Learning Analytics Tools

Professor Ramkumar Rajendran

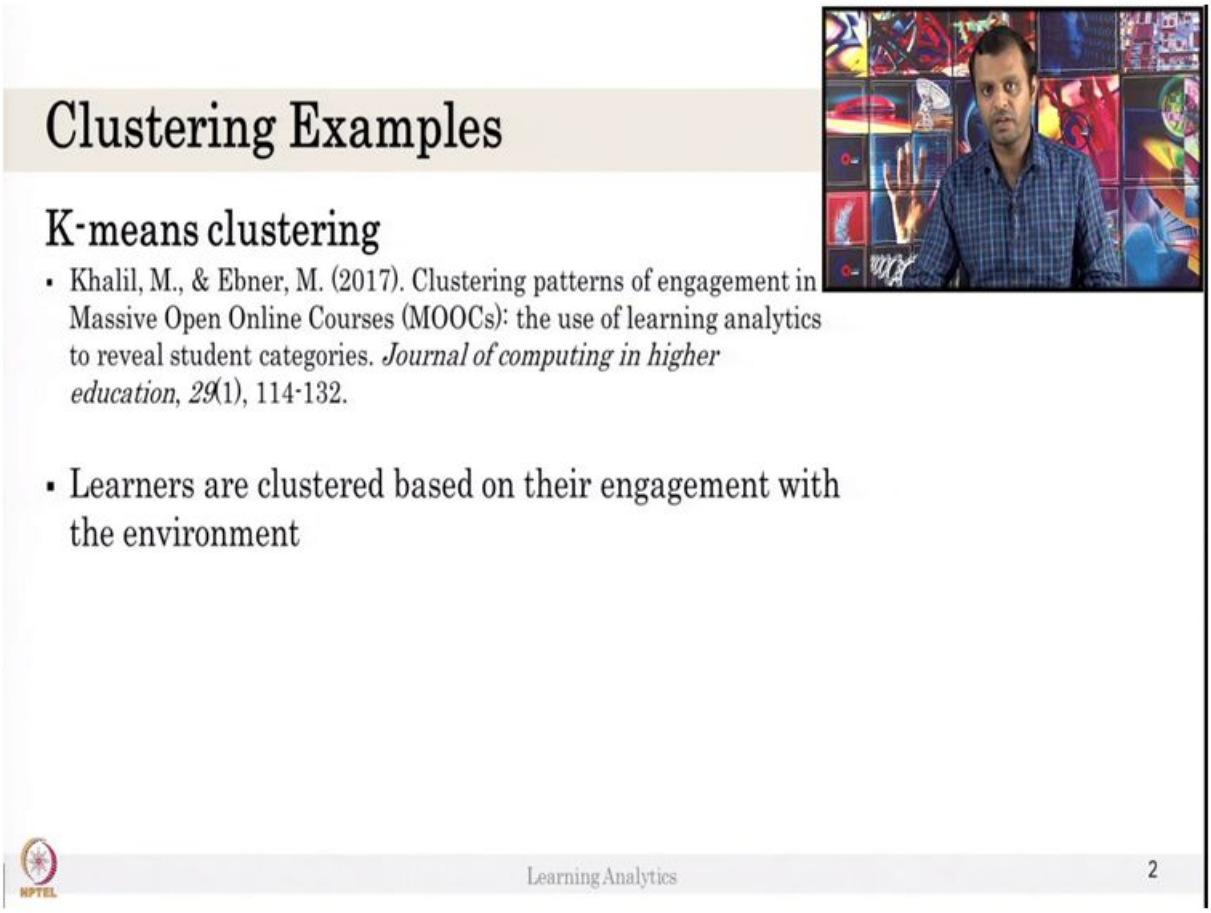
Educational Technology

Indian Institute of Technology, Bombay

Lecture 7.4

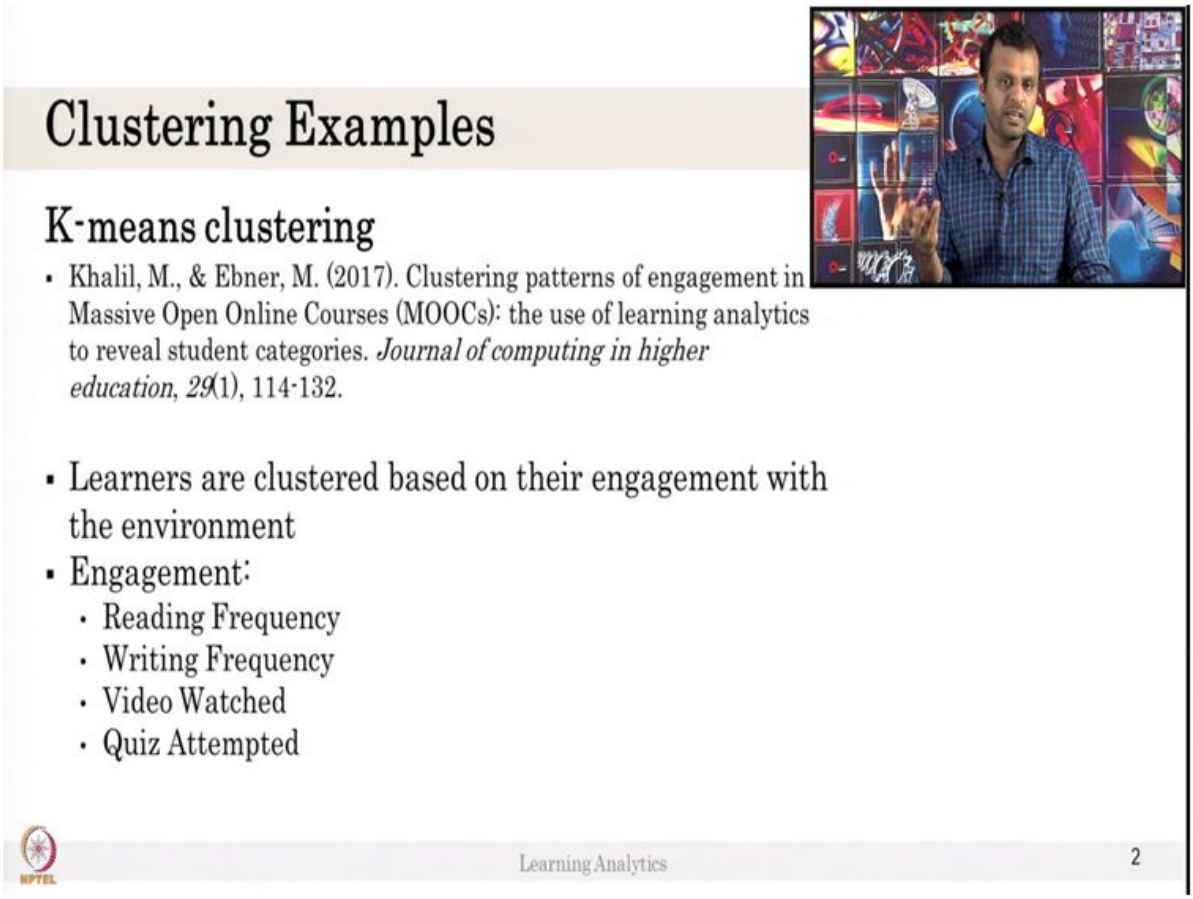
Clustering - Examples

(Refer Slide Time: 00:31)



So, this week we saw two clustering techniques from the diagnostic analytics, we will see where it is used in research papers. So, for K means clustering, I have selected this paper. This is a journal of computing in higher education. So, here the learners are clustered based on their engagement with the MOOC environment, okay? So that is a question of the clustering patterns of engagement in massive open online courses.

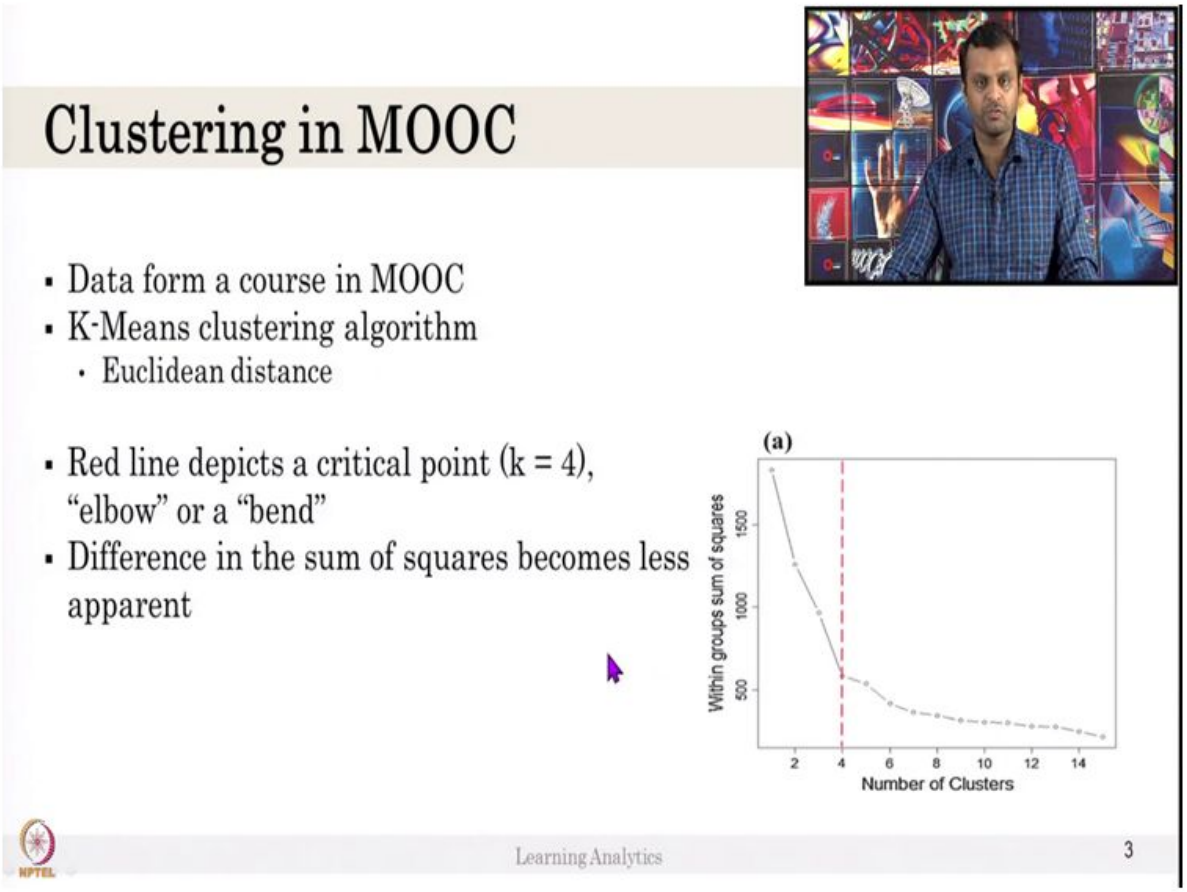
(Refer Slide Time: 00:48)



So, what is learner's engagement? They measure the learner's engagement in a MOOC environment using four metrics. What is reading frequency i.e. how many times a student read a certain text? What is writing frequency i.e. how many times a student writes in the forum, that is basically writing and reading in a forum, and how many videos a student has watched and how many times they attempted the quiz. So forum reading, forum writing, video watched and quiz attempted they computed the frequency of students interacting with the MOOC that is for each

student. Then if you have hundred students, their data is coming in. Now, there are four variables what we saw in last class in the clustering example is only two variables. So, we can think and look at it, but for four dimension we cannot do that. So, we can only apply the K means clustering algorithm.

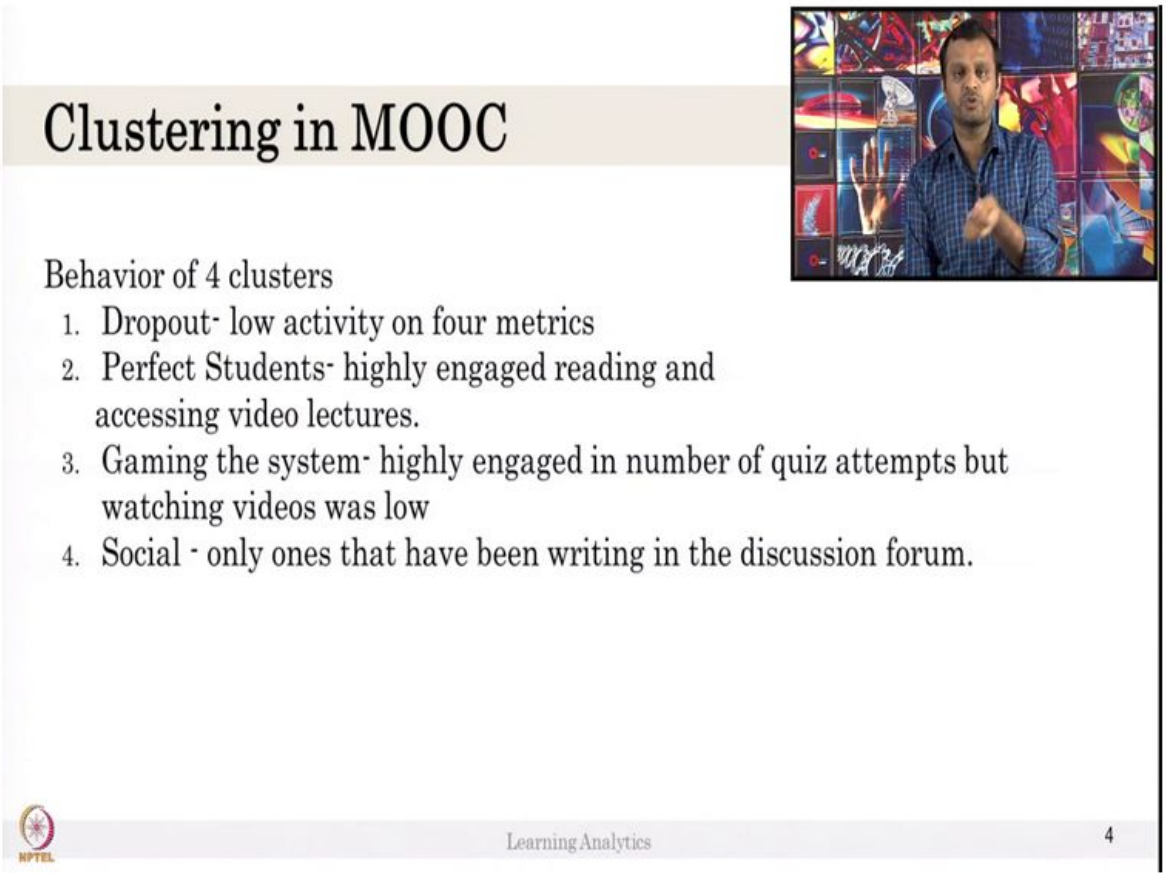
(Refer Slide Time: 01:36)



So, they applied K means clustering algorithm on all this data and they used Euclidean distance and they tried for k value from 1 to 15, remember? So, based on the K value, the objective function J value is reducing, right? The function is reducing, but they choose k equal to 4 the reason is we know the elbow is at this point. And also from this point (from 4 to 5), the error function reduction is very, very small.

Or you can choose k at 6, whether you can choose K at 4 or 6 based on your research questions and what the data is given to you. Here they choose k equal to 4.

(Refer Slide Time: 02:32)



Let us see what the k equal to 4 means. Here, they looked at the students in each cluster, okay, there are 4 clusters, each cluster gives you a set of students. Cluster 1 will have some 20 students, cluster 2 would have some students something like that. Looked at the data or looked at the behaviour of students in a cluster 1. Remember this is diagnostic analytics, this is not a predictive antics, so we do not know why a student was getting the lowest score? Or why the student is not able to clear the quiz?

So, we clustered and we see there is one cluster, all the students are doing less activity on all the 4 metrics like they are not doing any reading, post writing any forum messages or they are not watching the video, they are not taking attempting quizzes. So, that is why they are not really interested in the course, so they may dropout.

So, how do we give the name “dropout” or “perfect student”? It is up to you because you are a researcher based on the data based on your domain expertise, based on your understanding you can give the names. So, they give a name as “dropout”. So, because all these four metrics are low, so these students are mostly dropout so this we can check whether these students are going to drop out or not in predictive analytics, if you see that, that will be clustering can be classified as a cluster.

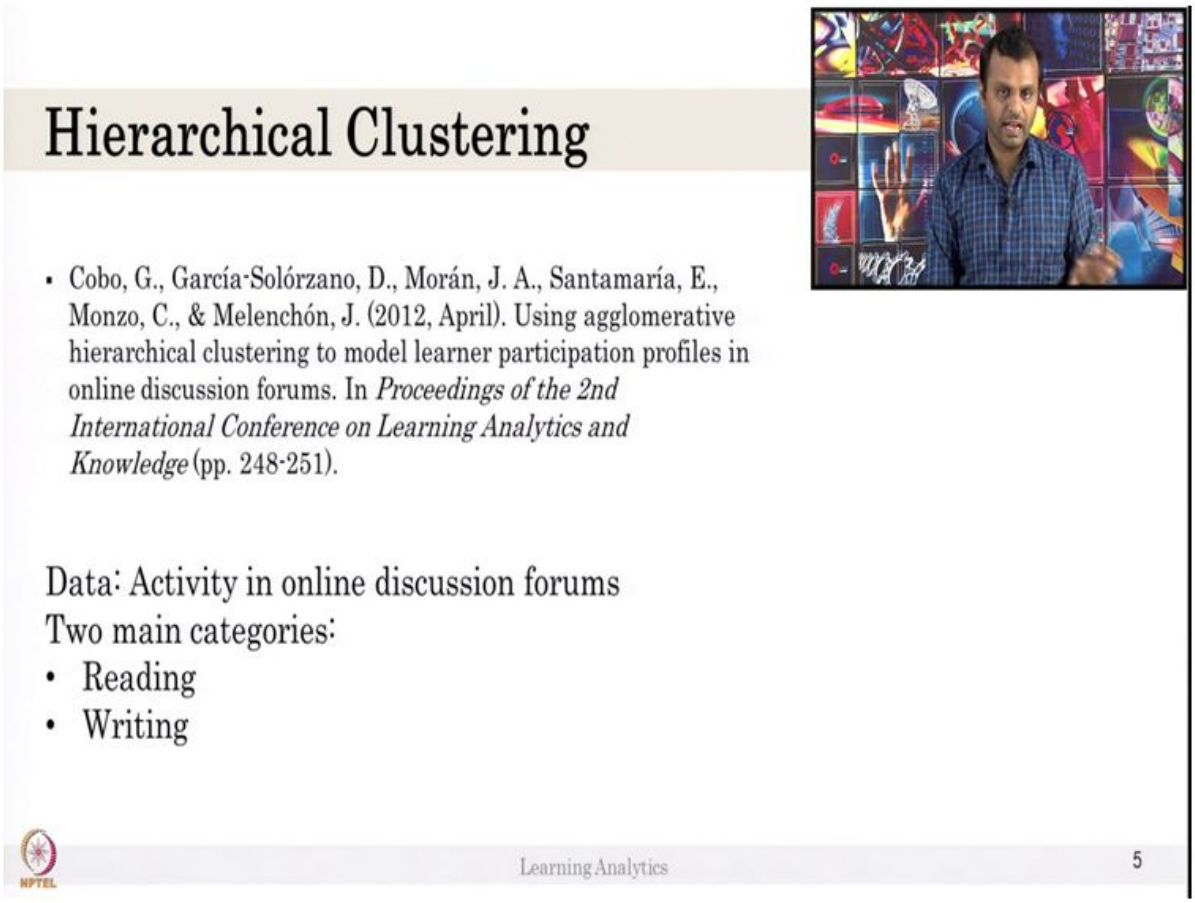
The second one is perfect student, they are highly engaged in all the four metrics especially accessing video lectures and reading. So, they are “perfect students”, they might do well and they might continue the course till the end. There are some students who were gaming the system, who are highly engaged in the number of quiz attempts, and very, very low on video watching.

Which means they do not care about what is delivered in the content. They were very confident, and think “I can solve the quiz I know all the content here”. So, they go and directly attempt the quiz, they are gaming the system, they might succeed, they might not succeed we do not know about that. See, we do not know about the success part here but based on interaction we know they are gaming the system.

There are some students who are social and were the ones who are highly writing in the discussion forum they do not care about other interactions like watching the video or reading. Instead, they are the one who are actually writing in the discussion forums. Others are even not even writing. So, these kinds of mostly social they are not watching video much.

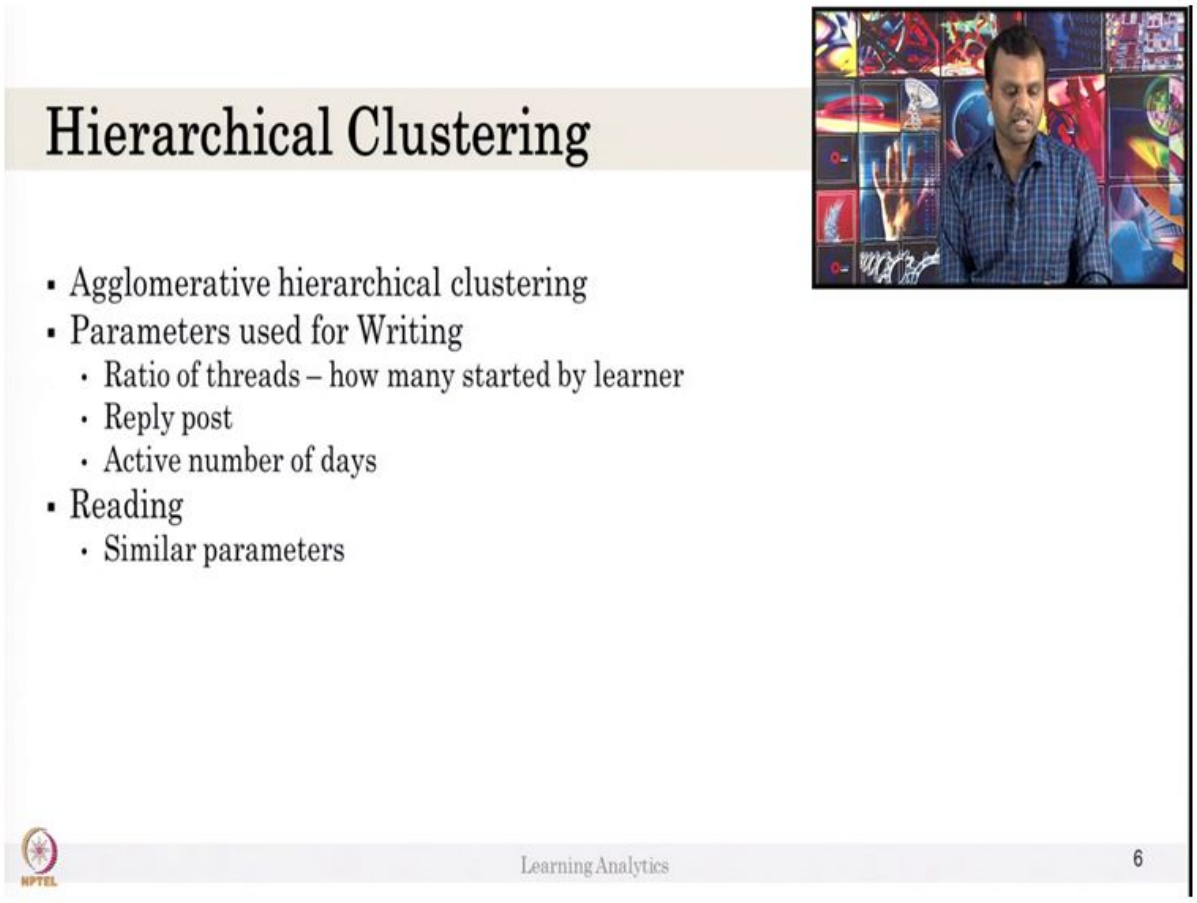
So it is up to the researcher to find out what are the metrics to consider for clustering, and how to make an inference from the clusters. So, the K means algorithm will help you to find the number of clusters, that is it. So, the technology will help you to find the number of clusters, how to group them from the data. But choosing the right parameters to create clusters and making inference from the clusters that is up to the researcher. That is why domain expertise is needed. So, in this course, I would like you to develop that expertise and not only to use the algorithm or apply the algorithm.

(Refer Slide Time: 05:33)



So let us look at the example for hierarchical clustering. In this paper, in the LAK 2019, the authors used an agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. Again, this is also on a discussion forms based on the student's participation in the forum they are creating the clusters(AGNES). They use data in discussion forums, they classify the data into two groups reading and writing. So, reading is a separate activity, writing is a separate activity.

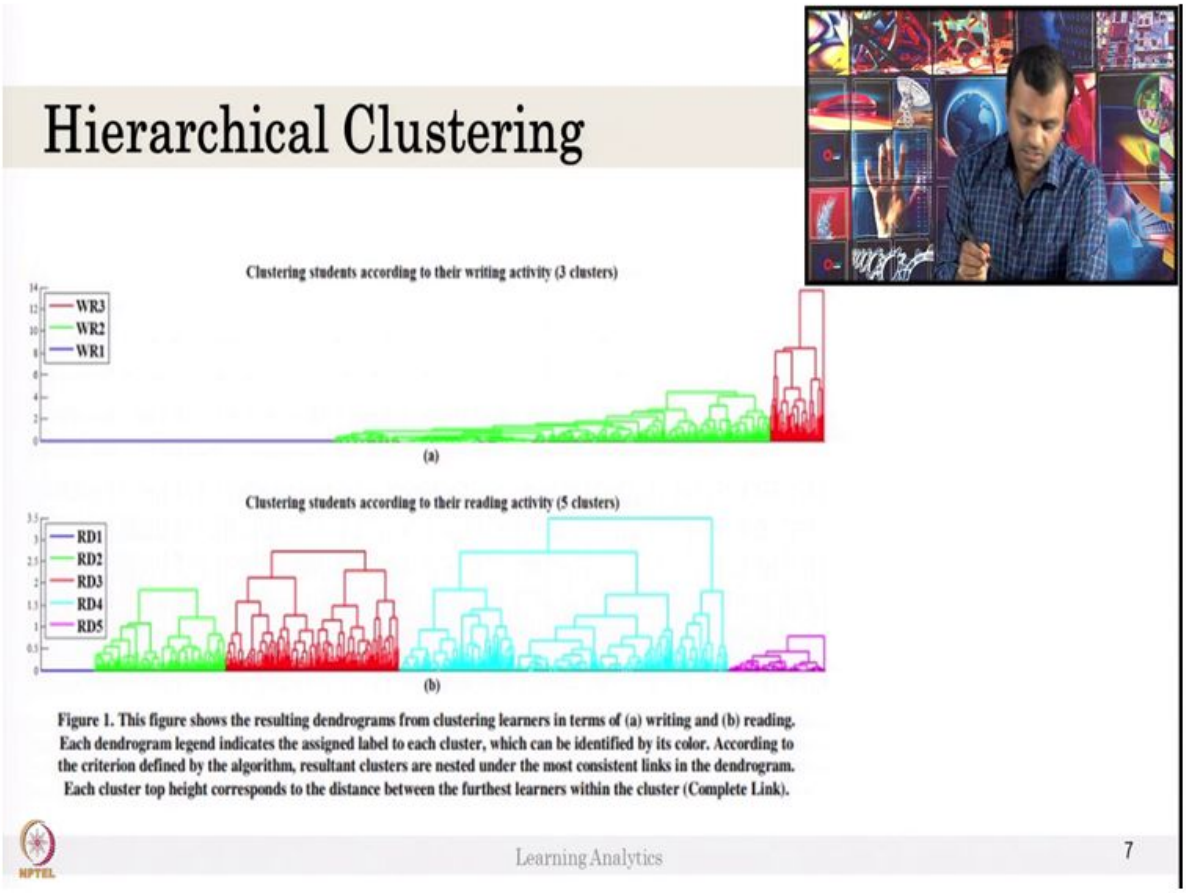
(Refer Slide Time: 06:06)



From reading, they selected some data points. For example, for writing, they used a ratio of the threads, how many times threads has been started by the learner compared to all the threads started in the discussion forum. How many times the learner replied to the post, active number of days, how many times has learner logged in, in the number of days how many days he really created a thread, or he replied to a post?

Similarly, they created four parameters for writing and similar parameters computed for the reading. So, now they have data for reading separately and writing separately, using these data for all the students, they computed the agglomerative clustering.

(Refer Slide Time: 06:57)

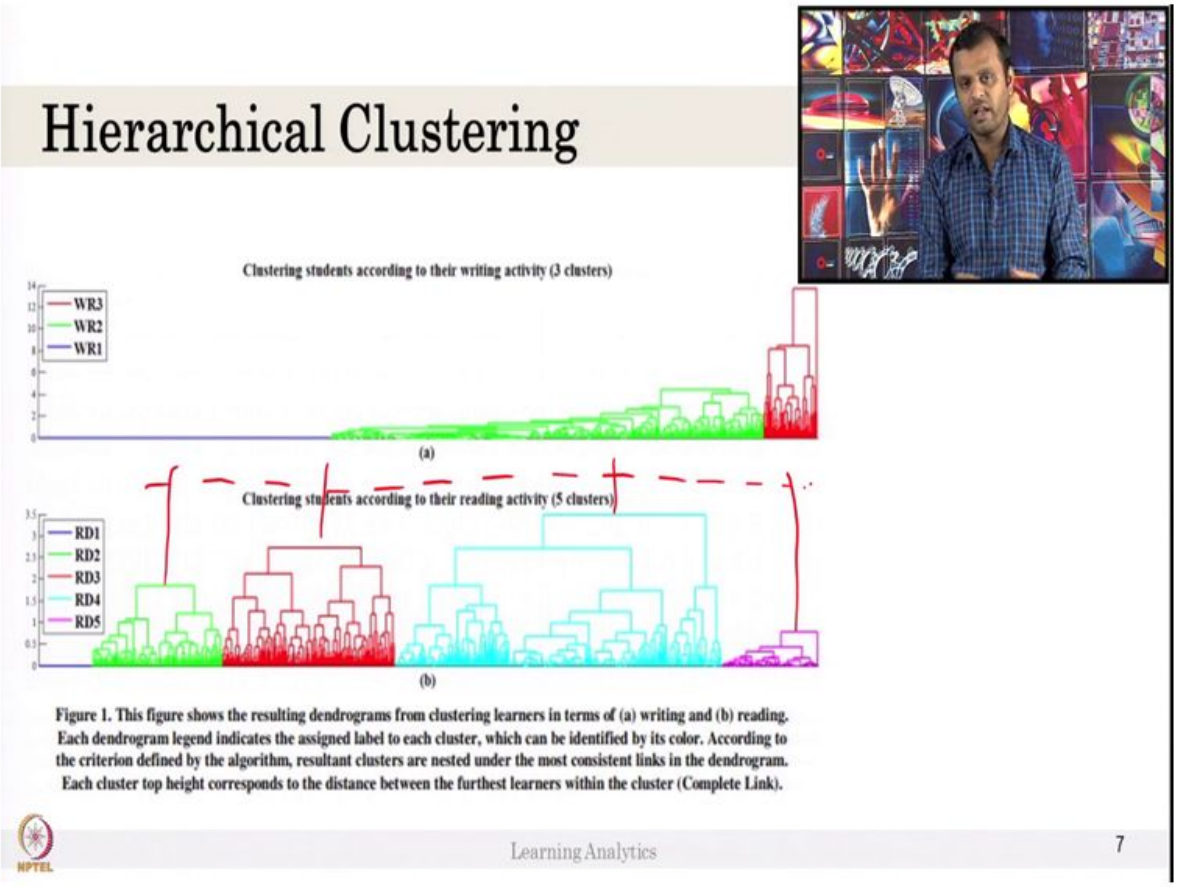


Let us look at the clustering values. So, what happened is they based on the writing activity, grouped into three clusters here, this can be combined further to make agglomerative clustering but they may leave it here actually.

Let us look at this, this height you know, the side distance actually tells you the dkistance between the farthest learners within the clusters. If you are finding the similarity measure between two points using one of the similarity measure function. The sum of function they have used here is a complete link. Complete link is finding the farthest point within the same cluster. So, if you use the complete link, the distance between the clusters is indicated by the height of this dendrogram.

You know, this height actually indicates how the distance from this cluster to other clusters. So, similarly, for reading behaviour, they computed the four clusters. See there are five clusters they created.

(Refer Slide Time: 08:16)



This might be combined again to two clusters, which again can be combined to form one cluster. But they picked five clusters, so they selected these students as behaviour 1, behaviour 2, behaviour 3, etc. So, the five clusters they wanted to analyze. So, this is an example of hierarchical clustering used on our data. So, you can also apply this kind of clustering algorithms on the data for diagnostic analytics to understand why student behaved like that.

(Refer Slide Time: 08:52)

Since we saw two papers in K means clustering also in hierarchical clustering, we asked the same question, we discussed at the beginning of this week's first video. Can you list down two applications of clustering in any learning environment? The same question we asked in the first video? Now you know, what is hierarchical clustering? What is k means clustering? Now you saw the application of them in two papers, can you list down? After listing it down, resume the video to continue.

There is no response to the previous activity. You have to compare your response to the first video the similar activity and the last activity and see there is an improvement or not. If you understood clustering and how clustering can be used in different learning environments. So, I request you to go and watch, read content regarding clustering in the online. Thank you