Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

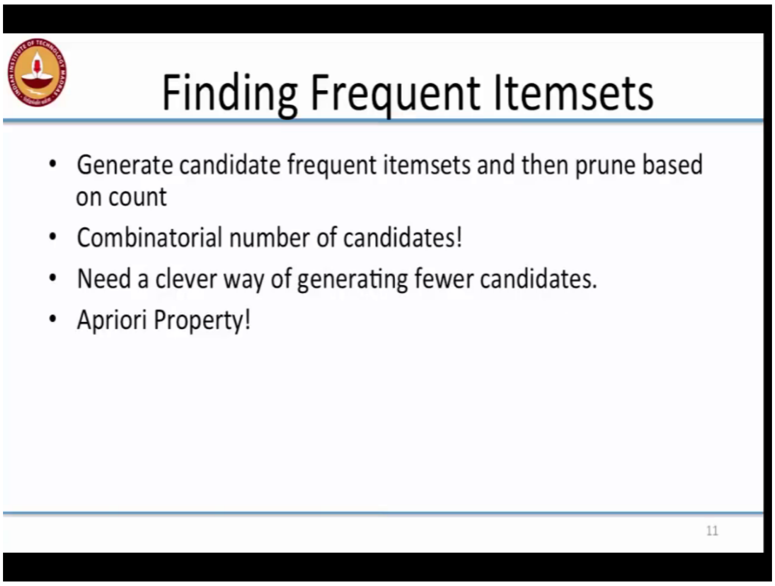
Indian Institute of Technology, Madras

Module - 07

Lecture – 38

Association Rule Mining Frequent Pattern Mining (continued)

(Refer Slide Time: 00:22)



Finding Frequent Itemsets

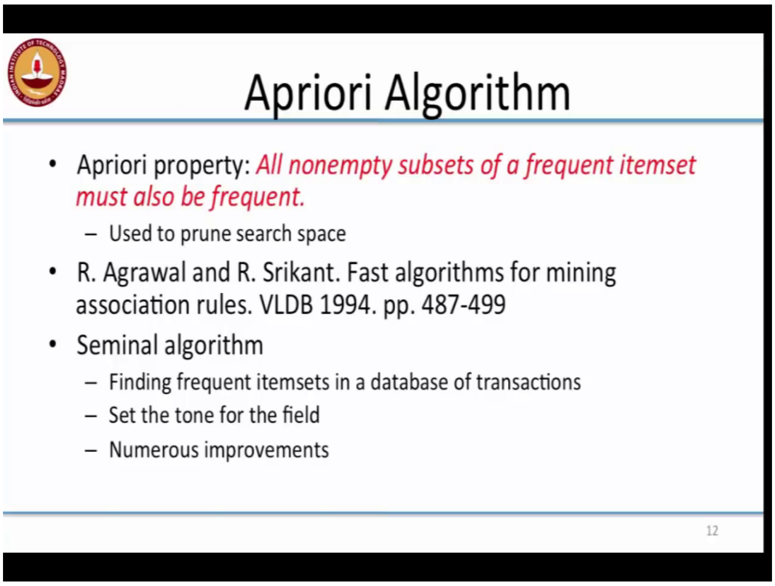
- Generate candidate frequent itemsets and then prune based on count
- Combinatorial number of candidates!
- Need a clever way of generating fewer candidates.
- Apriori Property!

Hello and welcome to the 2nd module of Association Rule Mining. In this module we look at a very popular algorithm that is used for association rule mining. As I said earlier in main problem in mining association rules is counting part. So, typically how you would do, it is that you generate candidate frequent item sets then, count how many times they occur. And then you prune the candidates based on the count. So, if their count is above certain frequency then, you are going to call them as frequent item sets and if they are below a certain frequency you are going to reject that.

But then, the problem with doing such a approach is that you have a combinatorial number of candidates. So, think about it. So, we have like 10 different items from which your transaction can be drawn; and that is say you are considering item sets. So, that essentially gives you 10 choose 2 candidates item sets; now this is a small number. Think about real applications where these candidate sets can be earlier. So, somebody like

Amazon or a Flipkart is great are like millions of candidates. And how would you even generate candidate item sets some more than is very small size of candidates. So, really need a clever way of generating fewer candidates. So, the very first approach for looking at generating fewer candidates came what late 90s which is the apriori property.

(Refer Slide Time: 01:54)



Apriori Algorithm

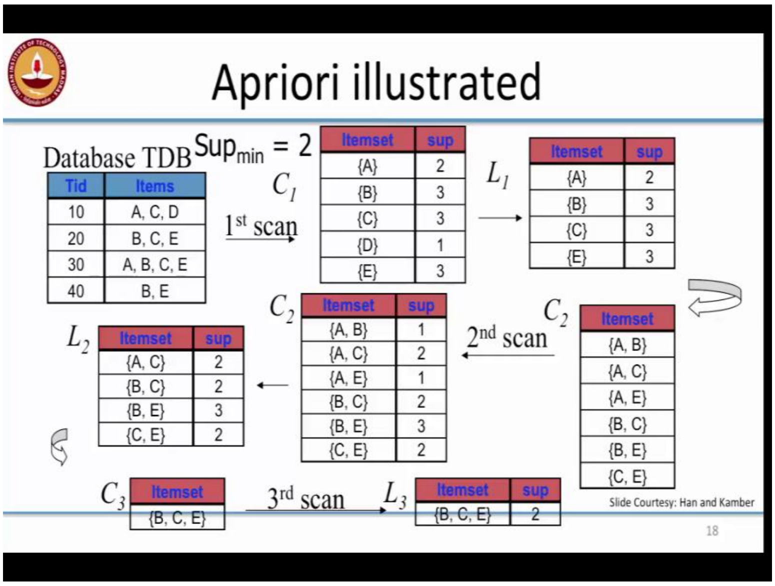
- Apriori property: *All nonempty subsets of a frequent itemset must also be frequent.*
  - Used to prune search space
- R. Agrawal and R. Srikant. Fast algorithms for mining association rules. VLDB 1994. pp. 487-499
- Seminal algorithm
  - Finding frequent itemsets in a database of transactions
  - Set the tone for the field
  - Numerous improvements

12

So, the Apriori property is very clever observations and just says all nonempty subsets of a frequent item set must also be frequent. And even if one of the subset is not frequent, then I do not have to consider the larger item set. You do not even have to count the larger item sets.

So, the first table that propose this Apriori property and came up with the fast algorithms for mining association rules was purposely 1994 and it is being a very similar algorithm. That they found first they introduce the problem of finding frequent item sets and data base of transaction. It is said the tone for the entire field. In fact, this use of the term mining association rules, almost was the recent for the whole sub field of data mining. And it since then there have been numerous improvements it have been propose on top of Apriori algorithm, that allowed to main really large data sets scale up to will be Millions and so forth. But still Apriori algorithm swap we have to start; alright explanation into association rules. So, in this module I will talk about the Apriori algorithm and give you more like introduction by an example and so, I will leave it you to look up other algorithms if you are interested in association rules.

(Refer Slide Time: 03:26)



Apriori illustrated

Database TDB $Sup_{min} = 2$

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

C_1

1st scan

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

Slide Courtesy: Han and Kamber

18

So, here we will start up with a very simple example. So, look at very small transnational data base. So, it is got only 4 transaction learning, so each of these have different id. And there are total of 5 different items which could be figuring in these transactions. So, we are going to call the A through E.

So, let us look at the set of 1item sets. So, we have A through E and this complete frequency of these item sets. Let us suppose that I have a minimum threshold for to support of 2. So, I have 4 transactions and item set can be called frequently appears in a least 2 of these 4 transactions. So, we can immediately see that item set D is not frequent. So, not only is the 1 item set D if not frequent, any larger item set that contains D as element in it can also not to be frequent.

So, when I am looking at the candidates, I have to use for generating 2 item sets I can completely ignore D. And now, all the candidate 2 item sets or those that do not have E as, I mean do not have D as part of it. So, essentially the 2 item sets that will have to consider are: A, B A, C A, E A, C and B, E and C, E. So, these are essentially generated by looking at all possible combinations of the frequent 1 item sets.

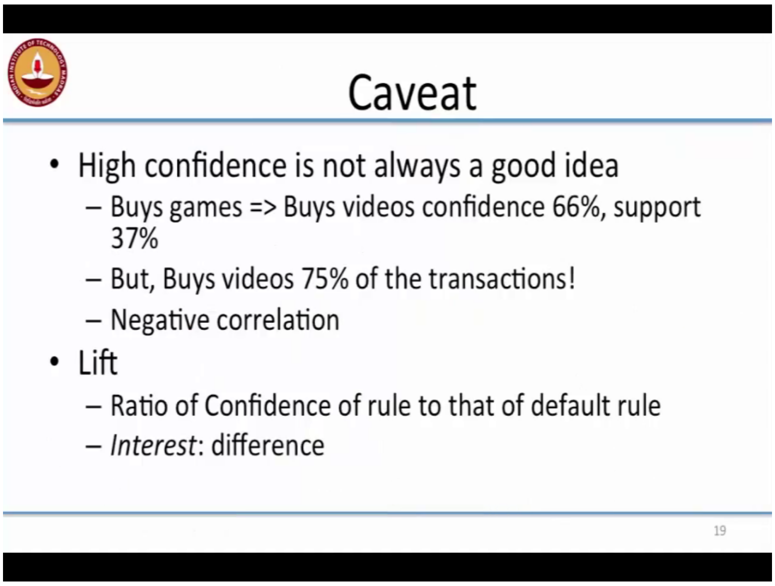
Now, that we have these candidates 2 item sets. So, we not the second scan through the data then, we count them frequency of the 2 item sets. Now, we can see effect A, B and A, E on actually not frequent among these and therefore, we can thrown those away. So, we have a list of frequent 2 item sets, which compares A, C B, C B, E and C, E. Is it clear so far? So, we had we started out of with all the 1 item sets, these are the candidates

1 item sets. Counter that frequency we left out the 1 item sets, that was not frequent. And then from the frequent 1 item sets, we generated a set of candidate 2 item sets which could be frequent. Then, we counter the frequency of these 2 item sets and from there we have proven the way the 2 that were not frequent. So, these are the frequent 2 item sets of on these we can generate candidate 3 item sets. So, if you think about it. So, the candidate 3 item sets could be ABC, ABE, and BCE.

So, can ABC be a candidate. It cannot be a candidate because the sub set AB is no longer, is not frequent. That sense the sub set AB not frequent, ABC cannot be a candidate item set; and likewise can we look at BCE. Can BCE be a candidate frequent item set? Yes, because, BC is frequent, BE frequent, CE is frequent. And what about ABE? Again that cannot be a candidate item set because AA, AE is not a frequent item set. Essentially we have only 1 candidate 3 item set. And when we count the frequency of the 3 item set and then we find that exactly frequent. Among we have the complete set of all frequent item sets. So, one 3 items sets which is BCE, four 2 item sets and then four 1 item sets which are all frequent. You do not have look for larger item sets, they both potentially possible because we do have 5 elements, 5 week events. But, we do not have to look for any further frequent item set because that is only one frequent 3 item set.

So, the in Apriori property allows us to proven the number of candidate item sets that we will have to generate. So, but still there is a problem in that. So, every time we generate candidate sets of larger size, we essentially do other scan over the data. So, the more recent algorithm tries to minimize the number of scans, you have to perform over the entire data sets; because that is very expensive operation.

(Refer Slide Time: 07:57)



Caveat

- High confidence is not always a good idea
  - Buys games => Buys videos confidence 66%, support 37%
  - But, Buys videos 75% of the transactions!
  - Negative correlation
- Lift
  - Ratio of Confidence of rule to that of default rule
  - *Interest*: difference

19

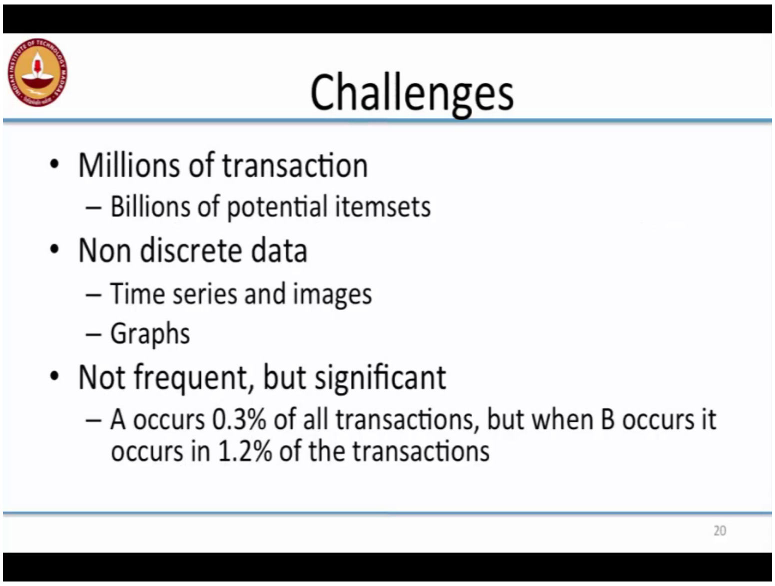
This is a big challenge but, there are one other Caveat which I want to point out. So, scaling up to large data set is a big challenge, but there is a one other caveat which I want to point out. So, high confidence is not always a good idea. So, we will have to be really careful about what we mean by high confidence. So, I can say that somebody buys games, implies they buy videos with confidence of 66 percent; the support of 37 percent. So, this is something that was actually observed in a real data set from video rental company, which also going selling video games with their shops. Because this, so from the real data they are found that somebody buys games, hence they will also buy videos with the confidence of 66 percent and support of 37 percent. So, essentially it means 37 percent of the people that came to their showroom, shop bought games and videos and 66 percent of the people to bought games also bought videos.

But then, if you just look at what fraction of their customers bought videos there is 75 percent of the customers actually bought videos. And so, even though this rule has a high confidence, you can see that if you buy a game, actually implies negatively on buying videos. This is the video store after all and so if we typically come into buy videos but, the occasional person who comes into buy a game, is not that interested in buying videos. So, it actually imply a negative correlation and if you had just blindly been using support and confidence to determine rules, then you actually trip this out as a important rule in terms of having a high confidence.

So, one measure with people gives instead of confidence and support alone, is known as lift. Lift is essentially the ratio of the confidence of the rule to that of a default rule. So, if you have a lift of 1 that means, that is really does not imply anything. So, whether A happened or not, B is always going to happened with same frequency. So, if I A implies B is the lift 1 and it is not a significant role, but if A implies B is the lift much larger than one that would be in that. So, the things 10 to truly indicator of B. But, again if you think about in this case of games and videos, were lift will be lesser than 1 and which case is indicates a strong negative corporation.

So, lift is a useful measure to have, and that lift is not only measure that people have to proposed the variety of different measures which people have proposed analyze association rules and association rules mining is a very very active area of research. And so this is essentially just an introductory module and if you are interested in association rule mining it is spent more time. And looking at the various modifications and auditions that people have come up with of viewers.

(Refer Slide Time: 11:08)



Challenges

- Millions of transaction
  - Billions of potential itemsets
- Non discrete data
  - Time series and images
  - Graphs
- Not frequent, but significant
  - A occurs 0.3% of all transactions, but when B occurs it occurs in 1.2% of the transactions

20

So, just to summarize we the major challenges in association rule mining is how do you extent to these millions of transactions. So, there could be billions and billions of potential item sets. So, how do you do the pruning efficiently? How do you minimize the number of parcels? How do you reduce the number of memory that is required when you are doing the counting? So, there are many many rules and many issues that we have to consider in claimed this large data sets.

So, we are talked about transnational data now, so some sets easy to count and it was discrete event that are happening; but, what about the continuous data like time series data and images? How would you go about even identifying what are appropriate items more which will be doing this counting. And data with rich structure like graphs. So, frequent pattern mining in graphs is a very important topic which is you used in diagnose area like graph discovery and social influence prediction and so on. So, how would you can you average a structure or can you make computation or efficient and it handle in see structure data. So, that is something which is a very active area of research and experimentation.

And one important twist to this whole frequent pattern mining issue is that, we are not sometimes not interested to in frequency within significance. So, if A occurs let us say in 0.3 percent of all transactions; but when B occurs it occurs in 1.2 percent of the transactions. So, it means B implies A and that is a significance effect, because it improves the frequency of A from 0.3 to 1.2. And A might be that in the A stack that we are looking for and finding B. And actually gives us greater evidence for the presence of A. So, such patterns or significant, but if you just go by the frequency a loan, these are very infrequent occurrences of a data. So, how do we actually look at such frequent, none frequent but significant pack makers. So, that is a big challenge in the association rule mining community, again like it is a very activity process and keep developing that brings us to the end of these.