IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10455

Sparsity and Matrix Factorization in Recommender System

Dilip Yadav[1], Tushar Vaghasiya[2], Savita Ravate[3], Prof. Vinit Raut[4]

Student, Computer Engineering, Viva Institute of Technology, Mumbai, India[1]

Student, Computer Engineering, Viva Institute of Technology, Mumbai, India[2]

Student, Computer Engineering, Viva Institute of Technology, Mumbai, India[3]

Professor, Computer Engineering, Viva Institute of Technology, Mumbai, India[4]

Abstract: Recommender systems play a vital role in engaging a user on a platform by recommending similar new items based on users interest. Some of the tech giants like amazon use recommendation systems to recommend new items based on users previous search or purchases, Netflix uses recommendation systems to recommend new movies to users based on their interest. In past few years due to the availability of huge amounts of data, it is difficult to find information that is useful and relevant. Collaborative filtering, Content-based filtering, Hybrid filtering are some of the techniques used by recommendation systems. Few issues that exist in recommendation systems are data sparsity problem and cold start problem. This paper proposes a method to address the issue of sparsity in SVD based approaches. Data sparsity refers to problems in finding similar users because users only rate a few items. Cold start refers to difficulties in generating accurate recommendations for users who have rated very small items.

Keywords: Matrix Factorization (MF), Singular Value Decomposition (SVD), Sparsity, Collaborative Filtering (CF)

I. INTRODUCTION

In recent years, due to the availability of enormous amounts of data, it has become increasingly challenging to find information that is appropriate and useful. Recommendation systems were needed to solve this problem. Recommendation plays a vital role in increasing a user engagement on a particular platform. Several popular techniques used in recommendation systems are Collaborative filtering, Content-based filtering, knowledge-based, and hybrid recommender systems.

In content-based methods, Item descriptions are used to make the recommender system, term content refer to descriptions. It used features of items to make similarities between items. the item descriptions, which are labeled with ratings and other futures, are used as training data to create a user-specific classification or regression modeling problem. Content based methods have some advantages for making recommendations for new items. It will use keywords in description for new item ones come in. For specific items, rating data is not sufficient and not available here content-based can be used. This is because other items with similar features might have been rated by the user. Content-based method is effective of providing recommendations for new items but not effective in user recommendation. This is because the training model for the specific user needs to use the history of her ratings. In fact, it is usually important to have a large number of ratings available for the specific user in order to make robust predictions without overfitting.

Collaborative-based filtering takes into consideration the collaborative ability of ratings provided by multiple users to make recommendations. The main challenge in building a collaborative-based recommendation system is the sparsity of the matrix. In a movie-based recommendation system, most of the users would have viewed a small fraction from a large universe of available movies. This results in most of the ratings being unspecified and causes the problem of sparsity. Sparsity leads to weak predictions. The unspecified ratings can be imputed as the observed ratings are highly in correlation between multiple users and items. Collaborative filtering is further divided into memory-based and model-based methods. Memory-based methods are also termed as neighbourhood-based collaborative filtering algorithms. The neighbourhood can be computed in two ways user-based and item-based. In user-based collaborative filtering, the ratings of like-minded users of target user u_i are used to make recommendations for u_i. The similarity are calculated between the rows of the rating matrix. In item-based collaborative filtering, similar items of Ii are determined to make predications. Similarity functions are computed on the column of rating matrix. The Advantages of memory-based techniques is that they are easy to implement.

Matrix factorization technique belongs to a class of collaborative filtering algorithms. Matrix factorization algorithms decompose the user-item interaction matrix into the product of two matrices that are lower in dimensions. Matrix

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10455

factorization generates the latent features when multiplying two different kinds of entities. Collaborative filtering is considered as the application of matrix factorization to identify the relationship between items and users entities.

II. PROPOSED METHOD

Collaborative filtering provides recommendations to the users but it suffers from data sparsity problem. To overcome sparsity problem, a new data creation method is being proposed. The Fig 1 represents the flow diagram of the system.

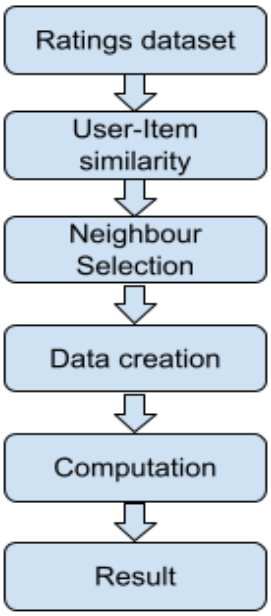

Fig. 1 System Flow Diagram

A. Ratings dataset

Ratings Matrix is created using the MovieLens 100k dataset which contains 706 users and 8572 movies. The dataset is converted into ratings matrix which contains m users and n movies and the value as ratings provided by the users. The dataset contains ratings in range 0.5-5.0. The ratings are provided from 1995 to 2016.

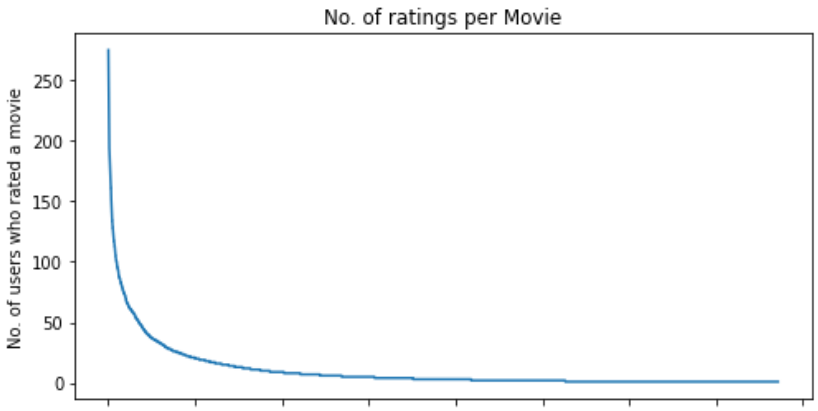

Fig II. Ratings Distribution

Fig II represents the ratings distribution of the ratings by the users. The main reason of the sparsity is that only few movies are rated by most of the users while many movies remained unrated. The dataset is about 99.91% sparsity of the dataset is calculated using the given below formula.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10455

$$\text{Sparisty} = 1 - \frac{Number\ of\ ratings\ provided}{Total\ Number\ of\ users\ and\ Movies}$$

B. Similarity Calculation

Cosine Similarity is used to calculate the user-user and item-item similarity. It measures similarity between the two non-zero vectors using inner product space. It equals to cosine of angle between them, which is same as inner product and normailzed to have lenght 1. Cosine of 0^o is 1, and less than 1 fot angle in interval $(0, \pi]$.

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \ \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2}\ \sqrt{\sum_1^n b_i^2}}$$

C. Neighbour Selection and Data Creation

To reduce the sparsity the data imputing method is created. After calculating the user-user and item-item similarity top 5 neighbous are selected for both the pair and median is calculated for that pair which is then imputed in the respective entity of the ratings matrix.

D. Computation

For Computation SVD(Singular Value Decomposition) is used. SVD is a matrix factorisation technique, which is used to reduces the number of features of a dataset by reducing the space dimension. In the recommender system, the SVD is used as a collaborative filtering technique. It uses a matrix structure where each row represents a user, and each column represents an item. It finds factors of matrices from the factorisation of a high-level user-item matrix. The singular value decomposition decomposes a matrix into three other matrices.

$$A=USV^T$$

Where A is a utility matrix of shape m x n which shows the relationship between users and items, U is a m x r singular matrix, which represents the relationship between users and latent factors, S is a r x r diagonal matrix, which represents the strength of each latent factor and V is a r x n singular matrix, which represents the similarity between items and latent factors. The latent factors describe the characteristics of the items, for example, the genre of the movie. The SVD decreases the dimension of the utility matrix A by extracting its latent factors. It represents the mapping between each user and each item into a r-dimensional latent space.

E. Result

To measure the prediction accuracy in CF root mean squared error(RMSE) is used.

$$\text{RMSE}=\sqrt{\frac{\Sigma_{(i,j)\in T} \quad (r_{ij}-\hat{r}_{ij})^{\ 2}}{|T|}}$$

Where $r_{u,i}$ is the actual user rating u for item i, r_{ui} is the expected user rating u for item i and T is the number of predicted values. The smaller the size, the greater will be the accuracy of the prediction.The square root of the ratio of the square of the deviation of the predicted value from the original value to the number of observations is known as RMSE. For large parts of the prediction RMSE needs high requirements on the stable of the experimental method.

F. Accuracy Analysis

Table I provides the analysis that without adding new data, the system performs low as compared to the method of adding new data into the system.

TABLE 1: COMPARATIVE RESULT

Approach	Accuracy
SVD(Without New Data)	RMSE: 0.8915163569066167
SVD(With New Data)	RMSE: 0.1337213941439405

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10455

III. CONCLUSION

In the past few years, recommender systems have made significant growth in various content-based, collaborative, and hybrid methods. Collaborative filtering is one of the outstanding strategies for recommendations, but it suffers from problems like data sparsity problems. The implemented system reduces the sparsity thereby increasing the accuracy of prediction. The sparsity is reduced by imputing the data generated into the sparse matrix, then the data is fed into a matrix factorization algorithm to generate the predictions. The implemented system outperforms the traditional approach of collaborative filtering when the dataset is sparse. The accuracy of prediction depends on the similarity measure chosen to calculate the similarity, using the more efficient similarity metrics may improve the accuracy of prediction.

REFERENCES

[1]. Charu C Aggarwal. An introduction to recommender systems. In Recommender Systems, Springer, 2016
[2]. Yehuda Koren, Robert Bell, Chris Volinsky, "Matrix Factorization Techniques for Recommender Systems", Computer 42 (8), 30-37, 2009
[3]. I. Mirza & Suharjito, "Film Recommendation Systems using Matrix Factorization and Collaborative filtering.", International Conference on Information Technology Systems and Innovation, 2014, (pp. 1-6).
[4]. Z. Sharifi, M. Rezghi & M. Nasiri, "A New Algorithm for Solving Data Sparsity Problems Based-On Non-Negative Matrix Factorization in Recommender Systems", 4th International Conference on Computer and Knowledge Engineering, 2014, (pp. 56-61).
[5]. A. Mochamad, H. Arief and Z. K. Baizal, "Preprocessing Matrix Factorization for Solving Data Sparsity on Memory-Based Collaborative Filtering", 3rd International Conference on Science in Information Technology, 2017, (pp. 521-525).