Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 05
Lecture - 21

Lecture 21: Introduction to content based recommender system: Foundations

Hello everyone. Today we are going to start a new approach for recommendation. So, this approach is called content based recommendation. And today in fact, for this week we are going to talk about few concepts which build the foundation of such recommender systems. And today is the introductory lecture in this regard. So, this is my content.

So, today basically we will be introducing the topic. So, this is a formal definition of content based recommender system. This system utilizes the content ah contents item contents to build model. So, the systems implementing a content based recommender approach analyze a set of documents or descriptions of the items.

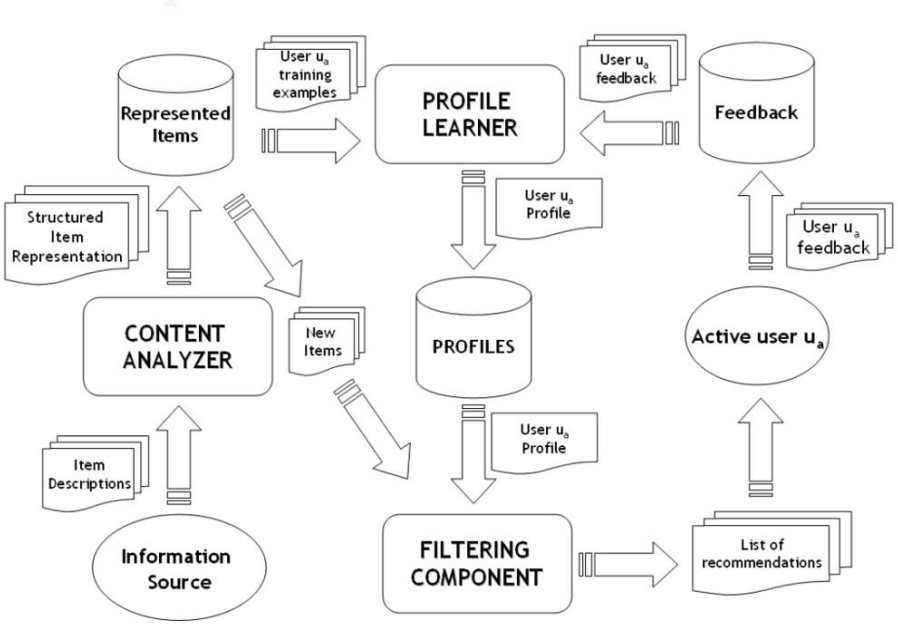
Why documents? Because mostly it is done with respect to document recommendation specifically news recommendation. Therefore, the term documents comes. Documents and descriptions of items previously rated by the user and build a model or profile of users user interests based on the features of the object rated by him or her. Now what is profile? The profile is a structural representation of the users interest. And what is recommend how the recommendation process takes place? By matching the attributes of the user profile against the attributes of the content of the object to predict the users interest level.

Now if a profile accurately reflects users performance, it is of tremendous advantage for the effectiveness of an information access process. So, for example, if you are you keep on watching comedy movies probably it is very probably that you will also watch the comedy movie. Sometimes the things may go wrong for example, in a online site you it may so happen that you are purchasing clothes, but it is not that all the time you will be purchasing the clothes you may purchase grocery items as well. So, it has its relative advantages and disadvantages as we are going to see. But before we move ahead and see its relative advantages and disadvantages these are the tasks to be performed in content based recommender system.

So, as we move ahead we will be going looking at the details of this, but here let us try figuring out what all happen. You have something called content analyzer, then you have something called profile learner, then filtering component these are the major components of this content based recommender system. In case of content analyzer what does it do? It takes the item descriptions and item descriptions are basically the they

come in the unstructured form. For example, if you are looking at a text document let us say this is my text document high level architecture of content based system this is my text document. So, do to develop any model do any kind of analysis I need a numerical numerical representation.

Because the models the machine learning supervised machine learning models are going to take inputs not in forms of text, but in terms of certain numeric values. Of course, in today's world there are the examples of deep learning models who take the text itself derive the features and based on the features to the prediction. But we will not be going to that direction we will be limiting to ourselves where feature extraction is a very important part of content based analysis. So, here you give the text document like this one and derive the derive the numerical representation. So, this structure representation is stored somewhere.



So, now, you have this all these documents like this which were text will be now represented in terms of a number of features. The features can take any numerical values let us say feature 1 for the item 1 feature 1 1 feature 1 2 feature 1 k then for the second item feature 2 1 feature 2 2 feature 2 k. So, similarly all the items in terms of such vectors will be in terms of such vectors will be getting stored here. Next is you have to get the training create the training example for creating the training example you have created the item features how many items all the items probably that is at your hand. So, if you have let us say certain 1 million documents you will have 1 million such features test sets 1 for each item.

Now for learning a user's model what essentially we need along with this feature you need a response variable ok. So, you have to have a response variable. So, for this one

there will be a response variable for this one there will be. So, how many such response variables you will have as many number of ratings available with the user because all the one it is impossible for one user or a group of user if at all you are adopting the idea of clustering these users beforehand and make one generalized model for that group if that is the case it is for the entire group otherwise it is for one individual. And it is very difficult that one individual has rated large number of items.

In fact, they will be rating very few may be few 100 plus in their lifetime. So, in that case if we have that less number of rating out of all the item features that you have derived take those item features for which a rating is available and that rating can become put a put together along with this can make a training data set for supervised learning. Now, this user profile learner makes use of this data this training data which is derived from the item feature and the users response taken together it can build one user profile. So, what is an user profile in this sense? It is a set of model parameter let us say for this learning you are using a neural network in that case or a regression model may be. So, in that case those regression coefficients will become the user profile and you keep these coefficients in the database this is the profile database when a new item comes it is features is going to be fed to that particular model ok.

Let us say for simply simplicity let us see it is only regression model. So, to the regression model you will be feeding this data as the input which data the new item which comes in. So, once you feed this data what is going to be the output a response variable. So, let us say your response variables were in terms of yes and no. So, either yes is 1 and no is 0.

So, some value will come up 1 or 0 ok. Let us say you are using a some kind of logistic regression or something. So, or even ordinary regression you have some kind of some kind of limit beyond which you consider it as 1 below which you consider as 0 you can adopt many methods you like, but this is just one example. And once you predict it is rating then for multiple items which are to be at your disposal let us say some 100 news articles have now come. On your smart phone how many will be shown not all top few will be shown.

So, which top for which the predicted rating is supposed to be very high and in case we have a binary rating kind of thing. We have to again decide some mechanism based on which even if the rating is binary you get the top few ok. So, those top few items will be displayed to the user. So, once you display those top few items to the user what the user does? User will either click on the news article or it it it would not. So, some new user feedback will come.

So, which means some of the recommended items user has not seen. So, we can say these are negative examples and there will be some positive examples as well. So, now,

you take this feedback give this to the profile learner update the profiles and you have your new profiles ready ok. So, with this basic idea now let us move ahead. So, these are the relative advantages and disadvantages.

So, it first advantage is it is not influenced by what other users are doing. Users rating is provided by himself to build his own profile. So, therefore, it is independent of other users. It is very explainable why something is happening unless otherwise you make your feature set certain certain make it very compact by using some kind of dimensionality reduction technique or selecting only few features and so on. But more or less it is very transparent and when a new item comes this cold start problem which would have been there in case of user user based because in user user based if you do not you if you get a new item nobody no rating is available for that then predicting rating is very difficult.

But here that new item cold start problem is over. So, it is capable of recommending items which are not yet rated by any users ok. Disadvantages, what are the disadvantages? Limited content analysis. The text document typical text document will have many features out of that you will be keeping only few. So, which means you are providing limited amount of information to your profile learner.

So, naturally it will be limited to whatever you give it give whatever features you give it whatever information you give it to the. So, the bottle may not be very accurate because of this lack of I mean the dropping of CPU features. Sometimes in fact, all the time it is over specialized for example, if you are giving high rating to comedy. So, you will always be getting comedy movies. So, there is so, this surprise factors like serendipity diversity etcetera are very low in this case.

Now, when a new user comes there is no training data set because for training data set you need user rating. So, it is not it cannot solve the whole stat user problem when the when a new user comes. So, with this we are moving ahead. So, these are the phases which we have already seen content analysis, content presentation, content based learning, user profile learning, then filtering and recommendation generation to. So, these are the phases as we have seen in this earlier diagram ok.

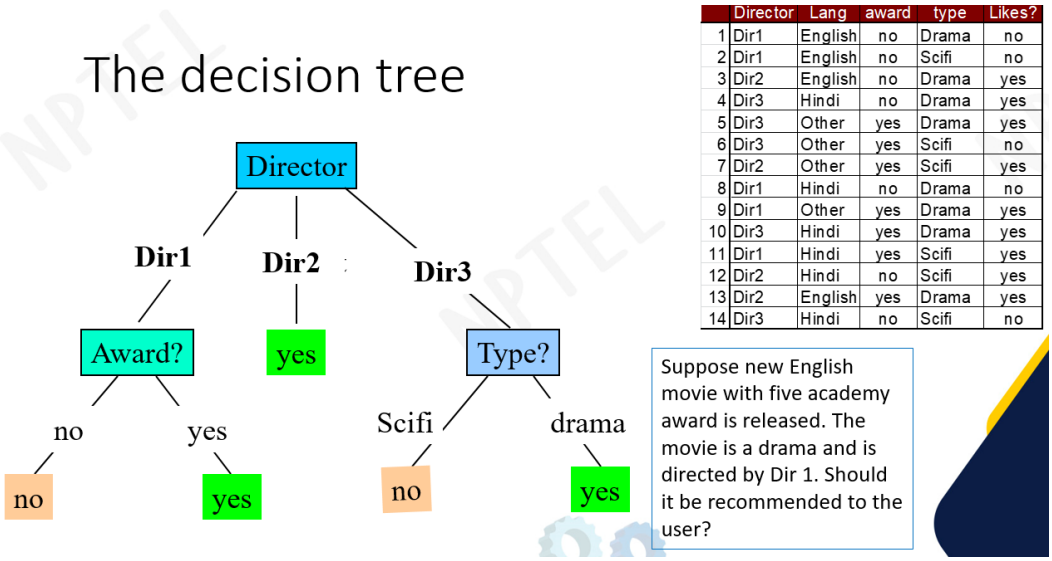
So, now one after the other we are going to look at this first one is content analysis. So, here this content analysis includes feature extraction and feature selection and typically dimensionality reduction as well. So, in this when we talk about the feature, feature can be either explicitly specified or extrinsic features for example, for movies this can be the features it is category the genre some ratings ok. Then awards, length, origin, director and so on this can be features which are available on a let us say on the movie or if you go to IMDB such IMDB site you will find those details. Next is feature extraction, it is not that everything about the item is explicitly stated.

Sometimes there will be summary that summary from that summary you can extract few keywords and if even if your genre says it is a comedy movie probably through this features you will be getting certain additional information about the movie. Let us say it belongs to a specific age group or something. It may so happen that just like in IMDB you have seen there will be many user viewer feedbacks. So, those viewer feedbacks probably you can extract certain text features and using use them. So, these techniques used for extrinsic features could be for selection of this extrinsic features could be no no what I want to say is that the number of features described in this manner may be very large specifically when it is text document.

	Director	Lang	award	type	Likes?
1	Dir1	English	no	Drama	no
2	Dir1	English	no	Scifi	no
3	Dir2	English	no	Drama	yes
4	Dir3	Hindi	no	Drama	yes
5	Dir3	Other	yes	Drama	yes
6	Dir3	Other	yes	Scifi	no
7	Dir2	Other	yes	Scifi	yes
8	Dir1	Hindi	no	Drama	no
9	Dir1	Other	yes	Drama	yes
10	Dir3	Hindi	yes	Drama	yes
11	Dir1	Hindi	yes	Scifi	yes
12	Dir2	Hindi	no	Scifi	yes
13	Dir2	English	yes	Drama	yes
14	Dir3	Hindi	no	Scifi	no

So, therefore, for choosing the minimum number of features you need some kind of feature selection procedure. So, for this either you can use some certain statistical analysis, you can use decision tree and so on. So, these are few examples as we move move ahead we will see few other techniques as well. So, all the steps that are involved we are now elaborating. So, the second steps is the sorry the first step is your feature is extraction and selection.

So, as I have already told you the text data has to be now converted to numeric form. So, when we convert it to numeric form either you can treat them as binary features or you can derive something called tf and idf rating and use them together. So, those things we are going to see shortly I mean the in next couple of lectures. Now representation what is this data representation? Now each item in the training data sets is labeled to indicate the preference as I told you already.



	Director	Lang	award	type	Likes?
1	Dir1	English	no	Drama	no
2	Dir1	English	no	Scifi	no
3	Dir2	English	no	Drama	yes
4	Dir3	Hindi	no	Drama	yes
5	Dir3	Other	yes	Drama	yes
6	Dir3	Other	yes	Scifi	no
7	Dir2	Other	yes	Scifi	yes
8	Dir1	Hindi	no	Drama	no
9	Dir1	Other	yes	Drama	yes
10	Dir3	Hindi	yes	Drama	yes
11	Dir1	Hindi	yes	Scifi	yes
12	Dir2	Hindi	no	Scifi	yes
13	Dir2	English	yes	Drama	yes
14	Dir3	Hindi	no	Scifi	no

So, you have the item features. So, from where it is coming this is from item description along with this you add some response variable user response variable or its users input which comes from the rating matrix. So, using this for all the available ratings you build your training data set ok. So, that is what it is saying independent variables can take the values binary continuous you can decide the right kind of scale. Dependent variables are ratings by a particular user or it can be a group if you decide to cluster similar users and make a model. So, it is your choice whether it is for a particular user or if you have enough ratings then particular user is always better because that user is going to talk about his own features is going to give better you can get better understanding of his choice by looking at his data, but if it is a group it is a group choice.

Next is user profile learning as I told you once your training data is ready training data is ready use any method of your choice. So, some of these methods could be multiple linear regression model, decision tree induction algorithm, back propagation or neural network based methods, support vector machines and so on. So, once you have the model trained then model parameters will be stored where certain database. Let us see one example situation here. So, this is a movie data available from one user.

So, these 3 4 independent variables are there director, language, award and type and one dependent whether the person likes the movie or not. So, basically there are 3 directors some 3 different categories of movies, award and so on. Now using this as your training example what are you going to do? You are going to build a model. How to build a model? Certain introduction we are given while talking about the supervised and unsupervised machine learning algorithms. However, when we look at specific methods few such methods we are going to elaborate further.

This is one example of application of this data to a methodology. This methodology which is called decision tree has some of the internal nodes and some of the leaf nodes

ok. So, now suppose this decision tree is made from this particular data. Now when the decision tree is how this decision tree is made we are least concerned right now. We will be talking about it we will deal about it how to make one such decision tree at a later stage.

But right now at least for observation purpose we can see that if the director is director 2 the choice is yes ok. What are the rows corresponding to director 2? This is 1, then this is second, this is third any other director 2 director 2 this is also yes ok. So, there were 4 examples for director 2 everywhere it is yes. Now for director 1 director 1 there are so, with this we now broke it into 3 parts all the movies belonging to director 1 are here all the movies belonging to director 3 are here and director 2 are here. So, if the director if the person if the director is director 2 person is definitely going to see the movie.

But if it is director 1 next variable to be considered is award. So, if the award is yes definitely he is going to watch the movie if it is no he is not interested. So, considering the fact that this is our tree and this is the optimal tree possible we will be discussing the algorithm in few in the in some of the other class. Now let us see if a new English movie with 5 academic award is released and if the movie is a drama and is directed by director 1 should it be recommended to the user we have to now travel through the tree ok. So, what is there in the root director? So, director is 1 in our case.

So, we move here what is the next variable to be considered? Award. So, it has 5 academic awards it has 5 academic award yes. So, he is going to watch the movie. Now what happens if you find that he is not watching the movie you generate 1 more data set here that 15 number of data that he has not watched movies even if it is a drama there is yes and it is Hindi no no no it is English and this is director 1.

So, created 1 negative example. So, as a result your model will not be this tree this model will be changing ok. So, once you change the model accuracy will be will be better. So, therefore, as we get more such feedback our accuracy is going to keep changing. So, with this example we stop the introduction to recommender system and we the there is a decision tree concept has been taken from here and all these 2 books I am using as reference. In fact, the process is from recommender content based recommender system process is from the first book.

So, with this we stop this lecture and these are our concluding remarks. We understood that content based recommender systems use item features and rating matrix to build user models. So, these user models are utilized for making the predictions. The phases in the contents of the recommender system are content analysis, feature extraction and selection, content presentation, content based learning of user profile and finally, filtering and recommendation generation. Thank you.