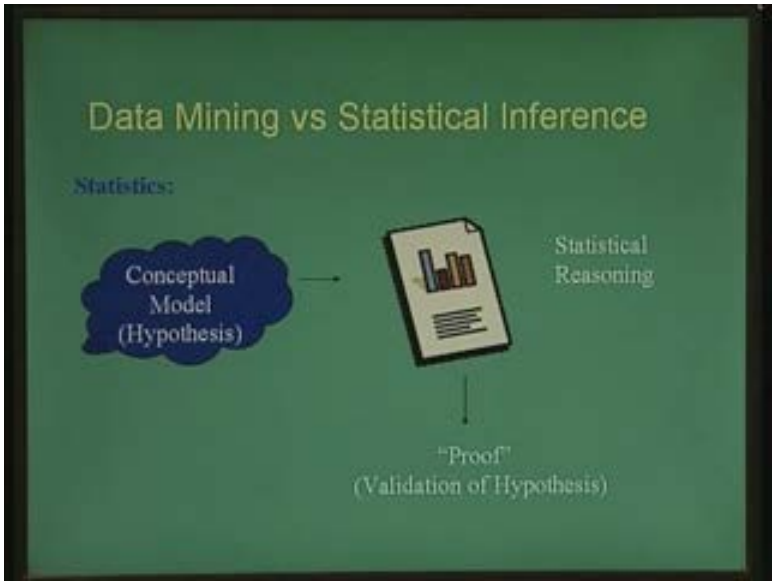
Similarly rarity, if something happens very rarely and we don't know about it or let us say rarity is again a very interesting pattern to be searched for when we are looking at say abnormal behavior of any system or abnormal behavior of network traffic and so on. So something that happens rarely that is away from the norm is again an interestingness pattern. Correlation between two or more elements and if the correlation being more than a threshold is again interesting or length of occurrence in the case of sequence or temporal data and so on.

And consistent occurrence, consistency that is consistency is different from frequency in the sense that overall in the set of all databases, overall for the entire database a given pattern may not be frequent enough. For example there could be one particular behavior pattern, let us say one particular customer comes to a bank every month at the tenth of each month. So if we are looking for frequently banking customers, let us say this customer would not figure out in this algorithm because this customer comes only once a month whereas other customers could be coming many times a month. However if we are looking for consistency in behavior then this customers behavior is far more consistent than someone who comes let us say arbitrarily 10 times the first month and once the second month and 50 times the third month and so on and so forth. So in terms of consistency in his behavioral pattern across different months, this pattern is interesting even though it's not frequent.

Then repeating or periodicity is slightly similar to consistency except that a periodicity is I mean consistency is across the entire set, across the entire set of months if you have divided our database into months but periodicity, the time interval could vary in in a periodicity of a pattern. If a customer comes let us say a 5 times to the bank every 6 month, we may not be able to catch it as part of a consistent pattern analysis but if we use an algorithm that detects periodicity of several occurrence of events, we will be able to detect it. And similarly there are several other patterns of interestingness that which one could think of.

Now when we talk about data mining, usually there is sometimes a misconception and not completely but usually there is a contention that data mining is the same as statistical inference. For many cases it is yes, the answer is true that is several concepts from statistics have been incorporated in to data mining and data mining software uses statistical concepts or many kinds of statistical algorithms comprehensively. However there is a fundamental difference between statistical inference and data mining which is perhaps the reason for the renewed interest in data mining algorithms. And here is the general idea behind the data mining versus statistical inference.

(Refer Slide Time: 17:30)

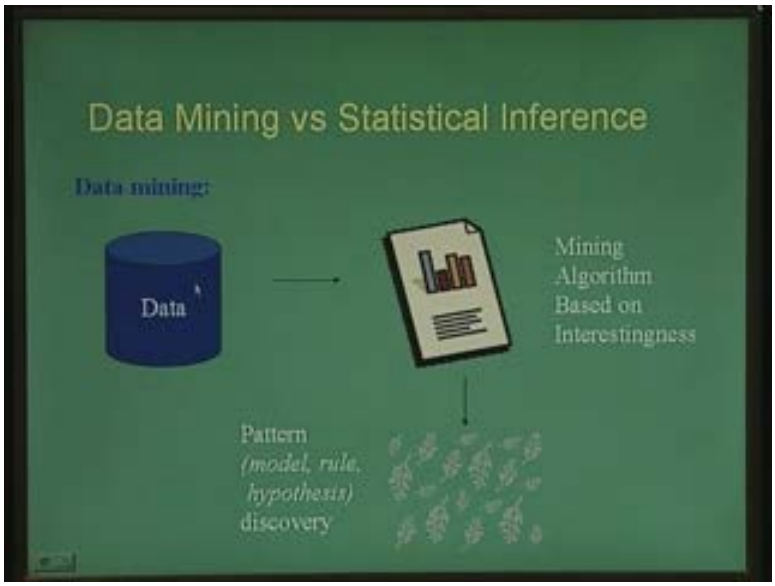


What do we do when we talk about statistical inference? Statistical inference in techniques, essentially have the following three steps as is shown in this slide here. In statistical inference, we start out with the conceptual model or what is called as the null hypothesis. That is we first of all present ourselves or perform a hypothesis about the system in concern. That is we make a hypothesis that if some something to the effect that if exams are held in the month of march then there would be I mean then the turnout would be higher than if it is held in the month of June or something like that. Now based on this hypothesis, we perform what is called as sampling of the data set or of the system.

Now sampling is a very important step in a statistical inferencing process. There is huge amount of literature in to what is meant by correct sampling or what is called as a representative sample and so on. Now based on the sampling of data set from the system, we either prove or refute our hypothesis. That is we show a proof saying, yes this hypothesis is true because statistical sampling of the system has shown that this is true otherwise it's false.

Now, when we sample for example if you are performing a statistical inference about user preferences or let's say some kind of market analysis, we present questioner to different users based on our null hypothesis or based on our conceptual model. Now it is this set of questioner, now this questioner has been created by our conceptual model. So this questioner already knows what to look for and the proof or the answers will either prove or refute the hypothesis but data mining on the other hand is a completely different process or rather it's the opposite process.

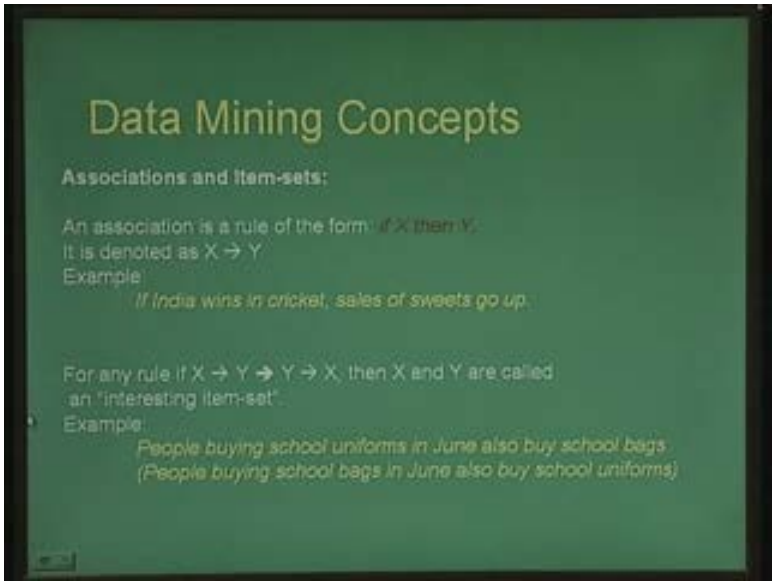
(Refer Slide Time: 19:57)



In data mining we just have a huge data set and we don't know what is it that we are looking for. We don't have any hypothesis, we don't have any null hypothesis to begin with. We just have a huge data set and we just have some notions of interestingness. Now we use this interestingness criteria to mine this data set and usually there is no sampling that is performed on the data set that is the entire data set is scanned at least once by the data mining algorithm in order to look for patterns. So there is no question of sampling and there is no null hypothesis to begin with. So we just have a weighed notion of an interestingness based on which we present an algorithm, data mining algorithm over the data set. Out of this comes out certain patterns, certain interesting patterns which form the basis for forming a hypothesis. So it's sometimes also called hypothesis discovery. Obviously, of course we cannot discover complete hypothesis using just data mining but we too discover patterns using which we can formulate a hypothesis. So in a sense it's an opposite process of statistical inference.

Let us look at some data mining concepts. Two fundamental concepts are of interest in data mining especially in the core algorithms of data mining especially the apriori based algorithms. These are what are called as associations and items sets. An association, when we say an association it is a rule of the form if X then Y as shown in this slide here and it's denoted as X right arrow Y.

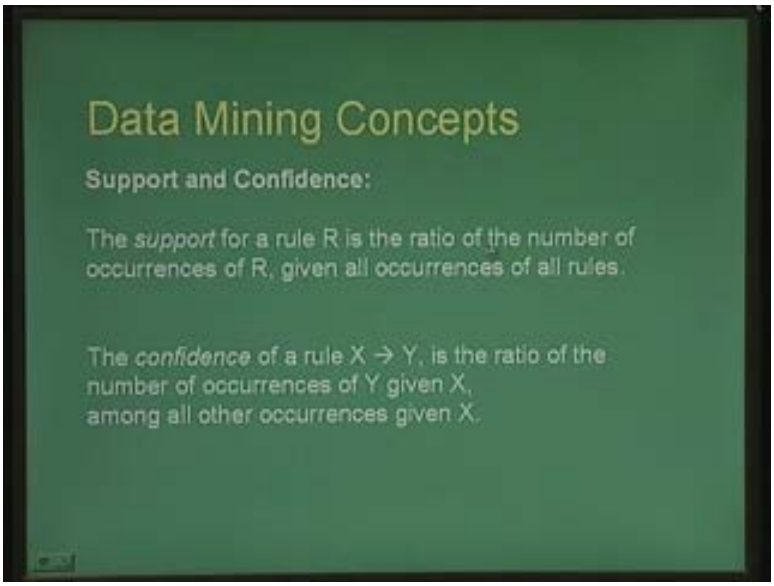
(Refer Slide Time: 21:23)



For example if India wins in cricket sales of sweets goes up, if India wins in cricket then sales of sweets goes up. So here X is India wins in cricket and Y is the predicate that sales of sweets go up. So we say that we discover such a rule if we are able to conclusively say based on analyzing the data that whenever India wins in cricket, the sales of sweets go up. And on other hand suppose if there is any rule of this form that is if X then Y then I can imply that if Y then X. (Refer Slide Time: 21:27) That is the ordering of this rules is not important. If India wins in cricket then sales of sweets go up, if sales of sweets go up then India has won in cricket and so on which may be true or may not be true but if that is the case then it is called an interesting item set. That is it's just a set of item. For example people buying school uniforms in june also buy school bags or you can also say people buying school bags in june also buy school uniforms. So it's just a item set that is school uniforms and school bags are a set of items which are interesting by themselves.

Once we define the notion of a association rule and an item set, we now come to the concept of support and confidence. That is how do we discover a rule to be interesting. We say that a rule is interesting in the sense of frequent occurrences of a particular rule, if the support for that rule is high enough. That is the support for a given rule R is the ratio of the number of occurrences of R given all occurrences of all rules. So we look into the exact or we will illustrate the notion of support in the next slide with an example where it will become more clear.

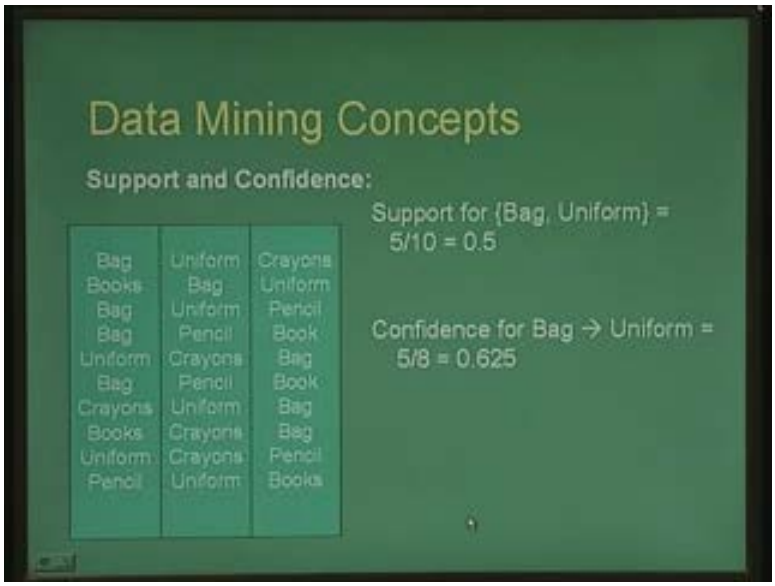
(Refer Slide Time: 22:41)



And when we say the confidence of a rule, suppose I have a rule if X then Y then the confidence of the rule is suppose I know that X is true, the ratio of all occurrences when Y is also true versus when for all other occurrences when X is true and something else is here (Refer Slide Time: 23:46). So that is it's a ratio of the number of occurrences of Y given X among all other occurrences given X. So if I know that X is true with what confidence, with what percentage of confidence can I say that Y is also going to be true?

Let us look at some examples here (Refer Slide Time: 24:04). Let us say these are some item sets let us say these are data that have been distilled from purchases of different consumers over a period of time over, in a given month let us say. So the first consumer has bought a bag, a uniform and a set of crayons, the second consumer has bought books and bag and uniform, the third one has bought bag uniform and pencil and so on and so forth. Now suppose I take the item set bag and uniform, (Bag, Uniform) what is the support for this item set. Now the support for this item set is look at all the transactions or the rows here in which bag and uniform occur 1 2 3 4 and 5 uniform and bag. Out of a total of 10 rows, 5 of them have bag and uniform occurring in that.

(Refer Slide Time: 24:34)

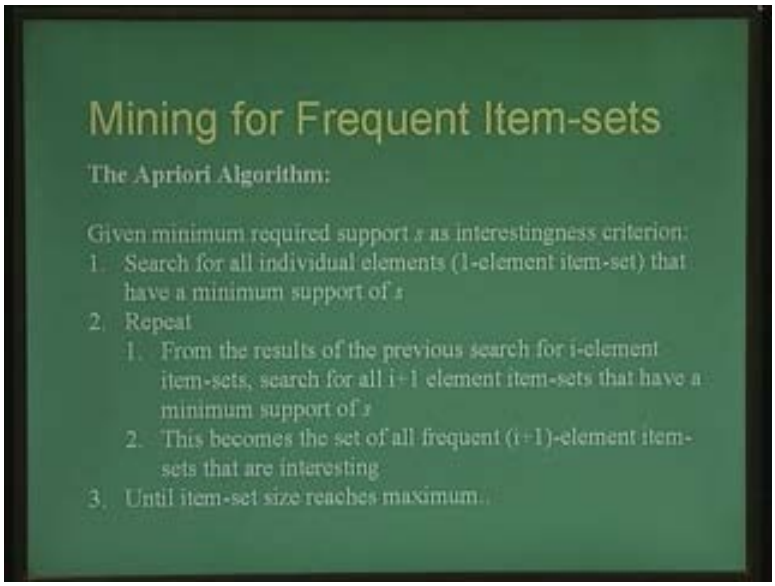


Therefore the support for bag and uniform is 5 divided by 10 which is 0.5 that is with a this dataset supports the assertion that bag and uniform will be bought together with 50% support that is 0.5 as its support. What is the confidence that, what is the confidence for the rule if bag then uniform? That is what is the confidence by which we say whenever somebody buys a bag, they also buy uniform. For this we have to look at the set of all item sets or the set of all transactions or rows here in which bag and uniform, bag occurs rather not just uniform in which bag occurs.

So bag occurs in 1 2 3 4 5 6 7 8 different rows, out of which bag and uniform have occurred in 5 different rows. Therefore the confidence for this assertion or this association rule is 5 divided by 8 which is about 62%. That means if some consumer has bought a bag then with 62% of confidence or 62.5 % of confidence, we can say that the consumer will also buy a uniform, a school uniform along with this. So the question now is how do we mine or how do we find out the set of all interesting item sets and the set of all interesting association tools.

Now have a look at this previous slide (Refer Slide Time: 26:50) once again. Now the association rule, when we talk about association rules we have just or rather when we talk about item sets first we just saw a single item set having two different elements here but that need not be the case, bag by itself could be an item set a single element item set, uniform by itself could be a single element item set, crayons could be a single element item set or let us say bag, uniform and crayons could be a three element item set and so on. So item sets could be of any size size 1, size 2, size 3, size n any set of elements. Now we have to find the set of all item sets that is the set of all items that are bought together and that have been together frequently as part of this transaction log here.

(Refer Slide Time: 26:48)



Now how do we do that? Now there is a very famous algorithm called the apriori algorithm which performs such a discovery process that is a discovery process for all frequent item sets in a very efficient manner. The simple idea behind apriori algorithm, it is shown in this slide here. However let us not go through the slide in a lot of detail, since it will be more easier to explain apriori through an example.

The idea behind apriori algorithm is that, the essential idea behind an apriori algorithm is that suppose I have any n element item set. Let us say suppose I have any 5 element item set, that is interesting or that is frequent. So if this 5 element item set is frequent then all sub sets of this item should also be frequent. This seems obvious but this is a very important conclusion or it's a very important observation in the apriori algorithm. That is if I discover the set of all one frequent item sets that is the set of all item sets of size 1 which are frequent then there is no need for me to look at other item sets when I am looking for two frequent item sets. That is the set of all item sets of size 2 which are frequent will be made up of combinations of set of all item sets of size 1 which are frequent.

So let us illustrate the process of apriori with an example. Let us take our consumer database again, the previous consumer database again where we have consumers buying several school utilities like bags and school bags and school uniforms and crayons and pencils and books and so on and so forth.

(Refer Slide Time: 29:22)

Now suppose we set when we say or when we ask the apriori miner to mine for all interesting item sets, we have to the interestingness criteria here is frequency that is frequent occurrence. Now frequency is or interestingness here is parameterized by a threshold parameter which is called the minimum support or min sup. So let us say minimum support is 0.3 that is we term an item set to be interesting if its support is at least 0.3 or greater.

Now given this what are all the interesting one element item sets? What is that mean to say what are all the interesting one element item set, which one element item sets occur at least at a rate of 30% or more. Now this database here or this data set here has a total of 10 rows therefore we have to look at all one element item sets which occur 3 or more times. So given this we see that all of these are interesting that is bag, uniform, crayons, pencil and books. Bag occurs much more than three times, uniform also occurs more than three times, crayons also occur more than three times and so on. So all of these elements here occur more than thrice which therefore all of this one element item sets have a minimum support of 30% or more.

Now from this, suppose we have to look at the set of all interesting two element item sets. Now how do we build the set of all interesting two element item sets? We just look at all possible combinations between one element item sets, therefore we have bag uniform, bag crayons, bag pencil, bag books, uniform crayons, uniform pencil uniform books and so on and so forth. Now out of this for each such two element item set that have been created, we have to see how many times they occur in this data set. Now we see that it's only these set of combinations which have a minimum support of 0.3 or more. So for example bag uniform, bag crayons, bag pencil and bag books all of them along with bag are interesting.

However let us say uniform and book is not interesting that is it doesn't occur more than thrice. So let us see how many times uniform and book occur? Uniform and books occur once and second one twice here, so they occur only twice but we need a minimum support of three times so that's not interesting. Similarly a pencil and uniform, so uniform and pencil is again is not interested. So therefore we have filtered away or we have thrown away certain item sets from our exploration here and identified only a smaller subset of the set of all possible combinations of one element item set.

Now from this if we have to look for all three element item sets, we have to generate the set of all candidate three element item sets. What are the candidate three element item sets? Perform a union across all possible combinations of these interesting two element item sets to create all possible distinct three element item sets and then look for those three element item sets which occur at least three times or more in this database. Given that we see that there is only one three element item set that is bag, uniform and crayons that is interesting that is that occur at least three times or more or that has at least, that has support of at least 30% in this in this data set.

(Refer Slide Time: 33:56)

So as you can see the apriori algorithm, you can visualize the apriori algorithm in the form of let us say an iceberg. Such queries are also called as iceberg queries when given on to databases that is at the base there are large number of one element item sets. But once we start combining them together, we start getting smaller and smaller numbers of combinations and we peak out at a very small of large item sets which are frequent. So the beauty of the apriori algorithm is that for every parse, it does not need to go through the entire data set. It does not have to parse through the entire data set, it only needs to consult results of the previous iteration or item sets that are of one element one lesser than the present iteration in order to construct candidates for the present iteration.

So given this algorithm here let us go back and look at the apriori algorithm. Given the explanation here with an example let us go back and look at the apriori algorithm which will now be a little more easier to understand. Initially we start with a given minimum required support s as the interestingness criteria. now given minimum support s as the interestingness criterion, first we e search for all individual elements that is one element item sets that have a minimum support of s. Now we start, we go into a loop where we start looking for item sets of sizes higher greater than 1.

So from the results of the previous search for i element item sets, search for all i plus 1 element item sets that have a minimum support of s. This in turn is done by first generating a candidate set of i plus 1 item sets and then choosing only those among them which have a minimum support of s. Now this becomes the set of all frequent i plus element item sets that are interesting. So this loop is repeated until the item set size reaches the maximum. That is there no more candidate elements to be generated for the next item set or there are no more frequent item sets in the current iteration.

Now that was about item sets. A property of item sets is that there is no, I mean you basically consider item sets as one entity that is there is no ordering between the item sets. that is it does not matter if somebody buys a bag first or a uniform first or a crayon first or whatever, as long as the, only thing that we are going that we infer from this is that the item set bags, uniforms and crayons are quite lightly to be bought together in in in one piece.

Therefore if I am let us say a super market vendor, I mean someone having a super market then it would make sense for me to place bags and school uniforms and crayons next to each other. So because there is a higher probability that all three of them are bought together. But when we are looking for association rules we are also concerned about the direction of association that is there is a sense of direction saying if A then B is different from if B then A. So association rule mining requires two different threshold, the minimum support as in the item sets and the minimum confidence with which we can talk about a, with which we can say or determine that a given association rule is interesting.

(Refer Slide Time: 37:22)

So how do we mine association rules using apriori. Again we shall do the same thing like we did in the past. We shall come back to this algorithm or the general procedure after we have illustrated an example by which we can mine apriori, using apriori algorithm by which we can mine association rules.

(Refer Slide Time: 38:47)

Now the main idea is the following. Now use the apriori algorithm and generate the set of all frequent item sets. So let us say we have generated a frequent item set of size 3 which is namely bag, uniform and crayons with a min sup or of 0.3 that is a minimum support threshold of 30%. Now this bag, uniform and crayons can be divided into the following

rules. If bag then uniform and crayons or if bag and uniform then crayons or if bag and crayons then uniform and so on so forth.

(Refer Slide Time: 39:38)

Now what is this thing mean? this thing means that when a customer buys a bag then the customer also buys uniform and crayons and this rule means that if a customer has bought a bag and a school uniform then the customer will also buy a set of crayons or if a customer has bought a bag and a set of crayons then the customer will also buy a school uniform and so on.

Now we have got all of these different association rules. Now each of these association rule has a certain confidence based on this data set. Now what is the confidence for each of these rules? What is the confidence for the rule if bag then uniform and crayon. That is if a customer buys a school bag then here she will also buy a school uniform and a set of crayons. In order to calculate the confidence of this, we have to first look at which are all the item sets here that have bags that is where the customer has bought a bag. So, there are 1 2 3 4 5 6 7 8 different entries where customer has bought a school bag.

Now among these 8 entries, in how many different entries did the customer also buy uniform and crayons? 1 and 2 3, so there are 3 different entries, 3 different instances out of 8 instances where this rule holds. Therefore whenever a customer buys a bag, one can say with 3 by 8 or 37.5% of confidence that the customer is also going to buy a set of school uniform and crayons. Similarly we can calculate the confidence for each of these other association rules like this is 0.6, 0.75, 0.428 and so on and so forth.

Now, given a minimum confidence as a second threshold and suppose we say that the minimum confidence is 0.7 then whichever the rules that we have discovered, every rule that has confidence of at least 70% or more.

That means we have discovered the following three rules, bag if bag crayons then uniform, uniform crayons then bag and crayons then bag and uniform. What is that mean in plain English? It means that people who buy a school bag and a set of crayons are likely to buy a school uniform as well that is bag and crayons implies uniform.

(Refer Slide Time: 40:47)

Similarly people who buy a school uniform and a set of crayons are also likely to buy a school bag that is here, somebody buys uniform and a set of crayons then they are also likely to buy a school bag. Similarly if somebody buys a set of crayons then they are very likely to buy a school bag and a school uniform as well.

(Refer Slide Time: 43:09)

So that is here, that is somebody buys crayons then with 75% confidence one can say that they also buy bags and school uniforms. So again it's a question of direct marketing or whatever. If somebody is interested in crayons then you might be reasonably sure that they are also interested in a bag and a school uniforms so on. Now so let us look at look back at the algorithm here (Refer Slide Time: 43:41) for mining association rules.

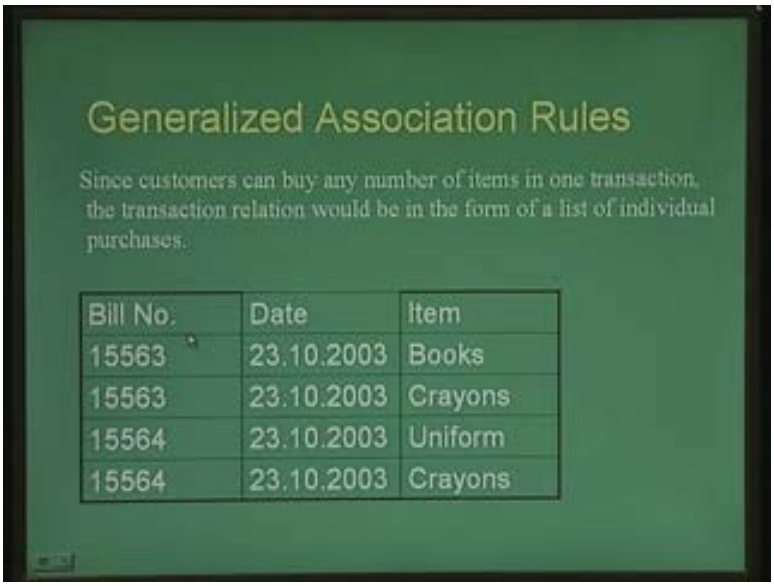
Simple mechanism for mining association rules is first of all use apriori to generate different item sets of different sizes and at each iteration, we can divide each item sets in to two parts an LHS part and an RHS part, the left hand side part and the antecedent and precedent that is the right hand side part.

So this represents a rule of the form LHS implies RHS. Then the confidence of such a rule is support of LHS divided by that is support of the entire thing divided by the support of LHS. That is support of LHS implies RHS divided by support of LHS will give us confidence of this rule. And then we discard all rules whose confidence is less than minconf.

So now let us look in to the question of how do we generate or how do we prepare a tabular data for association rule mining or let us say item set mining and so on. Now because we use let us say relational data set, relational database you might have observed that or you might have got a little doubt when we have been considering a data set like this. There is something peculiar about this data set. What is peculiar about this data set here? The peculiarity is that it looks like every consumer coming to this store is buying exactly three items which is very unlikely.

In fact what is more practical is that this set, this data set contains records of variable length. That is one customer may have bought just two different items whereas some other customer may have bought 10 different items whereas a third customer may have bought only 5 different items and fourth customer may have bought only one item and so on and so forth.

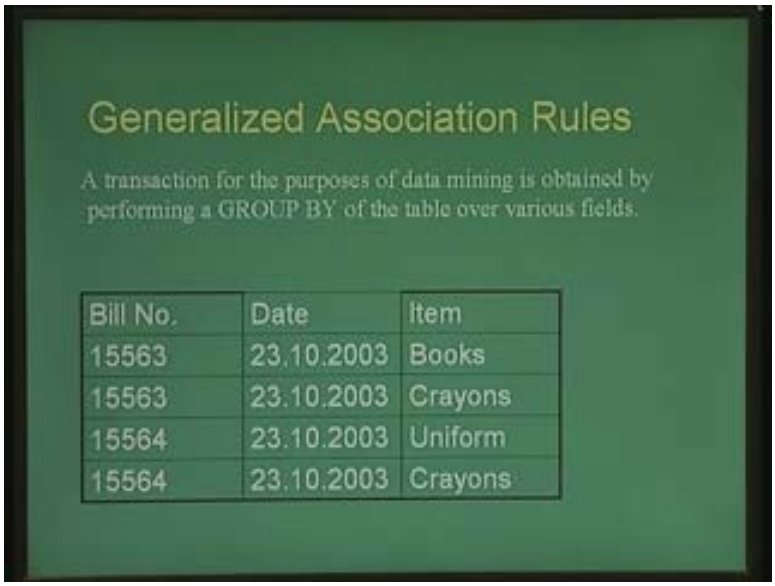
(Refer Slide Time: 46:34)



So it is not possible to represent this item set like a table, like a well form table like this because basically it is a set of all items of different lengths. In fact the best way to represent this would be in a normalized form let us say in a database where for example the same bill number here 15563 15563, both of this refer to the same customer. That is it's the same customer who has bought books and crayons and this is not completely normalized because date is not really necessary here but nevertheless here all of these records are of uniform length, if you order this based on the set of bill numbers then we get the set of all different transactions.

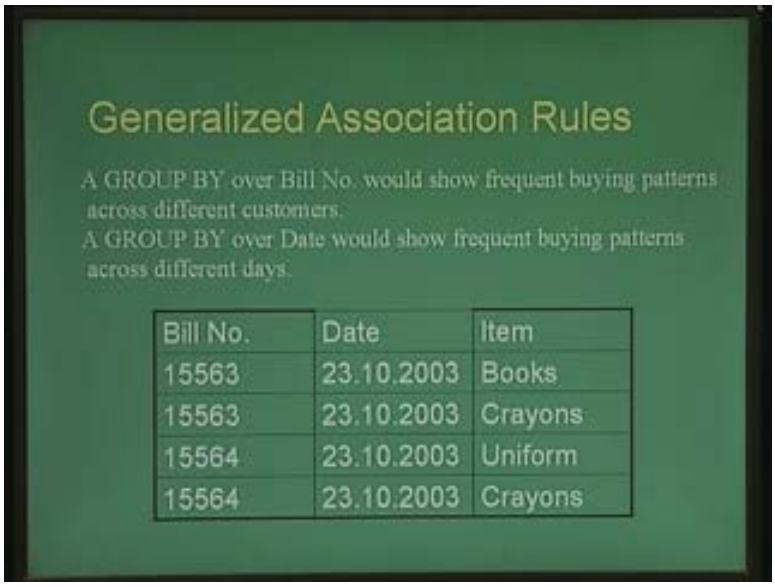
Now depending on what we are looking for this, this ordering might make a difference. How does this ordering make a difference here when we are looking at data set like this? Suppose given a dataset like this, here performing group by's on different fields will yield as different kinds of behavior data sets.

(Refer Slide Time: 00:47:19)



So what does it mean? Suppose let us say we perform a group by based on the bill number.

(Refer Slide Time: 47:37)



So suppose we perform a group by on the bill number on this table then each group will represent the behavior of one particular customer that is one bill represents one or one bill number represents one particular customer or one particular transaction. So suppose we group by based on bill numbers and then perform apriori across these different groups then we would be getting frequent patterns across different customers.

On the other hand suppose we group by over date, so rather than bill number. So all transactions happening on a given date will come in to one group and all transactions happening on another date will come in to another group but a given date may have transactions from several different customers but all of them are now grouped in to one single group. And suppose we run apriori over this set, over this different groups then we would actually be looking for frequent patterns across different days that is across the different dates. So we have to interpret what we mean by something that is frequent based on how we have ordered the data. If we have ordered the data over different customers then it would show aggregate behavior over the set of all consumers with whom you are interacting with.

On the other hand if you are running apriori or if you have performed group by over dates then it would show you aggregated behavior over a given time period rather than over the set of all customers. Well, it also includes the set of all customers but what is more important here is that how does the behavior or how has the behavior changed over time. So if something is frequent over time, it means that it is uniformly or in some sense consistent over this entire period of time.

So let us summarize what we have learnt in this session. We started with the notion of data mining and like I said in the beginning, data mining is a very interesting sub field of databases which has elucidated a lot of interest not just from researchers or and not just from the technology perspective but from several other perspectives like defense perspective or security perspective, commerce that is business perspective and so on. And there are several debate that have raged on whether it is right to use data mining to look for certain behavior pattern.

for example would it be right, if a government uses data mining over let us say the set of all different activities of people and find out the behavior pattern of any particular individual and so on. And their pros and cons on both sides of the debate, one would say for security reasons it is right to look for behavior patterns and one would say well for privacy reasons it's not right to look for behavior patterns and so on and so forth. so it's a topic which is very much pertinent and has spond a huge amount of interest from several different areas.

And data mining is in some sense, I called it as sub field of databases but that's not entirely true in a sense that data mining and knowledge discovery many would claim is a field in itself. That is it relies on database concepts as well as several other concepts like learning theory or statistical inference and several other concepts in order to perform data mine. So don't be really surprised if one would say that a data mining is a complete field in itself and its only associated with databases not really sub field of databases.

but anyway data mining as we said is the process of discovery of previously unknown patterns in the sense that we have not really sure what is it that database is going to give us or what new pattern or what new nugget of knowledge so to say is we are going to learn as part of the data mining process. As a result there is no query as part of a data

mining process that is a data mining algorithm is based around one or more interestingness criteria rather than a given query.

(Refer Slide Time: 50:11)

And we saw that in conceptually, it is in some way the opposite of statistical inference where we start with a null hypothesis and either refute or prove or hypothesis by sampling, statistical sampling of the population. While here we don't start with a hypothesis but the end result of the data mining process is the set of patterns which can help us in formulating a hypothesis. We also saw the notion of association rules and item sets as well and the concepts of support and confidence and two different algorithms the apriori algorithm for mining frequent item sets and from which we also saw the apriori algorithm for mining association rules. In the next session on data mining, we are going to look at several other algorithms like say classification or discovery. So that's brings us to the end of this session. Thank you.