DISTRIBUTED AND SCALABLE DATA ENGINEERING
FINAL PROJECT REPORT

On

Collaborative Filtering using the Netflix Data

By

Yashaswini Gouru (00667970)

Under the guidance of

Vahid Behzadan(Assistant Professor)

Tagliatela College of Engineering

Department of Computer Science
(Data Science)

300 Boston Post Rd, West Haven, CT 06516

COLLABORATIVE FILTERING

These days whether you look at a video on YouTube, a movie on Netflix or a product on Amazon, you are going to get recommendations for more things to view, like or buy. You can thank the advent of machine learning algorithms and recommender systems for this development.

Recommender systems are far-reaching in scope, so we are going to zero in on an important approach called collaborative filtering, which filters information by using the interactions and data collected by the system from other users. It is based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future.

RECOMMENDER SYSTEMS:

A recommender system is a subclass of information filtering that seeks to predict the "rating" or "preference" a user will give an item, such as a product, movie, song, etc.

Recommender systems provide personalized information by learning the user's interests through traces of interaction with that user. Much like machine learning algorithms, a recommender system makes a prediction based on a user's past behaviors. Specifically, it is designed to predict user preference for a set of items based on experience.

Mathematically, a recommendation task is set to be:

- Set of Users
- Set of Items that are recommended to the User
- Learn a function based on user's past interaction data that predicts the likeliness of item to user

Recommender systems are broadly classified into two types based on the data being used to make inferences:

- Content based filtering, which uses item attributes.

- Collaborative filtering, which uses user behavior in addition to item attributes.

COLLABORATIVE FILTERING:

Collaborative filtering filters information by using the interactions and data collected by the system from other users. It is based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future.

The concept is simple: when we want to find a new movie to watch we will often ask our friends for recommendations. Naturally, we have greater trust in the recommendations from friends who share tastes similar to our own.

Most collaborative filtering systems apply the so-called similarity index-based technique. In the neighborhood-based approach, number of users are selected based on their similarity to the active user. Inference for the active user is made by calculating a weighted average of the ratings of the selected users.

Collaborative-filtering systems focus on the relationship between users and items. The similarity of items is determined by the similarity of the ratings of those items by the users who have rated both items.

There are two classes of Collaborative Filtering:

- User-based, which measures the similarity between target users and other users.
- Item-based, which measures the similarity between the items that target user rate or interact with and other items.

COLLABORATIVE FILTERING USING PYSPARK:

The project is done in Jupyter notebook created on AWS EMR Cluster.

The implementation of project starts with creating an EMR cluster on AWS:

1. Choose Create cluster.

2. On the Create Cluster - Quick Options page, accept the default values except for the following fields:

- Enter a Cluster name.
- In Software configuration select the applications as spark.
- Under Security and access, choose the EC2 key pair that you created in Create an Amazon EC2 Key Pair.

3. Choose Create cluster.

Create EMR notebook using EMR Cluster:

EMR Notebook:

EMR Notebooks, a managed environment, based on Jupyter Notebooks that allows data scientists, analysts, and developers to prepare and visualize data, collaborate with peers, build applications, and perform interactive analysis using EMR clusters. EMR Notebooks is pre-configured for Spark. It supports Spark magic kernels allowing you to interactively run Spark jobs on EMR clusters written in languages such as PySpark, Spark SQL, Spark R, and Scala.

Steps to create EMR notebook:

1. Choose Notebooks, Create notebook.
2. Enter a Notebook name and an optional Notebook description.
3. select Choose, select a cluster from the list (cluster created above), and then Choose cluster. Only clusters that meet the requirements are listed.
4. For Security groups, choose Use default security groups.
5. For AWS Service Role, leave the default or choose a custom role from the list. For more information, see Service Role for EMR Notebooks.
6. For Notebook location choose the location in Amazon S3 where the notebook file is saved.

Open the Notebook created and select pyspark in jupyter notebook.

Implementation of Collaborative filtering on Netflix data in pyspark:

Analyzing the Netflix Data:

- Import pyspark
- The text data files are stored in S3. Import the text files and formed the Spark data frame.

(a) How many distinct items and how many distinct users are there in the test set?

```
df.distinct().count()
```

```
▸ Spark Job Progress
```

```
100478
```

```
df.select('CustomerId').distinct().count()
```

```
▸ Spark Job Progress
```

```
27555
```

(b) The collaborative filtering approaches lives from finding many similar users (for a user-user model) or many similar items (item-item model):

- **User-User collaborative filtering:**

  The method identifies users that are similar to the queried user and estimate the desired rating to be the weighted average of the ratings of these similar users.

- **ITEM-ITEM collaborative filtering**:

  ITEM-ITEM collaborative filtering look for items that are similar to the articles that user has already rated and recommend most similar articles. But what does that mean when we say item-item similarity? In this case we do not mean whether two items are the same by attribute like Fountain pen and pilot pen are similar because both are pen. Instead, what similarity means is how people treat two items the same in terms of like and dislike.

ITEM-ITEM Based Approach:

This method is quite stable as compared to User based collaborative filtering because the average item has a lot more ratings than the average user. So, an individual rating does not impact as much. To calculate similarity between two items, we look into the set of items the target user has rated and computes how similar they are to the target item i and then selects k most similar items.

To determine the most-similar match for a given item, the algorithm builds a similar-items table by finding items that customers tend to purchase together. We could build a item-to-item matrix by iterating through all item pairs and computing a similarity metric for each pair. However, many product pairs have no common customers, and thus the approach is inefficient in terms of processing time and memory usage. The following iterative algorithm provides a better approach by calculating the similarity between a single item and all related items:

Pivot table is taken by grouping the MovieID column:

```
pivot_Train = Train.groupBy("MovieID").pivot("CustomerId").sum("Rating").fillna(0)
```

Fill the null values with zero and compute the Correlation matrix. Correlation matrix gives the similarity between the items or movies in the train dataset and also obtained the k-most similar Users and k-most similar Items.

MovieID	8	28	43	...	17734	17741	17742
MovieID				...			
8	1.000000	0.026173	0.014027	...	0.041982	0.018298	0.003809
28	0.026173	1.000000	0.013081	...	0.026872	0.035921	0.003725
43	0.014027	0.013081	1.000000	...	0.067302	0.044276	0.036744
48	0.017317	0.081754	0.022921	...	0.052179	0.046530	0.005936
61	0.014884	0.004485	0.048389	...	0.042341	0.034853	0.040344
...	...	...	...	...	...	...	...
17725	0.016143	0.010176	0.051562	...	0.053199	0.036778	0.013022
17728	0.017812	0.007724	0.049892	...	0.059174	0.010200	0.050991
17734	0.041982	0.026872	0.067302	...	1.000000	0.065275	0.010723
17741	0.018298	0.035921	0.044276	...	0.065275	1.000000	0.004309
17742	0.003809	0.003725	0.036744	...	0.010723	0.004309	1.000000

Problem 3: Collaborative Filtering Implementation:

Step 1: Implementation: By using the Item-Item based approach predicted the ratings of the Test set.

Step 2: Execution and Evaluation:

compute the Mean Absolute Error and the Root Mean Squared Error for your predictions:

The Root Mean Squared Error-0.9438

Mean Absolute Error-0.7485

```
print( 'Summary results:')
Test['predictions'] = predictions
print ('  RMSE:' , round(np.sqrt( ((Test['Rating'] -
                                   Test['predictions'])**2).mean()),4))
print ('  MAE: ' , round(np.mean(abs(Test['Rating'] -
        Test['predictions'])),4), '\n')
```

```
Summary results:
  RMSE: 0.9438
  MAE:  0.7485
```

Step 3: Does your approach work for your own preferences

Added some movie ids with a customer id to my test set and predicted the ratings of movies. I have added 4 movie id's with a customer id with ratings. Out of 4 movie id's two ratings predicted match the true rating given.

Conclusion:

Recommendation algorithms provide an effective form of targeted marketing by creating a personalized shopping experience for each customer. For large retailers like Amazon, Netflix a good recommendation algorithm is scalable over very large customer bases and product catalogs, requires only sub second processing time to generate online recommendations, is able to react immediately to changes in a user's data, and makes compelling recommendations for all users regardless of the number of purchases and ratings. Unlike other algorithms, item-to-item collaborative filtering is able to meet this challenge.