BDA 6-3

1

Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering

School of Computer Science, in Beijing Institute of Technology

2

In this chapter, we introduced some useful platform， Spark MLlib and TensorFlow.

And in this session, we will talk about a typical big data analysis application, **Recommendation System**

3

A recommendation system is a subclass of Information filtering Systems that seeks to predict the rating or the preference a user might give to an item.

In simple words, it is an algorithm that suggests relevant items to users. Recommendation system is widely used in many fields, like Netflex, Amazon, JD, TaoBao

QQ music etc.

Recommendation Algorithms can be roughly divided into 3 categories. Collaborative filtering, content based and knowledge based. And the big difference between Collaborative filtering algorithm and content-based algorithm is that content based algorithm relies on the item features themselves, but the Collaborative filtering relies on how other user respond to the same item.

Collaborative filtering can be further divided into Neighborhood-based and model based, neighborhood-based algorithm includes user-based and item-based algorithm, and we will discuss in this session. Model based algorithms include Latent Factor model, Graph model and others. In latent Factor model, I will use matrix factorization as example to explain.

The content-based recommendation algorithm recommends items based on item features, which include structured features and unstructured features.

And the third category is knowledge-based algorithms.

4

Let ‘s look at the diagram,

on the left it is the Collaborative filtering, exampled by user-based filtering algorithm.

it shows that the similar users will have the similar preferences for articles,

and the similar users are defined based on the common articles (the green and red) they read.

Since they are similar users, the article read by her (the blue article) can be recommended to him.

On the right , it is content-based filtering recommendation , the red article and the green article are defined similar articles based on the article’s features, and when the red article is read by the user, the green article will be recommended to the same user.

5

This picture shows the user-based filtering and item-based filtering.

In user-based filtering, we find the similar users based on the common items they chose before,

here Tim chose Ice cream with Egg Roll, chocolate, ice cream cone and donut,

and John chose chocolate ice cream cone, Tim and John have certain items in common, then we define them similar users.

for the items Tim has chosen and John not, we can recommend to John, here are the Ice cream with Egg Roll and donut.

In item-based filtering, we define the Ice cream with Egg Roll and ice cream cone as similar items based on the user chosen behavior,

and John has chosen the ice cream cone, so we recommend the Ice cream with Egg Roll to him.

We will explain the user-based filtering and item-based filtering in detail.

6

In this session we will explain 3 main Recommendation System Algorithms,

User-based filtering, Item-based filtering, and Content Based Filtering.

7

First let ‘s look at User-based filtering Recommendation System Algorithms.

The steps for User-based filtering are shown in the diagram,

1） we collect the User purchase records,

2） we build Item to User table based on the User purchase records.

3） the User to User matrix is built based on Item to User table;

4） the User – User similarity matrix is built based on User to User matrix ;

5 ）then calculate the user's interest in different items.

6） Get recommendation result based on the user's interest in different items.

8

User-based collaborative filtering is an algorithmic framework where the similar users are identified based on the similarity with the active user, and then scoring of the items is done based on neighbor’s ratings followed by a recommendation of an item based item’s scores by the recommendation system.

Multiple users choose many of the same items, we can call these users **similar users.**

In the user-based collaborative filtering algorithm, we believe that: **a user will like items that his or her similar users like with higher****probability**.

So how to calculate the similarity of the users.

Usually use **Jaccard similarity,** **cosine similarity, Euclidean distance** and **Pearson distance** to calculate the similarity between two users.

Let **N(u)** be the set of items that user **U** likes, and **N(v)** be the set of items that user **V** likes.

**Jaccard similarity,** **cosine similarity calculation methods are shown here.**

9

Let us make an example to get recommendation result using User-based collaborative filtering.

Here the big ABCD are users, and the small a b c d and e are items,

Based on the User purchase records, we build the item- user table, which shows for each item, all the uses who have purchased this item.

For example, item a is purchased by user A and B, item b is purchased by User A and C, etc.

Then User - User matrix is built based on the item to user table, in the way that each element value is the numbers of common items purchased by the row user and column user.

For example, the element of the first row and second column is calculated by the common items number of user A and B, they have just one common item which is item a, so the value is 1.

The other matrix element value is calculated in the same way.

10

Based on the **User - User matrix we just calculated, use Cosine similarity we can calculate the User – User similarity matrix,**

**each element in this matrix is the similarity of the user in corresponding row and user in corresponding column.**

**Here N(u) is the set of items that user U likes, and N(v) is the set of items that user V likes.**

11

After we have the User – User similarity matrix,

For each candidate item i, the degree that the user u is interested in it P(u,i) can be computed using the above formula.

12

Suppose we want to recommend items to user A,

select K = 3 similar users,

and similar users are B, C and D, based on the User – User similarity matrix

then the items that B, C and D chose and A has not chosen are c and e,

then calculate user A 's interest in item c, P (A, c) and e, P (A, e) respectively,

they both are 0.7416, which means User A likes c and e equally.

13

Now let’s look at Item-based filtering

14

Now let’s look at the Item-based collaborative filtering.

Calculate the similarity between items and recommend items with high similarity to the items chosen by users.

In the item-based collaborative filtering algorithm, we believe that: the similarity between item A and item B is because users who like item A also like item B.

The calculation is also based on the user purchase records.

Item-item matrix is computed based on that.

In the Item-item matrix , each element is the number of the users who chose the item in the corresponding row and the item in the corresponding column.

For example, on the right, in the Item - Item matrix, the value of the element in the first row, second column, the number of the users who chose both item a and item b is one, just user A.

The other elements can be calculated in the same way.

15

Now let ‘s calculate the Similarity of

Let N(i) be the set of users who like item i, and N(j) be the set of users who like item j,

and Intersection of N(i) and N(j) be the number of users who like both item i and item j.

cosine similarity of item i and item j, Wij can be calculated as the formula,

the number of Intersection of N(i) and N(j) divided by the number of the union of N(i) and N(j) .

16

**Item - Item similarity matrix can be calculated using cosine similarity based on Item - Item matrix as showed on the right matrix.**

17

Now let ‘s compute the **user's interest in different items.**

For each candidate item j, the degree user u interested in item j, **P**uj can be calculate as the formula shows:

where**, N(u)** represents the collection of items favored by user U, **S (j, k)** is the collection of k items most like item **j**,

**W**ij is the similarity between item **i** and item **j**, and **R**ui represents user u's interest in item **i**.

18

Now we can get **recommendation result.**

Suppose we want to recommend items to user B, select K = 3 similar items,

so, the items user B has not chosen before are b, d and e.

then calculate P (B, d), P (B, b) and P (B, e) respectively and 1, 0.5, 0.5,

then we think **User B likes d most and User B likes b, e equally**

19

For user-based filtering

How to calculate the degree of user's interest in candidate items based on users?

We need to establish a connection between the user and the candidate item, and the connection is the **user** similarity.

For example on the left, we calculate the user A’s interest degree in Item c in P (A, c), the items have been purchased by users are B and D, so add the similarity of W.AB and W.AD.

And for item-based filtering

How to calculate the degree of user‘s interest in candidate items based on items?

We still need to establish a connection between the user and the candidate item, and the connection is the **item** similarity.

For example on the left, we calculate User B’s interest degree in Item d in P (B, d), the items have been purchased are item a and item c, so add the similarity of W.da and W.dc.

20

Now let’s look at Content-Based Filtering

21

Content-based Filtering Recommendations, is the first recommendation algorithm that **recommends items similar to a user's past favorite items.**

The key here is the measure of item similarity, which is the core of the algorithm application process.

Three main steps:

1. **Item Representation**: extracting some features (content) for each item to represent them
2. **Profile Learning**: Using the feature data of a user's past favorite (or not favorite) items to learn the user's preference features (that is user’s profile)
3. **Recommendation Generation:** By comparing the user profile obtained in the previous step with the features of the candidate items, a set of items with the highest relevance is recommended to this user.

22

In this session, we learned 3 main Recommendation System Algorithms,

User-based filtering, Item-based filtering and Content-Based Filtering

thank you for your attention, if you have any question, feel free to contact me.