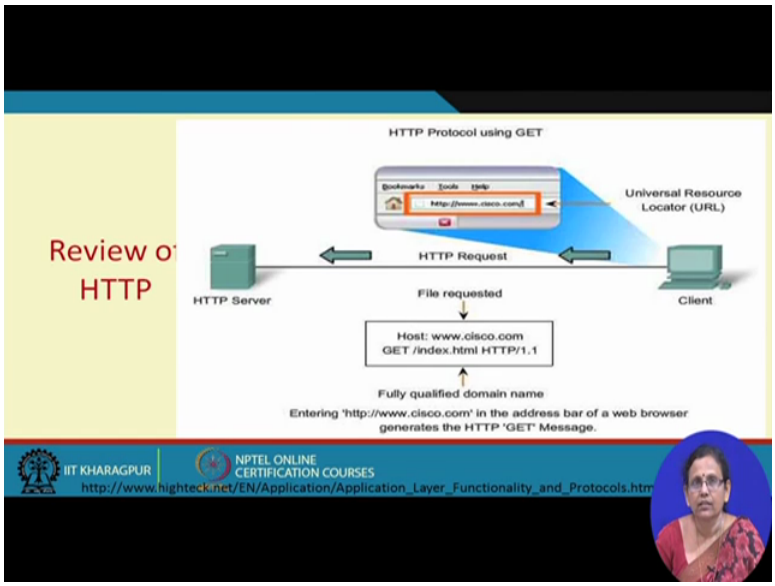
E-Business.
Professor Mamata Jenamani.
Department of Industrial and Systems Engineering.
Indian Institute of Technology, Kharagpur.
Lecture-45.
Understanding The Web Log.

Welcome back, now from today's class onwards we are going to see decision support systems at work. Let me tell you from this class onwards though I do not assume any background, I expect that at least very simple mathematical things you will be able to understand. And whenever possible I will be giving additional expression as well or directing you to sources using which you can learn more. So in this lecture we are going to come I am in couple of lectures, couple of, couple of few lectures, we are going to look at a new data source which has result because of the web transaction with the user.

How this particular data source is used in in understanding the user behaviour is going to, we are going to discuss about it. Not only user behaviour, and other allied decision-making situations, we are going to see that. So to start with, let us try understanding what is this data source. This data source we call as weblog. So let us try understanding what is a weblog. So in this lecture we are going to learn how the weblogs are generated, the structure of the access log, various preprocessing tasks and session identification. So these are the 4 things we are going to learn.

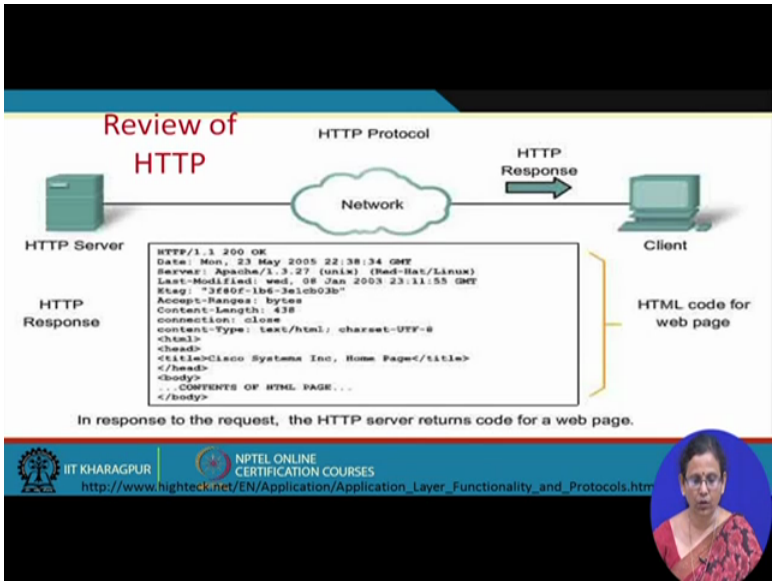
(Refer Slide Time: 2:25)



So to start with, we 1st of all need to remember the access log, the HTTP protocol that we have learnt while discussing about the technologies. So whenever any web transaction happens, as a client sitting in front of the browser, you send an HTTP request. So this HTTP request when you send, when you send this HTTP request like you generate, you send it in the form of either a get where you request the server will and get some data or it is in the form of a post, by post we mean along with the request for getting the data, you are also sending some parameters.

Think of well formed, we are filling up a form and this form we fill up, forget about the form, think of a, think of search engine, you send the search parameters, it is a 2 different situations, when you, let us say access IIT Kharagpur website, you simply write the web addresses and you get the content and everybody will be getting the same content. But in case officers did not like that of a Google, you have to provide your search parameter. So that is a kind of post but anyways we are not going to discuss the details but we, this discussion is primarily necessary for understanding the structure of weblog.

(Refer Slide Time: 4:42)

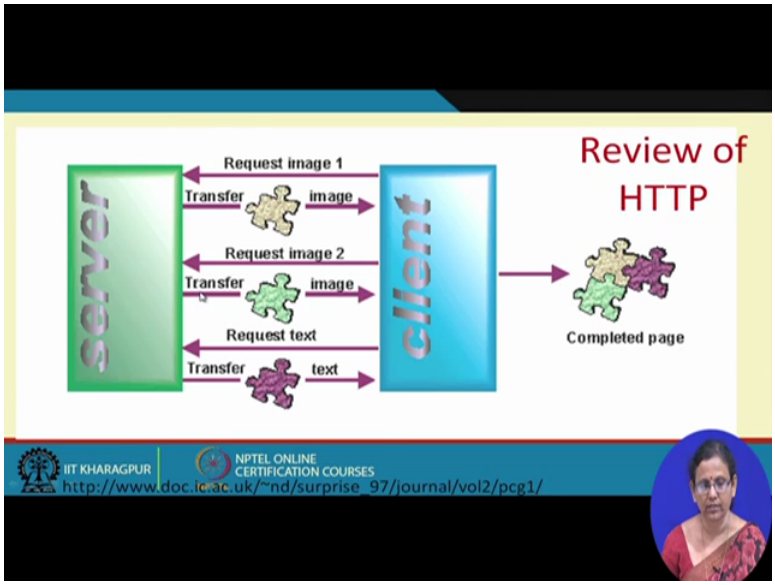


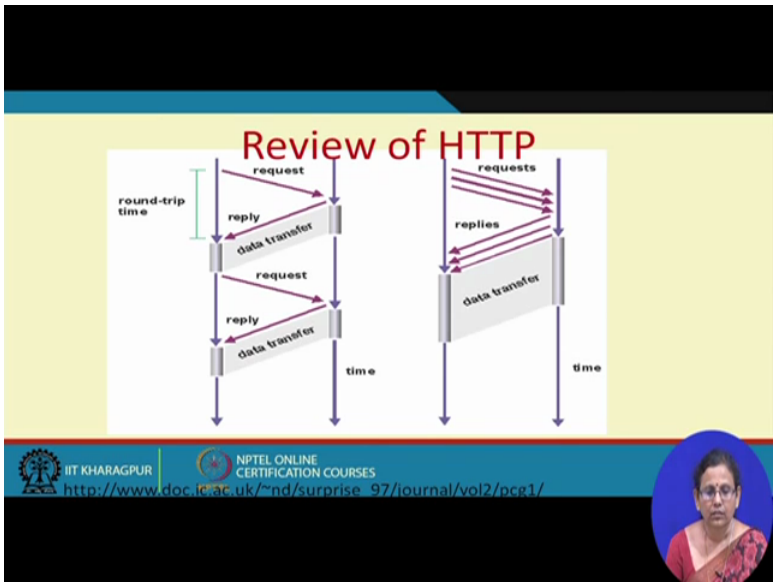
So when you send the request, you send it in the form of a get or a post, and when this particular request over the Internet is sent through this HTTP server, HTTP server generates a response back. Okay, so one it sends the response, when it gets the request and sends the response, in between it does a number of tasks, based on the user's query it tries to fetch a page, if the page is a dynamic page, it tries to connect to the database and try to fetch some data. If the page is connected, page has some embedded resources like that of your that of

your images or videos or something of this sort, audios videos, those things also are sent along with the page.

And all this information is sent back in the form of something called as HTTP response. The HTTP response uses the HTML file which come back, it is the HTML file that contains the content of the web page along with many other stuff. Like you, look at this, here HTTP 1.1, 200, okay. 200, it is a response code, it also sends one response code. Here the 200 me means tres, it is responding back and the page is there and it is sending the page and it is success. If the HTTP request results in a successful response, it is 200. You have, in fact you must have seen 404 page not found, which is also HTTP response code.

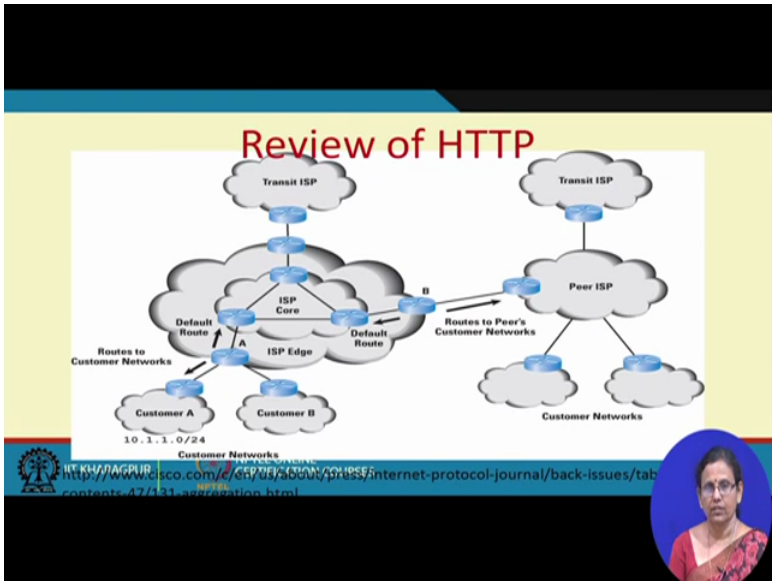
(Refer Slide Time: 6:31)





So after you get this, during this request response process, not only the, not only the data comes in but your request is broken into the sub requests for the embedded resources for, let us say for images, let us say page contains the text, image, 2 images, then all these are, all these comes to the client side and they are, browser combines them to make a complete page. So this is a complete request response model, okay. Now, in this time, in this process when the request and response comes, there are always certain time delay which is called the round-trip time.

(Refer Slide Time: 7:42)



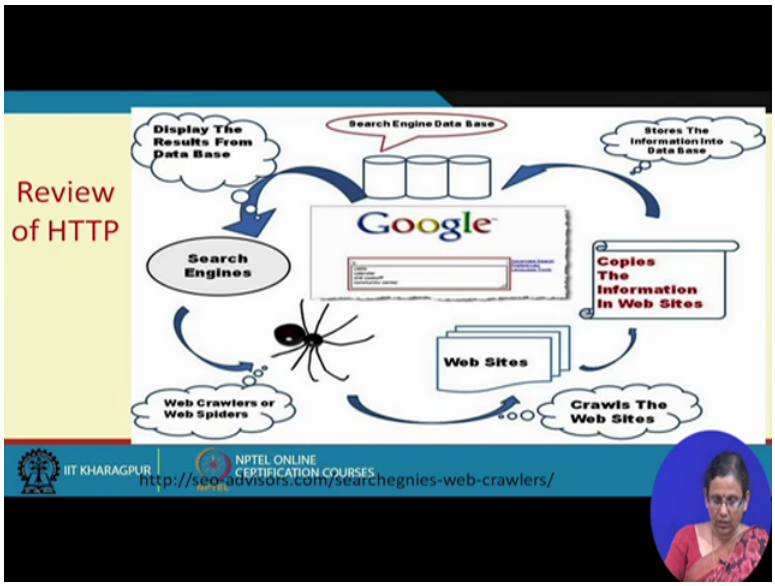
And if you remember we were telling HTTP is a stateless protocol, so before we even talk about the, whether initiative is a stateless protocol or not, before going to that discussion, let us try to understand the nature of this IP and this address, address that uniquely identifies each, in fact while talking about this we said IP address is something which uniquely identifies one entity over the Internet. Fine, it uniquely identified the entity but it is not that straightforward that we were talking about.

In fact all of us connect to the Internet using certain Internet service provider. As an organisation, we have our own LAN and when we connect to the outside world, which is beyond our LAN, outside network, which is beyond our LAN, for the security purposes of the whole organisation, all the computers are not exposed to outside. In fact all of us get connected to outside within an organisation through some proxy server. So when the data packets are routed from each customer, who maybe one individual customer or one organisation, all of us actually connect to the Internet service provider's router.

And ISPs themselves may be maintaining a number of Routers and they may be connecting to another ISP. And all this happen because through, might be happening to certain proxy server. Now, in this process, the IP address that from which you have set the packet, let us say sitting in my department computer I have sent some request, it is going to 1st my Institute proxy, then it is going to one of the ISP servers, and the ISP has its own proxy, through a few proxies of ISP, it is going to some another transit ISP and going through many search transit ISPs, finally the packet reaches at the destination.

So which means the IP of my actual IP which is visible to me in my own network is not the IP because the address translation takes place, the Times in between, by the time it reaches the, reaches the host, reaches the web server from which I have requested a page, my IP address will be different. So when the response comes back, everytime, the average translation takes place in corresponding proxy servers. And the address from which the data request has come to that the response is sent back.

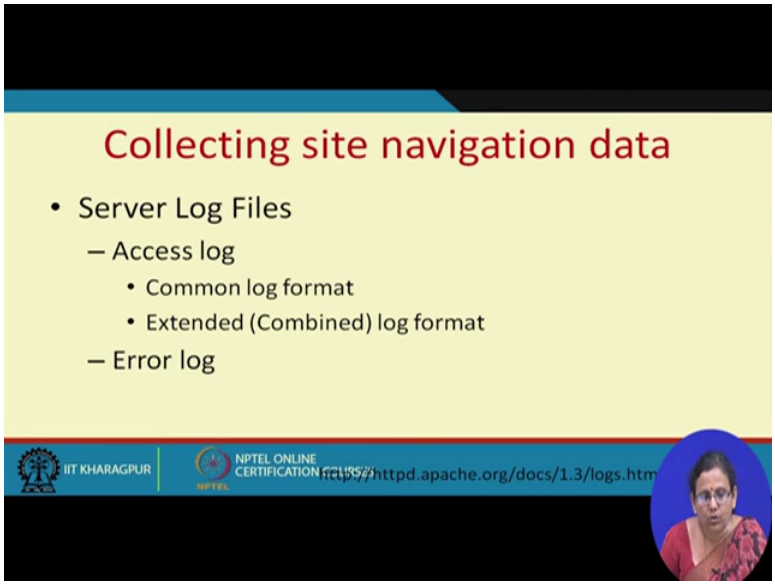
(Refer Slide Time: 12:21)



So in this reverse process, I am definitely getting the response but my exact IP which is supposed to be my, supposed to be my unique identification is not exactly known to the server. In fact when, what are we talking about, when talking about the server log files, server log files get generated due to user interaction. So what we have understood so far is the server log files do not reflect the exact IP of the user. It shows the IP of the last ISP through which the request has come. So this is somewhat little problematic.

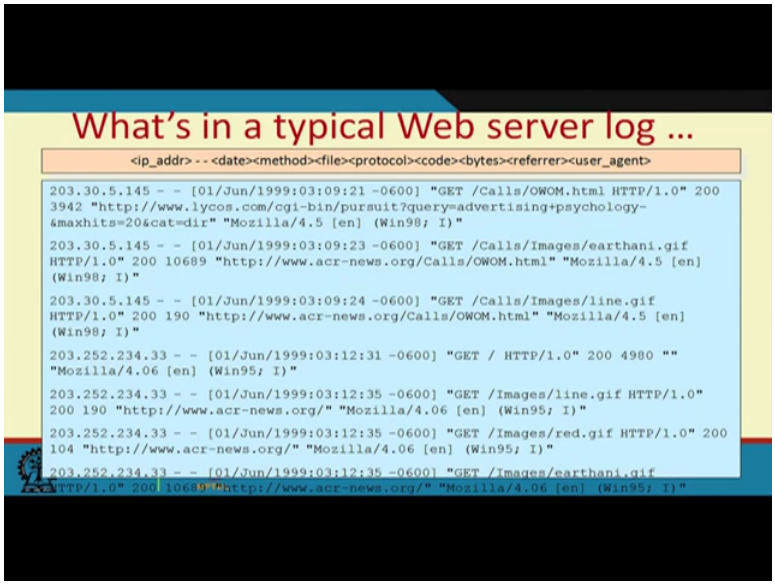
2nd is a website is not always accessed by real human users, there are entities called software agents and crawlers or spiders who may be accessing your website and getting the data back. In fact if you look at search engine like that of Google, it will be actually employing a number of, huge number of web crawlers to download the data, to collect the webpages, prepare the webpages from the web server. So if you are, if your intention is to find out the user activity from the log files which contain these HTTP request and response, instead of human users, this crawler request may surface.

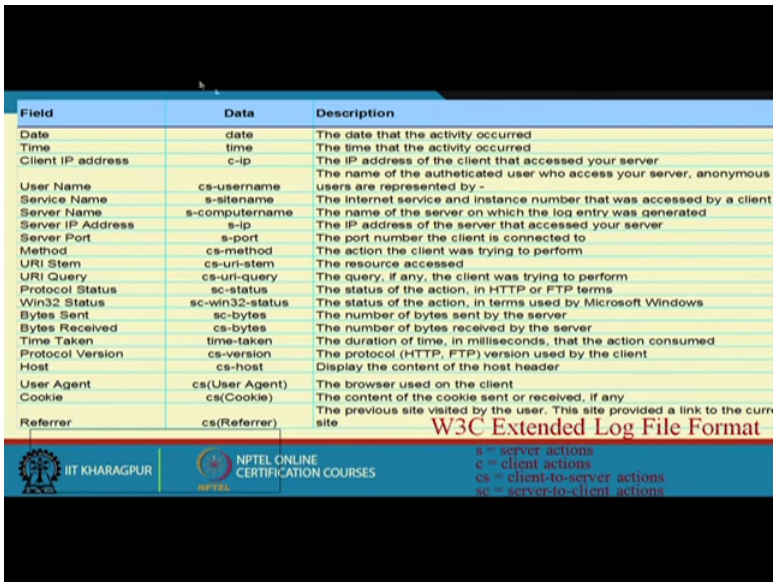
(Refer Slide Time: 14:15)



So if you are really interested to know the activities of your human users, you have to get rid of these crawler activities, this also makes your access log data erroneous. Now this as I told you, this site navigation data of the real users or the scrollers is stored in the server itself in some log files. There are many log files and important among these are your access log and error log. This access log has different formats, different kind of formats which your web server, you have to customise within a web server. But we are going to look at a few common type of log formats.

(Refer Slide Time: 15:05)

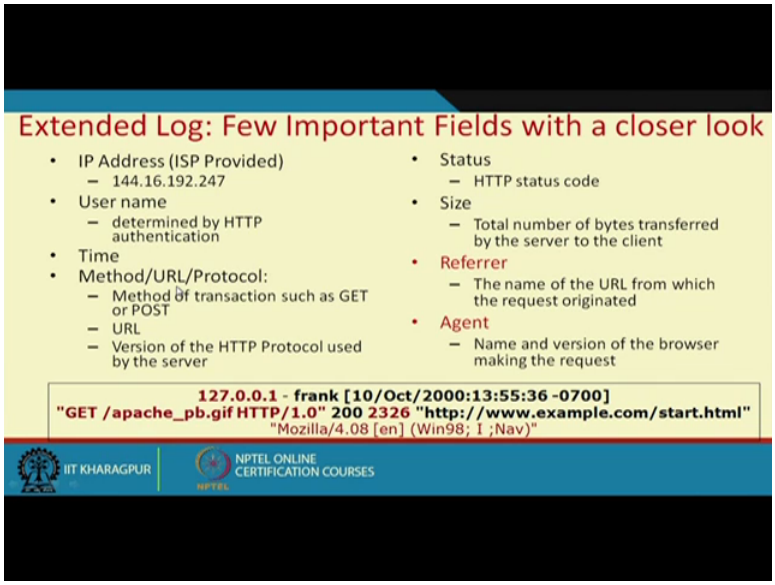




A typically log file will have a look of this type. Where each entry for each of the request, there will be each of the page request that will be more than one entry. By more than entry we mean when a request is sent for a specific page, the request is also sent for its embedded resources. So naturally in response, not only the text file which contains the data pack comes, is downloaded, your other embedded files also get downloaded. So one typical request can result into a number of, large number of responses depending upon the embedded resources.

This is it, this is the typical fields in a web server logs. It shows the date that that activity occurred, the time when the activity occurred, the IP address of the client, which in fact we discussed that this is may not be the true IP. The name of the authenticated user to accessed your server and so on, you can read them out.

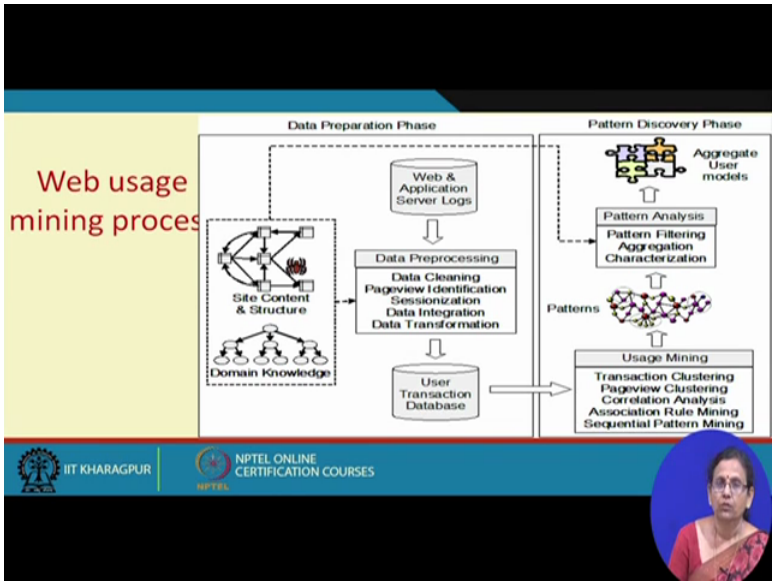
(Refer Slide Time: 16:16)



In fact of this, few important ones we are going to look at. In fact, 1st is your IP address, this IP address is actually provided by the ISP, then you have this username, it is determined by, if there is HTTP authentication by chance, this name is determined, then you have date and time. Then you have method, I was telling you about this get and post business, that the method can be either get or post, then the URL which is getting downloaded, then version of the HTTP protocol used. Then HTTP status code, in fact we saw in one of the examples, earlier examples, status code basically indicates whether it is a success or a failure.

If it is a success, what kind of success, if it is a failure, what kind of failure. In fact, if after the status code, here which is in this case, in this particular example your status code is again 200, it tells us the size, the total number of bytes transferred by the server to the client, the number of bytes transferred. Then it also shows the referrer, by referrer we mean the web page from which the request has come. This webpage can be one of the pages in your website, let us say on page A you have a link to page B, so if you click the link, you will be sent page B but your referrer will be page A.

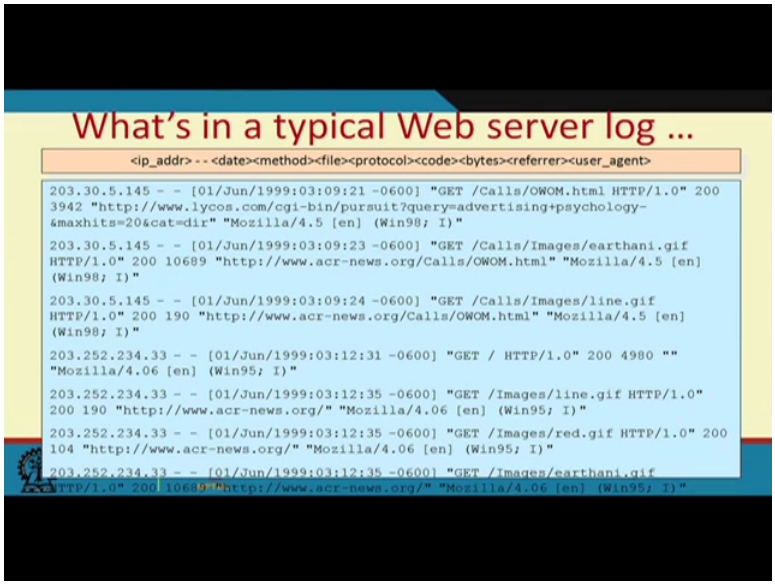
(Refer Slide Time: 18:46)



Then the user agent, this user agent, this user agent is basically the type of browser, type of browsing agent that you are using, for example here it is Mozilla. Now once these access logs are collected, in fact they are automatically collected, see this is we were talking about the big data, right, big data is something where, which is generated very frequently, so if you have a very, if you have a website, your company has a website which is very frequently used by your combine your visitors, the visitors will continuously come to your site and with each request the data is going to get generated.

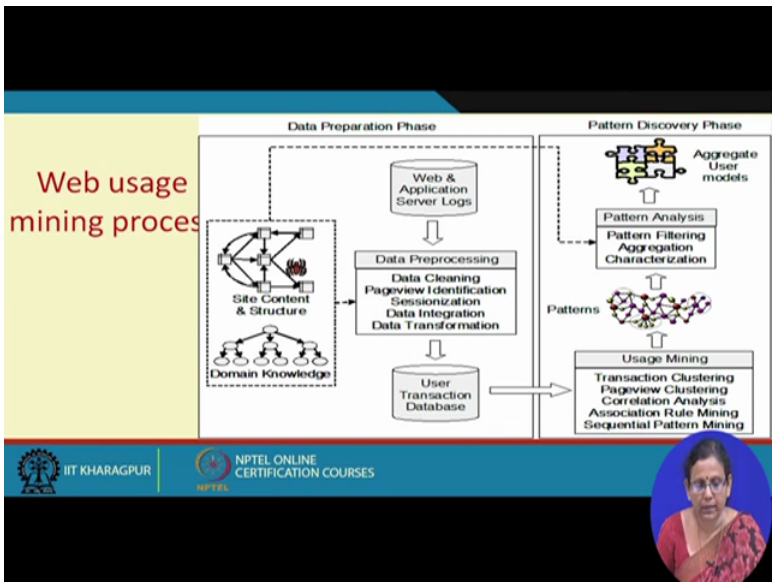
So you have huge amount of data getting generated and stored in these text files which are called access log, error log, etc. So these, once you collect this data, what to do with this data? This data is huge amount of data, data gets generated continuously, they contain all the details that we discussed but what to do with this data, what insights you can get out of this data and how to, what activities are to be performed to make this data usable?

(Refer Slide Time: 20:09)



Because as we saw, this data looks like this, then the text files will look like this, so naturally it is not expected that any human user with millions of such entries will be able to get some meaningful insights out of this. But data is there, so you can, you should try your level best to get some insights. In fact this kind of analysis of, 2 important terms are associated with this, 1st one is your web usage mining, this is, the weblog basically expresses how the website is being used. So mining the data, analysing this data using various statistical and optimisation tools is called Web usage mining.

(Refer Slide Time: 21:37)



And another term is also associated with this, this data gets generated by the browser clicks, if the user clicks on links, it gets generated. So this kind of analysis is also called clickstream

analysis. Now let us just have a look at the Web usage mining process. What is this web usage mining process? You have this log data, maybe coming from multiple sources, by multiple sources we mean your web server, your organisation's web server you may not use, usually companies will not be maintaining a single web server.

Nobody wants a singlepoint failure, if the web server fails, you are disconnected. So naturally people will be having alternative arrangements. And to avoid failure you will be having a number of maybe mirror servers. Okay. So from getting the data application, web and application server data from all these mirrored locations is the 1st task. Okay. Then this data gets preprocessed, so this preprocessed data, preprocessing involves data cleaning, identifying page views, sessionizing the data, we will be knowing about sessionization later on in a little bit more elaborate manner.

But right now that me tell you one thing, 1st in the context of web browsing, sessions can be considered in 2 ways, 1st client sessions, then server sessions, both of course with respect to the client only. By client session we mean, from the beginning of very open your web browser till you close your web browser, whatever activities you do, it may not be restricted to one side, it can be, you can open multiple tabs, then each tab you do various activities, so from the opening instance of the browser till you close the browser, whatever activities you do is termed as a client session.

And out of this client session, if a particular website, that resets a cookie, that cookie can capture some details about the client session. Next when you make multiple requests to a server, you make something called a server session from your 1st request till your last request, that is called your server session. So when we are talking about the sessionization , it is with respect to server session. Sessionization is a difficult task because HTTP is a stateless protocol.

If you connect to HTTP for the 1st time, when you come out , the server actually suggests you. It does not have any data regarding you, when you again go for the next time, you are recognised as a new requests. In fact to overcome this, cookies are a method to overcome this but basically taste, sessionization is about uniquely identifying each session through cookie or through something else, that we will be shortly seeing. There are many, it will be done heuristically only.

Then data integration, as I was telling, from various sources data will come, then is your data transformation, bringing the data, the text data that you saw just now, this text data into a form so that some processing is possible. So to make this processing possible, you have to bring it to some, some form and put it in a possibly put it in a database. And while doing these activities of preprocessing, the knowledge about your site structure and domain knowledge about the area in which you are trying to conduct your experiment with are input to both data preparation phase and pattern discovery phase. Okay.

So after this data preprocessing is over, during pattern discovery, many on this clean data, many Web usage mining techniques are applied like transition clustering, page view clustering, correlation analysis, association rule mining, sequential pattern mining, those tools and techniques are applied to generate the pattern. These patterns are further analysed to build user behaviour models. Okay, so not only these patterns but in general your domain knowledge and the site, your idea and the site structure also helps you in Web usage mining sources. Thank you very much.