

Applied Optimization for Wireless, Machine Learning, Big Data
Prof. Aditya K. Jagannatham
Department of Electrical Engineering
Indian Institute of Technology, Kanpur

Lecture - 77


Introduction to Big Data: Online Recommender System (Netflix)

Hello. Welcome to another module in this massive open online course. So, in this module let us start looking at yet another innovative application of optimization and this is in the field of Big Data and in fact, Big Data is something that has gathered significant amount of attention of late because of the tremendous rise in the amount of data that is being generated each day in various websites or various on or various online services that are there and Big Data several applications.

Of course, we are not going to get all possible applications of Big data or all possible how do you put it all possible techniques for a Big Data. But we are going to look at one very specific and a very relevant technique for Big Data known that is very popular and very relevant in the current scenario and that is for the design of what are known as a recommend Recommender System.

So, we are looking at. So, this is basically an application you can just treat this as in application or a technique, you can think of this as a technique for Big Data and what this is I am going to describe this more as we go along.

(Refer Slide Time: 01:39)



The slide features the IIT Kanpur logo on the left. The title 'Recommender Systems' is underlined. A handwritten note 'Purchase Viewing History' is written above the first bullet point. The first bullet point is 'Recommend other items of interest based on Purchase/Viewing History', with 'Purchase/Viewing History' underlined and a handwritten question mark. The second bullet point is 'Extremely critical for Online Retail and Social Networks'. The third bullet point is 'Examples -', followed by three sub-bullets: 'Amazon's product recommendations' (checked), 'YouTube videos', and 'Pandora music selection'. At the bottom, the citation 'Mung Chiang, "Networked Life: 20 Questions and Answers", Princeton University Chapter 4, "How does Netflix recommend movies?"' is shown, with the chapter title circled.

Recommender Systems

- Recommend other items of interest based on Purchase/Viewing History?
 - Extremely critical for Online Retail and Social Networks
- Examples –
 - Amazon's product recommendations ✓
 - YouTube videos
 - Pandora music selection

Mung Chiang, "Networked Life: 20 Questions and Answers", Princeton University
Chapter 4, "How does Netflix recommend movies?"

Now, what is a Recommender System? We need to understand first what is the concept of a recommender system, a Recommender System is something that is very simple that is it recommends other items. So, recommender system as an implies, it recommends other items based on your purchase or viewing history. So, this is based on your purchase or viewing history ok. And for example, you might have seen it in several various popular online websites.

For instance, if you go to any E-commerce websites which is Amazon; you have several product recommendations, for instance in any commerce site based on a based on your viewing history of the items that you have browsed or based on even your past purchase history of a set of items that are recommended that all that the website things are most suitable all right or I think would be in your interest or would be of high interest to you alright.

Or YouTube videos, even when you go to a web site like a video streaming site like YouTube, when you look at the different videos or when you are watching current video, the website come automatically comes up with a recommendation of other videos that you might find a lot of similarity or that you will be highly interested in.

And similarly music websites are giving their Pandora for instance is a music website and then, it comes up with a set of music videos or music albums or songs that would be of very high interest to you and some of these you might not be where and interesting about these is this thing is this is very helpful because some of these things you might not be aware exists.

So, by coming up with this highly specific set of recommendations, it is a win-win situation both for you. Because you cannot browse the infinite number of products on an E-commerce website and similarly, it is also beneficial for website because enticing the customers to this possible set of objects that the customer is interested in they are increasing their business.

So, it is a win-win situation for everyone. It saves your time, increases the business of the website and so on and for this part this module will be referring in particular to this very interesting book and in particular the chapter on the book by Professor Mung Chiang titled "Networked Life: 20 Questions and Answers" by Princeton unit Princeton University and the chapter that we are talking about is Chapter 4, which is "How does

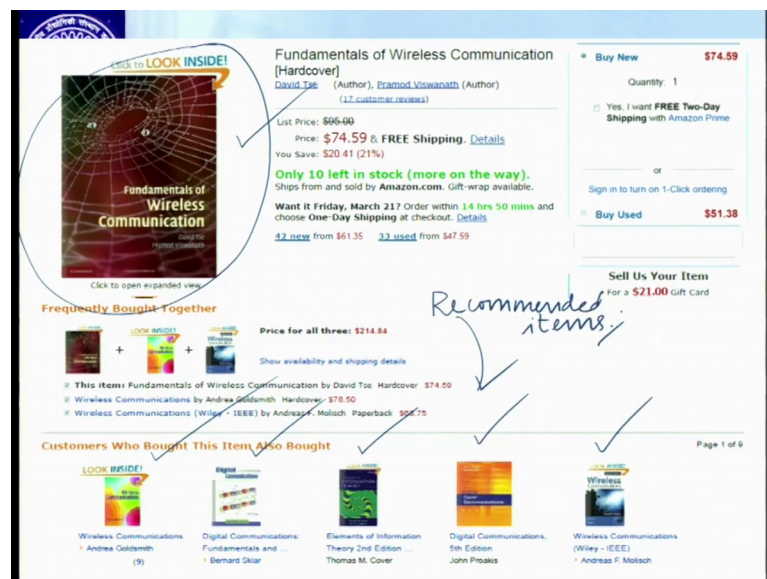
Netflix recommend movies?” We are going to talk more about this. Netflix is one particular website.

So, all such systems which basically recommend various options for you, various other product various movies or videos or products for you to buy these are known as recommender systems that is they are recommending a set of products or objects to ok. Of course, no, you cannot come up with arbitrary recommendations; the more pointed or the more specific your recommendations are for instance an E-commerce website.

You can just recommend an arbitrary set of products. But that is not going to help much. That was specific in the sense that the more interest, the more the more closely your recommendations match the interest of the consumer, the better your recommenders recommender system is, then the better is going to be the efficiency of your website alright.

So, the idea is the goal is to design the best recommender system. It comes with a very specific and a highly interesting set of recommendations.

(Refer Slide Time: 05:04)



For instance, is a simple snapshot from one of the E-commerce websites? You have a book that you are interested in buying alright. This is the book you are interested in and the website comes up with an alternative set of recommendation. So, you look and look at this these are your alternatives or these are your set of recommendations alright or

your recommended items based on which is either based on your viewing history or interestingly also this might be based on the viewing history of someone who has a similar viewing history or who has an interests that are very similar to yours.

So, the recommender systems they are very interesting. So, they look at patterns of different users their purchase histories mind these patterns and come up with recommendations based on what other people who had similar interests have purchased or have viewed and so on ok.

(Refer Slide Time: 06:02)



And for instance again on another E-commerce website, you are looking at an object or you are browsing at an object and these are your recommended set of objects. These are your recommended objects ok.

(Refer Slide Time: 06:22)

Netflix

- Online DVD rental site
 - Started in 1997
- No late fees – Retain DVDs as long as you are subscribed
 - The catch – **Cannot** rent new DVDs till the old ones are retained

order online

send DVDs by mail

And no, coming to our original problem I mean recommender systems are everywhere and if you go to any E-commerce website any video website or so on, Recommender system are everywhere ok. One particular interesting application that we are going to talk about is that of Netflix. Netflix is an Online DVD rental site ok. Started in 1997 and the model of Netflix is that you send DVDs by mail that is you send DVDs by mail or regular post which you can order online ok. So, DVD is rental.

So, you can order these order online or basically you can request these online and they will be sent by mail to you for a few of course. Now, of course, once you send in again once you watch the DVDs; you send them back. You get a new set of DVDs alright, that specific to that particular website.

(Refer Slide Time: 07:18)

The slide features the logo of Anna University on the left. The main content is titled 'Netflix' and lists three key innovations:

- Key **innovation** in Netflix
 - Generate personalized movie recommendations for each user
 - Based on past viewing history
 - And collectively mining viewing patterns across an extremely large number of users

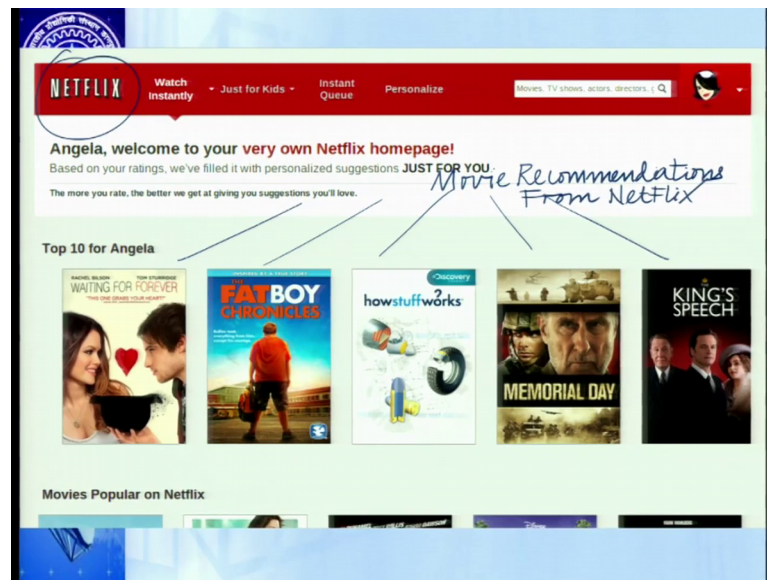
Handwritten notes in blue ink are present: 'specific to each user' is written above 'Based on past viewing history', and 'Based on movies already seen' is written below it, with a line connecting the two phrases.

Now, the key innovation here that we are talking about is basically Netflix the website, it generates personalized movie recommendations for each user, based on your past viewing history. So, based on the movies that you are already seen ok, so, we generate movie recommendations for each user specific to each user; specific to each user based on movies that you already seen or based on your past viewing history all right. So, what do you do for this is basically you collectively mine the viewing patterns of an extremely large number of use, that is users that is all the moves you have seen and the movies that a large number of users have seen.

Come up with these, so mine these patterns to extract information and then based on this mining of this collective data of users and movies, you come up with the specific set of recommendations for a particular reason for that matter for each user. Which is in fact, which you consider this problem by itself is very fascinating. So, it comes up with a set of recommendations or comes up in essence it comes with a set of ratings for you for the movies which you have never seen. That is in essence what the problem is.

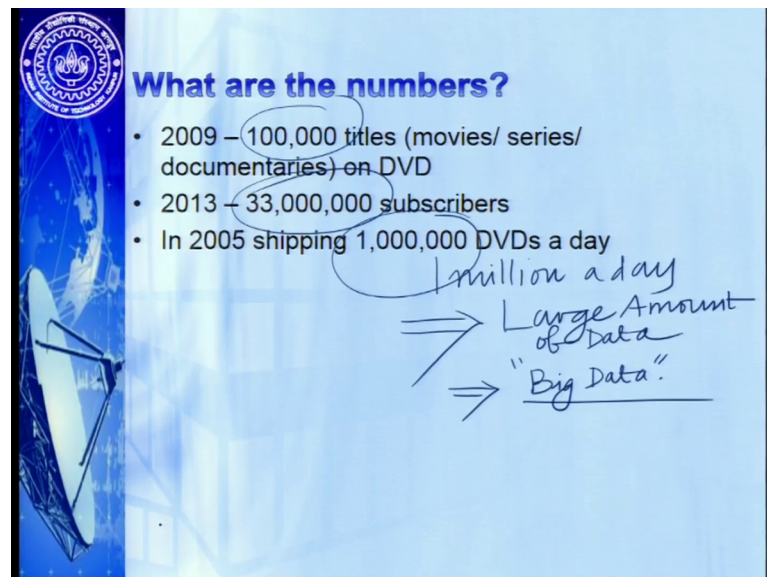
So, it recommends movies to you based on what it thinks what Netflix thinks are movies that you are going to rate highly which means it has to generate a predicted rating for you for a set of movies that you have not seen and then, choose a set of movies based on what Netflix or that website thinks you would rate very highly. So, this problem in itself is a very interesting problem ok.

(Refer Slide Time: 09:15)



And what is interesting is that Netflix, for instance is simple. If you look at a Netflix website, it is a simple snapshot of Netflix. These are some movie recommendations from Netflix. These are some movie recommendations from Netflix, based on what it thinks that particular user and you can see there is a set of movies alright.

(Refer Slide Time: 09:52)



Based on what it thinks that particular user would of course, as I have already said the more relevant these movies are or the more the more interest these movies generate from

that particular user or the more close or the closer these movies are to that particular users interest the better is the recommender system.

So, what are the number in 2009, Netflix had about a 100000 titles and in 2013 about 33 million subscribers ok. So, there is 33 million subscribers and in 2005, it is close to shipping about 1 million DVDs a day. That is a large number; so, you can think of that 1 million a day implies you can think of the large amount of data that this generates implies this is a Big Data problem. We are generating a large amount of data which is basically nothing your but your Big Data problem.

So, from this large an amount of data, how do you mine the patterns ok?

(Refer Slide Time: 11:00)

Netflix Challenge

- In 2006 Netflix rolled out a challenge to the research community
 - \$1,000,000 prize
- Publicly made available 100,000,000 ratings
 - Roughly, could fit into a standard desktop memory (around 2006)
- For 480,000 users and 17,770 movies
 - Observe max number is $4.8 \times 10^5 \times 1.77 \times 10^4 = 8.53 \times 10^9$
- Only a small fraction of ratings available ie $1 \times 10^8 \sim 1\%$
 - available! $\frac{10^8}{10^{10}} = 1\%$
 - possible 2×10^4 movies
 - # ratings = $5 \times 10^5 \times 2 \times 10^4 = 10^{10}$

Handwritten notes on the slide include: $100 \times 10^6 = 10^8$, "Big Data", "100 million ratings", " $\sim 20,000$ movies", and " $\sim 5 \times 10^5$ ".

Now, in 2006, Netflix rolled out an interesting challenge. This is termed as the Netflix challenge to the research community and had a huge price and what happened in that challenge is it made available 100 million ratings. So, 100 million ratings were made available. So, you can see that is basically your Big Data large amount of data which is if you look at 100 million that is 100 into 10 power 6 equals 10 power 8 ratings which could fit into a standard memory of a standard desktop.

Now, of course, at that point it at about 480000 users and now of course, these ratings were about 480000, that is roughly half a million users, approximately half a million users for 17770 that is approximately 20000 movies. That is if you look at this, this is 5

into approximately 5×10^5 that is half a million users and this is 2×10^4 movies which implies approximately number of ratings equals $5 \times 10^5 \times 2 \times 10^4$; number of possible ratings, I will clarify this in a moment. $5 \times 10^5 \times 2 \times 10^4$ which is basically approximately again this is 10^{10} .

So, that approximate number of possible ratings equals 10^{10} . Now, this is the number of possible ratings, but the actual number of ratings is only 10^8 . Now the reason for that is very simple because remember you have half a million users, you have twenty thousand movies. But not each user, but each user has not seen every movie right. So, each user has probably seen a fraction of the movies; correct, each user has probably seen 100, 120 or 200 movies or so on and so forth.

So, if you look at all the users and all the movies the total number of possible ratings is 10^{10} . If each user watches every movie and rates every movie, but that is not possible all right. So, the total number of possible ratings, total number of actual ratings that are available are only close to 100 million that is 10^8 which means only 1 percent of the ratings are available.

(Refer Slide Time: 13:37)

Netflix Challenge

- In average terms,
 - Each movie was rated by approx 5000 users
 - Each user rated approx 200 movies
- Develop a program to predict ratings and provide recommendations to users

Handwritten annotations:

Each movie ~ 5000 users

$$\frac{200}{20000} = 1\%$$

Netflix challenge!

So, which means only a small fraction at this is the important point. So, only a small fraction of ratings are available that is 10^8 which is 1 percent. So, 10^8 by total possible 10^{10} , if you look at that is basically 1

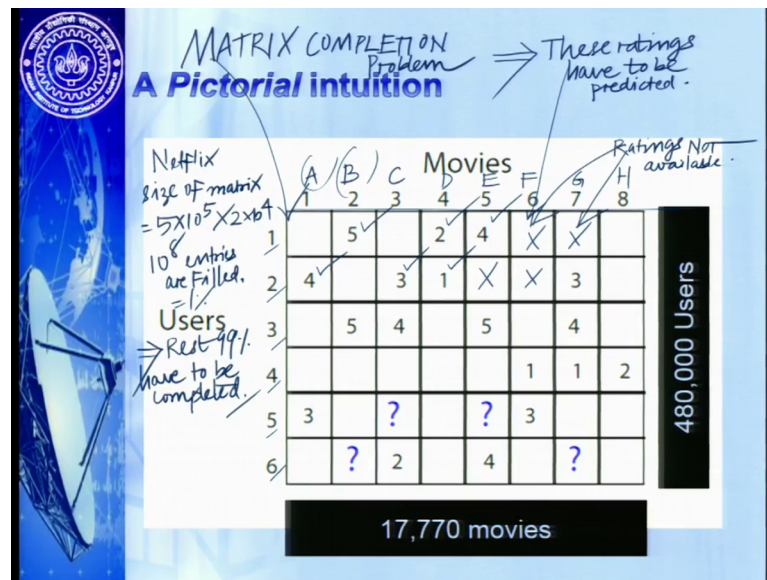
percent ok. 1 percent are available; only one percent, which means what do we have to do? We have to predict the rest of the rating ok.

So, rest of the ratings have to be predicted and based on these predictions, you come up with the movie recommendations for each user. Remember if each user, you have seen every movie then there is no need to come up with a recommendation right. In E-commerce website, if every browser every person has what every item then of course, there is no set to cover. The challenge is because few people have bought few items and it is not even that few people have bought the same items; different people have bought a different item. So, from this checker kind of matrix, we have to come up with the ratings and recommendations for each user. So, each movie was rated by approximately 5000 users.

So, each movie approximately 5000 users and each user rated approximate. So, you can see approximately 200 movies. So, approximately 200 means 200 divided by 20000 total number of movies. So, this you can see is approximately 1 percent. So, that is where you get the number from all right. So, each user is seen about a percent of the movie. So, 2000 titles are available each user seen roughly about 100 to 200 movies. Of course, over a long period of time maybe or 2 years, 3 years couple of years.

So, you have seen 200 movies rated them all right. So, each user roughly seen 1 percent of the movie. Each movie of course rated by again roughly about 5000 users ok. Now, what we have to do is as I have told you to develop a program to predict the ratings and provide recommendations for you. So, this is the challenge. This is the Netflix challenge or this was the Netflix challenge rather all right. So, you look at all these ratings that are available and then, you come now how do you test this?

(Refer Slide Time: 16:05)



Now, if you look at it this, this problem looks something like this. So, you have a set of movies here movies 1 2 3 4 5 6.

So, this is for the sake of this thing to avoid confusion, let us call it movies A B C D E F G H. So, you have movies A B C D E F G H and you have users 1 2 3 4 5 6. For instance user 1 has not seen movie A, but he has seen movie B and rated movie B. User 1 as seen movie D, rated movie. User 2 has seen movie A, movie C, movie D rated. So, these are the ratings that are available. Now these ratings the empty blocks are rating.

So, for instance user 1 has not seen movie 6 not seen. So, these ratings are not available which implies these ratings have to be predicted, these ratings have to be predicted. And now, if you look at this you (Refer Time: 17:21) matrix, for instance this is a matrix is nothing but rows and columns. You have this rows, rows are the users; each row corresponds to user each column corresponds to you.

So, now some users have rated some movies; therefore, some entries of this matrix are filled. The rest of the entries of this matrix are weakened. And therefore, we have to complete predicting the ratings means, basically we have to complete this matrix. This problem is a fund fundamental importance in signal processing and various other or big data. This problem is known as a Matrix Completion Problem. This is known as a Matrix Completion Problem.

This problem is known as a Matrix Completion Problem. So, you have about 480000 users ⁷ that is a half a million users correct in terms of the example that we just seen or the Netflix problem, you have half million users, 20000 movies which means the total size of the matrix number of elements in the matrix is half a million into 20000 ok.

So, in a Netflix program size of matrix is this is a matrix of size 5 into 10 power 5 half a million times 2 into 10 power 4 and out of these, 10 to the power of 8 entries are filled. It is like a puzzle implies the rest equal to 1 percent implies rest 99 percent have to be completed or filled and this is basically a Matrix Completion Problem that is the interesting aspect of this is a Matrix Completion Problem and the size of the matrix is humongous as you can see, it is a huge.

(Refer Slide Time: 19:20)

Algorithm

100 Million
Training Set
Public

1.4 M
Probe Set
Hidden

1.4 M
Quiz Set
Hidden

1.4 M
Test Set
Hidden

- Design scheme on the publicly available *Training Set*
- However, the solution is finally tested on the hidden *Quiz* and *Test* sets.
- Netflix own algorithm *Cinematch* had an average error of 0.9514.

And this is because there is a problem big data. The way the contest was organized was very simple, you have 100 million or training set that were made available to the public and there is a probe set and the probe set all right.

So, the probe set the user can test because obviously, you do not because one has to of course, eventually test what is the performance of the algorithm that is being proposed. So, there is a probe set, there is a final quiz set and the these are hidden. Now the quiz set and the test set these are hidden and finally, the algorithm is tested on the quiz and test set.

So, these are hidden and finally, these are tested on the hidden intersect to find different researchers submitted different algorithms the performance of the algorithms is tested on the quiz sets or the test set and whichever algorithm performs the best that is it gives the recommendations on the test set correct which are best or which are closest to remember the ratings that are given by the users all right.

So, the performance is testing, it is actual ratings given by the users because these are hidden and that algorithm is obviously, best because it is best able to predict the ratings of the different the different users all right and that is it. So, that is basically the interesting problem and this is a very interesting problem and one of the most elegant and one of the most interesting applications of Big Data and in fact, one of the most powerful applications of Big Data as I have already told you is in the context of this recommender system. Because they are used everywhere, they are used in online websites, video streaming, music website. They have a very big application in E-commerce.

In fact, one of the key forces or you can say one of the key components of any E-commerce website is the recommender system because the more robust your recommender system is, the more the better your recommendations are and the more closely your recommender system is able to predict the ratings for different users for the objects which they have not purchased based on their browsing or purchase history, the better is going to be the business of the E-commerce website and the more efficient it is going to be all right.

So, this is a very important point a very key component or very important application of Big Data and we convex optimization has a very important role to play this. In fact, optimization has a very important role. So, we will stop here and look at how to handle this problem; how can convex optimization help in addressing this problem, we look at that in the next module all right.

Thank you very much. Or basically how do you use convex optimization to handle this problem of Big Data that is what a Big Data and specifically this problem of matrix completion which is closely tied to a recommender system alright. So, will stop here and look at start looking at this exploring this problem in the subsequent modules.

Thank you very much.