CHAPTER 7

RECOMMENDATION FOR MOVIELENS

7.1 INTRODUCTION

The web is a variety of information, changing rapidly with high volume and velocity, evolving big data. This information source makes overloading and requires new form of processing in knowledge discovery, decision making and process optimization. Big data analytics find hidden, useful patterns, unknown correlation and search trends. Information retrieval systems handle big data scenario by following user centric approach with personalized recommendation techniques. Personalized recommendations opens a new set of challenges and opportunities in Big data technology that exposes user with a huge collection of items. Recommender Systems (RS) implementation increases rapidly in E-commerce. The RS problem addresses information overloading issue by suggesting similar items to similar users. The abundant application domain causes several consequences, and the ever growing big data encompasses specific algorithm for a specific domain. This chapter presents a discussion of information abundance in RS perspective.

The RS is now found in many modern online applications to provide users with a list of recommended items they prefer. In general, the RS tales item’s detail, user’s details and rating details as input to recommend new items to the user. The RS must be able to infer new user’s requirements from past interactions. The recommendation task is designed as a prediction task which includes user interest, item’s details, their ratings and search context.
The RS system maps user with item by measuring similarities as utility. The similarity indexing and dimensionality reduction techniques reduce search cost. In the early era recommender system is started with a content based model in which past purchased items are considered for prediction. The present generation of RS moves towards a context based approach which makes use of integrated devices such as sensors for location based prediction. This chapter discusses about recommendation problem, general recommendation architecture with its algorithm. Moreover, this chapter discusses about item based CF technique with MovieLens 100K dataset (http://grouplens.org/datasets/movielens/100k/).

7.2 RECOMMENDER PROBLEM

The recommender problem defines a utility function that predicts the rating of items for the new user from the ratings of existing users. The utility function is defined on the rating matrix $U \times I$. In real time this matrix is sparse, even the user rates for a small subset of available items. The definition of recommender problem is given as follows.

Let $U$ as set of users such as $U = \{u_1, u_2, u_3 \ldots u_n\}$ and $I$ as set of items such as $I = \{i_1, i_2, i_3 \ldots i_m\}$ then

Find a utility function $UT$ that stands for the usefulness of $i_1 \ldots i_n \in I$ to user $u_a \in U$, such that

$UT$ finds $\{r_1, r_2, r_3, \ldots r_n\}$ for items $i_1 \ldots i_n \in I$

Where $\{r_1, r_2, r_3, \ldots r_n\}$ are the ratings for items $i_1 \ldots i_n \in I$

Which derived from user_item matrix

Hence, for each user $u_a$, Recommender system needs to choose items from $i_1 \ldots i_n \in I$ that maximize $UT$. 
A good RS is to be scalable in terms of user and items. The RS needs to cover entire items for recommendation, diverse and privacy preferable. The CF techniques are followed by RS in order to find recommendations from rating dataset. Similarity metrics are used to find correlation between users and items. Pearson correlation coefficient and spearman correlation coefficient are commonly used to construct rating matrix.

7.3 GENERAL RECOMMENDATION ARCHITECTURE AND ALGORITHM

The RS suggests items that best match with user need. Some RS are designed with the combination of content based and CF based techniques that are referred as Hybrid Recommendation Systems. Content based filtering method follows the relationship between content of the items and user preferences and it is suitable to recommend texts and web pages whose contents are abundant and easy to analyse. CF technique suggests recommendation based on the interest of other like minded user. The general architecture for RS is given in Figure 7.1.

![Figure 7.1 General RS architecture](image-url)
The general RS algorithm takes standard users data and their ratings and incoming new users rating and predicts best top n recommendations to the new user. The general form of recommendation algorithm is as given below.

Input:  
New User u, ratings of New User ur (rating data)  
Standard users data U, rating data UR (rating dataset)  

Output:  
Top N Recommendation for new User  

Begin:  

// clustering: Model based approach  
if ( ur > Threshold)  
User Profile Set S = \{ identify similar profile of u with every Ua , where Ua ⊆ U \}  
Form Clusters of S  
For each item of u find Average Ranking score of UR in same cluster  
Find Top N Recommendation  

// Collaborative Filtering: Memory based approach  
Take rating data of New user ur, Standard User UR  
Construct Rating Matrix ( User User or Item Item)  
Apply Similarity Measures (Pearson correlation / Cosine / Adjusted cosine)  
Find Top N Recommendation  

End
The RS follows memory based or model based approach to find recommendations (Cheung & Tian 2004). The CF is a memory based approach which tries to find active users for the new user. This algorithm is not as fast to generate real time recommendations due to scalable issue for very large datasets. Hence, the model based approach builds a data mining model such as clusters, bayesian networks for the dataset. This model offers the benefits of speed and scalability since this model avoids referring of dataset each time.

7.4 MOVIELENS 100K DATASET

MovieLens 100K dataset consists of 100,000 ratings of 943 users on 1682 movies. This dataset is downloaded from http://grouplens.org/datasets/movielens/100k/. In this dataset each user has rated at least 20 movies. This dataset consists of movie details and user’s demographic details such as age, gender and occupation. This dataset is available in different sizes such as 100K, 1M, 10M and 20M. MovieLens dataset are collected by GroupLens Research Project at the university of Minnesota. This dataset is collected during September 1997 to April 1998 in MovieLens website.

7.5 ITEM BASED COLLABORATIVE FILTERING

CF is one of the most important techniques used in RS. The CF refers to a class of techniques used in RS that recommend items to users that other users with similar tastes have liked in the past (Xiaoyuan & Khoshgoftaar 2009). There are two types of CF techniques such as user based and item based. The user based technique works by comparing user similarities based on the pattern of rating of items whereas the item based technique works by comparing item similarities based on their pattern of ratings across users (Morid et al. 2014). The CF algorithm makes predictions based on neighbors of relevant users. The basic idea is that people who agreed
in subjective opinion in past behavior are likely to appear similar in future. 
The knowledge of choosing efficient and effective recommendation technique is important for a RS to provide good useful recommendation to individual users. The item based CF works such a way that each the items that have not rated by u are taken as y. The items which are already rated by u are taken as x. The similarity between x and y are measured by pearson correlation coefficient (Owen et al. 2011). Then, based on the similarity score find the rating for items y by averaging the rating of x. This technique works based on the similarity between users to be predicted with the items which are rated by training users. Items average rating does not change frequently in this approach. Pearson correlation is the most commonly used similarity score in RS construction and it is represented in Equation (7.1).

\[
sim(x, y) = \frac{n\hat{A}_{xy} - (\hat{A}_x)(\hat{A}_y)}{\sqrt{n(\hat{A}_x^2 - (\hat{A}_x)^2)(n\hat{A}_y^2 - (\hat{A}_y)^2)}}
\]  

(7.1)

Where

\[n\] = number of ratings

\[\sum xy\] = sum of products of x and y ratings

\[\sum x\] = sum of x’s ratings

\[\sum y\] = sum of y’s ratings

\[\sum x^2\] = sum of squared x’s ratings

\[\sum y^2\] = sum of squared y’s ratings

Pearson correlation is a statistical formula that measure similarity between two items. The variables x and y stands for the ratings of item x and y. The item y is suggested to the user if the similarity score of x and y are positive where x is a personalized item for the user. Here, the meaning is user has already rated x. The system recommends y to him.
7.6 RESULTS AND DISCUSSION

This section presents the results that are obtained by experimenting item based collaborative filtering technique on MovieLens 100K dataset. The pearson correlation is used as the similarity measure. The experiment is done in two levels. The first level uses entire items in the dataset for rating matrix construction. The second level finds similar active users U for the new user u by checking their gender and occupation. The grouping of users based on gender and occupation provides a utility to reduce the size of the rating matrix. The RMSE comparison in presents of grouping and in absence of grouping is shown in Figure 7.2. This grouping reduces RMSE value for the items which are rated by more than 300 users. The promising results show the benefits of grouping in recommendation and the reduction of RMSE which is an evaluation score to prove the efficiency of recommender system.

![Figure 7.2 RMSE comparison](image)

Figure 7.2 RMSE comparison
In case of items whose ratings are more than 300, the prediction of RMSE for items in presents and absence of grouping shows similar results. Their RMSE value does not have appreciable decrease. When the number of ratings are less than 300, the grouping approach shows appreciable variation in RMSE value.

The sample of predicted rating scores for new user u and their corresponding Movie IDs is presented in Table 7.1. The dataset which is stored in knowledge base collector in Chapter 6 can be used to make the proposed dynamic personalization system as dynamic personalized recommendation system in future.

**Table 7.1 Prediction rates of item based collaborative filtering**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>3.0105</td>
<td>3.2224</td>
<td>3.1094</td>
<td>2.3764</td>
<td>3.2722</td>
</tr>
<tr>
<td>679</td>
<td>2.7546</td>
<td>3.4627</td>
<td>3.4801</td>
<td>3.2145</td>
<td>3.8063</td>
</tr>
<tr>
<td>500</td>
<td>3.6436</td>
<td>3.5101</td>
<td>3.6429</td>
<td>2.5042</td>
<td>3.2675</td>
</tr>
<tr>
<td>700</td>
<td>3.0351</td>
<td>4.1525</td>
<td>4.1979</td>
<td>3.5</td>
<td>4.3954</td>
</tr>
</tbody>
</table>

The sample of predictive rating for UserID 237 for 1682 movies is given below.

2.8744, 3.1651, 2.8593, 2.9796, 2.5836, 2.4399, 2.7813, 2.4805, 2.7183, 3.0321, 3.0471, 3.0985, 2.8401, 3.0065, 3.105, 2.8798, 3.1887, 2.9805, 3.3219, 3.0419, 2.2515, 2.4587, 2.1087, 2.3424, 2.5276, 2.248, 2.0546, 2.2752, 2.7662, 2.3986, 3.0642, 2.0171, 2.6226, 2.7516, 3.1939, 2.9519, 3.0664, 3.0713, 2.4534, 3.0376, 1.7314, 3.3252, 2.1765, 2.46, 2.9417, 2.7827, 3.1514, 2.6921, 2.8067, 2.9545, 1.1167, 1.9969, 2.4498, 1.6271, 1.7517, 1.9928, 1.1711, 1.5733, 1.647, 1.4756, 2.4906, 2.743, 2.9863, 2.6885, 2.4953,
These values are the predicted ratings of 1682 movies for UserID 237. This section details the principle of item based recommendation and presents the prediction result. Similarly the recommendation systems can be built with user based, Singular Value Decomposition (SVD) based CF approaches.

7.7 SUMMARY

This chapter discusses the overview of recommender system with RS problem, general architecture and RS algorithm. Performance tuning of recommender system is an upcoming research attempt to find better algorithm and to design crucial characteristics of them. This chapter presents the building of RS with CF and grouping technique with results. This chapter addresses the robustness and transparency of RS technique with appreciable results.