1

00:00:00,000 --> 00:00:05,275

[SOUND] This lecture is about

2

00:00:05,275 --> 00:00:10,784

Collaborative Filtering.

3

00:00:10,784 --> 00:00:16,032

In this lecture, we're going to continue

the discussion of Recommender Systems.

4

00:00:16,032 --> 00:00:20,665

In particular, we're going to look at

the approach of collaborative filtering.

5

00:00:20,665 --> 00:00:25,161

You have seen this slide before

when we talked about the two

6

00:00:25,161 --> 00:00:30,126

strategies to answer the basic

question will user U like item X.

7

00:00:30,126 --> 00:00:33,723

In the previous lecture,

we looked at the item similarity,

8

00:00:33,723 --> 00:00:35,988

that's content-based filtering.

9

00:00:35,988 --> 00:00:39,332

In this lecture, we're going to

look at the user similarity.

10

00:00:39,332 --> 00:00:42,804

This is a different strategy

called collaborative filtering.

11

00:00:42,804 --> 00:00:46,043

So first of all,

what is collaborative filtering?

12

00:00:46,043 --> 00:00:48,971

It is to make filtering decisions for

13

00:00:48,971 --> 00:00:53,504

individual user based on

the judgement of other users and

14

00:00:53,504 --> 00:00:57,944

that is to say,

we will infer individual's interest or

15

00:00:57,944 --> 00:01:01,925

preferences from that,

of other similar users.

16

00:01:01,925 --> 00:01:04,350

So the general idea is the following.

17

00:01:04,350 --> 00:01:09,601

Given a user u, we are going to

first find the similar users,

18

00:01:09,601 --> 00:01:16,247

u1 through **uN** and then we're going to

predict the used preferences based on

19

00:01:16,247 --> 00:01:21,520

the preferences of these similar users,

u1 through **uN**.

20

00:01:21,520 --> 00:01:26,711

Now the users similarity here can be

judged based on their similarity.

21

00:01:26,711 --> 00:01:29,799

The preference is on

a common set of items.

22

00:01:29,799 --> 00:01:35,836

Now here you'll see that the exact

content of item doesn't really matter.

23

00:01:35,836 --> 00:01:39,691

We're going to look at the only,

the relationship between the users and

24

00:01:39,691 --> 00:01:40,353

the items.

25

00:01:40,353 --> 00:01:44,788

So this means this approach

is very general if it can be

26

00:01:44,788 --> 00:01:49,234

applied to any items not

just with text objects.

27

00:01:49,234 --> 00:01:53,436

So this approach, it would work well

under the following assumptions.

28

00:01:53,436 --> 00:01:59,010

First users with the same interests

will have similar preferences.

29

00:01:59,010 --> 00:02:03,293

Second, the users with similar preferences

probably share the same interests.

30

00:02:03,293 --> 00:02:08,765

So for example, if the interest of

the user is in information retrieval,

31

00:02:08,765 --> 00:02:12,929

then we can infer the user

probably favor SIGIR papers.

32

00:02:12,929 --> 00:02:17,758

And so those who are interested in

information retrieval researches probably

33

00:02:17,758 --> 00:02:21,648

all favor SIGIR papers,

that's something that we make.

34

00:02:21,648 --> 00:02:23,473

And if this assumption is true,

35

00:02:23,473 --> 00:02:27,335

then it would help collaborative

filtering to work well.

36

00:02:27,335 --> 00:02:32,712

We can also assume that if we

see people favor SIGIR papers,

37

00:02:32,712 --> 00:02:38,016

then we can infer the interest is

probably information retrieval.

38

00:02:38,016 --> 00:02:42,254

So these simple examples,

it seems what makes sense.

39

00:02:42,254 --> 00:02:47,967

And in many cases such as assumption

actually does make sense.

40

00:02:47,967 --> 00:02:52,763

So, another assumption you have to make

is that there are a sufficiently large

41

00:02:52,763 --> 00:02:55,826

number of user preferences

available to us.

42

00:02:55,826 --> 00:03:00,260

So for example, if you see a lot

of ratings of users for movies and

43

00:03:00,260 --> 00:03:03,255

those indicate their

preferences in movies.

44

00:03:03,255 --> 00:03:05,367

And if you have a lot of such data,

45

00:03:05,367 --> 00:03:09,153

then collaborative filtering

can be very effective.

46

00:03:09,153 --> 00:03:14,505

If not, there will be a problem and

that's often called a cold start problem.

47

00:03:14,505 --> 00:03:18,485

That means you don't have many

preferences available, so

48

00:03:18,485 --> 00:03:23,805

the system could not fully take advantage

of collaborative filtering yet.

49

00:03:23,805 --> 00:03:29,441

So let's look at the collaborative

filtering problem in a more formal way.

50

00:03:29,441 --> 00:03:36,081

And so this picture shows that we are in

general considering a lot of users and

51

00:03:36,081 --> 00:03:39,999

showing we're showing m users here.

52

00:03:39,999 --> 00:03:45,935

So, u1 through **uN** and we're also

considering a number of objects.

53

00:03:45,935 --> 00:03:50,267

Let's say,

n objects denoted as o1 through on oN and

54

00:03:50,267 --> 00:03:56,312

then we will assume that the users will

be able to judge those objects and

55

00:03:56,312 --> 00:04:01,285

the user could for example,

give ratings to those items.

56

00:04:01,285 --> 00:04:05,923

For example, those items could be movies,

could be products and

57

00:04:05,923 --> 00:04:10,260

then the users would give ratings

one through five, let's say.

58

00:04:10,260 --> 00:04:14,859

So what you see here is that we have

assumed some ratings available for

59

00:04:14,859 --> 00:04:16,203

some combinations.

60

00:04:16,203 --> 00:04:21,521

So some users have watched movies,

they have rated those movies.

61

00:04:21,521 --> 00:04:25,893

They obviously won't be able

to watch all the movies and

62

00:04:25,893 --> 00:04:29,763

some users may actually

only watch a few movies.

63

00:04:29,763 --> 00:04:34,581

So this is in general a response matrix,

right?

64

00:04:34,581 --> 00:04:39,104

So many item many entries

have unknown values and

65

00:04:39,104 --> 00:04:44,234

what's interesting here is

we could potentially infer

66

00:04:44,234 --> 00:04:49,360

the value of a element in this

matrix based on other values and

67

00:04:49,360 --> 00:04:55,317

that's actually the central question

in collaborative filtering.

68

00:04:55,317 --> 00:04:59,648

And that is,

we assume an unknown function here f,

69

00:04:59,648 --> 00:05:03,999

that would map a pair of user and

object to a rating.

70

00:05:03,999 --> 00:05:09,027

And we have observed there are some

values of this function and

71

00:05:09,027 --> 00:05:13,086

we want to infer the value

of this function for

72

00:05:13,086 --> 00:05:18,615

other pairs that we,

that don't have values available here.

73

00:05:18,615 --> 00:05:24,401

So this is ve, very similar to

other machine learning problems,

74

00:05:24,401 --> 00:05:31,067

where we would know the values of the

function on some training there that and

75

00:05:31,067 --> 00:05:37,168

we hope to predict the the values of

this function on some test there.

76

00:05:37,168 --> 00:05:40,142

All right.

So this is the function approximation.

77

00:05:40,142 --> 00:05:47,129

And how can we pick out the function

based on the observed ratings?

78

00:05:47,129 --> 00:05:49,287

So this is the, the setup.

79

00:05:49,287 --> 00:05:53,828

Now there are many approaches

to solving this problem.

80

00:05:53,828 --> 00:05:57,287

And in fact,

this is a very active research area.

81

00:05:57,287 --> 00:06:02,970

A reason that there are special

conferences dedicated to the problem

82

00:06:02,970 --> 00:06:06,390

is a major conference

devoted to the problem.

83

00:06:06,390 --> 00:06:16,390

[MUSIC]