Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur
Lecture 34 - Recommendation Engine and Retail Analytics

Hello everybody, welcome to Marketing Analytics course. This is Dr. Swagato Chatterjee from VGSOM, IIT Kharagpur who is taking this course for you. We are studying week 7, session 1 and in week 7, we will discuss about Recommendation Engine and Retail Analytics. So this is our topic. So majorly, we have, in the next couple of weeks we will talk about what kind of problems that we face in retail and we, I will at least for the first two weeks means week 7 and week 8, I will majorly focus on e-commerce.

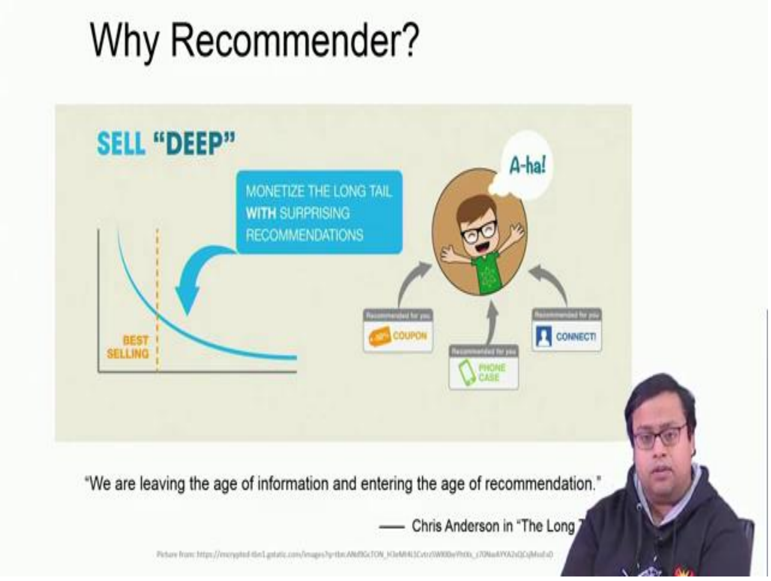
Now why e-commerce? E-commerce is something that is coming up in the whole world actually in a quite a big way. So Internet is available with pretty much many people and let us say this Jio is giving this much offer and the other like Vodafone, Airtel all are giving various offers in Internet. Internet is becoming cheap. At our time when let us say around 10 years back when we were youngsters our the Internet was not very much available and the cost of having Internet in your mobile phone was very high but now Internet has become cheap and people who are in rural areas also can access Internet because availability of internet is also there means the connection is also good.

Giving that as the background there are lots of e-commerce firms that has come up, so Flipkart, Amazon are there in India. Probably other in the B2B sector also there are Alibaba and let us say there is Blackbuck and there are various kinds of aggregator firms that is coming up and then also there are aggregator firms who works as a platform, for example, they are actually connecting the customer requirements and the, I would say the service provider firms, they are creating a platform where these two people can come together.

So those kind of platforms are also coming up and they are giving services in urban areas, semi urban areas, rural areas. So in short, e-commerce is becoming an important aspect of retail business. Still in India, I would say that majorly Indian retail is unorganized, probably majorly the whole world's retail is unorganized. A very small part of that is organized and again very smaller part of that is e-commerce but the growth rate of e-commerce is pretty high. So though it is still in terms of size probably not significant if we compare with the total retail industry, the growth rate is something which is very much higher for e-commerce firms.

Given that as a background we have to focus on how Marketing Analytics can help retail firms as well and to be specific e-commerce firms because that is where the application is pretty high. Now why a recommendation engine, is something that I will discuss now.

(Refer Slide Time: 03:18)



Why Recommender?

"We are leaving the age of information and entering the age of recommendation."

— Chris Anderson in "The Long

Picture from: https://encrypted-tbn1.gstatic.com/images?q=tbn:[illegible]

So here you will see in this particular slide, you will see that there is something called you have, you can sell deep. Sell deep means getting the first initial sells is pretty easy but monetize the long tail with surprising recommendations is something that is this particular picture is saying.

It is saying that up to that, up to this line this is easy to get, no issues. But after that whatever sells happen in this area which is the minute sells or which is the incremental sells but that is a long tail is there, means you can do incremental sells, incremental sells slowly slowly as much as possible and as that incrementation can go up to infinite you can actually sell, upsell or cross-sell little little products. When you do it in a large scale, that contributes to a revenue in a very huge matter.

So if I let us say all these recommendation engine and etcetera is incremental sells. For example, if I go to an e-commerce firm in a website, I have I will have something in my mind to buy. Let us say I want to buy a mobile phone. I have, I will probably do all these, I have already talked with my friends, I have talked with other people who has idea about mobile phones, I have searched in websites and in online media and etcetera and then I have created a consideration set. Now with that consideration set I have let us say probably I have done a little bit more analysis and created a choice set.

Now I want to make a choice. So when I go to the e-commerce firm's website, I want to buy something with very high probability, whatever I have planned for I will buy that. But if some recommendations come, I might probably switch from whatever I wanted to buy to whatever these guys are recommending. So little bit or probably I will buy an extra item. For example, let us say somebody you are buying a mobile phone.

They give a recommendation for cover, supportive cover for the mobile phone or protective cover for that mobile phone. Now you might not have an idea that okay, you will buy that. You want to buy a mobile phone but because that particular thing was recommended and that was a very small price in comparison to the mobile phone's price and you thought that okay, I will buy that and when you thought that okay, I will buy that, you bought that and that is a cross-selling that happened.

Now this cross-selling is incremental sells that does not contribute towards the majority of your sale volume. But that incremental sale, let us say there are 10 million mobile phones are being sold in a particular year and out of those 10 million 10 % guys actually do a cross. So 90 % case it is a failure, only 10 % case that recommendation leads to some successful conversion. 10 % is a very big number.

It probably does not happen that way. But even if it is 10 % then 1 million cases are sold and each case is let us say 500 rupees. So 1 million × 500 rupees you are getting 500 million extra revenue from probably if you have not done that, probably most of the people will not buy. You will say that how many people if you go and check your friends and how many people actually buy a supportive case. Not even 10 % probably, probably less than that. But this recommendation might boost a little bit up.

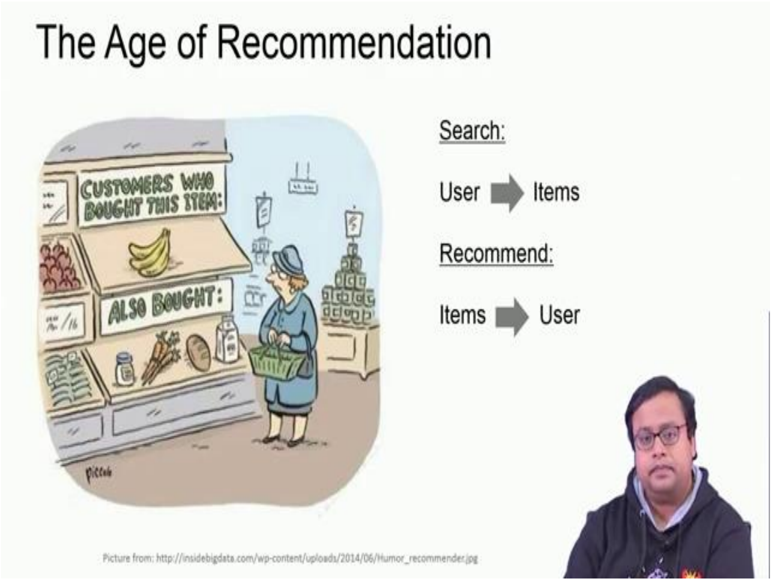
So that is something that we are trying to see that how I can capture the long tail. So recommendations for you can be the coupons, recommendations for you can be the phone case as it is told. It can be, recommendations for you can be whether you want to connect with other people or not. So there are various types of recommendations that comes and these recommendations creates ‘aaha’ feeling. It is a surprise that comes into your mind.

That aaha feeling is something that is important and we are leaving the age of information and entering the age of recommendations. So please understand the "The Long Tail" he told that recommendation works lot. So often times we actually want recommendation from a legitimate source of information but sometimes even if the information is not legitimate in terms of you do not know who is recommending it, it is the platform that is recommending you. You still probably would like to have a little bit of positive association towards this recommendation because without recommendation without any suggestion to you your chances of buying is much lower than with recommendation irrespective of from whom this recommendation is coming.

For example, let us say you are walking in the road. You are hungry, you want to have some food. One thing is that let us you ask somebody, so why will I get a good food? Anything, good fresh food. Whatever he says even if you do not know that person, you might probably whatever he says you will have a positive association towards that particular thing. Let us say he says okay, go to that restaurant, it is good I know. So the moment he says that, even if you do not know the person, you might say that okay, why will he give a bad recommendation to me?

He does not know, me I does not know him so I do not know him. So that will not create a problem. So I do not think that he is recommending to create a harm to me. So in general, we trust recommenders. So based on that I am saying that the moment is a simple nudge, a small recommendation comes up, your chances of buying goes up and that is why recommendation engine is important.

(Refer Slide Time: 08:44)



The Age of Recommendation

Search:

User → Items

Recommend:

Items → User

Picture from: http://insidebigdata.com/wp-content/uploads/2014/06/Humor_recommender.jpg

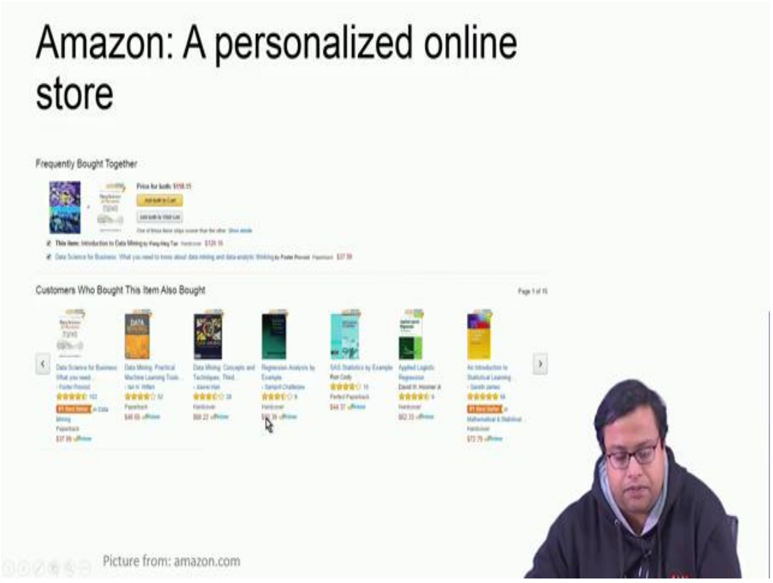
So there are search and so there are two types of products. So customer, so this is something that we see. So customers who have bought this item has also bought these items and it is a special case of in a brick and mortar retail store. But in a brick and mortar retail store, actually, it is not written like this. In brick and mortar store also recommendation engine works but they just put the product.

For example, you will see that just beside utensils or just beside crockeries, there will be a Scotch bite or the powder, washing powder, dish washing powder will be there, just beside the crockeries because it is often be the case that when you are purchasing this, you might have in your mind that okay, I might have to wash these dishes as well and then they create a special kind of dish washer, dish washing soap and etcetera, which is required for specialized some kind of crockery.

So you might have a feeling that okay, for my crockery which is of high value, I should buy a liquid dish washing soap rather than a solid cake kind of soap because we have a perception that cake kind of soap is probably not for high quality good crockery. It might create a damage. It is there is too much of pH content high or low or whatever and then that will probably create a corrosion on my utensils and etcetera. So that kind of feelings might come and that is why you might prefer a liquid thing.

That will be good to your hand also because you or whoever is dish washing and it will be good for the crockeries also. So those kind of feelings are there and because those kinds of feelings are there people actually put those items beside the crockery items. So in natural cases, users search items, so you find out the user and you see that what kind of items are good for him and in case of recommendations you find out the items and you suggest that the items are suggested to the users rather than users are searching the items. So this is something that is also coming up very commonly in both in e-commerce, mainly in e-commerce but also in smaller in brick and mortar retail firms as well.

(Refer Slide Time: 11:03)



Amazon: A personalized online store

Frequently Bought Together

This Item: Introduction to Data Mining by Pang-Ning Tan

Data Science for Business: What you need to know about data mining and data-analytic thinking by Foster Provost Paperback $37.99

Customers Who Bought This Item Also Bought

Data Science for Business: What you need... Foster Provost 102 Paperback $37.99

Data Mining: Practical Machine Learning Tools... Ian H. Witten 52 Paperback $40.65

Data Mining: Concepts and Techniques, Third... Jiawei Han 28 Hardcover $60.22

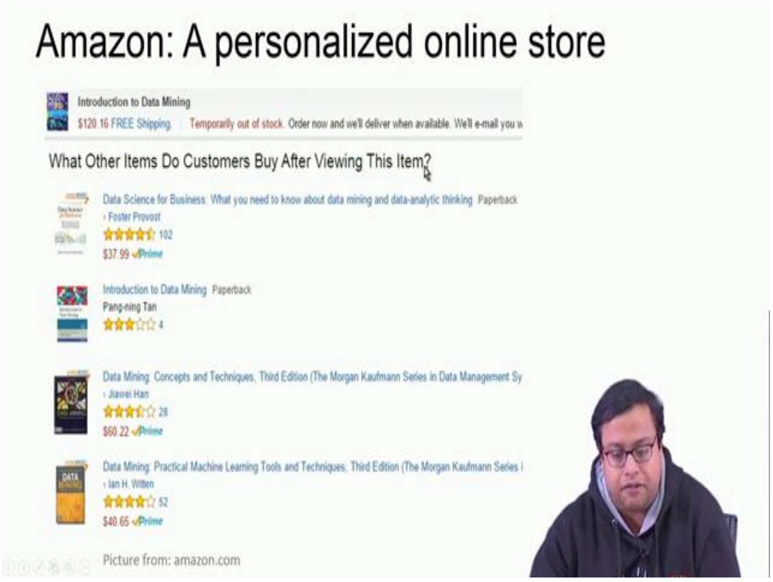
Regression Analysis by Example Samprit Chatterjee 9 Hardcover

SAS Statistics by Example Ron Cody Perfect Paperback

Applied Logistic Regression David W. Hosmer Jr. Hardcover

An Introduction to Statistical Learning... Gareth James Hardcover

Picture from: amazon.com



Amazon: A personalized online store

Introduction to Data Mining

$120.16 FREE Shipping. Temporarily out of stock. Order now and we'll deliver when available. We'll e-mail you w

What Other Items Do Customers Buy After Viewing This Item?

Data Science for Business: What you need to know about data mining and data-analytic thinking Paperback
› Foster Provost
102
$37.99 Prime

Introduction to Data Mining Paperback
Pang-ning Tan
4

Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Sy
› Jiawei Han
28
$60.22 Prime

Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series i
› Ian H. Witten
52
$40.65 Prime

Picture from: amazon.com

Now going ahead this is some of the examples. For example, frequently bought together. So these are the two books, for example, book one and book two. These two books are frequently bought together. So these kinds of suggestions comes up in Amazon or customer who bought this item also bought these items.

So customer who bought this first book, so this is the page of the first book and that is why the other book is tagged along and they will also buy these books as well. So these books are also there, so because they know that okay, this guy is searching for data mining books and there are lots of other data mining books, so those books will also come in the recommendation of Amazon. So this is one type of recommendations.

Another kind of recommendation is more personalized. So Amazon is a personalized opening online store where the introduction to data mining is what you are seeing. What other items do customers buy after viewing this item? So, see, they are different. These two are bought together. These two are when some of the item one is bought, the second item is also bought.

And here when one item is seen, second item is bought. So the recommendations also when you create these recommendation, you can have different kinds of recommendation. One is that both items should be in the same basket. Second is both items in past history might be in different baskets but both items has to be bought. So you buy apple and you buy orange. You buy apple and orange together that is the first case, this case.

You buy apple and orange differently but in a time period, in 7 days time period you buy in different trips you buy but you buy them both apple and orange, this is your case and you have seen apple but bought orange. You have contemplated that or you have delicated with yourself that okay, this apple and the orange I have seen it but then I bought orange, that is this case. So different kind of purchase situation, in a customer journey there can be different kind of situations where this will work.

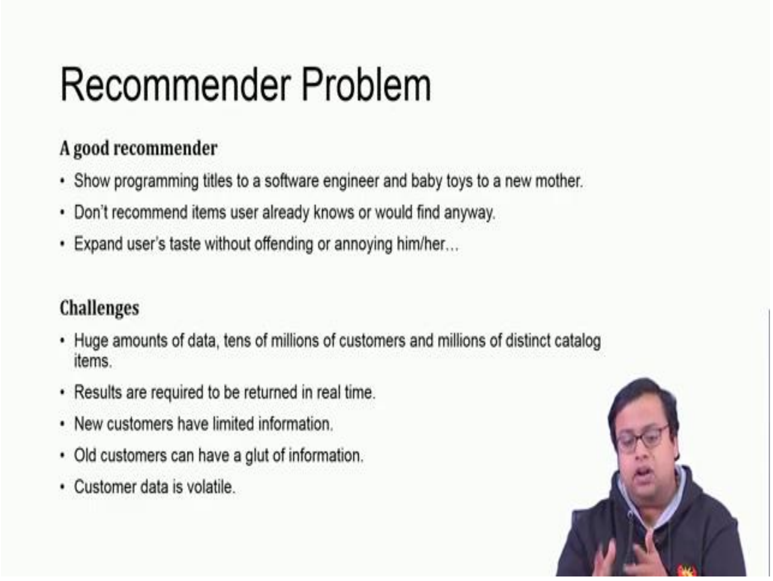
So this is probably applicable when you are let us say in information search stage in your customer journey. On the other hand, this is applicable when you are actually purchasing, you are actually making the purchase and this is post purchase. Customers who bought this item also bought is a post purchase situation. So you have bought it, you got to the probably it will be more applicable when you went to the purchase you got the purchase order number and etcetera, payment successful, that particular page and then it comes in the below that okay, you have purchased this and now there are lots of customers who already purchased this, also buys this.

Probably also it can come when you open up your Amazon. You say that okay, so that is the last purchase that you have made then people who have bought these products also have bought this kind of products. So that is also one kind of recommendation engine. So different kinds of recommendation engines can be created in terms of when the recommendation is coming. So based on that the backend calculations might be a little bit different, different in terms of what source of data you are taking.

If you are taking this thing then you have to also consider what customers are seeing. Customer viewing data should also be collected because customers who have seen this then they bought this one. So customer viewing data and there customer buying data has to be matched. On the other hand, in this case only customer basket data in the same basket will be matched. Two different baskets will not be matched.

So you bought product A and product B in the same basket then only it is 1, otherwise 0 but here product A and product B can be bought in different baskets but both has to be bought not viewed both has to be bought then only 1. So the first hand calculations the initial I would say contributing data set, the primary raw data set will be different in these three different cases but more or less the algorithm will remain same.

(Refer Slide Time: 15:18)



Recommender Problem

A good recommender

- Show programming titles to a software engineer and baby toys to a new mother.
- Don't recommend items user already knows or would find anyway.
- Expand user's taste without offending or annoying him/her…

Challenges

- Huge amounts of data, tens of millions of customers and millions of distinct catalog items.
- Results are required to be returned in real time.
- New customers have limited information.
- Old customers can have a glut of information.
- Customer data is volatile.

Going ahead, what is the recommender problem? A good recommender is which shows show programming titles to a software engineer and baby toys to a new mother. In other words, there has to be a fit. Whatever you are recommending, the customer’s requirements and your recommendation should have good fit. You will not show a programming title, means that means a book for programming to a mother and baby toys to a software engineer. So you might show baby toys to a software engineer as long as he is a father or mother but you will not show a baby toy to a software engineer.

So mother or father irrespective of his or her profession. So programming language you will show it to a programming language's books you will show to a software engineer. So in other words, you have to have a good fit of whatever you are recommending and whoever is the recipient of the recommendation and do not recommend items users already knows or would find anyway. So see, we are doing incremental sales. This is something that is very very important when you are doing. So it this anyway recommendation engine will require a little bit of calculation.

The more you can reduce the calculation the better because it in often case in e-commerce firms there are lots of products probably how many? Millions or if not millions lakhs of products that are there in the e-commerce firms these days. Now if I have to create some kind of matrix similarity and etcetera for 1 lakh products, that becomes a very difficult calculation. Lots of, not difficult I would say very time taking calculation, heavy job. Now I have to find out ways to reduce that heavy job, as simple as that. If I, and because that is costly, so any calculation see when it is 1 million by(/) 1 million matrix you are calculating that will take a huge space in your in whatever server or whatever you are doing the calculations.

That will take more time also. So 1 million, so 10 to the power 6 into 10 to the power 6 (power)that is 10 to the power 12 cells you have to calculate, so that will also take lots of time. So time and resources cost money as simple as that. So if it costs money it is better to find out ways to reduce that cost. So I will only incur cost when that incremental benefit, the marginal benefit of this recommendation is pretty high. If you anyway buy it, so I will not recommend anything which you will anyway buy and I will not recommend anything which will you never buy.

So these are the two things. Feed is important because you will never, otherwise you will never buy. For example, a mother I will recommend him, I will not recommend him let us say something which is related to let us say yeah, programming titles or books or probably items which she cannot use right now. I will not recommend them those kind of stuff. For example, a pregnant lady. You find out a pregnant lady. Will you recommend probably anti-pregnancy pills or let us say some pregnancy test to that pregnant lady when you know that this person is already pregnant?

For long period of time probably that will not be needed but you will probably recommend baby toy not baby toys but probably books which is related to baby's upbringing. You can recommend like folic acid or different kinds of medications which helps her in keeping things okay and etcetera. So those kind of stuffs you will recommend. So you have to find out products which are fit to this particular person who is getting the recommendation, that is number one. Number two is you will not recommend such kind of products which he will anyway buy. You do not have to, so because if you have to take those kind of products also in the recommendation platform then you have to also calculate, do some calculation for them.

Now that some calculation will give you no result because this person will anyway buy, irrespective of the nudge he will buy or she will buy. Now the nudge will create some cost in your resource, in your time allocation and etcetera, it will create some cost. So it will be a basically a loss for you. So do not recommend items user already knows or find out anyway and expand user's taste without offending or annoying him her. So this is also important decision. So there is some time you have to find out that this is also actually associated with the fit case that I will not recommend something which is very far from the fit which has to be reasonably far.

Now how far you will be allowing is something that is a very crucial decision. What are the challenges? So huge amount of data I told, tens and millions of customers and millions of distinct catalogue items. So if there is millions of products and probably 10 million 20 million customers then this becomes a huge calculation and those such a size of calculation becomes problematic. What are the challenges?

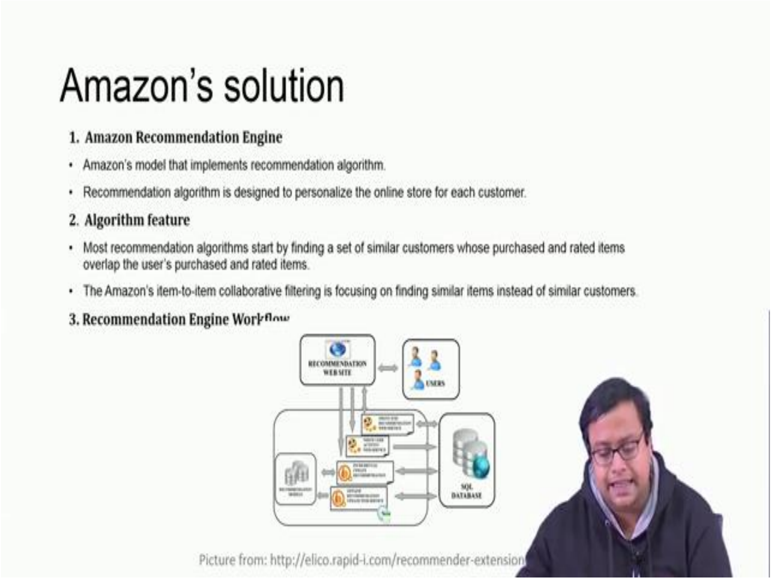
Results are required to be written in a real time. So many result you have to, so I will, let us say I am seeing a website in Amazon and you give me recommendation probably half an hour later that does not work. You have to give me recommendation then and there. So you have to do all the high end-calculation beforehand so that at the time of the recommendation it works real time. New customers have limited information that is number thing, so it requires a little bit of customer purchased data previously. You will be giving me a good fit only when you know what is fitting for me, what kind of person I am, what I like, what I dislike.

If you do not know that information you cannot recommend. Now if to get that information you have to have be old customer. I cannot be a new customer because for new customer those data is not available. Old customers can have a glut of information. So old customer have sometimes have huge information, you do not know which one to use, which one not to use. A customer data is volatile.

Customer change is overtime. The same customer might be different so the same person me let us say before my child was born and after my child was born are two different people. The same person before marriage after marriage we say for female, there are lots of in a customer segmentation video I told that women life changes and that is why woman's lifestyle not lifestyle say woman's life period, life span is a very interesting case for marketers because that changes overtime like anything.

So the same woman who is in adolescent versus who is in let us say in her youth age and then after a mom and probably after the second child and etcetera, the same person changes a lot. So customer data is volatile, how you can handle that? So all of these things are challenging which we will try to solve.

(Refer Slide Time: 22:36)



Amazon's solution

1. **Amazon Recommendation Engine**
- Amazon's model that implements recommendation algorithm.
- Recommendation algorithm is designed to personalize the online store for each customer.

2. **Algorithm feature**
- Most recommendation algorithms start by finding a set of similar customers whose purchased and rated items overlap the user's purchased and rated items.
- The Amazon's item-to-item collaborative filtering is focusing on finding similar items instead of similar customers.

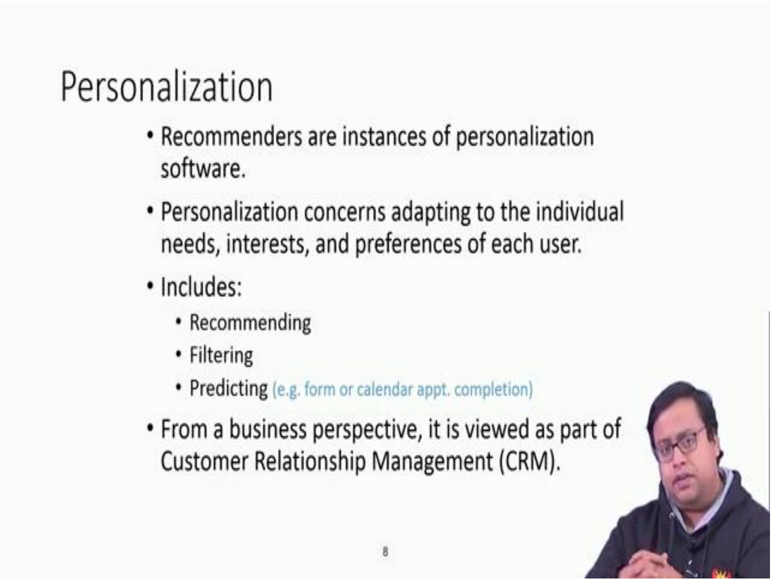
3. **Recommendation Engine Workflow**

Picture from: http://elico.rapid-i.com/recommender-extension

So Amazon has created a recommendation engine where Amazon's model that implements the recommendation algorithm and algorithm feature is most recommendation algorithms start by finding a set of similar customers who purchased and rated items and overlap the user's purchased and rated items. So they find out that whether customer A and customer B are similar. How they are similar?

Based on whether they are purchased the similar kind of items or they have given ratings to similar kind of items. So ratings, why? Because sometimes you might not have purchased in this platform but you might have purchased in some other platform but you have rated means probably you have used that product. So based on that I can give a similarity and the Amazon's item to item collaborative filtering is focusing on finding similar instead of similar items instead of similar customers. So I will talk about that what is collaborative filtering and recommendation engine work flow look like this.

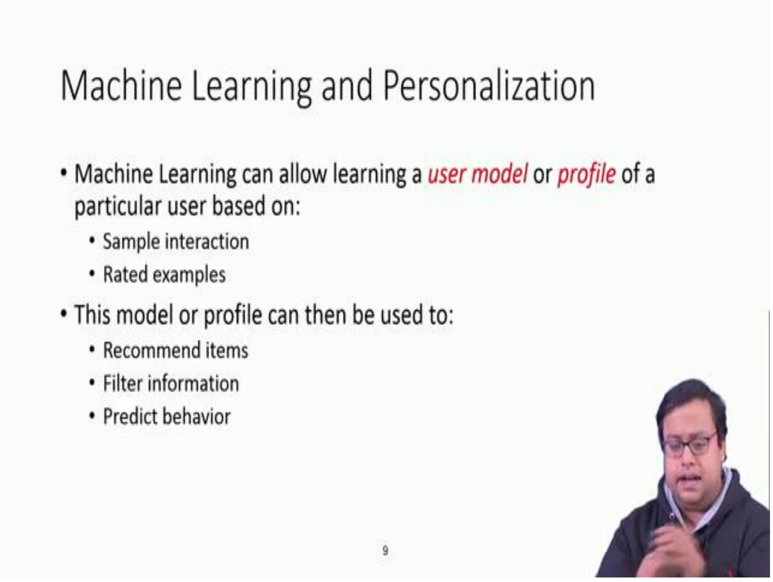
(Refer Slide Time: 23:35)



Personalization

- Recommenders are instances of personalization software.
- Personalization concerns adapting to the individual needs, interests, and preferences of each user.
- Includes:
  - Recommending
  - Filtering
  - Predicting (e.g. form or calendar appt. completion)
- From a business perspective, it is viewed as part of Customer Relationship Management (CRM).

8



Machine Learning and Personalization

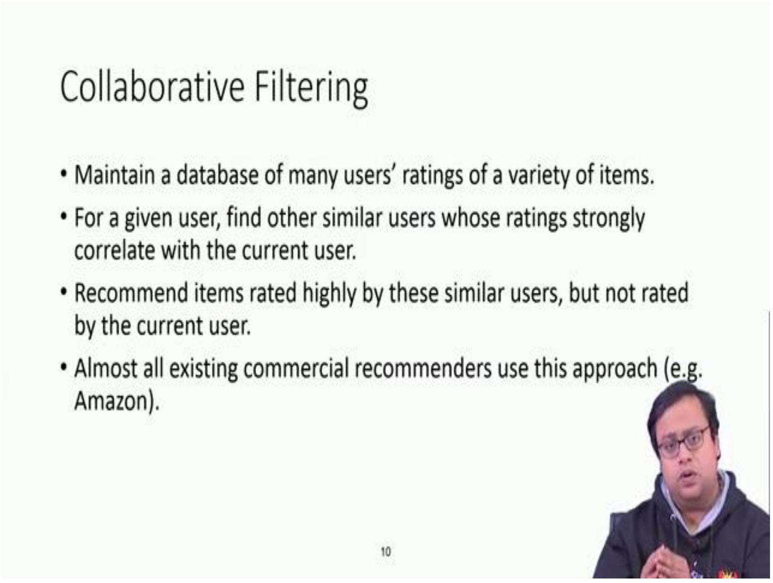
- Machine Learning can allow learning a *user model* or *profile* of a particular user based on:
  - Sample interaction
  - Rated examples
- This model or profile can then be used to:
  - Recommend items
  - Filter information
  - Predict behavior

9

The major focus here is personalization. So what is personalization? Recommendations are instances of personalization software so basically this is in a whole world of retail, retail personalization is a topic under which this recommendation comes. Personalization concerns adapting the individual needs, interests and preferences of each user. So you actually try to adapt with customer's needs and some of the areas are recommending, filtering or predicting.

So recommending is something that we are focusing on. In the broad domain of marketing, it comes under CRM and when we focus a little bit more, machine learning can allow learning a user model a profile of a particular user based. On the other hand, the model or profile can be used to recommend items, filter information's or predict customer behavior.

(Refer Slide Time: 24:25)



Collaborative Filtering

- Maintain a database of many users' ratings of a variety of items.
- For a given user, find other similar users whose ratings strongly correlate with the current user.
- Recommend items rated highly by these similar users, but not rated by the current user.
- Almost all existing commercial recommenders use this approach (e.g. Amazon).

10

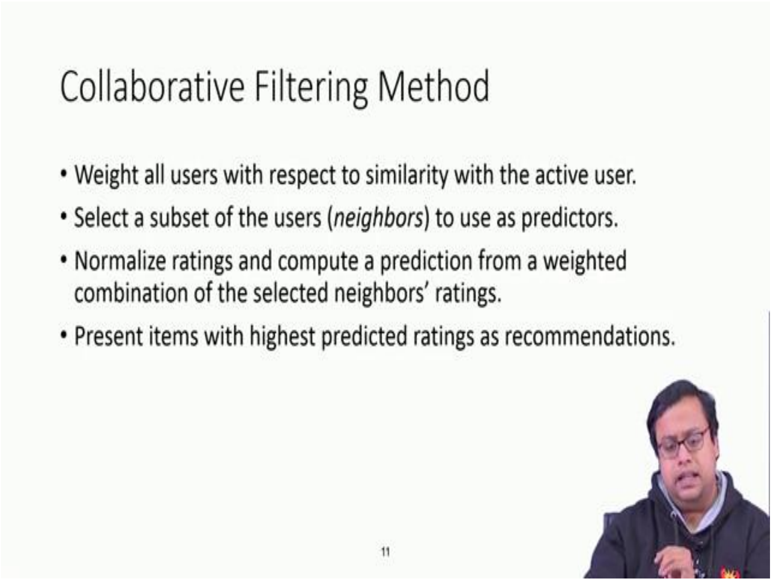
One of the major thing that we use, I told that Amazon use collaborative filtering. So what is collaborative filtering? Collaborative filtering requires maintaining a dataset of many users' ratings of variety items, sometimes ratings sometimes purchased decisions. From a given user find out similar users through those ratings.

So you have to find out which two users are similar and recommend items rated highly by these similar users but not rated by the current user. So I will recommend such item which I have not used till now means I will be recommended such items which I have not used till now or not rated till now but similar customers who are similar to me have rated it pretty high. So for example, let us say I and one of my friend both like Shahrukh Khan and both have seen Kuch Kuch Hota Hai, Kal ho na ho, DDLJ and etcetera. These guy has seen let us say Main hoon na and I have not seen it.

So when you see that okay, so these three four movies I have this guy have seen and this particular person have seen Main hoon na and I did not see it. He will recommend it to me because both me and that person is similar depending our taste. How that comes up? Based on our past purchase history of Shahrukh Khan movies.

Now I have not seen a Shahrukh Khan movie, he has seen a Shahrukh Khan movie, let us recommend it to him. So that is the basic crux. Almost all existing commercial recommender using this approach. So this is the most common approach and that is why we will majorly focus on collaborative filtering.

(Refer Slide Time: 26:07)



Collaborative Filtering Method

- Weight all users with respect to similarity with the active user.
- Select a subset of the users (*neighbors*) to use as predictors.
- Normalize ratings and compute a prediction from a weighted combination of the selected neighbors' ratings.
- Present items with highest predicted ratings as recommendations.

11

Now weight all users with respect to similarity with the active user. So you have to find out that, so I will if there are 4 guys who are similar to me whose recommendation whose behavior will be most matched to me? So after the calculation of similarity, we should also find out that who is active and who is not active. We will majorly focus on the active user's purchase behavior and based on that you will recommend somebody else than by inactive users and select a subset of users because that we will discuss later. Normalize ratings and compute a prediction from a weighted combination, for example, let us say do you think that all when you calculate distance between me and some other user, that distance is measured based on as I told the purchase ratings or purchase behaviors.

Now do you think that all purchase behavior should be similarly treated? For example, let us say you are thinking about movie recommendations. In Netflix, you are talking about movie recommendations. Now he also watches one series, I also watched one series one-one but he has watched pretty high and I watched pretty low. So in a series he watched let us say Grey's Anatomy is a series. He watched all the seasons, I watched 5 seasons and did not watch the rest of the seasons.

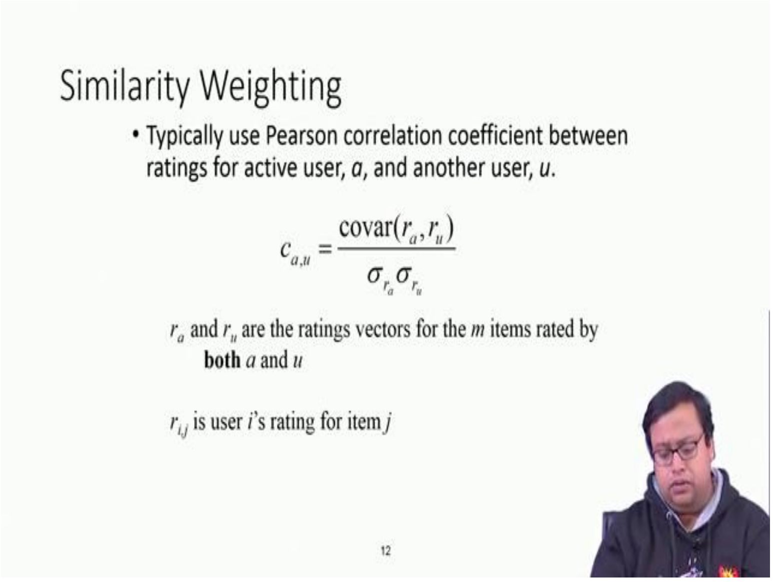
On the other hand, in case of movies, he watched the full movie, I watched also the full movie. So in first case it is in movies case it is both is 1, in Grey's Anatomy case also both is 1 because I have not considered the how much I have watched. Should we give same weightage to these two cases when we are discussing similarity? So I do not know. So that is something that is a difficult question.

So when you measure similarity, forget about purchase behavior’s data. Let us say two persons you have demographic data about them. Let us say I and one of my friend I am a male that person is also a male and I am a Bengali and that person is let us say an Odia, let us assume versus another of my friend I am male she is female but I am Bengali, she is also Bengali. In what kind of situation which attribute the gender versus your native which attribute becomes more important?

When it is food items probably, let us say Bengali foods, probably Odia foods are equally probably very close Bengali foods but still I will say and if it is food item, probably our food habits match based on the native. So native will get more importance. But let us say if it is apparel then probably the gender matching is more important rather than the so the male of Bengal and male of Odisha will probably wear similar kind of products in comparison to male of Bengal and female of Bengal. So all I am trying to say that the which attribute you give more weightage depends on the situation, the product purchase situation that you are talking about.

So present items, so normalize ratings and compute a prediction from a weighted combination of selected neighbor's ratings and present items with highest predicted ratings as recommendations. So this is something that we are trying to do.

(Refer Slide Time: 29:34)



Similarity Weighting

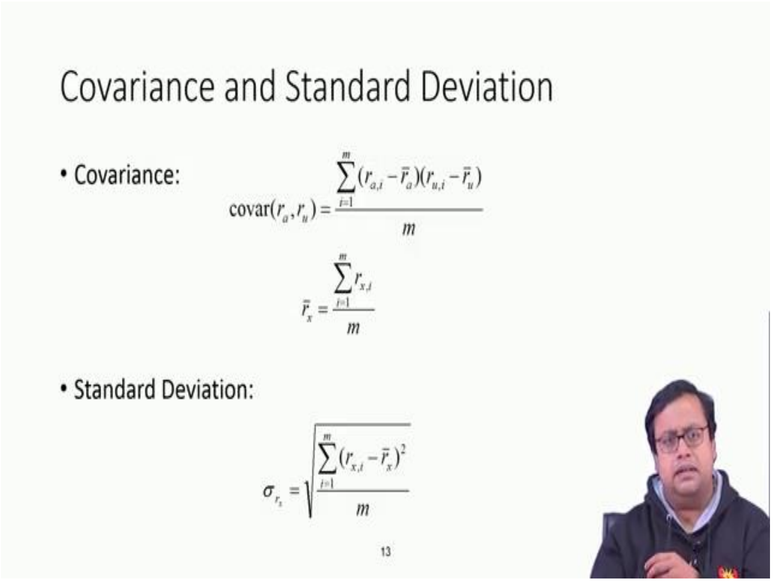
- Typically use Pearson correlation coefficient between ratings for active user, *a*, and another user, *u*.

$$c_{a,u} = \frac{\text{covar}(r_a, r_u)}{\sigma_{r_a}\sigma_{r_u}}$$

r_a and r_u are the ratings vectors for the m items rated by **both** a and u

$r_{i,j}$ is user i’s rating for item j

12



Covariance and Standard Deviation

- Covariance:

$$\text{covar}(r_a, r_u) = \frac{\sum_{i=1}^{m}(r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{m}$$

$$\bar{r}_x = \frac{\sum_{i=1}^{m} r_{x,i}}{m}$$

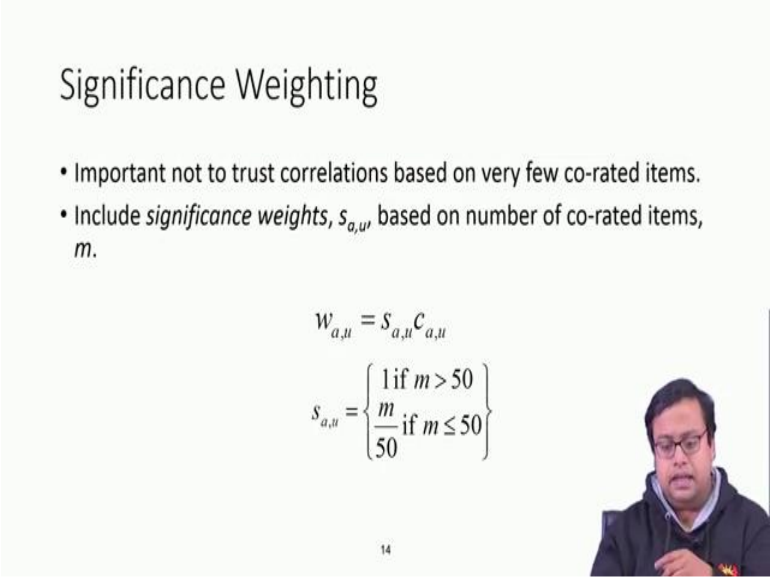
- Standard Deviation:

$$\sigma_{r_x} = \sqrt{\frac{\sum_{i=1}^{m}(r_{x,i} - \bar{r}_x)^2}{m}}$$

13

Now how to find out the similarity? There are different ways. One is correlation. This is a very simple formula of correlation. So Pearson correlation you can find out. We also use cosine covariance, standard deviations. These are also measures. So these measures will be used to calculate the correlation. I will not go into details. You know what is covariance, what is standard deviation.

(Refer Slide Time: 30:02)



Significance Weighting

- Important not to trust correlations based on very few co-rated items.
- Include *significance weights*, $s_{a,u}$, based on number of co-rated items, *m*.

$$w_{a,u} = s_{a,u} c_{a,u}$$

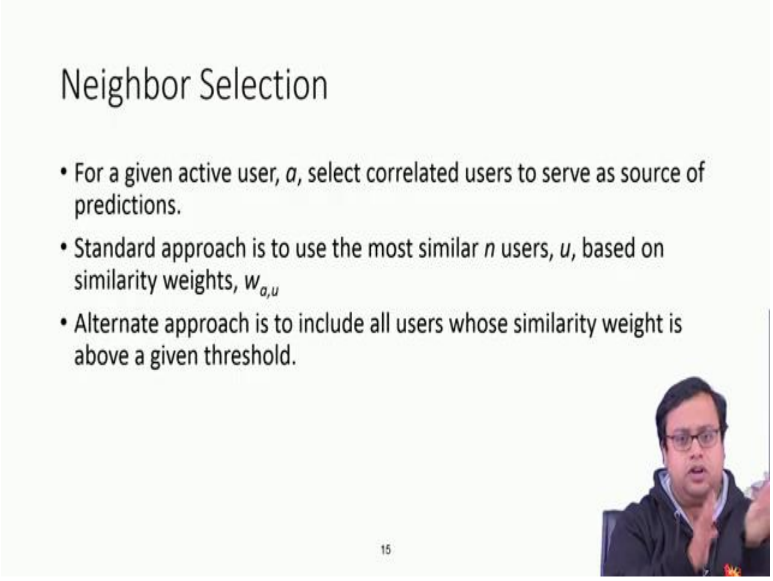
$$s_{a,u} = \begin{Bmatrix} 1 \text{ if } m > 50 \\ \frac{m}{50} \text{ if } m \leq 50 \end{Bmatrix}$$

14

And then there are other measures like cosine matrix we create. Cosine is also a measure. We will talk about that in a different video and then we created a weighted sum. So the significance weight is S a,u and C a,u is the correlation value and W a,u is S a,u into C a,u.

W a,u =(S a,u ×C a,u) Now what is S a,u? S a,u is 1 when m is greater than (>)50 and m by 50(m/50) if m is smaller than(<) 50. So something like that is what we are trying to create.

(Refer Slide Time: 30:32)



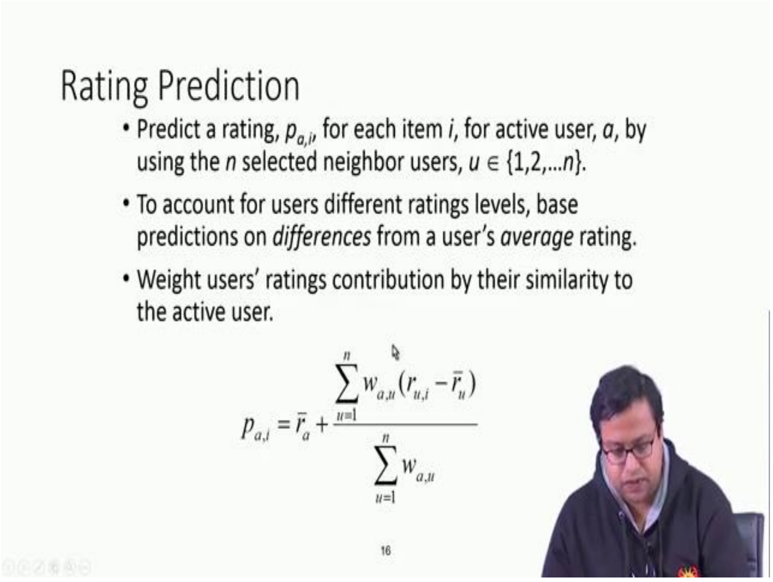
Neighbor Selection

- For a given active user, *a*, select correlated users to serve as source of predictions.
- Standard approach is to use the most similar *n* users, *u*, based on similarity weights, $w_{a,u}$
- Alternate approach is to include all users whose similarity weight is above a given threshold.

15

And for a given active user a, select correlated users to serve as source predictions, that is number one. Standard approach is to use the most similar n users, most similar users based on similarity matrix W a,u. So whoever is the most similar you find out those people. Alternate approach is to include all users whose similarity weight is above a given threshold. So if that similarity W a,u is above certain cut-off, you consider them or you consider the top 50, top 100 whatever.

(Refer Slide Time: 31:12)



Rating Prediction

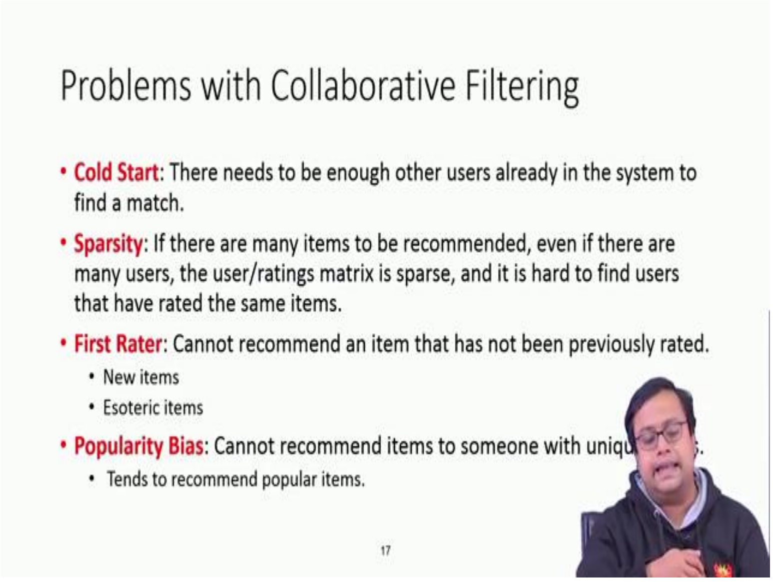
- Predict a rating, $p_{a,i}$, for each item *i*, for active user, *a*, by using the *n* selected neighbor users, $u \in \{1,2,\ldots n\}$.
- To account for users different ratings levels, base predictions on *differences* from a user's *average* rating.
- Weight users' ratings contribution by their similarity to the active user.

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} w_{a,u}(r_{u,i} - \bar{r}_u)}{\sum_{u=1}^{n} w_{a,u}}$$

16

Now in case of rating prediction, how do you predict the rating? Your prediction rating is basically ra + the weighted sum, this is the distance, you will see this is the distance of your rating with the all the average ratings and then × the weight divided by the summation of weights. This is the formula that we use for rating prediction and there are other collaborative filtering methods for other kind of problems. We will talk about that.

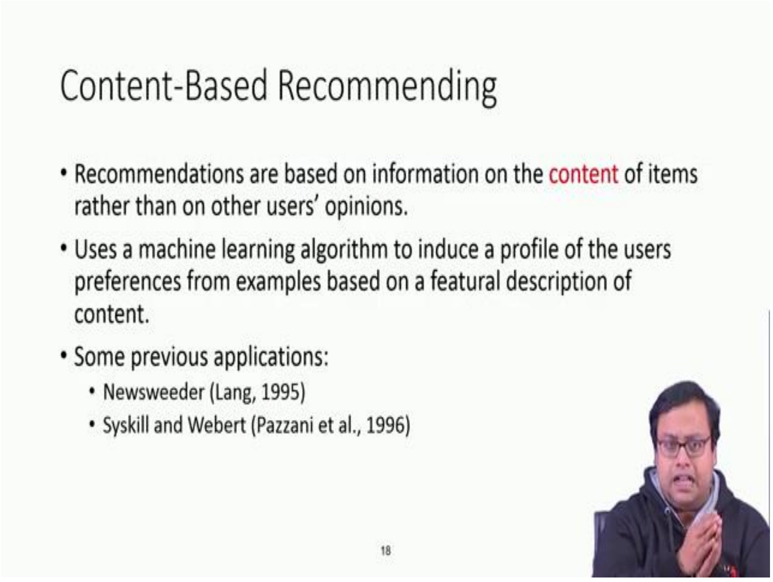
(Refer Slide Time: 31:38)



Problems with Collaborative Filtering

- **Cold Start**: There needs to be enough other users already in the system to find a match.
- **Sparsity**: If there are many items to be recommended, even if there are many users, the user/ratings matrix is sparse, and it is hard to find users that have rated the same items.
- **First Rater**: Cannot recommend an item that has not been previously rated.
  - New items
  - Esoteric items
- **Popularity Bias**: Cannot recommend items to someone with uniqu .
  - Tends to recommend popular items.

17



Content-Based Recommending

- Recommendations are based on information on the content of items rather than on other users' opinions.
- Uses a machine learning algorithm to induce a profile of the users preferences from examples based on a featural description of content.
- Some previous applications:
  - Newsweeder (Lang, 1995)
  - Syskill and Webert (Pazzani et al., 1996)

18

The basic problems are same for all cases. There is a cold start that means there needs to be enough other users already in the system to find a match. Sparsity is also a problem. If there are many items to be recommended, even if there are many users, the user rating matrix is sparse and it is hard to find users that are rated same items. So that is so lots of zeroes will come in the matrix. First rater cannot recommend an item that has not been previously rated. If one item is very new, it is not only the customer is new will be create a problem even the item is new will create a problem.

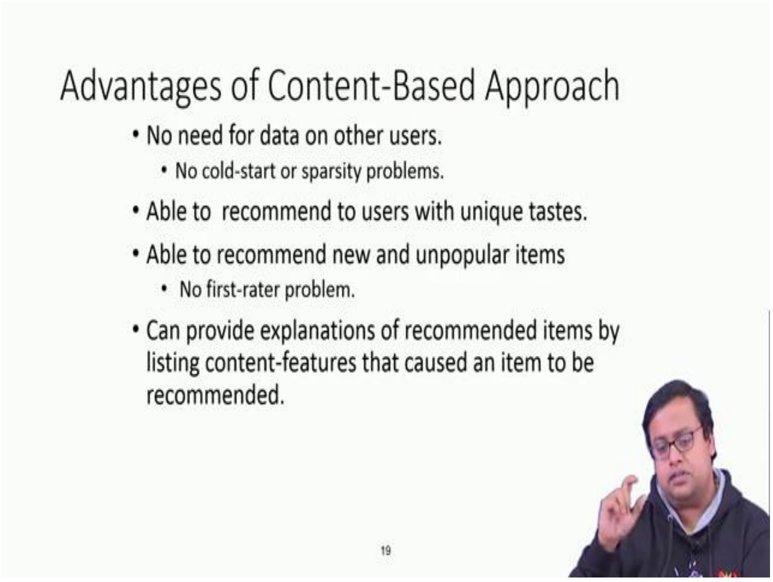
So if the item is new, nobody has rated nobody has seen it, I do not know whom to recommend. So that creates sometimes a problem and we will try to solve that and popularity bias is cannot recommend items to someone with unique tastes. So certain niche products you cannot recommend. So that is also a problem. To solve the first problem so to solve this problem where the first rater, sometimes we do not believe on customer's purchased data or rating data. We believe on the content of the product.

So let us say a new movie of Shahrukh Khan has come up. I do not know what is the quality of movie and etcetera, who has seen, who has not seen, whether those who have seen are similar to me or not? I do not know those details because this particular movie is new in the Netflix platform or whatever platform and then I do not have past purchased data for that particular item. But what I will do is I will find out which of the products I have similar content.

Now content can be the genre of the movie, can be the time period of the movie, length of the movie I would say, can be the how many heroes, how many heroines, who is the major figure, who is the director. All of these details if you match then that is a content based recommendation. So recommendations are based on information of the content of items rather than other's opinions. So not based on product ratings, not based on customer purchased data but based on the content.

Uses a machine learning algorithm to induce a profile of the user’s preferences from example based on feature description of the content so this is straightforward. Some previous examples are newspaper, newspaper items or let us say Syskill and Webert, you can check these papers.

(Refer Slide Time: 34:04)



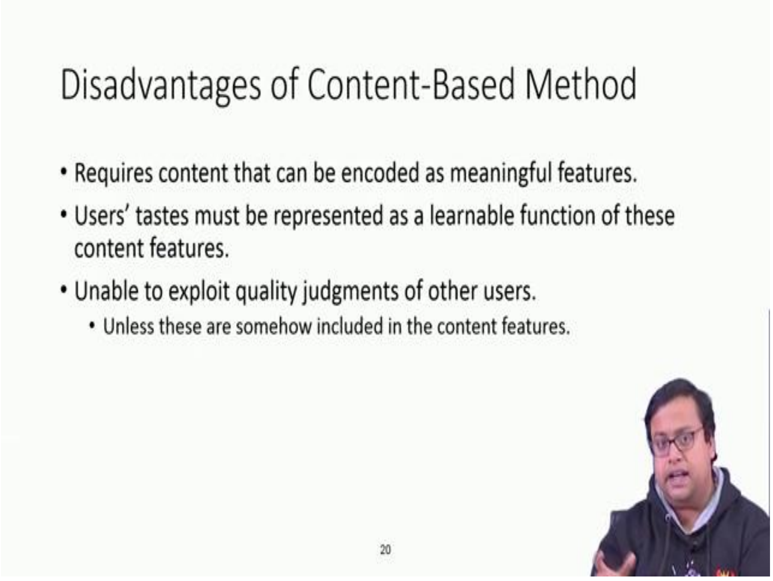
Advantages of Content-Based Approach

- No need for data on other users.
  - No cold-start or sparsity problems.
- Able to recommend to users with unique tastes.
- Able to recommend new and unpopular items
  - No first-rater problem.
- Can provide explanations of recommended items by listing content-features that caused an item to be recommended.

19

What are the advantages? No need of data of other users. So whatever was the disadvantage of collaborative filtering becomes an advantage here. Able to recommend to users with unique tastes. So we will solve one collaborative filter one content based model also. Able to recommend new and unpopular items also. So no first rater problem is there and can provide explanations of recommended items. So there it is not like somebody else bought that is why you bought. Which particular feature is actually contributing to this recommendation of this product to you can also be tracked. So that is content based.

(Refer Slide Time: 34:42)



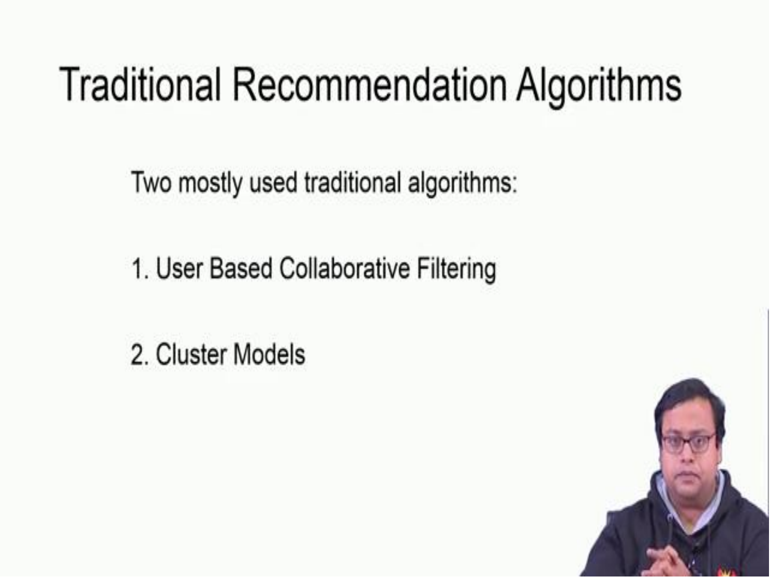
Disadvantages of Content-Based Method

- Requires content that can be encoded as meaningful features.
- Users' tastes must be represented as a learnable function of these content features.
- Unable to exploit quality judgments of other users.
  - Unless these are somehow included in the content features.

20

What is the disadvantage? Requires content that can be encoded as a meaningful feature. User’s so not all content is can be converted to a so it is a mathematical calculation. It has to be quantitative. So not all content can be converted to quantitative values. Users’ tastes must be represented as a learnable function of these content features and unable to exploit quality judgements of other users. So the judgement the ratings is not considered here which is also a disadvantage.

(Refer Slide Time: 35:10)



Traditional Recommendation Algorithms

Two mostly used traditional algorithms:

1. User Based Collaborative Filtering

2. Cluster Models

So majorly, two recommendation engines that we use, collaborative filtering and clusters method. In the next part of the slide, we will discuss about these two methods in a little bit depth. I will see you in the next video in terms of discussing about user based collaborative filtering. Thank you.