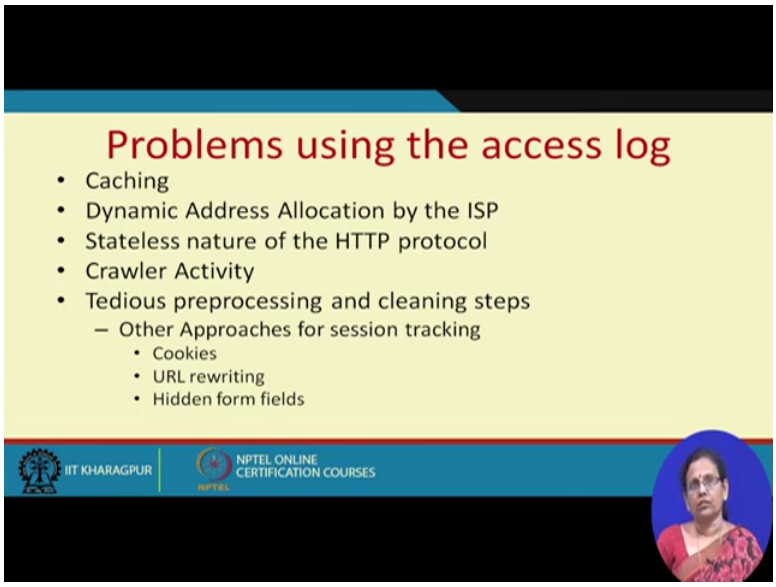
E-Business.
Professor Mamata Jenamani.
Department of Industrial and Systems Engineering.
Indian Institute of Technology, Kharagpur.
Lecture-46.
Understanding The Web Log-II.

(Refer Slide Time: 0:27)



So we will continue with our effort on understanding the weblog, this is a part 2 of that. In fact out of this we already know how weblogs are generated and what is the structure of access log file and the various preprocessing steps we just named. And let us now try to know little bit more about each of the preprocessing steps. Okay, before we know about, discuss about the preprocessing steps, these are some of the issues that we have to resolve during

preprocessing activities. 1st, all the, the everybody who is accessing my website, their entry may not be available in my access log.

Now what is the problem if it is not available? The model that I will be making about the user may be erroneous. Let us try to understand why exactly it is so. 1st the problem of caching, the caching, you know that this, whenever, you must have experience at least this thing, even if your browser does the caching. Suppose you are not connected to the Internet, your Internet connectivity is currently not there and you are requesting for a page, many times the browser will render the page, show the page and will tell you that whether you would like to see it off-line which means it is no more online.

So if it is not online, if it is not connected to the Internet, wherefrom it is getting that web page? That web page was already there in the browser's cache memory. So not only your browser has this memory, browser has this cache memory, this cache will be there in your, from browser where do you go, you go to your proxy, your proxy maybe having the replica of the page. You are connected to, do you think that everytime you, you know the process of finding, resolving this and this domain and IP issues.

Each IP you know that IP is associated with the domain name and once the domain name is given in text form in the browser, this is also in your proxies, in your proxy as well. So let us say you are connecting to Google, do you think that everytime you are, for address resolution you are going to the dot com server or the root server? No. You will be, 1st the page will be, this address resolution, DNS entry in your nearby system will be 1st resolved, if it is found in the nearby cache, it will be rendered to you. Unless otherwise it is specifically mentioned in the web page, in fact you need to know more about the this HTTP header, HTML header tag which talks about how to control this cache.

So unless otherwise nothing is mentioned, this web page which is in the nearby cache will be delivered to the user. So which means the request to a web server may not ultimately go to the web server, it will be rendered from the nearby cache, even from the your browser's cache, if it is available. So therefore, in the access log is erroneous because it is not containing all the user request entries, that is fine, you cannot do anything about it. If you have other mechanisms available and a few are connecting to, know how to collect this this thing from the cache of the proxies which is quite fairly not possible, but still that is, can exploit that.

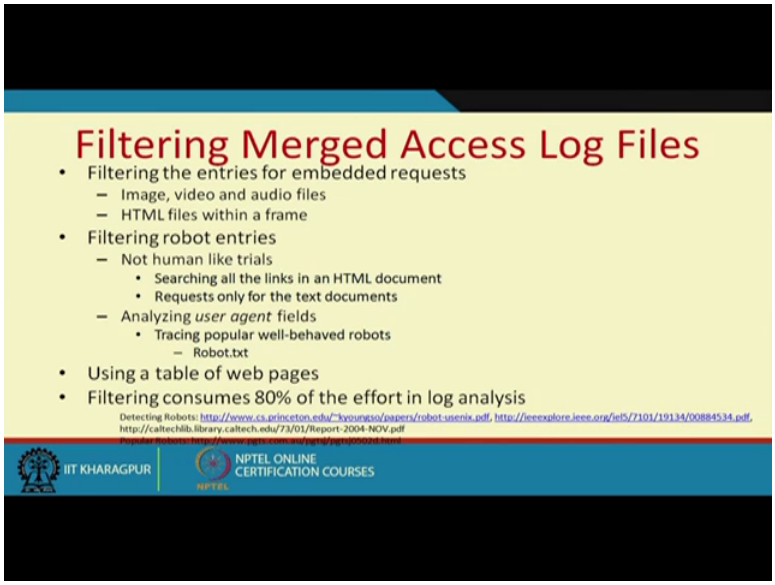
Then, dynamic address allocation by ISP, by this we mean we were just looking at when the webpage, in fact we can, in fact in the earlier lecture what we did, we saw is that, whenever request is going, it is going through a number of ISPs in between. So everytime the ISP, dynamically ISPs in the root, dynamically allocate one IP address. And while the response from that, it gets resolved and finally our response comes to your own browser, that is fine. But the point that I am going to make here is the address that appears in the access log in the place of the IP address of a specific request, it is that of the last ISP through which the request has come, so it is erroneous.

Then sessionization is not possible, we were talking about 2 types of sessions client session and server session, so seran ver sessionization is not possible because of HTTP protocol, stateless nature of HTTP protocol. Then resolve this, then you have crawler activities, in fact last class we saw that crawlers are all these software agents or soft bots who will be sent from the search engines and other such, let us say advertising agencies etc. to your phone company's server.

So their interest will also be, because they also make request, so their interest will also be also appear in the access log, so that makes the access log erroneous as well. So therefore it requires tedious preprocessing steps but somewhat because of this use of cookies, URL rewriting etc., Stateless nature of HTTP protocol to some extent can be resolved. So therefore the 1st task in preprocessing is filtering the entries for embedded requests.

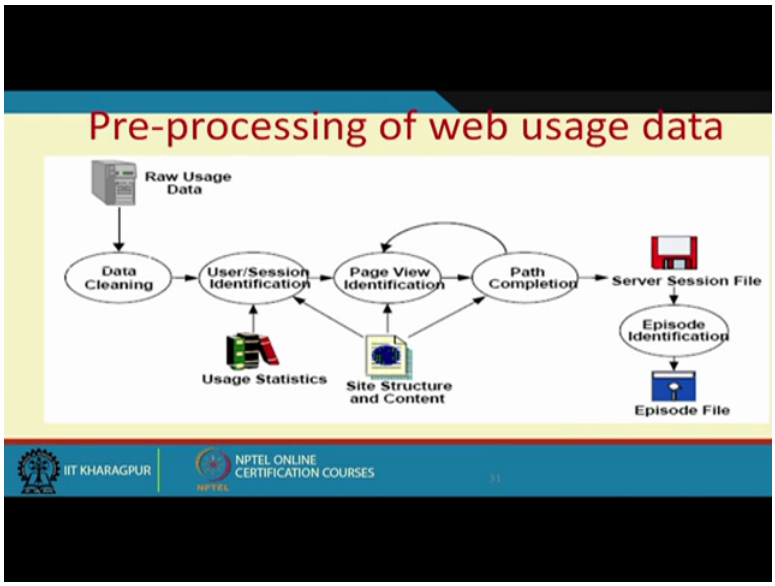
By embedded requests we mean the HTML file to which to which you place the request along with the HTML file, many embedded resources like images, audio, videos etc. also get downloaded. So removing those embedded requests is the 1st task, you keep only the HTML pages. Again there is a problem, there are certain HTML pages which are like frames, which then once HTML page, another HTML page will be embedded, so that again creates little problem and if you have the knowledge of, domain knowledge of your site, you can even straighten this up.

(Refer Slide Time: 8:15)



2nd is removing the robot entries. You have to have appropriate algorithms for removing these not humanlike trials and analysing the user agent field, you can identify these robots, algorithms do exist for this and you remove. If a site wish to control the robot activities, they publish a text called robot.txt and the requests which come from robots like spiders, crawlers, etc., they will be going through these entries, owing through this robot.txt and if it is a well-behaved crawler it will learn the instructions and work accordingly. Again if you maintain a table of true pages in your site, then keeping those aside remaining things, you can simply delete to get a reprocessed access log.

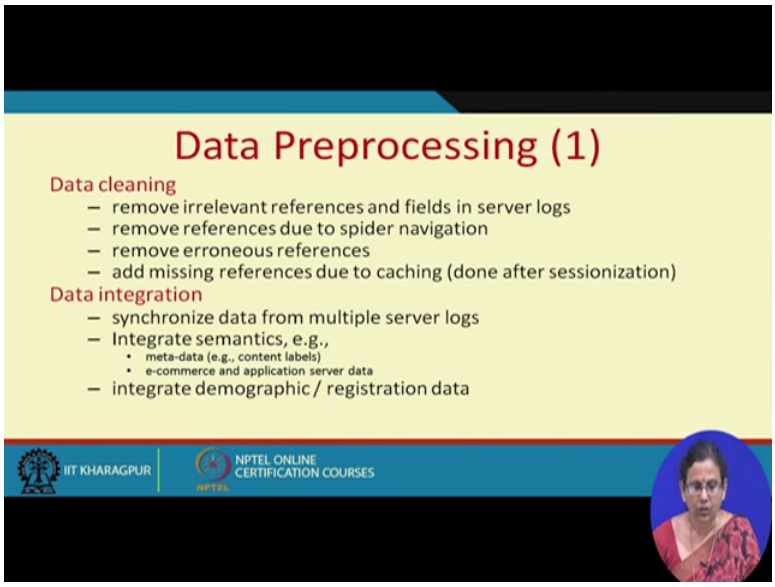
(Refer Slide Time: 10:03)



In fact these filtering activities consume 80 percent of the log analysis. Now this is the part of data preparation about the steps we learned, earlier we just discussed, let us see each of this space, each of this step little bit more elaborate way. So you have, get the raw usage data, you clean the data, you have to identify the session and get the usage statistics, identify the page views, then you complete the path, my path completion we mean that you complete, that we are going to discuss little later because it will be difficult for me to explain it right now.

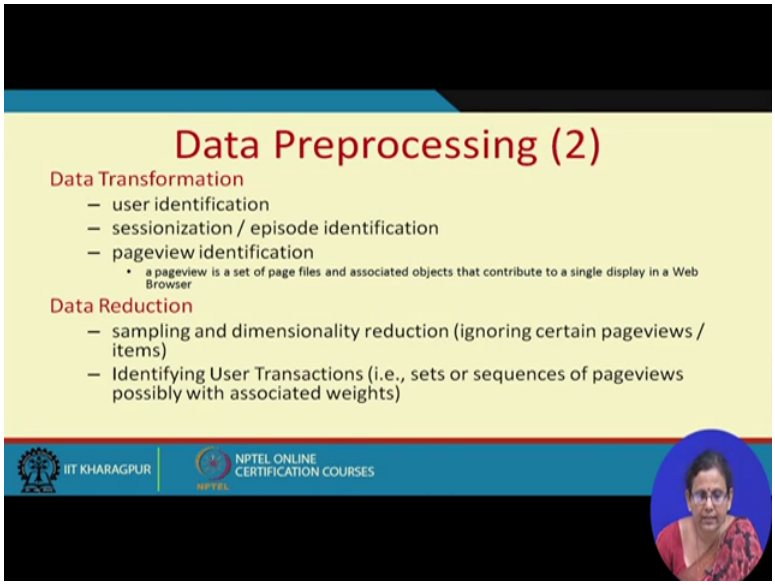
And input to this is your site structure and content. Then finally you make something called a session file. From a session you identify various names for the session which are meant for specific purposes. Some people come for product search, some people came to know about your company, details of your company, so various types of sessions can be identified to make something called an episode file. Okay.

(Refer Slide Time: 11:31)



Now, data cleaning in the reprocessing is about removing irrelevant references and fields in the server logs, removing references due to spider navigation, remove erroneous references, add missing references due to caching because if you try to sessionize, it is which is about finding a sequence of pages, related pages, you may find out some of the things are missing. So you may try to reconstruct the sequence. Then it is about data, next thing is about data integration which is about getting the data synchronised from multiple server logs and integrate the semantic level detail by which gives more meaning to the data that you collected.

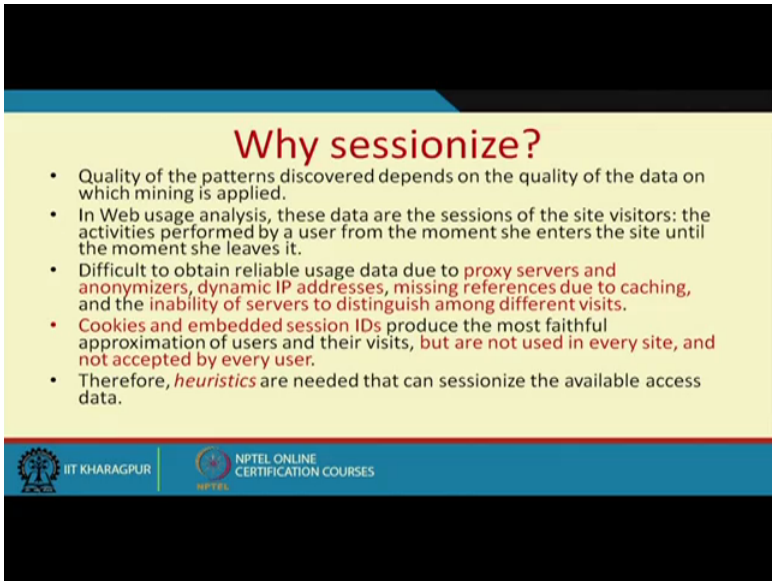
(Refer Slide Time: 12:41)



Then integrate demographic and registration data if available for customer behaviour modelling. Then next preprocessing task is about user identification which I told you is not possible unless otherwise user specifically logs into the site. Then next is sessionization and episode identification. And finally page view identification where a page view is a set of pages and associated objects that contribute to a single display in the web browser. So deleting those those embedded requests, request to the embedded resources is a part of this.

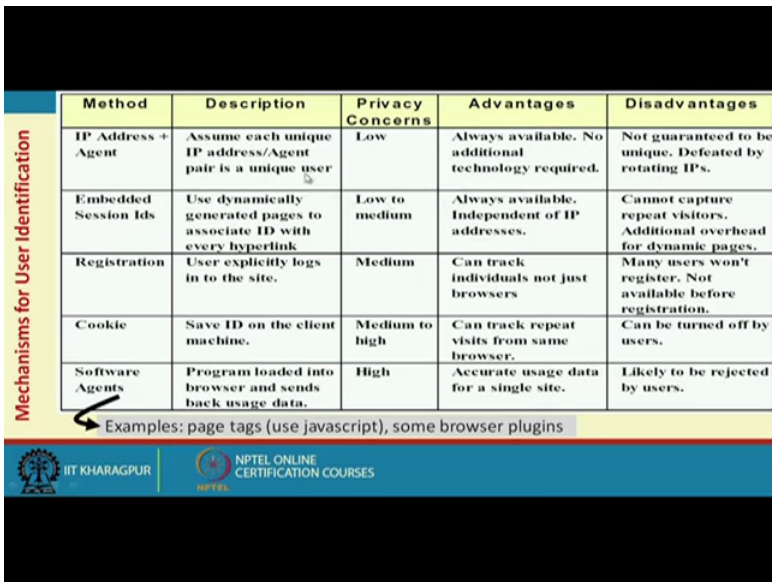
Then data reduction, now here, because these data files will be generally very huge, maybe you will be sampling a part of the data, you may be trying to reduce some of the page views and so on. And some of the you know because after all you are using some heuristics to do all this, so sometimes you may encounter with some sequence of page views which probably are not related, so you may have to drop them using your domain knowledge.

(Refer Slide Time: 14:10)



So most important thing here is to sessionization are bringing all, because HTTP is a stateless protocol, bringing all the HTTP requests together coming from one user makes a session. Now why this identifying the sessions are reported because quality of the pattern discovered depends on the quality of the data on which the mining is applied. Now in Web usage analysis, these data are sessions or site sessions of the site visitors that which indicates the activities that are performed in the site and we have already discussed how exactly, why exactly it is difficult.

(Refer Slide Time: 15:49)



Then using these cookies etc, it can be resolved up to some extent but these cookies etc are not foolproof and their own accepted by every site, it is under the control of the user. If the

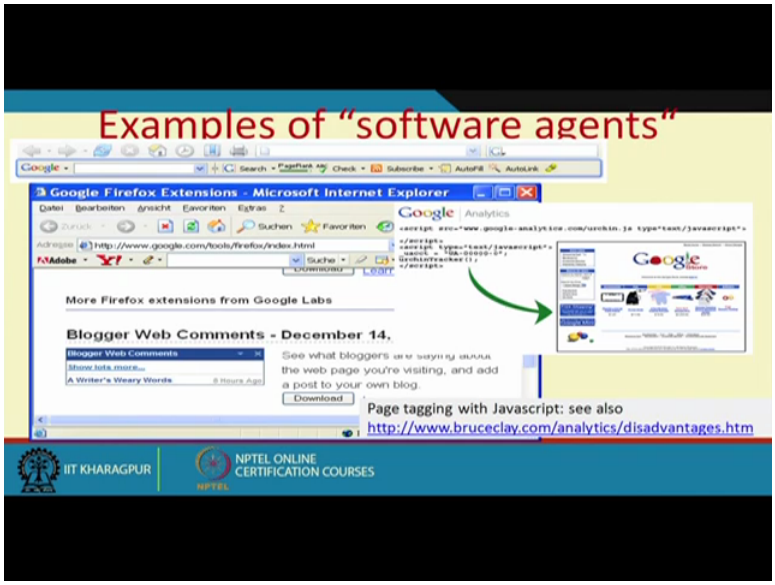
user wants, the user will disable the cookies. So therefore there are many heuristics are used for sessionization. For user identification, again many methods are used, use of IP addresses + the agent, the agent is the referrer agent, the house from which that request has come, that browser agent, that pair may be unique for each user.

It is always available, so there is no additional technology required and what are the disadvantages, it may not guarantee to be unique. Then unique in the sense, next time when the same request comes from the user through that proxy or different IP will be combined with the specific browser type. But it is one of the options. Then embedded session ID, the use of dynamically generated pages are to associate an ID with for every hyperlink. So by this embedded session ID we mean this ID will be embedded along with the page request that you send and next time when the 2nd request comes from the user, this is again embedded and comes back. Okay.

So now they are independent of IP of course, because they are going, whatever server they are going, from that, whatever client they are going, from that client they are coming back. However, they cannot capture revisits, within a specific client-browser instance, if the browser, if the client access at the same website from another browser instance, they are not going to work. Then registration, it is quite straightforward, the user is asked to register but most of the time if just to browse the web page if you are asked to register, do you, will you be registering, no, you will not.

So in fact it is again depended on the user, if the user wants user can make the registration and can get tracked but otherwise it is not possible. Then cookies, it is also cannot be guaranteed because the user can stop using cookies but if the user uses, this is one of the best way to not only to track, track the client within a session but in multiple sessions when the repeat visit happens, then also the user is tracked.

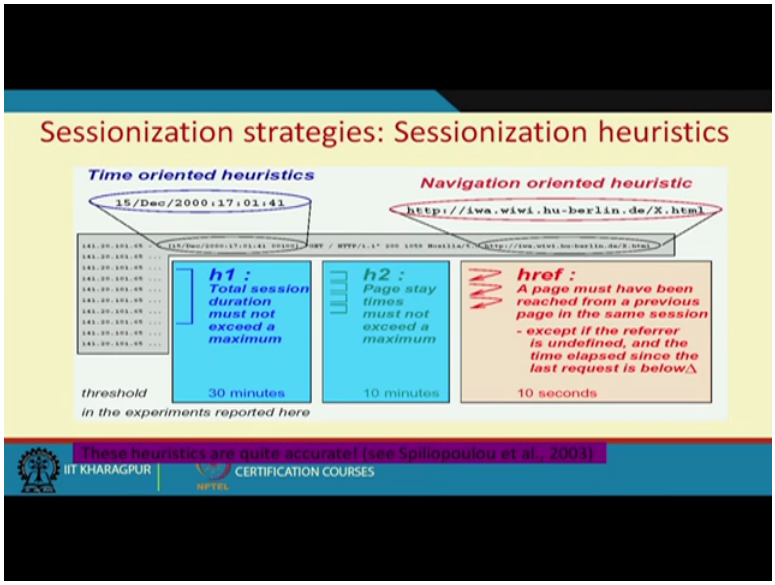
(Refer Slide Time: 19:20)



Next is software agents, these software agents can be loaded into the browser and send back the usage data. The example of these software agents are some of the plug-ins. So here we show one example of a browser login in terms of where some you know some JavaScript from the, what is JavaScript, JavaScript we have already discussed while talking about the dynamic webpage generation. JavaScripts are the specially designed functions written on in a scripting language called JavaScript and they are part of your HTML file and gives dynamic client side dynamic effects.

So some JavaScript files provided by, about Google analytics we are going to talk after sometime, they can be embedded within the webpages and they collect and send back the usage data.

(Refer Slide Time: 20:26)



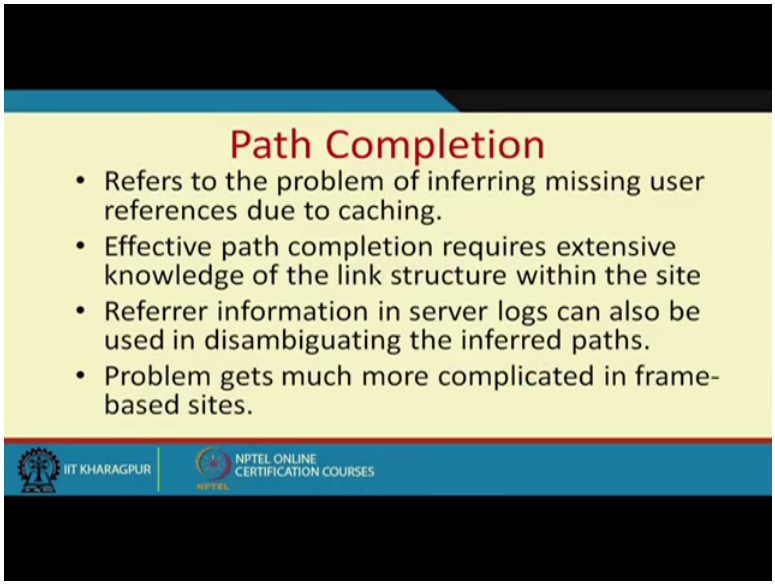
Then these sessionization heuristics can be broadly classified into time oriented heuristics which count upon the time when the last page was sent and as you know that is a rule, industry rule called 30 seconds rule. So it says that if the 1st request comes now, then after 30 seconds if the request, within the 30 seconds the next request does not come, then probably it is the end of recession, but it may not be so. But anyway it depends on, what this is not the golden rule or anything, these 30 seconds can be adjusted but using, looking at this time, the your, what are we talking about, we are talking about how to clean and preprocess the access log.

And from the structure of the access log, we know that it contains a field which gives the date and time. So looking at this field, the requests which are close enough in time can be put together, this is one time oriented heuristics. The 2nd group of heuristics for sessionization are Navigation oriented heuristics. They just look at how the webpages are actually connected so that if somebody is browsing lets a page A and there is a link to another page B within that, then if there are 2 nearby requests in the access log to page A and B, then probably they are from the same user, that is how they are put together.

In fact both these time oriented heuristics and navigation oriented heuristics are combined many times to generate the sessions. Path completion, see where trying to sessionize, by sessionize we mean collecting, why are we doing all this, where trying to, we are doing all this because we will be cleaning the access log and the access log after cleaning will be using to understand user behaviour, user's navigation behaviour, user's activities, user's search behaviour within the website that is our intention.

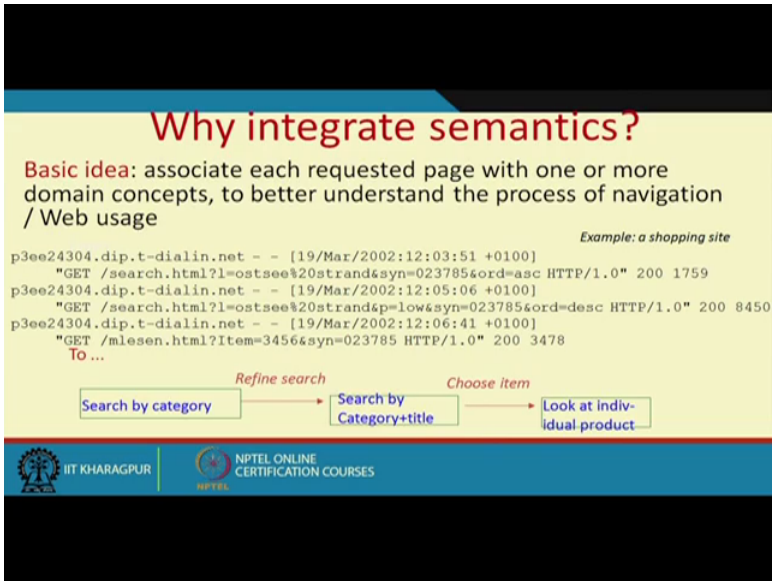
So for this purpose, we have understood the structure of the access log and what problems can be encountered while using it and we are right now they are talking about how to clean the access log. In the cleaning process, one of the activities is to identify the sessions. If we group all the user activities together, then only we will be able to give some generic idea about the behaviour, right, otherwise it is not possible.

(Refer Slide Time: 23:45)



So therefore next task after sessionization is Path completion. It refers to the problem of inferring the missing user references due to caching. And then following again certain heuristics to combine them to create a complete path of activities. Now effective Path completion requires excessive knowledge of Link search within the site, referrer information in the server logs can also be used to decide be great the inferred path. Then in case the site contains frames, frames are basically embedding HTML pages within another HTML pages, Path completion events becomes difficult.

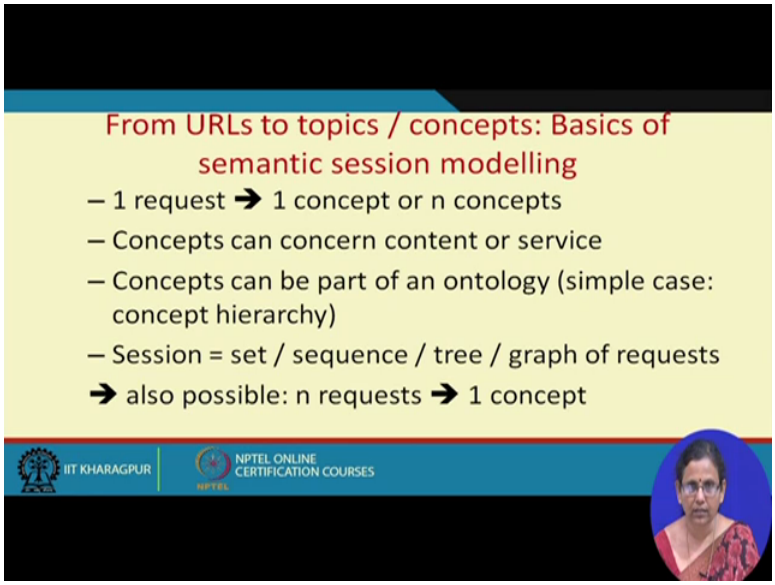
(Refer Slide Time: 24:52)



Because always the entries in the access log will come in the name of the parent frame. Now in order to, for this Path completion, how the webpages are semantically connected through the meaning of the, the meaning of the through the what you call, the content of, content of the web page, if we can associate it with the pages with respect to that, then even sessionization and Path completion becomes and develop heuristics based on that, even sessionization Path completion becomes possible, at least can be done in a better manner.

So here the basic idea is to associate each request page with one or more domain concept to better understand the process of navigation or web usage. For example, you can see 3 consecutive 3 requests which may, which you identify the session and you say that 1st the person searched the items through category, then he searched all the items through category and title and finally he has chosen the items looking at the individual items. So such kind of activities about the user can be traced back, if you semantics.

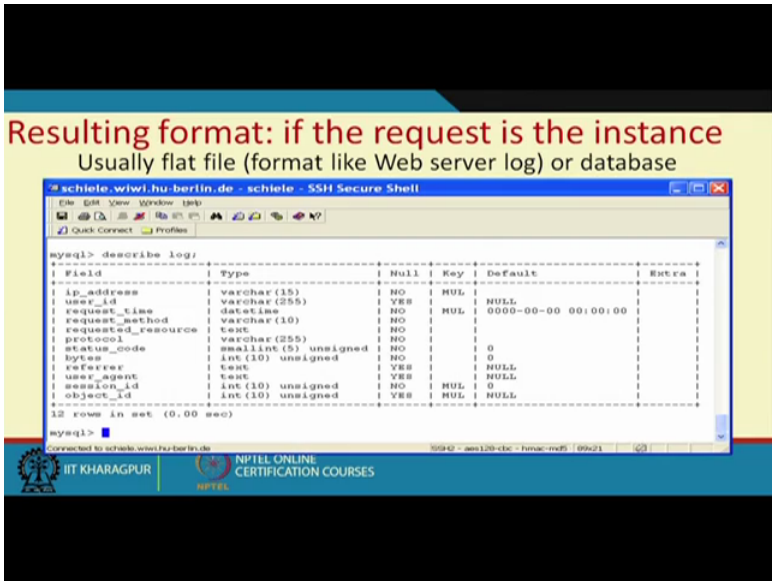
(Refer Slide Time: 26:32)

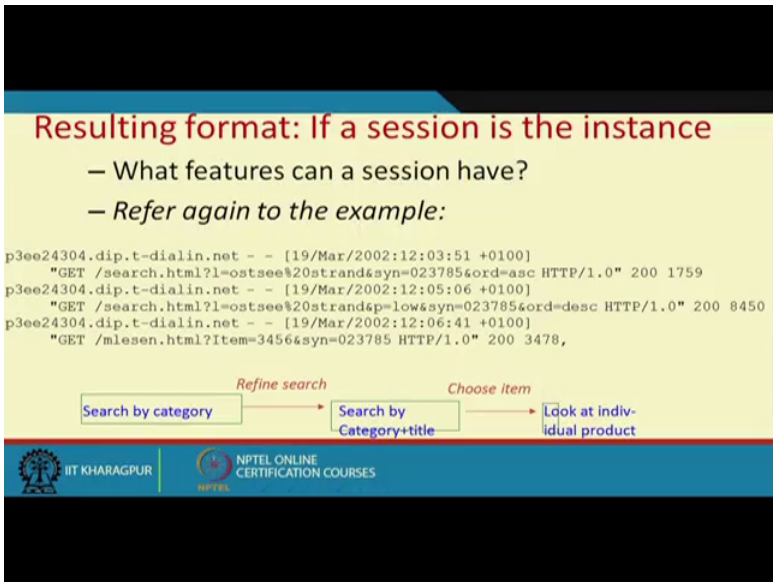


Now for this, associating the semantic or the meaning, you need to connect the topics from the URLs that is in your website. So each request can have, can be connected to one concept or more than one concept. Now this concept, this is some fundamental idea about the basics of semantic session modelling which is a part of preprocessing task only in which using the semantic, using the, using the knowledge of the semantic more better sessions can be formed, better Path completion algorithms which can be designed and so on.

So these concepts can concern the content, specific content or service, these concepts can be part of a bigger ontology, ontology is basically a concept hierarchy, how the concepts are linked with each other. And then a session can be viewed as a set or a sequence or a tree or a graph of request of possible M requests which ultimately leads to one concept. So using all this to try to identify the sessions but after you identify the sessions etc, and filter all the erroneous entries, all the undesirable entry from the website, finally you bring the data into a flat file form.

(Refer Slide Time: 28:32)





In this flat file, it can be now stored into a database, for example here there is one example from some source which shows the fields like IP address, user ID, request time, request method, request resource, protocol, status code, bike, referrer, user agents, session ID which you created following all these , one creates following all these heuristics that we just love discussed. Then object ID, the one, the item that is get downloaded. Now the resulting format, if the sessions cannot be resolved, if the session is the instance, then the features, you have to now identify what kind of session it is because you are going to identify the episodes, right.

So what kind of session it is, I was, in fact we are coming back to the same example again. If one group of users are there who search by category, then search category + title, then search

by items, they are, let us say half specific features of browsing. But there will be another category, who probably maybe your old user who will be directly looking at individual products because they are already acquainted with how you have organised this you are main vendor. So if you, this is for instance I am telling, if you can actually distinguish these 2 groups, maybe 2 different behave as you can observe for the repeat visitors and for new visitors. So with this we end this particular and from next lecture again we will continue with this topic, thank you very much.