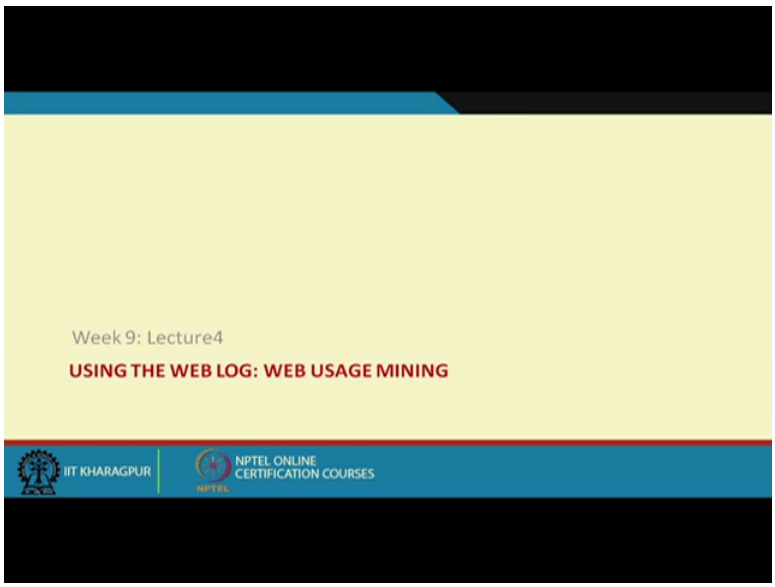
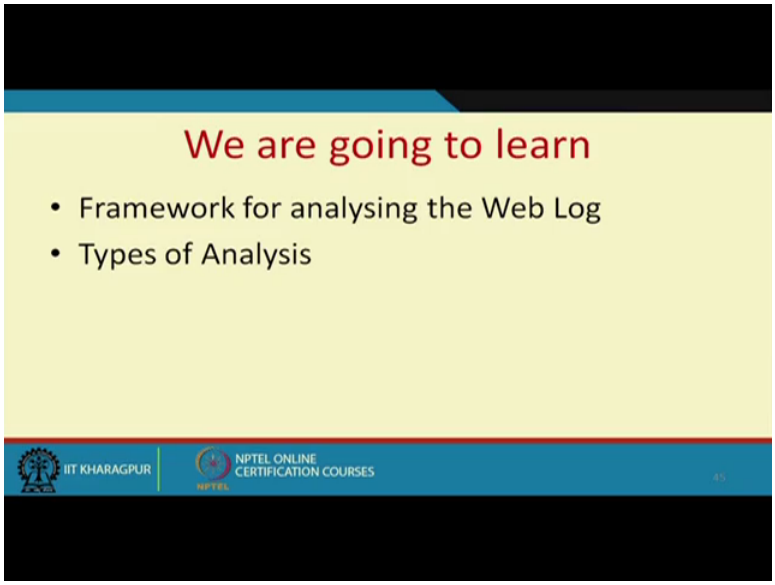
E-Business.
Professor Mamata Jenamani.
Department of Industrial and Systems Engineering.
Indian Institute of Technology, Kharagpur.
Lecture-47.
Using The Weblog: Web Usage Mining.

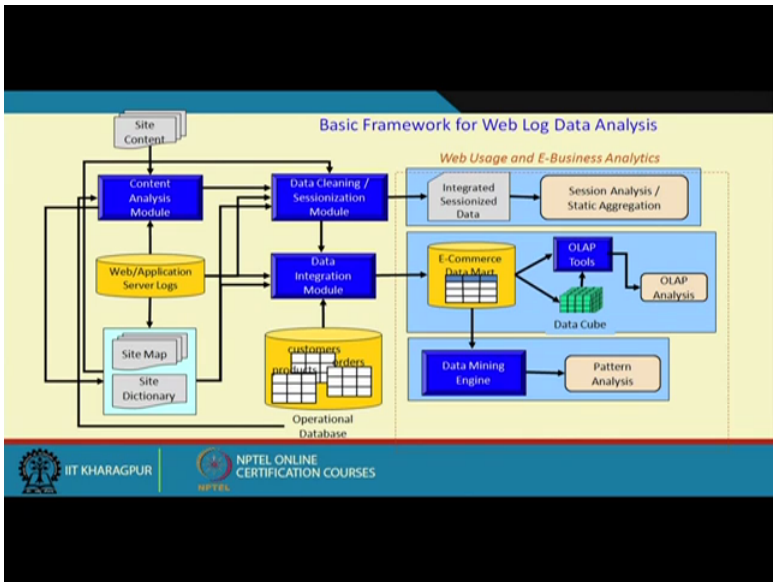
(Refer Slide Time: 1:06)



Welcome back, so far we have been talking about how to a specific data source called Access log file and from that we understood that what all problems, though this file automatically collected by the web server and collect the user's navigation details in order to identified user's behaviour from this and get some meaningful insights, we 1st have to remove the problems associated with this. We have to find out the ways and means to get rid of the problem that is it has and in this context we saw that various free processing tasks that one conducts to finally bring it into the session form and find out the complete part of the user's activity and so on.

(Refer Slide Time: 1:31)





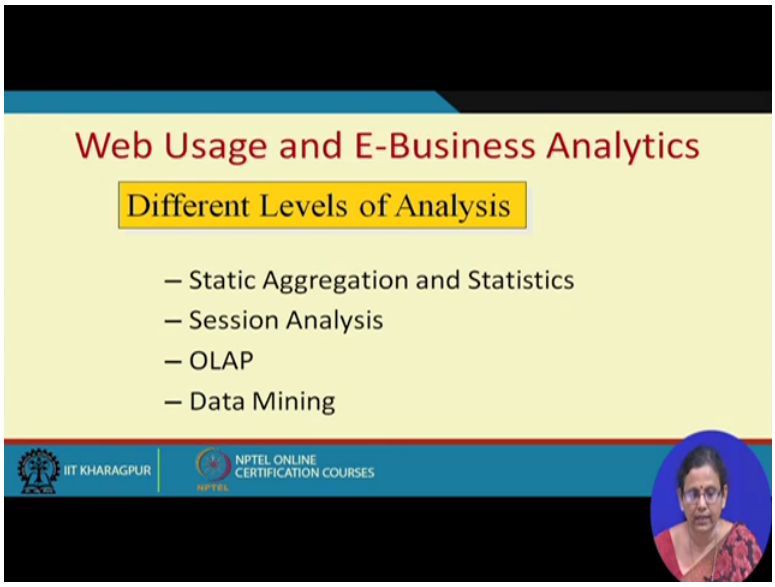
So today's class we are going to look at how this, once this is cleaned and all, how this weblog is used for various activities. So here we are going to talk about the framework for analysing the server log and different types of analysis possible thereof. So this is the basic framework for weblog data analysis. Every site has contents in various contents in webpages, log captures this and you get the and this weblog captures the details from the site map or site additionally if it will you maintain any, then the site content.

Then this, look at this site map is also gives you, gives the input during data cleaning phase, this content gives the input for creating a site dictionary and using all these details, taking input from a site map, from content, from Access log, finally you have this data cleaning

sessionization process after which you generate that flat file that containing various feel that we discussed and then we try categorise this data to learn about the customer behaviour, product access pattern, ordering, ordering nature of ordering various products, the process of ordering various products and so on, which can be used for various other purposes.

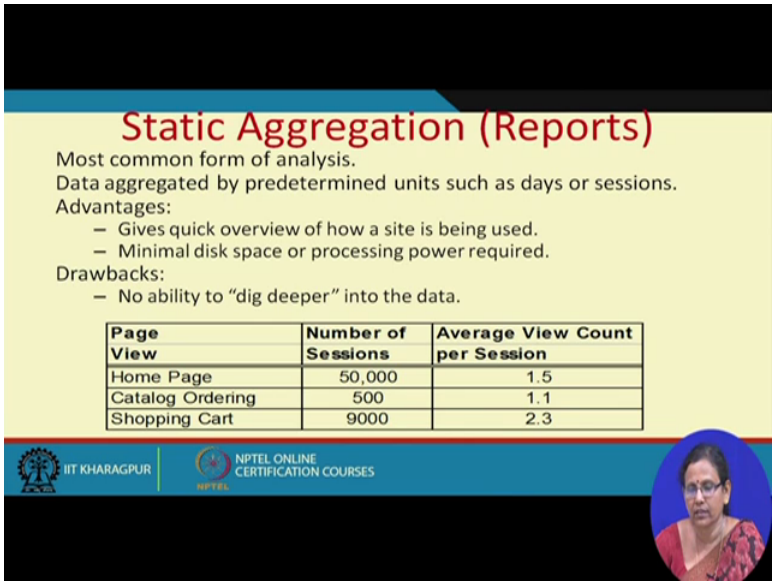
Then this sessionized data can be used for various session analysis and static aggregation. The data which combines both this session data as well as other content and other detail data, they can be kept in a data warehouse, sorry data mat, data mat is a specific, is a multidimensional view of the data belonging to a specific category. So very can use OLAPP tools to have a multidimensional view of the data, you can apply various data mining algorithms when this data comes to determining engine, the algorithms can be applied to generate patterns.

(Refer Slide Time: 4:46)



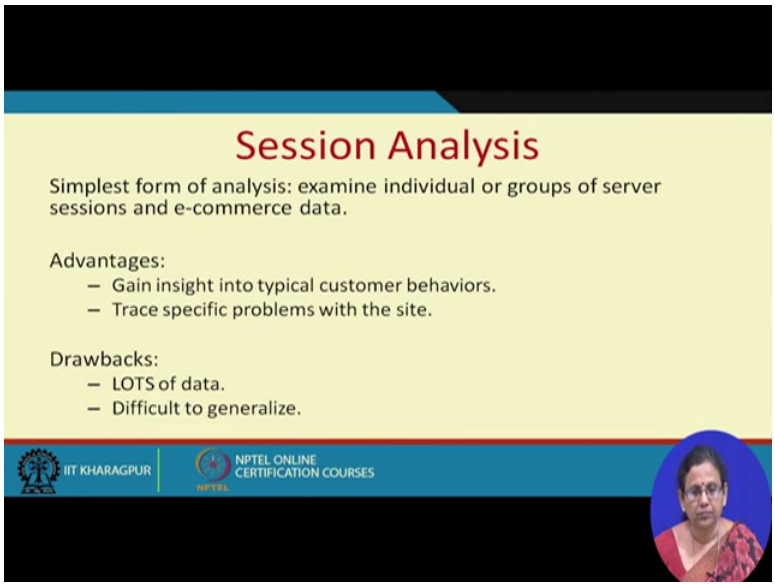
Again this level, there are different levels of analysis, 1st off is static aggregation and generation, finding the generic statistics, session analysis and online analytical processing which is for for the data warehouse or the data mat where the group of data from multiple sources have been kept, then you also apply the data mining. Now 1st task is using this data you can generate static aggregate, through the process of static aggregation, you can generate various reports.

(Refer Slide Time: 5:15)



So this is the most common form of analysis, this data is aggregated by predetermined units such as day or session. The advantage here is this can give you quick view of how the site is being used. And you do not, because you do not involve any analytical process, analytical you know modelling here, it does not require much processing and it also minimal disk space it requires. Then the drawbacks, it cannot give any deeper insight into the data.

(Refer Slide Time: 5:59)



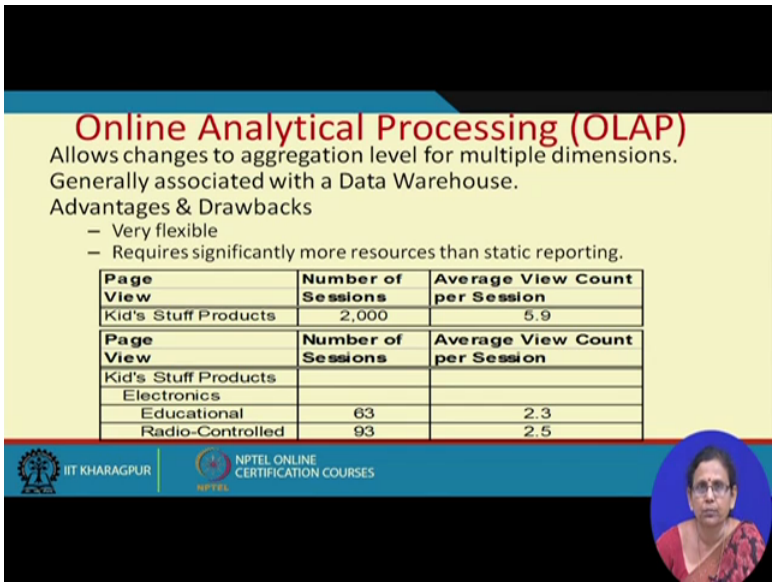
For example, here you can say like if your page views are for homepage, how many number of sessions contained this homepage viewing and average number of counts for sessions which work for this viewing homepage and other related pages, how many number of sessions were for catalogue ordering, how many sessions people use shopping cart, doing this

shopping cart operation, what was the average number of page views and so on. During session analysis, which is also quite simple, you come, of course, building the session requires a plot of, involves a lot of heuristics that we have learnt.

There are certain time based heuristics, there are certain you know not only time, looking at the content, how the contents are related, it is another heuristics, so there is groups of heuristics. But after using those heuristics, once you identify the sessions, from the sessions, examining individual groups , group behaviour in each session is also quite important. So analysing each of the group of sessions can give better insight into customer activities and it can trace specific problems with the site like if some of the pages which people are trying to browse together, whether they are linked or not properly and so on.

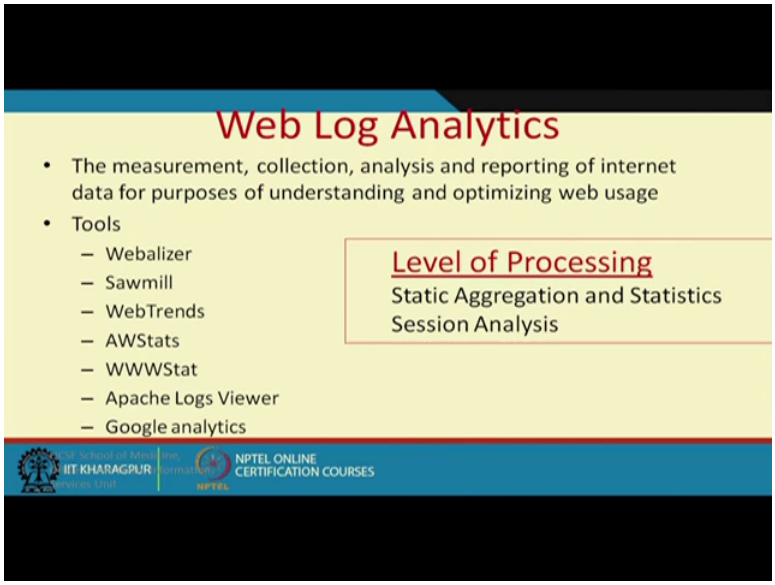
But anyway if you simply group them in a very crude banner, group the sessions in a very crude manner, then it is very difficult to generalise this data. So you would need specific user behaviour models in here.

(Refer Slide Time: 8:20)



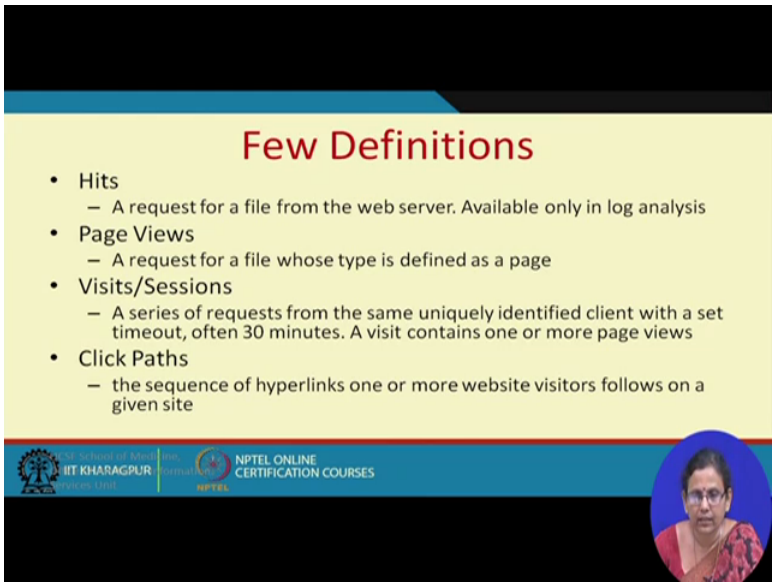
Next is you can have a multidimensional view of data using OLAP for example you can dig little bit deeper. For example let us say some sessions are associated with browsing kids' products. And within that session again some people only browse educational products, I mean browse only the electronics products and within electronics of people browsed only educational products and so people browsed some radio control, radio controlled like your toys or something products. So we have already discussed about this OLAP queries while we talked about the data resources so that time it is about digging down the data further.

(Refer Slide Time: 9:36)



Then there are certain specific software tools which help giving such static reports. So these tools are called weblog analysers and the activity they do is called weblog analytics. So these weblog analytics is the process of, is the measurement, collection, analysis and reporting of this weblog data for the purpose of understanding and optimising the web usage. These are few tools webalizer, sawmill, web trends, AW stats, WWW stats, Apache log viewer, Google analytics and so on. So they, such log analytics tools either do some static aggregation and give you some genetic statistics or sometimes using their own heuristics, they do session analysis.

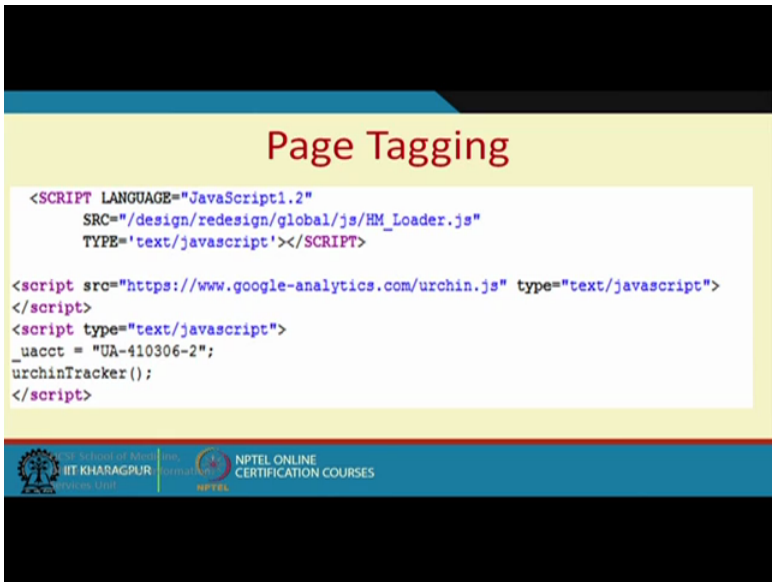
(Refer Slide Time: 10:25)



In this process, many terms are used to describe how a website is being used, you can say these are few metrics for using the, to understand the web usage. 1st is hit, which is if a request is made to the server and the server is able to find the corresponding resource and it sends back, then it is called a hit. Then you have something called page views, if a request for a specific file whose type is defined as a page, not the embedded resource, then visits per session within a session, definition of session have already told you several times and we are talking about server sessions only.

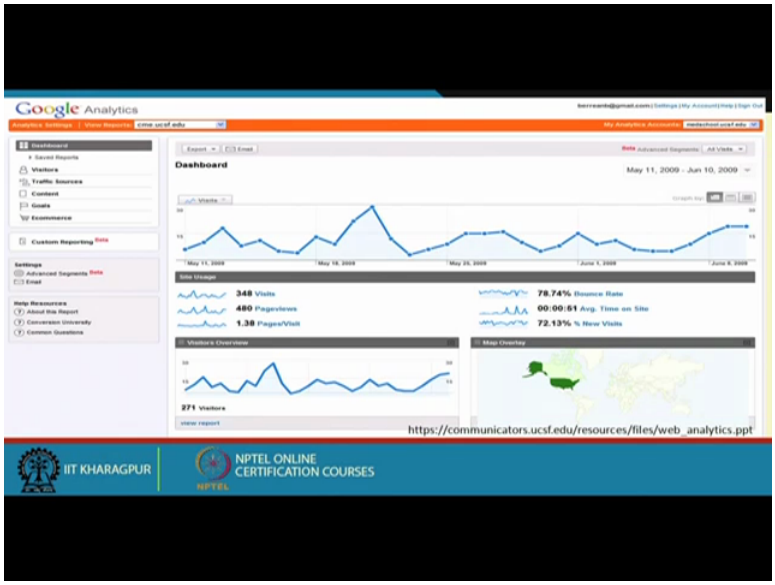
So it is a series of requests from the same uniquely identified client, again I am reminding you, identifying a client uniquely is a very tough job, but it is and requires a lot of heuristics. But the visits per session, once it is heuristically done, visits our sessions are called as a request for, request coming from the same uniquely identified client with a set of timeouts which is of 30 minutes, a visit contains one or more page views. Then click path, see this access log data is also called clickstream data.

(Refer Slide Time: 12:43)



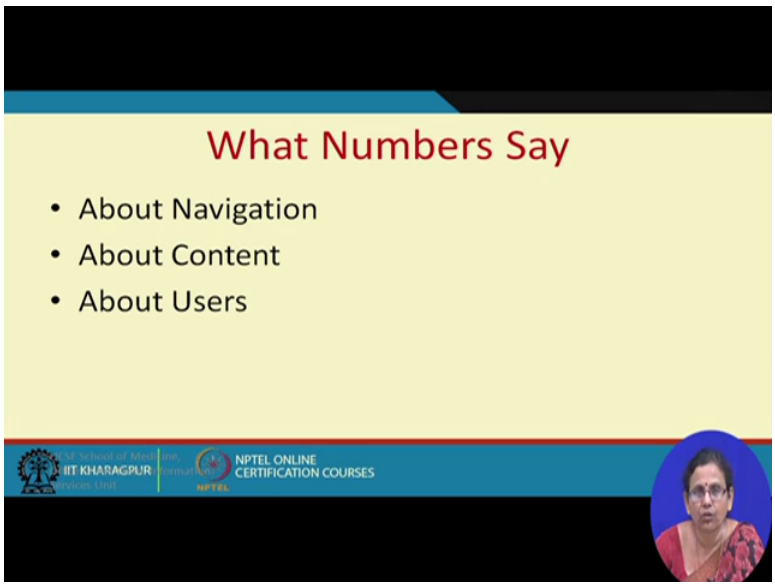
So click path is basically a sequence of hyperlinks, one or more website visitors follow a specific path, then it is called a click path, how many uniquely parts are there and so on. Now as I told you there are specifically Google analytics provides you with some JavaScript which acts like an agent and it is once it is embedded within your webpage, they start sending your user's data, to where, to your Google server so that you log into your Google account and can see the performance of your web server. Now Google has your, all your data, page view data.

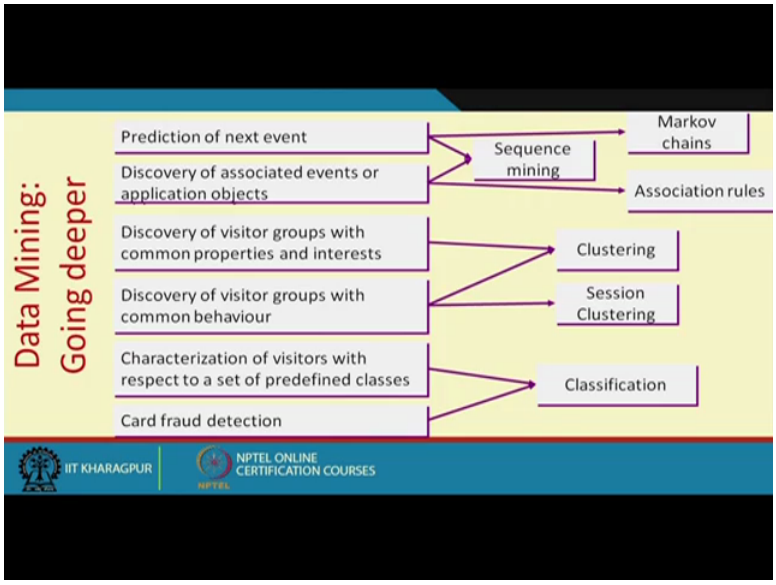
(Refer Slide Time: 13:00)



So Google, this is a snapshot from Google analytics collected from someplace and this basically shows how the page views happen across the region for example it shows how the site is used, then on various dates, then what are the geographical regions from which you have got the requests, then and once again all these reports which you get uses the pattern of, the access pattern of that this thing but in this, in case of Google analytics, Google analytics directly connects the data. But this Apache log viewer etc, the log file if you give them input, they will be generating search reports.

(Refer Slide Time: 14:02)





So these numbers can tell how the site is being navigated, whether, how exactly the people use the content of the website, it can also tell you about the site, website users. In fact to understand more about these categories, you have to take various, you have to use various tools like data mining tools to get certain insights. For example, if you like to predict what the user is going to next do, you may be using sequence mining, you may be doing Markov chain analysis.

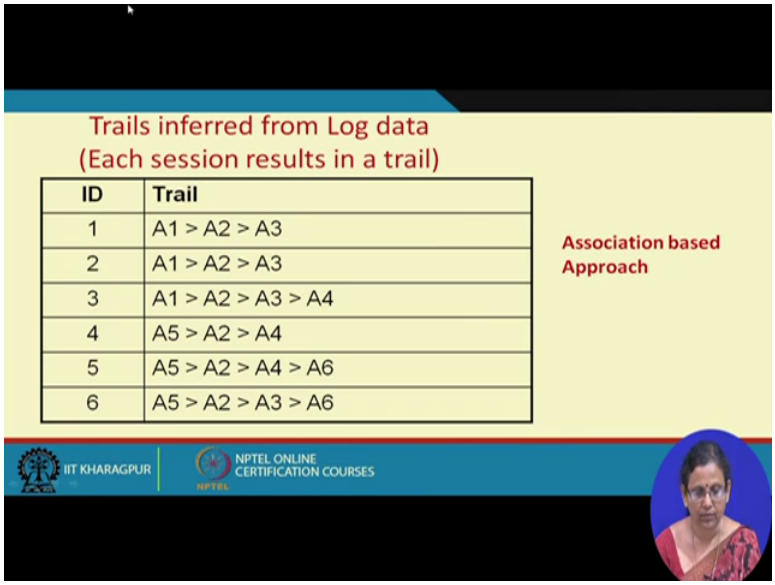
Similarly if you are discovering the associated events or application objects, you might again be doing association rule mining or sequence mining, discovering of visitor groups within the common properties of interest, you may be doing clustering, similarly discovering the visitor groups with common behaviour, you may be doing clustering or session clustering. Then characterising the visitors with respect to a specific predefined classes, you may be doing certain classification, then for example if, detecting your credit card fraud, you might be doing this classification.

(Refer Slide Time: 15:49)

So in fact the left-hand side shows the activities that you may be doing and the right-hand side shows various data mining tools that is frequently used for this purpose. Now for mining this navigation pattern, we have already talked about the sessions, the sessions usually give you a trail of webpages, user trail through the website, a set of sequence of webpages a user has browsed through. So a trail is a sequence of webpages followed by user during a session ordered by time, ordered by time of access.

For example if you would like to find out which kind of such trails are often followed by the users, maybe you can do some kind of frequent pattern analysis. So in order to understand the cooccurrence of the webpages which are happening together, you may be using either association rule mining or Markov chain modelling.

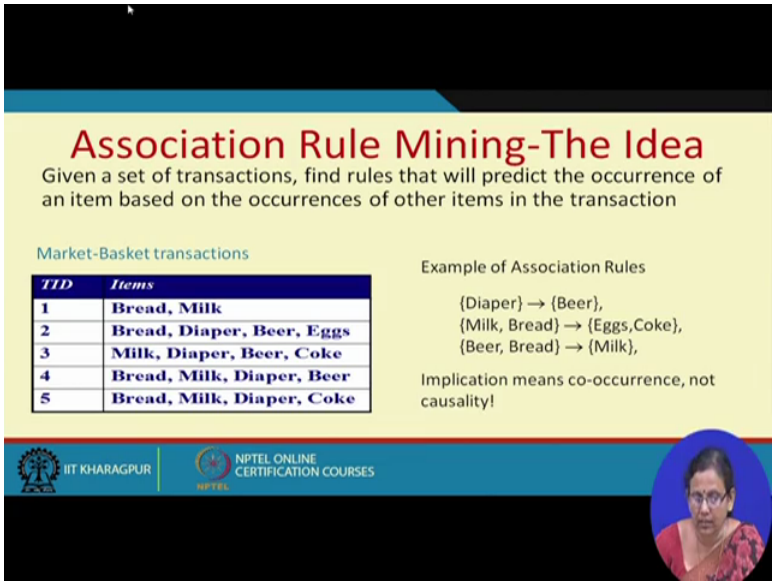
(Refer Slide Time: 17:00)



For example, if you are using association-based approach, suppose these are your trails A1, A2, A3 visited together, A1 visited together, then A1, A2, A3 visited together and so on. Suppose you have these 6 things. Now out of these how can you say which are the webpages which are visited together. For this purpose, again a new, data mining tool is used which is called association rule mining. In fact while talking about recommended system, we will be probably we will be talking title bit more about association rule mining. But right now let us try to understand that this association coalmining is not limited to this particular application of finding the related webpages which are often browsed together.

It has many more applications, it has in fact one of the very primary application of this association coalmining is your market basket analysis, this because they are not, right now we are not going to go deeper into Association rule mining and talking about recommended systems, we will be talking little bit more about this association rule mining.

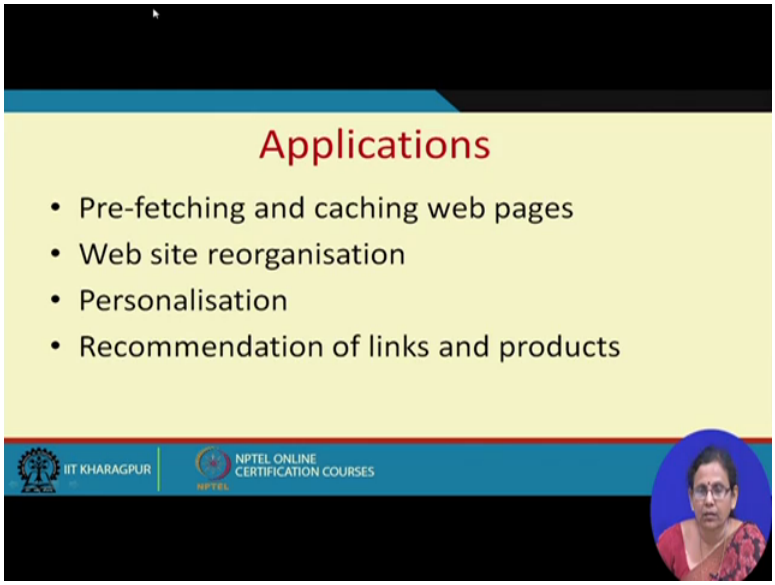
(Refer Slide Time: 18:23)

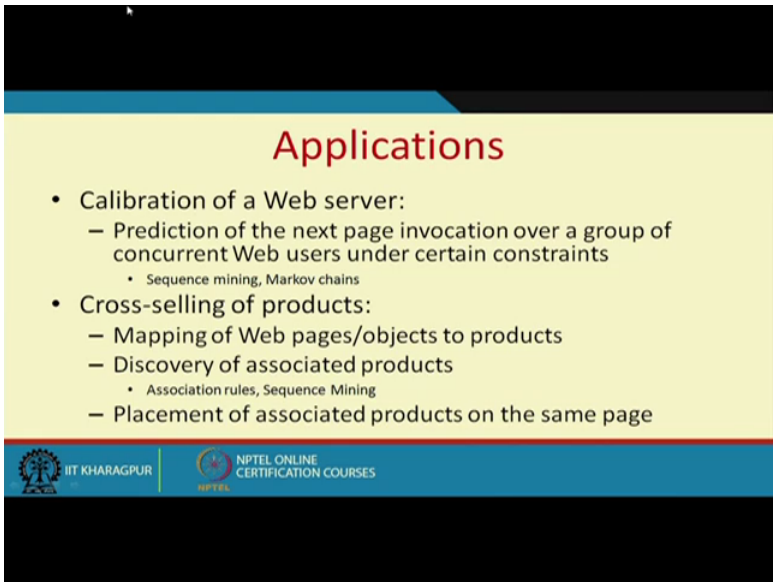


But right now let us try to understand, it is about finding out the frequent patterns where 2 pages were together, in fact this example is from that famous diaper beer example where people have, people have discovered that even unrelated products are often purchased together. So anyway this has nothing to do with our webpage access but this is a very generic example people always give that even unrelated, this is about knowledge discovery, right. So you do not know what kind of pattern you are going to get, after the analysis only you will be knowing about the kind of pattern you are getting.

But there are certain, there will be number of algorithms for this but right now anyway we are not going to talk about the algorithms. There are many applications of this web usage mining, let me remind you, this web usage mining is about mining the access loss or the sessions resulting therefrom. There are algorithms to clean that, that we have already discussed. Assuming that such a clean form it exists, many applications can be done on this particular thing.

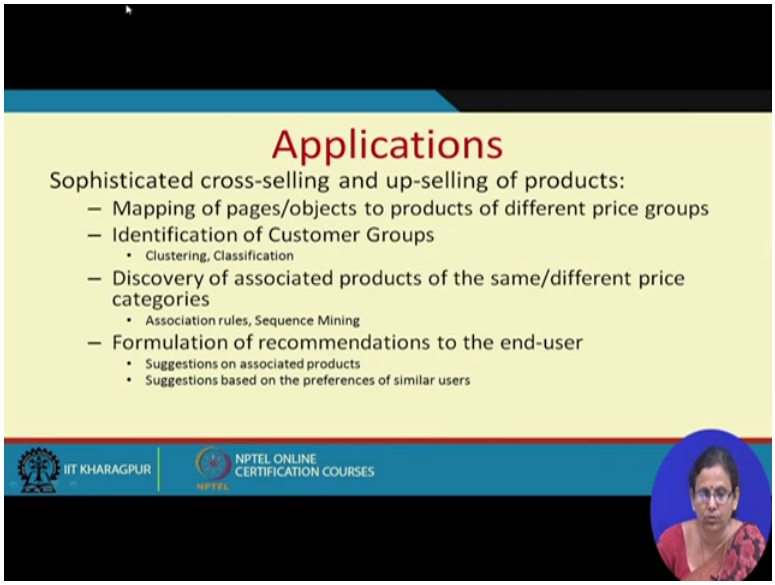
(Refer Slide Time: 19:46)





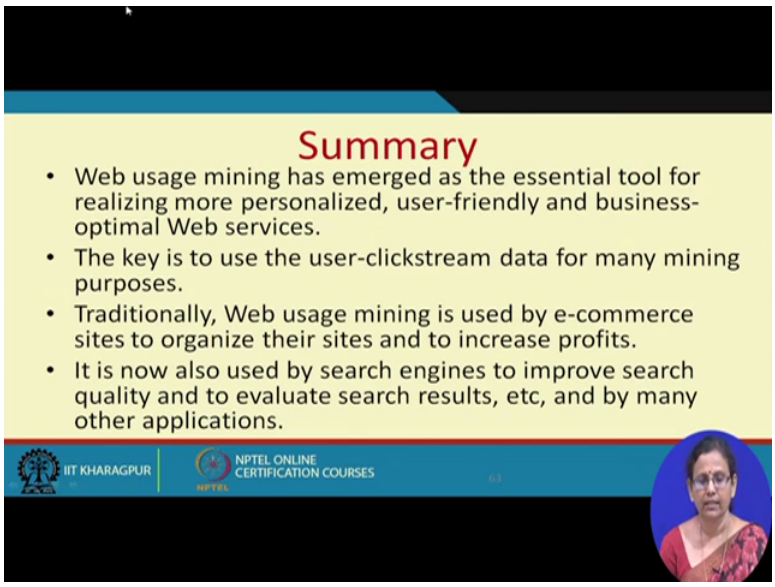
1st of all prefetching and caching webpages, website reorganisation, website personalisation, recommendation of the links and products. Various applications like calibration of web server which has, which is about prediction of the next page invocation over a group of concurrent web users under certain constraints which involves sequence mining and Markov chain. Cross selling of the products, about the cross selling and up selling, we discussed during our CRM, when we discussed about the customer relationship management. So that time we had these terms defined.

(Refer Slide Time: 21:05)



Anyway this is about mapping of webpages and objects to products, then discovering associated products which are often purchased together, then placement of associated products in the same page using, of course here you will be using the Association rule mining again. Then for the sophisticated cross selling and up selling of products, you require mapping of pages or objects to products of different price groups, identification of various customer groups, discovery of associated products under same or different category and finally formulation of recommendations for the end users.

(Refer Slide Time: 21:35)



In summary web usage mining emerges as the essential tool for realising more personalised, user-friendly and business optimal websites. The key is to use the user clickstream data, that

is access log data for many mining purposes. Traditionally web usage mining is used by e-commerce sites to organise their sites and to increase the profits. Now it is used by search engines to improve search quality and to evaluate search results and many more applications also happen on this. With this we finish this particular lecture, thank you very much.