Graph based Recommendation System

Movie Recommendation System

Lipsa Mishra
MMATH (MDSAI)
University of Waterloo,
Waterloo, Ontario, Canada
lmishra@uwaterloo.ca

ABSTRACT

In this project, I have implemented a movie recommendation system using User-Based Collaborative Filtering technique. The Movielens dataset is taken from the grouplens site. The recommendation system is implemented using GraphX, which is Apache Spark's API for graphs and graph-parallel computation. The goal of this project is to implement collaborative filtering using GraphX to recommend top 'n' movies to the user. Movielens dataset is represented as a weighted bipartite graph where edges from user to movies are weighted by rating. A graph based resource allocation algorithm is followed to produce similarity measures between every pair of users. Then the top 'n' movies are recommended to the user with the highest average rating.

1. INTRODUCTION

The amount of content in the internet like social media sites, shopping sites, movies, TV shows and so on is growing rapidly. Hence, it is very difficult for a user to decide what to select from these endless choices. We need a filtering system that would filter out the contents based on the user's preferences and choices, and help the user to decide what they want to watch. Recommendation systems are built to solve this problem. It is used in many applications to suggest products, services, and information to potential customers. It is mostly embedded in the business web-based system. It collects the key information (purchases, ratings, browsing activities, and so on), from the users, either from the sales transaction database or through a real-time system, to make recommendations. The data collected is analyzed by the recommendation system's engine, which designs a customer preference model. The recommendations are then presented to the customers to support marketing and purchasing decisions [1].

In this project, a User-Based Collaborative Filtering algorithm is implemented that uses existing data to give better suggestions to a user, based on other users having similar choices. A bipartite graph is built from the Movielens dataset to support graph traversal for collaborative filtering system.

KEYWORDS

Recommendation; collaborative filtering; similarity; graph representation; graph algorithms

2. DATA

The dataset is obtained from the website http://www.grouplens.org/. GroupLens Research has collected and made available rating datasets from the MovieLens website (http://movielens.org). The datasets were collected over various periods of time, depending on the size of the set [2]. In this project, the two main files that are used to build the

recommendation system using collaborative filtering are movies.csv and ratings.csv.

Movies.csv

The movies.csv has the following structure:

Movie ID	Title	Genre
1	Toy Story (1995)	Adventure | Animation | Children | Comedy | Fantasy
2	Jumanji (1995)	Adventure | Children | Fantasy
3	Grumpier Old Men (1995)	Comedy | Romance
4	Waiting to Exhale (1995)	Comedy | Drama | Romance
5	Father of the Bride Part II (1995)	Comedy

Ratings.csv

The ratings.csv has the following structure:

User ID	Movie ID	Ratings	Timestamp
1	1	4	964982703
1	3	4	964981247
1	6	4	964982224
1	47	5	964983815
1	50	5	964982931

For testing purposes, the model is first implemented on the smaller dataset which consists of 100,000 ratings of 1700 movies given by 1000 users. Finally, the model is tested on a larger dataset which consists of 1000,000 ratings of 4000 movies given by 6000 users.

Below are the two datasets for the testing purpose:

Dataset size	Number of users	Number of Movies	Number of Ratings
100k Dataset	1000	1700	100000
1M Dataset	6000	4000	100000

2.1 REPRESENTATION OF THE DATA

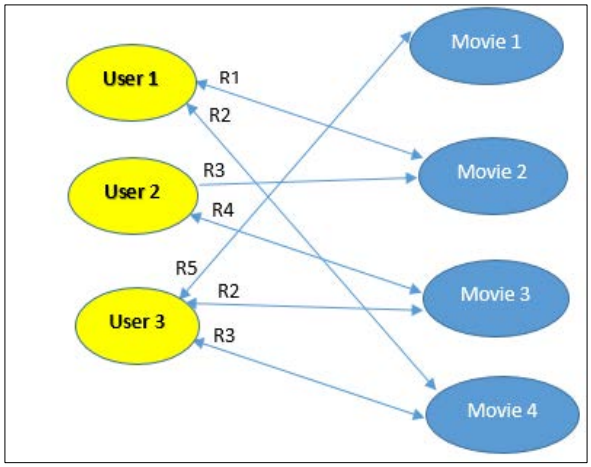
The Movielens dataset is transformed into a bipartite graph to implement collaborative filtering. GraphX is used to transform the given dataset into a bipartite graph. GraphX extends the Spark RDD by introducing a new graph abstraction: a directed multigraph with properties attached to each vertex and edge [3]. The bipartite graph is divided into two distinct set of vertices (Users and Movies). There can be a link or edge (ratings) between the vertices. But the link between the same set of vertices is inadmissible. In this project, the vertices are Users and Movies, and edge is the rating provided by the user to a given movie **[4]**. Below are the tabular and graphical representations of the data:

Tabular Representation of Data:

User / Movie	Movie 1	Movie 2	Movie 3	Movie 4
User 1		**Rating 1**		**Rating 2**
User 2		**Rating 3**	**Rating 4**	
User 3	**Rating 5**		**Rating 2**	**Rating 3**

The values in the above table represents that a rating is provided by the user for a given movie. The zero in the table represents that that User has not rated the given movie.

Graphical Representation of Data:



An edge between a user and a movie in the above graph represents that a rating is provided by the user for a given movie. If there is no edge between the user and a given movie, then the user has not rated the given movie [3].

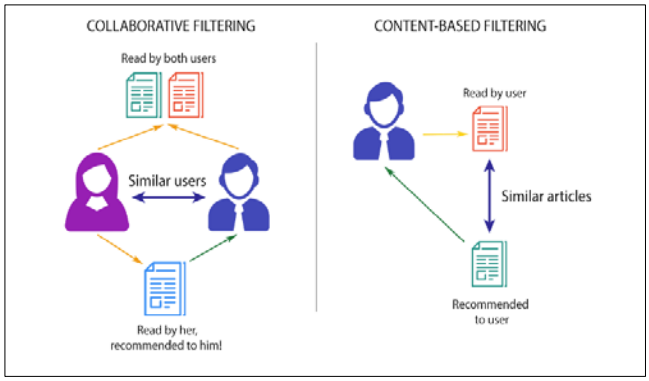
The actual data set has 4000 movies and 6000 customer. So the bipartite graph matrix is of size (4000 * 6000).

3. RELATED WORK

Below are the Filtering Methods that are used to build a recommendation System:

A. Content Based filtering

This filtering method is based on the description of an item and a profile of the user's preferred choices. In this system, keywords are used to describe the items. Besides, a user profile is built to state the type of item the user likes. In other words, the algorithms try to recommend products which are similar to the ones that a user has liked in the past. The idea of content based filtering is that if a user likes an item then he/she would also like a 'similar' item. This approach has its roots in information retrieval and information filtering research. But this filtering technique is not applicable to this project.



B. Collaborative Based filtering

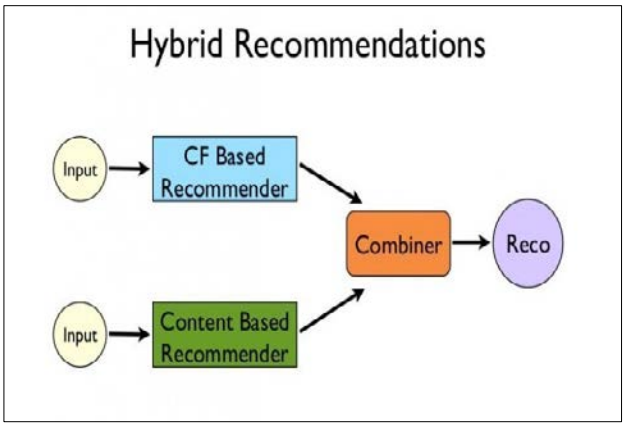
This filtering method is usually based on collecting and analyzing information based on users' behavior, their activities or preferences. It then predicts items based on the similarity with other users. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content, and thus, it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself. Collaborative filtering is based on the assumption that people who agreed in the past will agree in the future, and that they will like similar kinds of items as they liked in the past. In this project, collaborative filtering is used to recommend movies.

For example, if a person A likes item 1, 2, 3 and B likes 2, 3, 4, and if they have similar interests, then A should like item 4 and B should like item 1. This is the basic concept behind collaborative filtering.

C. Hybrid Based filtering

It is a combination of Content and Collaborative filtering. Hybrid approaches can be implemented by making content-based and collaborative-based predictions separately and then combining them.

For example, Netflix is a good example of the use of hybrid recommender systems. The website makes recommendations by comparing the watching and searching habits of similar users (i.e., collaborative filtering) as well as by offering movies that share characteristics with films that a user has rated highly (content-based filtering).



4. METHODOLOGY

In this project, Collaborative filtering is implemented on the Movielens dataset. It is basically of two types:

(1) **User based collaborative filtering:** User-based collaborative recommendation aims to calculate some similarity metric between all pairs of users and then predict a particular user U's rating for an item 'i' by collecting and processing the ratings of U's "neighborhood" (all the other users with high similarity as compared to user U) for item 'i' [5].

(2) **Item based collaborative filtering:** Item-based collaborative recommendation seeks to calculate some similarity metric between all pairs of items and then predict a particular user U's rating for an item (i) by collecting and processing the ratings of item(i)'s "neighborhood" (all the other items with high similarity as compared to "i" that user U has rated) [5].

In this project, user-based collaborative filtering is used. In this type of collaborative filtering approach, when a user tries to recommend a movie, then the algorithm tries to find other similar users who have watched almost the same movies as the current user. Here, cosine similarity metrics is used to find such similar users.

Cosine Similarity:

It is a method to measure the distance between two non-zero vectors in an inner product space. For user-based collaborative filtering, the two users' similarity is measured as the cosine of the angle between the two users' vectors [6].

For example, if two people b and c have similar movie preferences and we have ratings of two movies as below,

User/Movies	Iron Man (2008)	Pride and Prejudice (2005)
b	4	3
c	5	5

then each user's ratings can be represented in a separate vector:

$$\vec{b} = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \qquad \vec{c} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$

Cosine Similarity helps to measure the similarity between the two vectors b and c with the below formula:

$$similarity = \cos\theta = \frac{b.c}{\|b\|\|c\|}$$

$b.c \Rightarrow Is\ the\ Dot\ product\ of\ the\ two\ vectors$

$\|b\|\|c\| \Rightarrow Is\ the\ product\ of\ each\ vector's\ magnitude$

RECOMMENDATION ALGORITHM

1. Find the list of movies 'M' rated by a user U1 in the Movielens dataset.

2. Find the list of all users who rated at least one of these movies M.

- Using "aggregateMessage" property of GraphX

3. Find "P" different users, who have the most similar choice in movies as user U1.

- Using Densevector() to convert the ratings of the users to a vector.

- Using Cosine Similarity to calculate the similarity between the two vectors

4. Find the list of movies M, and it's count that user U1 has not rated yet but "P" users have rated them.

5. Finally, recommend the top 'n' movies by the "P" users to the user U1 with the highest average rating.

- Print score, movieId, title and genre, for the top 'n' movies

5. RESULTS

For 1 Million Movielens dataset:

Below are the top 10 movies that are recommended for User U1 in the Movielens dataset of 1000,000 ratings:

Top 10 recommended movies

```
#-recommended: 10
···· recommended movies ····
(4.609) [2905] Sanjuro (1962)  Action Adventure
(4.561) [2019] Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)  Action Drama
(4.555) [318] Shawshank Redemption, The (1994)  Drama
(4.525) [858] Godfather, The (1972)  Action Crime Drama
(4.521) [745] Close Shave, A (1995)  Animation Comedy Thriller
(4.517) [50] Usual Suspects, The (1995)  Crime Thriller
(4.510) [527] Schindler's List (1993)  Drama War
(4.508) [1148] Wrong Trousers, The (1993)  Animation Comedy
(4.491) [922] Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)  Film-Noir
(4.478) [1198] Raiders of the Lost Ark (1981)  Action Adventure
```

6. EVALUATION

As our recommendation system algorithm recommends top 'n' movies, there is a need to evaluate both the relevance and the order of the movies. In the Movielens dataset, the ratings given by each user isn't available for every movie and hence the recommended movies might not have a target label present to evaluate its relevance. I came up with an alternate approach to evaluate, in which I used the additional attribute of genre, which wasn't used by the algorithm

to validate the claim. The following steps are required to evaluate the metric "Precision of Preferred genre":

1. Find the genre most frequently rated by the user i.e his/her preferred genre.

2. Out the top P movies recommended to the user, find the percentage of the movies which are of the preferred genre.

The percentage relevance for the Movielens data was 40%.

```
percentage relevance of user is (40.000)
```

7. CONCLUSION

The evaluation metric used has some limitations such as not considering contribution of all genres that are rated by the user and being immaterial to priority of recommendation. These above limitations along with smaller values of P chosen (10), are the cause for the lower value obtained. The primary focus of this project was on design decisions of framework chosen and learning of that framework (GraphX). Even after the choice of a strict evaluation, the recommendations were still relevant indicating the efficiency of the algorithm. Future work along this would focus on coming up with a more relevant and fair evaluation framework to measure the performance of the algorithm, which considers the prioritized vector of genre and relevance of recommendation discounted similar to that of NDCG@K.

REFERENCES

[1] Zhou, T., et al., Bipartite network projection and personal recommendation, Physical Review E, 2007. 76(046115)

[2] GroupLens Movielens Dataset (2019) Social Computing Research at the University of Minnesota. https://grouplens.org/datasets/movielens/

[3] ApacheSpark – GraphX- GraphX Programming Guide: http://spark.apache.org/docs/latest/graphx-programming-guide.html

[4] Shang, M., Fu, Y., Chen, D., Personal Recommendation Using Weighted BiPartite Graph Projection, Apperceiving Computing and Intelligence Analysis, 2008. ICACIA 2008. International Conference on vol., no., pp.198 - 202, 13-15 Dec. 2008

[5] Pinela, Carlas (2017), Recommender Systems — User-Based and Item-Based Collaborative Filtering https://medium.com/@cfpinela/recommender-systems-user-based-and-item-based-collaborative-filtering-5d5f375a127f

[6] Bellogin, A., & Parapar, J. (2012). Using graph partitioning techniques for neighbor selection in user-based collaborative filtering. In Proceedings of the sixth ACM conference on Recommender systems (pp. 213–216).