

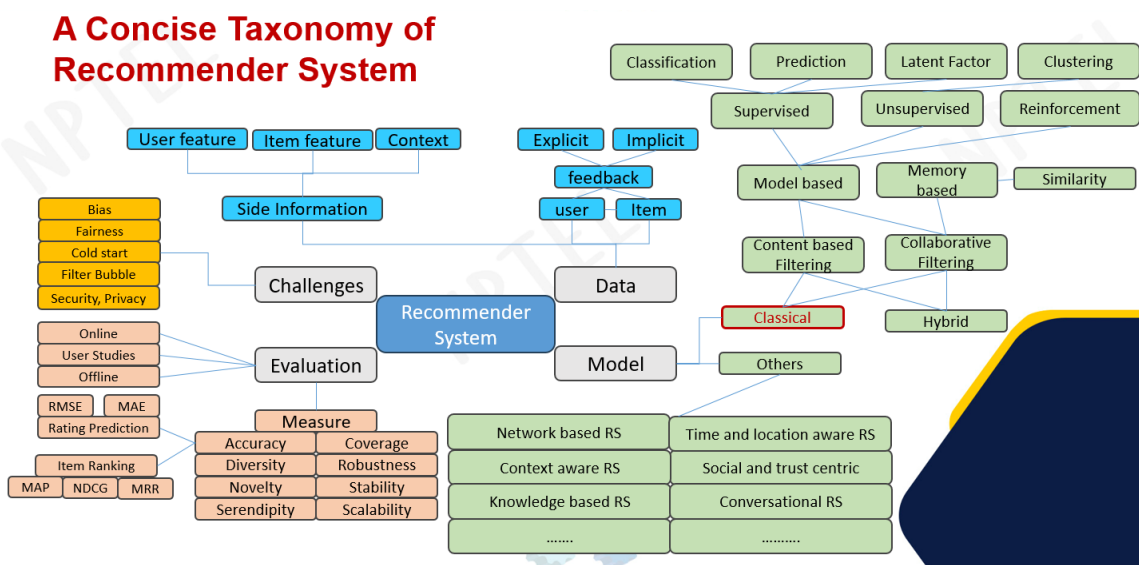
Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 01
Lecture - 04

Lecture 4: Data Collection

Hello everyone. We are again continuing with introduction and specifically we will be talking about the data collection in recommender system setting. Moving ahead these are the concepts which we are going to cover data collection strategy, measurement scale and understanding the nature of the data. Then I come back to this slide. Here we are going to cover this part in detail. In last lecture we also discussed little about little bit about the data, but how to collect this data and how to represent this data we did not discuss on this.

So, we are going to deal with this in a bit more detail. This is again that old figure and we have two strategies to collect data. They say if you do not have right kind of data even if your algorithm is good there will be garbage in and there will be garbage out. So, the data is extremely important.

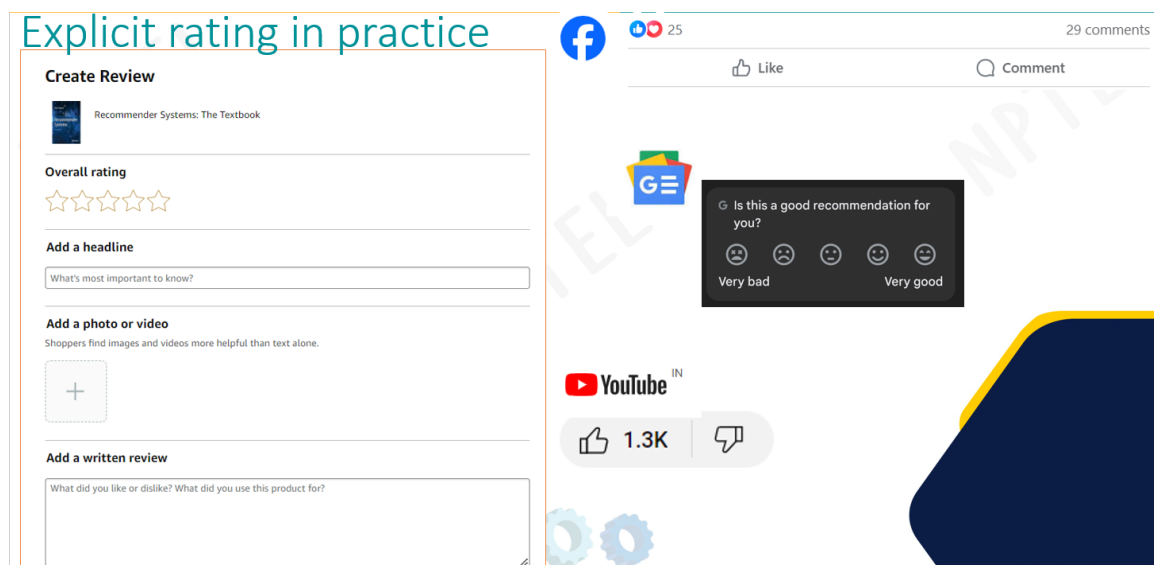
So, how do we get this data in case of recommender system? The first one is by asking directly by asking questions. Second one is by interacting indirectly by observing ok. So, which data we are going to collect? These are the three basic data that we discussed besides that the side information if we get that also can be explicit or implicit. So, coming to this slide do you think all the items that you have seen in a ah in a typical recommender system have you explicitly erected them? No you did not. Then still you are able to get the recommendation.



In a site like that of your let us say Amazon while searching for a book, suppose you are not a registered user still then also you get some recommendation. So, how exactly this these things are happening? So, you would like to see that. So, this user demographics data when it comes to user demographics data you have to ask the user how many times you have been you are asked and you are giving feedback about yourself never try to remember you are trying to login into some site and that site is asking whether you would like to your use your Google information for that. So, probably they are also collecting some data from Google where you have actually registered yourself. So, if you are giving then you are providing your detail.

So, if we this data need not be coming from one source probably they are collaboratively trying to get the data from different sources, but whatever may be the case you are not directly giving the data. So, implicitly it can be collected. Besides this your activities online can be observed how many times you are looking at which part of the screen and what is being displayed in that screen, what is your exposure level, did you click on the recommended item, did you refer the recommended item. So, all these things are marked and because all these things are marked some numeric value can be extracted for this. So, with this idea let us move ahead look at this explicit data.

So, here the users are directly asked. So, Amazon's star rating system, YouTube's thumbs up thumbs down sometimes they provide also provide some kind of graphical bar you move on that sliding bar and select one continuous value. But most of the time the data is collected in a something called a Likert scale. This scale is named after some after a very famous American psychologist Rhenis Likert. So, this in this scale this scale is typically an odd scale where for example, in this 7-star scale you strongly disagree, disagree, slightly disagree, undecided, slightly agree, agree, strongly agree.



So, these numbers 1, 2, 3 etcetera they basically indicate your level of agreeing or disagreeing. Think of Amazon's 5 star rating scale. So, if you give a star what it indicates Amazon provides this detail and while filling up you may be looking at this and you fill up certain value. So, now this data we understood can be in continuous or in sometimes in in in in case of non-numeric

form. Non-numeric in the sense it if it is in star rating as showing this thumbs up or thumbs down it has to be now brought into numeric form for processing in the algorithm.

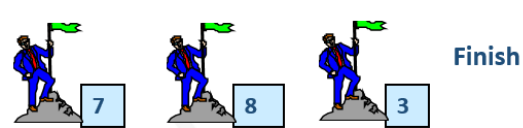



So, these are the problems with collecting the explicit data. First some additional effort is required by the user. Most of the times users are not ready. So, they may not be willing and they may not want to put this additional effort. Personal biases while expressing their preferences.

What is personal bias? If I find everybody is liking certain movie and I do not like it personally probably I will think that probably everybody is liking and I have not seen it properly. So, when I give rating I will be influenced by other people. So, I can so, I got biased by others. Similarly, I can be biased. Today I am not in a good mood.

I am still I am watching some kind of comedy movie and I am not very happy. So, I give a very low rating. So, this can be my inconsistency. So, this bias can be either because of noise due to personal inconsistency or by the preference given by other users. Besides this rating systems are often prone to attack.

So, what do you mean by they are prone to attack? By prone to attack we mean they will be misused by the commercial entity who whose item you are trying to rate or by his competitor. The commercial entity whose item is being rated can use certain spurious raters and they can give very high rating. So, as a result the rating increase. So, similarly what the competitor can do? Competitor again can create some spurious users who will be giving bad rating. So, how do we know? This is this data is not under attack.

Primary Scales of Measurement: Example

Scale	Example	
Nominal	Numbers Assigned to Runners	
Ordinal	Rank Order of Winners	
Interval	Performance Rating on a 0 to 10 Scale	
Ratio	Time to Finish in Seconds	

This again can be a problem. This is one example how we give rating. This is your Amazons star rating. You also give some kind of text value. Probably based on that text something can be

generated.

So, this is Facebook like, this is YouTube like, there is difference Facebook only allows you to like not to dislike. In case of YouTube you have dislike option as well. This is the data collection strategy by Google news. So, they provide you certain recommendation. Is it very bad or it is very good? This is again a Likert scale even if the faces are appearing. This is a Likert scale. This is a binary scale. This is a binary scale and here both up and down make sense. So, this is a symmetric binary. Here only up making sense.

So, which means if you do not like that not necessarily that means that you. So, you did not press it that dot not necessarily mean that you have actually did not like it. You simply decided not to press like, but you might be liking you never know and nowadays they have come up with all these signs. So, which means you have different types of ways in which you can express your how you prefer. Look at this 5 point you gave 1 in the Likert scale it is 1, 2 rating Likert scale 2, 3, 4, 5.

So, based on this, this is probably you did not like it, this is your neutral, this is you liked it very much. I am a spurious user. I can give some good rating if I am on behalf of the actual person who is marketing this product, who is dealing with this product, whose over is selling this product and if I am from the competitor side I will give on giving this one. That is how I can now make the data which is generated for the recommender system I can make it bias. So, the way out is besides my explicit expression some implicit data looking at my other behaviors should be collected and may be merged with this one.

So, before we start looking ahead let us look at how this implicit data gets collected. It is based on the observable behavior exhibited by the user. This can be categorized into this 4 categories. What are the observable behavior here? You select something, how much time, with what frequency you repeated this particular examination, whether you clicked on this many times, whether you actually made the purchase. Second is retention, whether you saved this reference for future, you selected a movie and kept it for future viewing, selected an item and kept it in your shopping cart, but you have not paid for it.

Are you printing it by chance? Are you referring? How you are referring? From object to object, object, portion to object, object to portion. So, object to object, are you forwarding it to someone, replying, posting, reposting it, you are following up and so on. From that item are you clicking and following another hyperlink? Are you cutting and pasting it somewhere? All your activities are observed. Are you trying to rate it and publish it elsewhere? So, all these things which are very subjective in nature has to be now quantified. So, the quantification is the quantification is the key.

So, how do we quantification? Depends on your domain expertise and depends on companies. The people who are dealing with this on behalf of the company from their perception. So, this data has to be generated based on their perception. So, which is written here, it has to be converted into user preference by this quantification process. The simplest strategy to make the value is 0, if the behavior does not exist and 1 if it exists.

So, it can be a 0 1 value, but let us say duration how much time you have seen it, it can be some kind of continuous value as well. So, once with this understanding of how to get this data, next is how to characterize this data. Look in a typical recommender kind system there are many other ways in you can get exposed to recommender system through other lectures and all. Most of the time it is assumed that you are already acquainted with the nature the statistics and machine learning ideas behind this. This course is unique in the sense I will try to infuse the ideas which I am going to use in future.

Suddenly telling that assuming that you know everything may not be a good idea. So, therefore, some of the things which you already might be knowing I will be telling little bit as a part of my introductory activities. So, characterizing data knowing about different types of machine learning algorithms representing the data etcetera are part of this initiative of mine. So, let us look at characterizing the data. As we know the data that is collected from the user is one random variable and this random variable can be either continuous or it can be discrete.

So, a continuous random variable is something which can take any value in a continuum. It is a time spent by the user on a particular page it is a continuous value. I told you can also give your rating in the form of using a sliding bar. So, there a joke for example, in jester you rate a joke interestingness of a joke it is a continuous value. Any discrete values we saw star rating of Amazon opened thumbs up and thumbs down and so on.

Now, when we look at this nature of all this data are basically different. Different in the sense they are in different measurement scales. This measurement means assigning numbers or other symbols to characterize to characteristics of objects according to certain pre specified rules. Star rating 5 star 1 star number 1 second 2 stars number 2 and so on.

Thumbs up thumbs down certain value. Let us say thumbs up certain you expressing your love or something that are some other value. So, you have to assign numbers to those subjective things. How did you like how much did you like? So, you have to give certain numbers to this subjective thing which is shown to you. Now, one to one correspondence between the numbers and the characteristics are to be measured in this case. The rules for assigning numbers should be standardized and applied uniformly so that algorithms do not produce biased result.

So, this rule of assigning numbers to this subjective thing should not change over the time. So, this scaling involves creating a continuum upon which the measured objects are located. So, the scales determine the type and amount of information contained in the data. Scales indicate the data summary indicates the data summarization and statistical analysis that are most appropriate for the nature of the data. There are basically 4 types of measurement scales.

One is nominal, ordinal, interval and ratio. So, when you describe a scale for you characterize a scale 4 important things you have to keep in mind. Description of the scale by description we mean unique labels or descriptors that are used to designate each value of the scale. All scales have to have certain description. Then order it is the relative size or position of the descriptor. This order is denoted by descriptors such as greater than, less than and so on.

Distance can we find out the absolute difference between the scale descriptors that makes the distance. If we can then the distance is permissible. Is there any origin based on which you are trying to put your scale? So, existence of origin also is a different characteristics of the scale. This is the example of this primary scales of measurement come to the nominal. Here this is in this example there are few runners and you are assigning them.

These runners have some kind of you know number on them first one, second person, third person and so on. Suppose these are the 3 runners who reach the finished line and on their back there is a number. Do you think if we add these numbers does it make any sense? No because they simply are some alternative to the name of that person and there comes the word nominal. It is replacement for the name category category of a name in name of a person name of a category and so on.

Next comes the ordinal. Now look at these 3 persons who are numbered 7, 8 and 3 how did they reach the finish line? The third person reached first, eighth person reached second and seventh person reached third in terms of order. So, the order in which they are reaching decides how they are winning. So, now the ordering makes sense ok. So, now the data is now ordered. So, if I now call the first person 1, second person 2, third person 3 and their original numbers were 3, 7 and 8 that time this 7, 8 and 3 were numbers.

Now this 1, 2 and 3 are the order. If I arrange them that first person is better than the second person and second person is better than the third person it makes sense. In a Likert scale if I give 1 and someone else gives 2, 2 is better than 1 then comes the interval scale. In this in the context of this example this interval scale let us say the first person has some performance rating in some of how well he reached and all if there is some kind of mechanism to determine this performance rating in terms of speed or something. So, if I make this then it is again ordered, but the difference between them let us say the velocity in which they were.

So, the difference actually makes sense now. Then the last one is ratio let us say the time in which they finished. Time which in the finished means the time started with 0, you started with your stop was 0. The first person reached in 13.4, 13 minutes 4 second let us say. So, the difference between this along with a 0 is now becoming relevant.

So, this is called ratio scale. We will be doing more details about them now. So, this is what I have already said. Nominal scale in this scale the numbers serve as tags and labels only to identify or classify an object. So, this measurement normally deals with non-numeric quantitative variables where the numbers have no value. Gender of a user, genre of a movie in context of recommender system are examples of such nominal scale.

Now, permissible descriptive statistics for this. Now what is descriptive statistics we are going to see shortly, but at least you I assume that you know all these basics. So, I will be refreshing your memory after a while. So, here percentage makes sense and mode. What is mode? Mode is highest number of times. Let us say in your movie base you have more number of comedies.

So, you count. So, counting you counted there are 5 comedy movies, there are 2 action movies, there are 1 something some other kind of genre and so on. So, which movie you have in highest number? Comedy movies now look at the how your users are watching the movies are the users watching the comedy movies more. So, more number of users will be the number of users if you take. So, wherever the mode number of users mode is highest the highest number of users watch that is the category where people are preferring. So, here some permissible inferential statistics are also given here I would not be talking about them because this is just for your reference.

So, this is the liquor scale. Here also the feedback rating that we saw in liquor scale makes comes into this category. So, here it relates to order that is how the name is ordinal. Here the numbers indicate the relative position of the object, but not the magnitude of the difference between them. So, if I say I have given once rating 1, I have given rating to some other product 2, some other product 3. If I make the absolute difference between these 3 and 1 and so on, this is just my perception.

This cannot be quantified the difference cannot be quantified. So, typically it is a kind of qualitative data that groups variables into ordered categories. So, this one along with the other two descriptive statistics we saw here percentile and median also makes sense and there are few inferential statistics as well. If we have time possibly at the time of when we discuss about how to find out the quality of algorithms in the form of some kind of matrix, if possible I will try covering inferential statistics a bit, but it is desirable that from some statistics course you know all this. Then comes interval it is a quantitative scale here the difference makes sense, but the zero point is arbitrary. Now if I say how much time the user got exposed to the advertisement for that duration let us say 15 second, 10 second, 5 second.

So, that exposure time is independent of when it started. I am only interested in this. So, in that sense time now becomes one item one attribute with interval scale. So, along with the other descriptive statistics we saw here also range, mean, standard deviation those things make sense and they are quantitative. So, they can be added, subtracted and so on. Next is your ratio scale it is a quantitative scale as well, but here zero makes sense.

Now, let us come to the same example of time when we were talking about the time of that runner in that running example, we use the stopwatch to measure the time and we started from zero and first person went for some 13 minutes or something he took to reach the destination. In that sense now time is in ratio scale, age of a customer age starts from zero always. So, this is also in ratio scale. So, here the zero point is fixed and ratio of the scale can be compared and that is how the name is the ratio scale. So, here besides your arithmetic mean, geometric mean, harmonic mean etcetera also make sense in ratio scale and there are certain in additional inferential statistics as well.

So, with this we try summarizing the primary scales of measurement. This I have already told and you can see and here I have simply put them in a table so that you can easily compare them. And these are my conclusions for today's lecture. We saw that data can be collected explicitly by asking the user or by implicitly observing the behavior of the user. And we also discussed about

four primary scales of measurement nominal, ordinal, interval and ratio. And we specifically saw nominal is associated with name and ordinal is associated how the items can be ordered.

And the rating scales typically that we see like five point rating scale thumbs up, thumbs down etcetera they come under this ordinal type. And besides that this interval and ratio scale as we move ahead to item features, user features and other side information they will also make sense. But in case of rating matrix probably this nominal and ordinal if the rating matrix is discrete these two scales make sense right now. These are some of the additional references I have used and let me remind you I have already told you about my guest my textbooks and all. So, those textbooks anyway you are supposed to refer and besides that whenever I have any additional resources I am going to let you know.