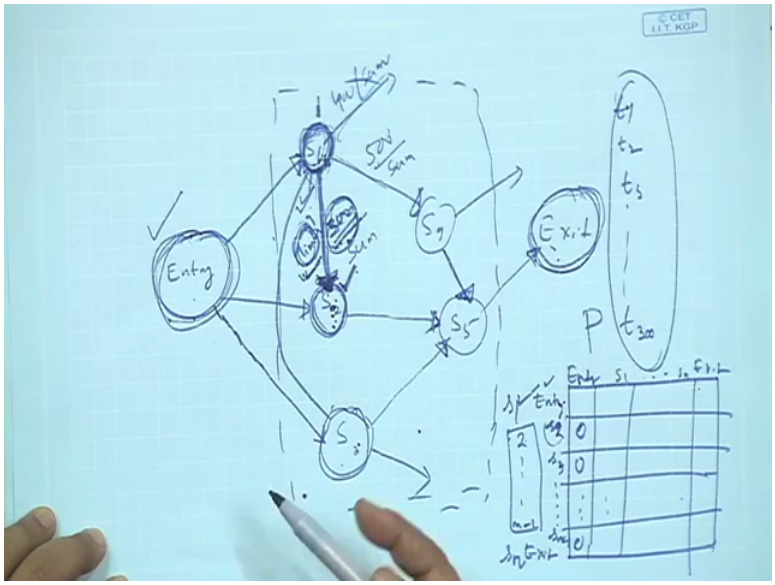
E-Business.
Professor Mamata Jenamani.
Department of Industrial and Systems Engineering.
Indian Institute of Technology, Kharagpur.
Lecture-49.
User Behaviour Modelling From Web Log (Continued).

(Refer Slide Time: 0:50)



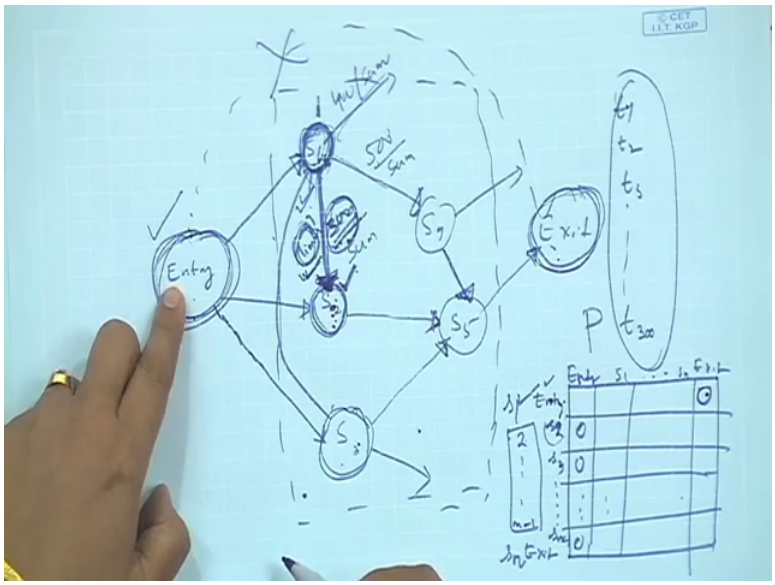
Welcome back, we continue our discussion on building a customer behaviour model, customer's navigational behaviour model and we thought it as a Markov process and we are proceeding. So now let us have a look at what other properties of this transition probability matrix. After you make these transition probabilities, you calculate these transition probabilities, you can now represent it in the form of a matrix. Then of course you will have an twisted, you have exit state and you have other states, okay. Similarly here you have entry, exit and other states.

And the transition probability values that you calculated, so this is my transition probability matrix, this is my transition probability matrix. So in this transition probability matrix, my 1st state is the entry state and my Nth state, lasted, let us call it from S2 onwards, so that it becomes consistent, this is the S1 and this is my SN, okay. Now the properties if you look at the properties, the properties, they say is, from Ith state, I is any other state, you have 2 to total N -1 number of other states, S2 to N -1 number of other states.

So this one, 1st property tells that from state I, the probability of going from state I, that is any other state to the 1st state, that is the entry state is 0. Of course this is true because this is a

dummy state and you have assumed that all the visitors entering into the site, before entering into your site stay in this particular state. So therefore, therefore from any other state you cannot go to entry state, so from any other state, any other state, the this is your matrix P, P I1 going to entry state from any other state is 0, going from S2 to entry is 0, this is 0 and so on.

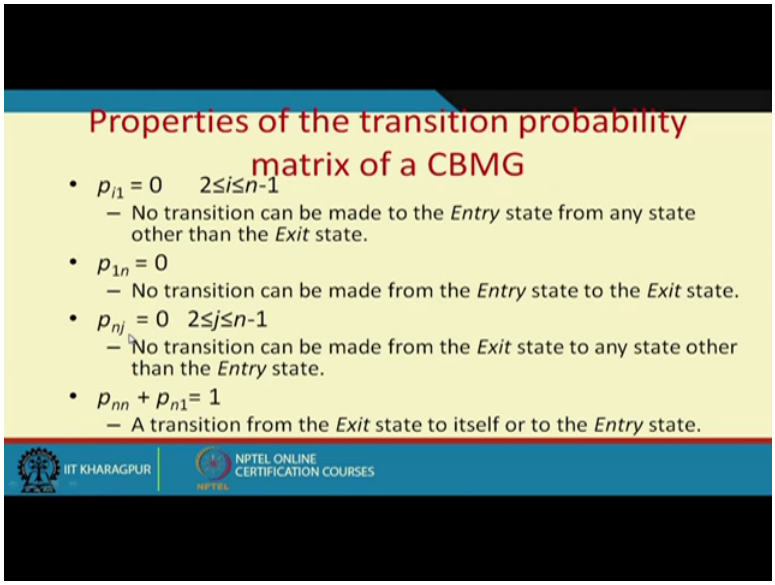
(Refer Slide Time: 4:17)

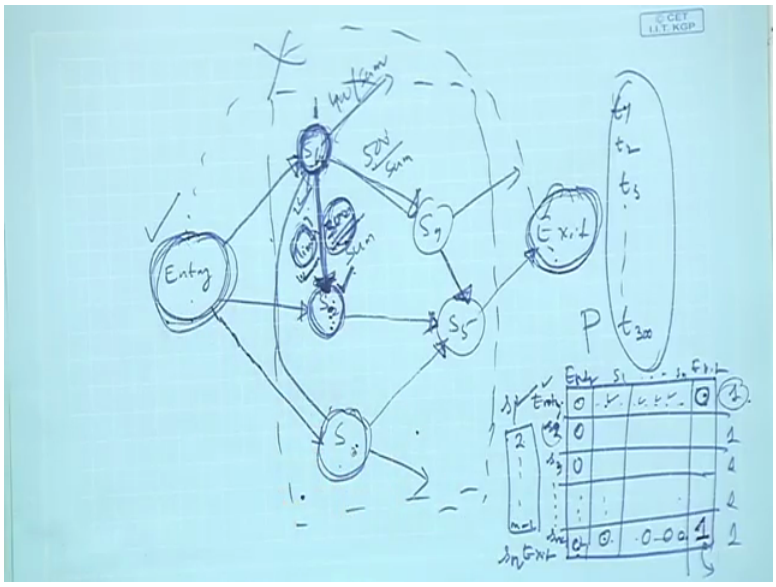


So now going from 1st state to the last state, from entry to exit, nobody, because, because in the model people will be visiting some of the, at least one state before entering into the exit state, otherwise he is not your customer. So therefore going from the 1st state, that is entry state to the Nth state, that is exit state is 0, no transition can be made from here to here, this is not possible, no transition can be made. If you assume there is a transition from here, then he is not your customer, he is not entering into your site, so therefore this is 0. Then coming from, we are considering a model in which return of the visitors is not possible.

We are not considering that, we are assuming that we are not considering returning visitors. Look, we are doing a model and we are defining our model assumptions, you can even extend this model to consider re-entry of the visitors if we have the additional data source from a cookie, where the past visit details might also be written, might also be captured. But anyways we are limiting ourselves to access log only, so therefore we do not have any knowledge that somebody is revisiting a specific page.

(Refer Slide Time: 5:32)



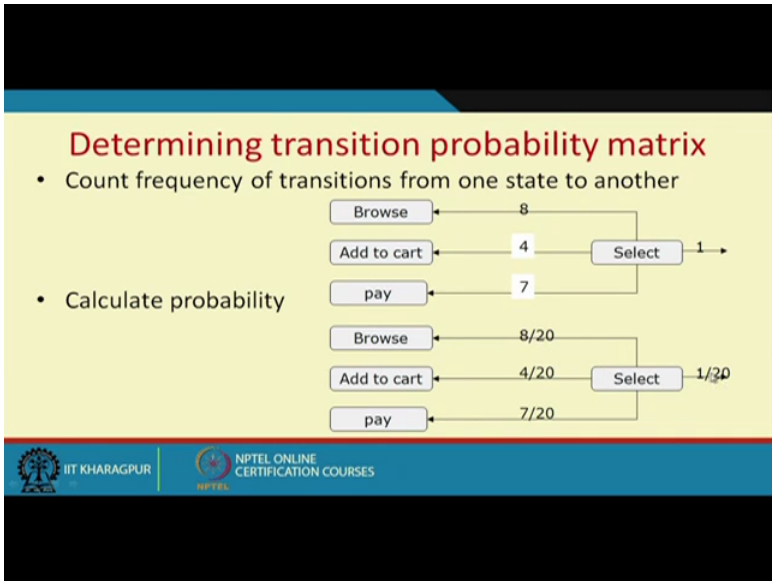


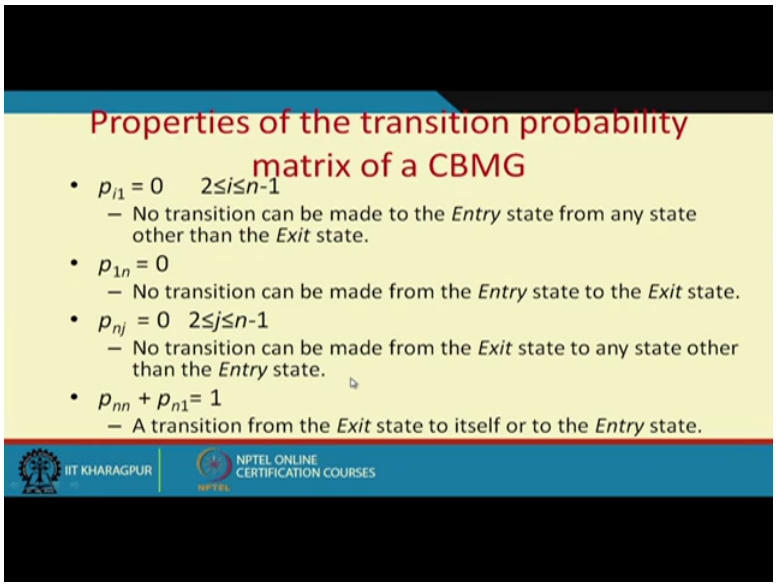
So therefore, therefore going from page N, that is exit state, from exit state to J, that is any other state is not possible. So going from exit state to any other state is not possible, so all these values are also 0. Now this is after all a transition probability matrix of a Markov basis, so sum of all these need to be one. Okay. Now from entry state there will be some values everywhere but from exit state, all these values are 0, it from exit said you are going nowhere.

So we can assume that from exit state, somebody can come back to entry state, though we are not capturing this detail, we are simply assuming that customers always returns to the site, okay. And here from whatever, because to make this sum 1 possible, whatever values you have here, you add it up and make entry to entry as 0. And that all of them will add up to 0

because after all we are taking all these values then dividing by these total number of visits, total number of visits from all the States.

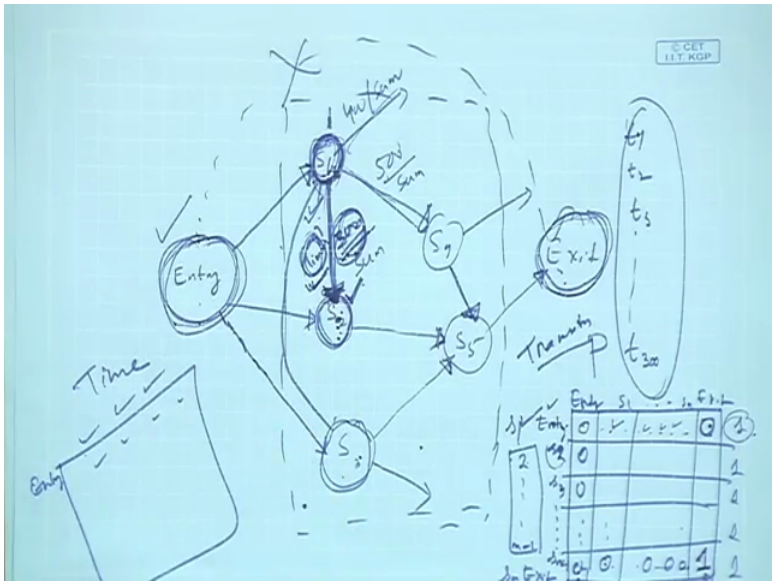
(Refer Slide Time: 7:02)





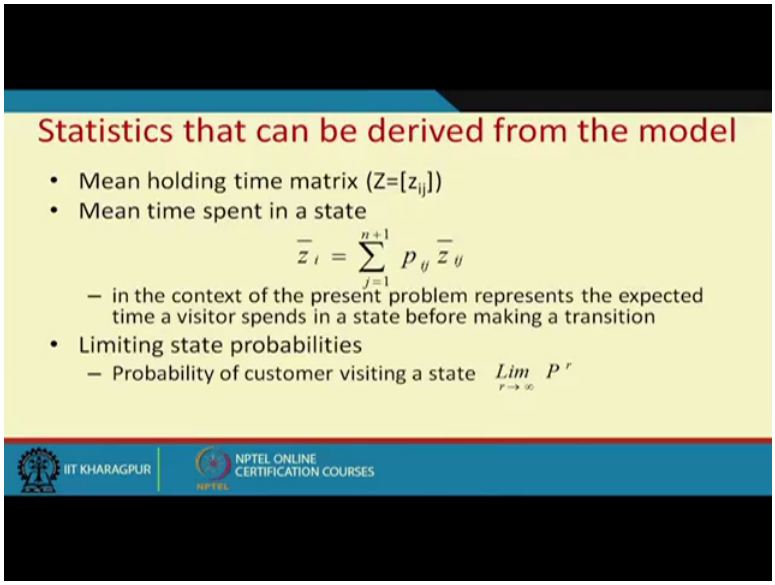
Let us go back, here it is not clear. Here we are dividing it by this total number of visits out of this particular state. So it is always going to be 1, so you do not have to worry about it. Okay. So, now this is a property of a customer behaviour model graph when we, and this customer behaviour model graph is nothing but a state transition details of the Markov, corresponding Markov chain. Now, I have assumed so far that you do not have any background here and I have tried explaining in a very lucid manner what exactly we mean by state transition, how to capture it, how to capture think time and so on.

(Refer Slide Time: 8:29)



We made 2 matrices so far, one that transition probability matrix, another is the time matrix by that total time in the, here in the grass, this total time T1 to T300 for this particular page, for this particular transition we took the sum of all these and divided by this 300, so that is how we got the time. And we have, we can have a time matrix, this is a transition matrix, similarly we can have a time matrix. In the time matrix again you will have this entry, exit and all other states and every where you can have the time average time spent. Okay.

(Refer Slide Time: 8:48)



Assuming that these matrices are ready and we are using this Markov process theory which we have not discussed here but at this theory at the end suggests few statistics that can be derived from the model. So the formulas of that we will be now directly using and will be

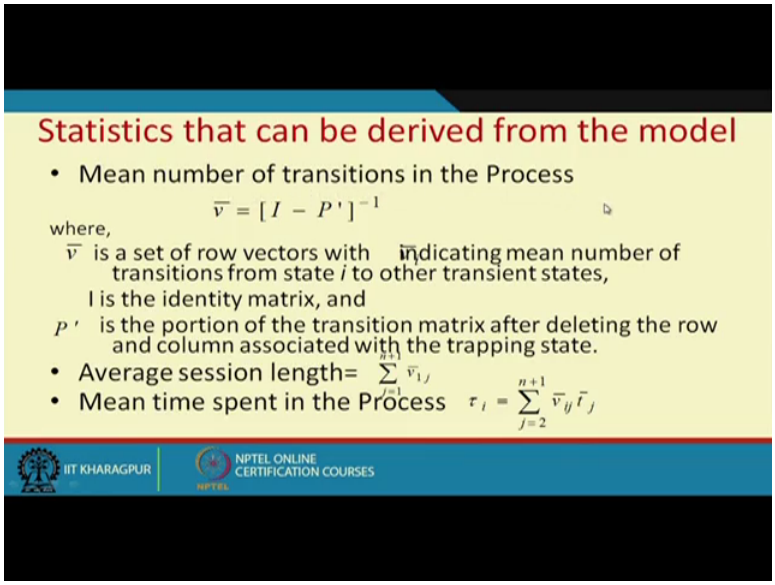
trying to interpret it. So we have the main holding time matrix, the time matrix we are calling as the holding time matrix where the mean values of this time is represented, mean value of this time is represented. So here you have this Z, Z 11, Z 12, Z 13 and so on. So these are the small z values and this is your matrix capital Z.

So your main holding time matrix is this one, mean time, the average time, the mean time spent in a state, how much time is spent on a specific state, on a specific page now can be computed from this model. So this is basically your Z I bar, the average time, the mean time spent in a state Y is equal to sum of J equal to 1 to N +1, first state to, up to the exit state, in fact the 1st, the exit state so we have last one, we have considered exit state to be the Nth state probably, yes we have considered exit state to be Nth state.

So therefore please consider this as N, instead of N +1 in the slide I will be making it N, N modification, right now I am not able to make it because equation editor is not installed here. So so this is the expected time, please consider this is from your starting state till the exit state. Now, this is how much mean time spent , typical visitors spend on a particular state I. Now we have another statistics called limiting state probability, which is the probability of weather, what is the probability that a customer will visit a specific state. So this will be calculated by raising this transition probability matrix to some very large value.

So if you raise the power of this matrix to a very high value, here of course we write R tends to infinity but actually depending on the matrix with, not infinity number of transitions will be actually required. Theoretically this is something described this way but if you keep on increasing R value, at some point of time you will see that this matrix, this matrix P to the power R is becoming stable, there is no changes happening to this particular matrix. So that will give you the probability of customer, individual entries will give you…

(Refer Slide Time: 13:20)



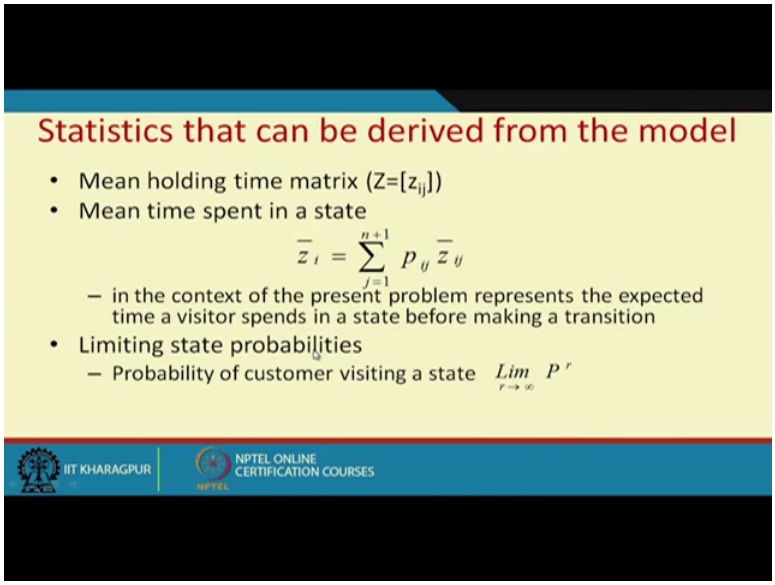
Now see what will happen if you make P to the power R, after some point of time, all the roles will be same. So any entry will give you, so whatever will be this entry, this entry will be same, this entry will be same and so on. So at this point, this one will give you the probability of visiting, let us say this is state I, this is from your visiting state I, so this is what is P to the power R. Then may number of transitions in the process. This can be calculate it using this formula where this nu bar is a set of row vector is with nu I indicates the mean number of transitions from State I to other transient States, transient states are the one from which if some visitor enters the state he is definitely go out of the state.

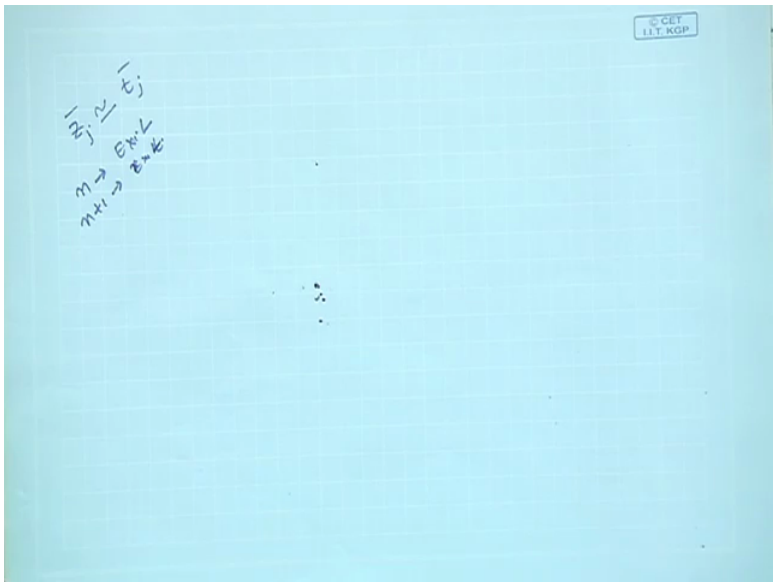
This capital I is the identity matrix and P dash is the portion of the transition probability matrix after deleting the row and column associate it with the trapping state. The trapping state is the one from which once the user enters can never come out. So in this case the trapping state is actually your exit state. So once this nu bar vector is known, the mean number of transitions in the process, these row vectors are known, then every session length will be J equal to 1 to N +1 where N +1 is the last state.

Please do not get confused because in the 1st slide I have made the Nth state to be the last state but here in the slides which I will be giving to you, it will be this modification will be made, so N +1th state which is the last state, if you add up all these nu bar values study from your entry state to all other states, that will give you the average session length. So once you know the average session length, you can find out the meantime spent in the process, that is meantime spent in the website.

So that this meantime spent in the website in a particular session, that we can say as the session length is Tao I, which is sum of the time, this venue we have already found out. This is the time spent on each of the pages, the average time spent on each of the pages and if you multiply it with this nu J I, the number of transitions, number of transitions from State I, previous state to the State J, then this is the time spent. So there is again one more inconsistency in the symbol, so here I consider this meantime as Z I and here in this slide I will be making the corrections, this is T J bar.

(Refer Slide Time: 16:52)





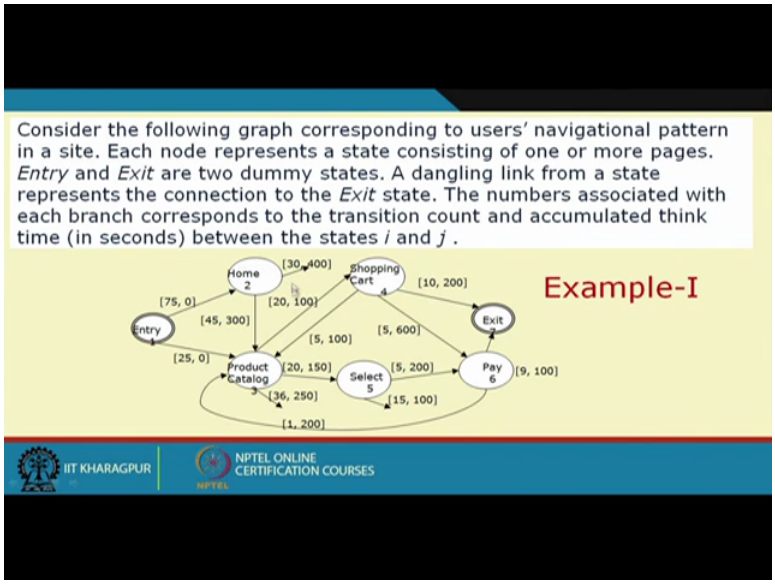
So the mean holding time, this Z I is considered synonymous as ZJ as TJ here and also another thing, the transition, the last transition was, I called as N as the exit state but this formula N +1, I call as the exit state, please note these 2 changes here. Okay. So now let us

see one application, the 1st example, we assume that in this particular website, there are some 5 states home, product catalogue search, shopping cart, select and pay. These are the, these can be individual webpages are these can be group of webpages under this common theme.

Now entry and exit state have their usual meaning, this is your 1st and this is your 7th, so there are total 7 states in this particular example. So and what will be the size of your transition probability matrix, it will be 7 cross 7 matrix and what will be the size of your average Time matrix, that is again 7 crore 7 matrix. So now consider the following graph corresponding to uses navigational pattern in a site, each node represents a state consisting of one or more pages, entry and exit are 2 dummy states.

These dangling links, these are I call as dangling links, I did not connect it is this exit state but actually it would have made this graph look little more not readable, so just to increase the readability, these are just left free. But please assume that these links which are not connected to any state are actually connected to exit state, which means the user can go out of this state. Okay, the number associated with each branch corresponds to the transition count, so transition towns are, here from state 2 to state 3, transition count is 20, from State 3 to state 4, transition count is again 20 and so on.

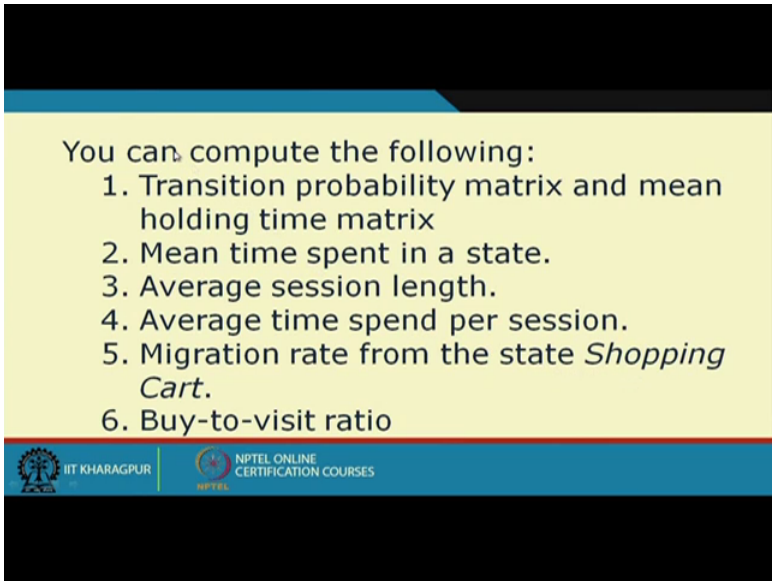
(Refer Slide Time: 18:05)



And from entry, let us say total 75 entries have been made to the home state and 25 links are made to the product catalogue state. This, the numbers associated with each of the branch corresponds of the transition count, this is the transition count and this is the accumulated

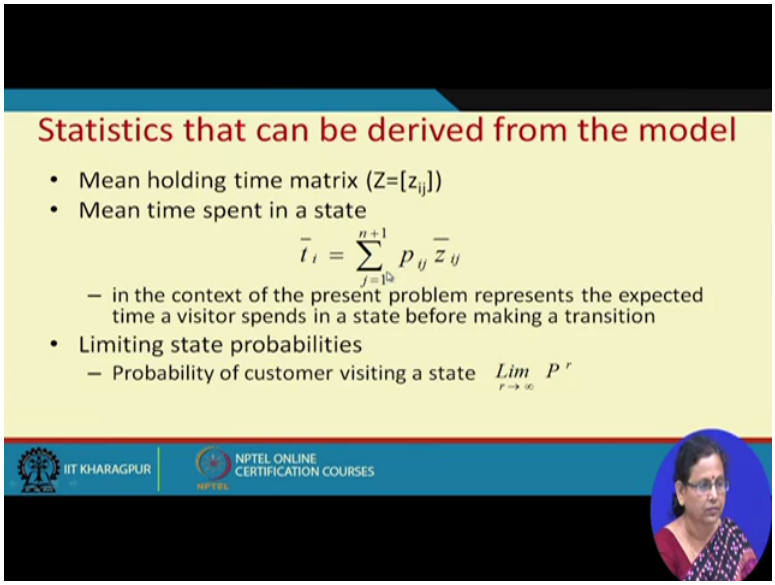
think time in seconds, accumulated think time. So to to find out the average think time, you have to divide it with this number, this frequency number. Okay.

(Refer Slide Time: 20:27)



So now we can compute this, I expect all of you to compute these things and when we gave you the questions, we will be giving the questions before I attempted to solve the problems that will be given to you. This particular example problem, you do it yourself and if you have any doubt, please ask. So the transition part is, so you can compute the following transition probability matrix, you can compute them mean holding time matrix, then after you compute these 2, then you can find out the mean time spent in a state, this is the mean time spent, this is the mean time spent, okay, it is possible.

(Refer Slide Time: 21:51)

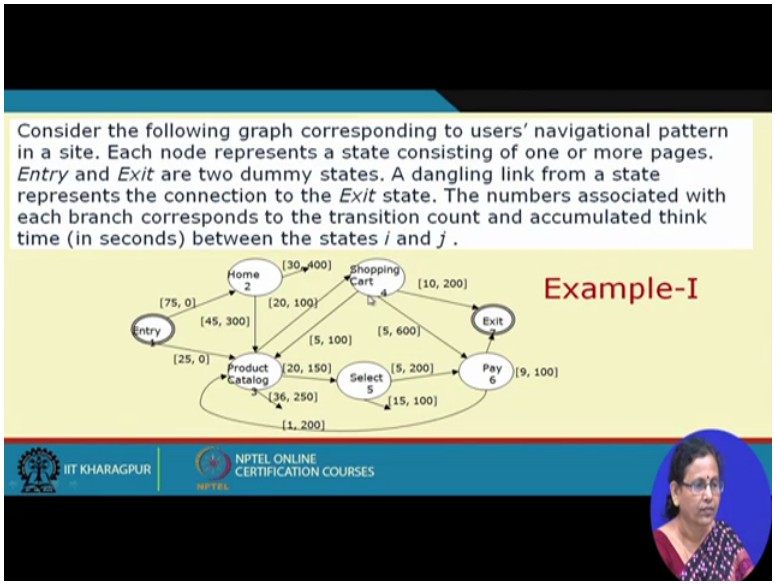


So this is the mean time spent in this state. So the mean time spent in this state can be found out from this formula and once you find out the mean time spent in the state, you can find out the average session length. Now how can you find out the average session length, average session length is this one. Okay. So here also, let me make one more correction, let me make one correction so that the corrections are, you are no more confused on this because we have assumed from the beginning there are total N number of steps and here also we have assumed that there are total N number of steps and here also we assume that there are total N number of steps.

And what is this Nth state, Nth state is the exit state. Okay. So now the slides are corrected here itself, so you do not have to worry about it anymore. So the average session length, average session length can be found out from this formula. Look, doing breeze preliminary calculations, these calculations are quite primitive because what is complicated here is finding this formula. We are not going to talk about how to get this formal because this thing is already available in any textbook, follow let us say Hillier Lieberman on stochastic processes and this formula you can find out from there itself, or any other standard textbook.

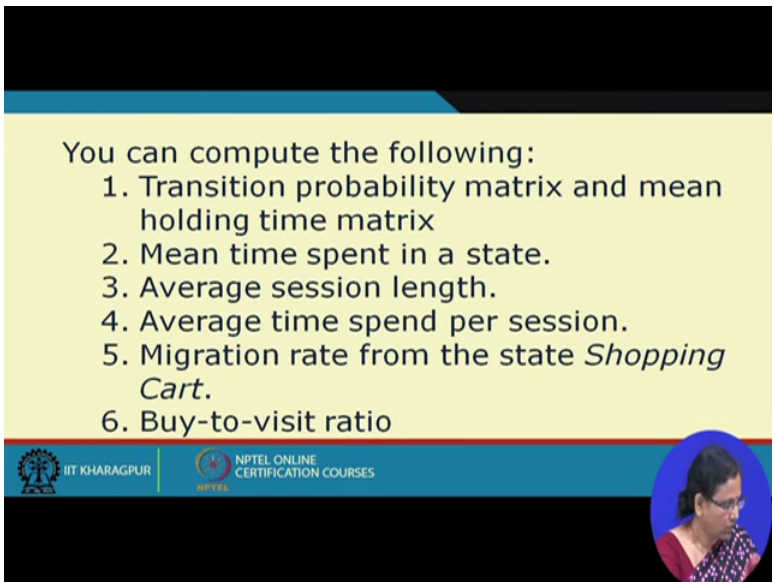
So after you find out the session length, the question asks average time spent per session, so average time spent per session can be found out from here, this is average time spent in the process which means in the in the entire session. For your convenience, because session is something in which the user is responsible for keeping, for being within the site. So this is the average time spent in the session. Then migration rate from State shopping cart, now what is the migration rate from shopping cart?

(Refer Slide Time: 24:31)



Look this is your shopping cart, how many people went out of the site from the shopping cart, which means they have put their products in the shopping cart but they did not come to paste it, they actually straightaway exited from the site, which means they do not ultimately buy the product. Okay, So how many such people, may be many people are, this can be interpreted in this way, many many people are coming, they are selecting the product and they are keeping it in the shopping cart, they are knowing the total price etc. but while buying the product they are not buying it from your online store, they are going to maybe physical store to buy it.

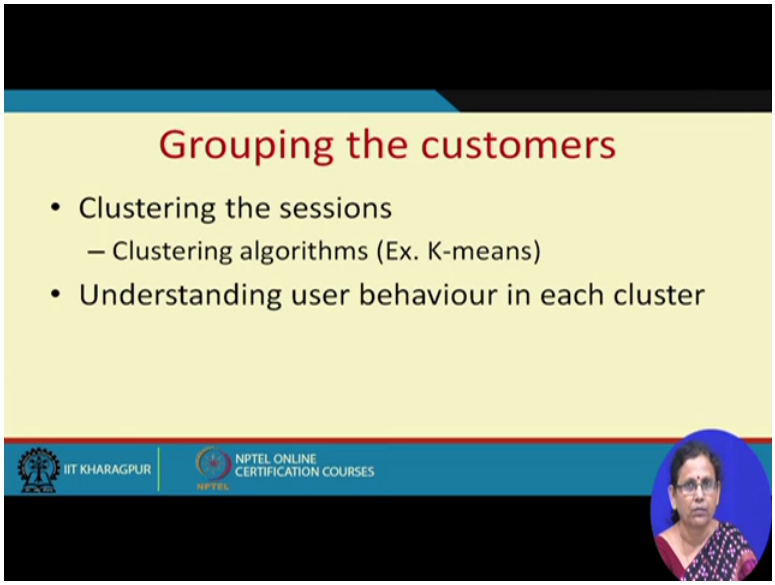
(Refer Slide Time: 25:51)

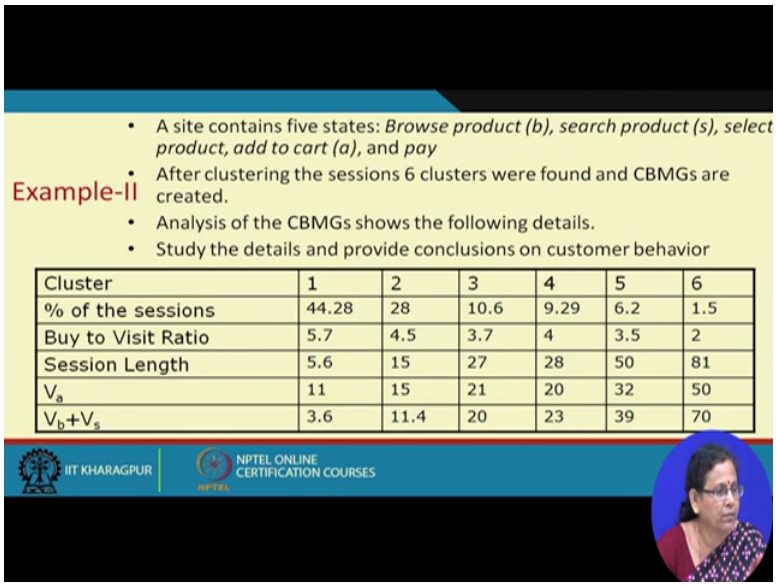


So what is that migration rate, what is the rate at which people are leaving from this particular state, that also you try to find. What is your buy-to-visit, so how many people came to the site

and out of that how many people actually come to the payment sorry payment state, your, this is your payment state, this is where, this is you know finally people who have made payment, which means they have made the purchase. Then how many people entered into the site, lets a total 75 + 25, 100 transitions were made to the site, out of that how many people actually completed this pay process and left the site, it is only 9. So this is your buy-to-visit ratio.

(Refer Slide Time: 26:38)





Cluster	1	2	3	4	5	6
% of the sessions	44.28	28	10.6	9.29	6.2	1.5
Buy to Visit Ratio	5.7	4.5	3.7	4	3.5	2
Session Length	5.6	15	27	28	50	81
V_a	11	15	21	20	32	50
V_b+V_s	3.6	11.4	20	23	39	70

Now so far we have considered that entire access log can be thought of as, taking it as a whole, we have calculation of the customer behaviour. However, to make it more refined, this process, we can now cluster the centre, we can use for clustering the centre, we can use any standard clustering algorithms like K means, etc. and after clustering the sessions for each cluster, we can think of understanding the customer behaviour, that will give few more

insights. For example, consider this 2nd example, here the site contains 5 states browse products, search products, select product, add to cart and pay.

And we do not discuss any clustering algorithm but sessions can be clustered, just assumed this. Now this is your 1st person, 2nd, 3rd, 4th and 5th, you made 6 clusters. Look the cluster types can be understood from these values that you have calculated. For example, look buy-to-visit ratio for the 1st cluster is much more higher than that of the 6th cluster, which means more, in this cluster more number of people actually bought. To explore this further, look, here total number of sessions devoted for this cluster is these many and total number of clusters diverted 4, total number of sessions devoted for this cluster is 1.5.

Which indicates more sessions, people in this cluster where more number of sessions are devoted, which means more sessions means, each session is for captures the history of user for that period. So these many number of sessions, these many fractions of the session, the buy-to-visit is more of this number is much higher than this, so which means more number of people who are coming to the site are interested to buy. It is a good thing that people who are coming to your site is, more people are interested to buy from your site, look at again another interesting phenomena.

The session length here is much less than the session length here, this category of people were not, they, they are, their percent, very few people are of this category. All, their buy-to-visit ratio is very less and they are, they are session length is very high, which means they have continuously browsed through your products, maybe they are the newcomers who browse a lot and purchase less. Maybe this is a group which you should target, this kind of, if you, online environment you can track the properties of this group of visitors, then you can put more online effort like showing advertisements or doing some kind of online activities.

In fact during marketing, we were discussing about how to convert a regular visitor finally to your buyers. So you should be added, you should be considering if they can actually group and target this customer, we can say we can provide the facilities which will make them more interested and finally, possibly they can become a more aggressive buyer like this group from your site. Now this is tells you VA, how many people have added the products to the cart, V A is the add to cart.

Here 11, around this had to cart, this is, what is this V, V is the average number of visits, so have a number of visits of this group to product add to cart category is 11 whereas this is 15.

So this group of people will be adding more to the card but they are not purchasing, they are adding to the cart but most of them are actually purchasing. Then what is Vb, here this browsing and product search is very less because they know where the product is and they have started becoming and buying. Here VB + Vs which is actually browsing and searching the product is too high, these are very inexperienced group of people within your site.

So the point that I am, this is just one example, where the parameters are assumed to be derived from a customer model, customer behaviour model graph which is nothing but a state, discrete time Markov, which has been modelled as a discrete time Markov process and after we derived these parameters, the point that is being made here is you can get many business insights from this data. So the data that we had we saw in the beginning, that is your access log data, which was looking quite unreadable now has been, through this modelling procedure has been brought to a form so that many business insights can be derived.

Now what do you do with these business insights is again a matter to be considered but anyway we are not going into it but at least we know that so many things can be derived and can be, can give useful insights about your customer behaviour. Okay, thank you very much, again we would continue with this discussion in the next class.