IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 2, February 2019

Movie Recommendation Engine using Collaborative Filtering with Alternative Least Square and Singular Value Decomposition Algorithms

Rohan Mhetre[1], Dr. Priya G[2]

School of Computer Science, VIT University – Vellore[1,2]

Abstract: Recommender system is a process or approach used for filtering information from a vast dataset and predicting the choices to the users in the areas they are mostly interested in. This system nowadays is the backbone for the commercial aspect of an industry and is established in a variety of areas including movies, music, videos, web-pages, e-commerce, services etc. In this paper the focus is on movie recommendation and the technique currently present for this is the collaborative filtering technique. Of the collaborative filtering techniques, the matrix factorization algorithms namely Alternative Least Square and Singular Value Decomposition are implemented to predict or recommend the movies. Further to improve the processing and time computation for a large dataset we have used Apache Spark along with Elastic search and the accuracy is compared between the two algorithms for different values of testing subsets.

Keywords: Collaborative Filtering, ALS, SVD, Apache Spark, Elastic Search

I. INTRODUCTION

Recommendation systems have been on the peek nowadays as they suggest relevant information which the user typically desires from a vast number of choices**.** The main purpose of the recommendation system is to predict or provide suggestions to the users according to their previous history for a particular attribute. This gives user satisfaction and with that. These systems are usually built considering the user or customer perspective providing them the suggestions for the items they will be most interested in and eliminating the choices or items which the user or customer will be least interested. These recommendation systems work for a particular domain such as music, movies, news, web pages, e-commerce, services, etc. Number of choices are available for use on the Internet and there should be something to filter or prioritize the relevant information or content and reduce the problem of information overload which worries the Internet users. [1] Recommender systems tend to solve this issue by prioritizing only the needed information from a large dataset. This system minimizes the set of choices for a user and allows them to discover choices that could very well suit them. On the larger picture this provides customer loyalty and satisfaction and the industries get to know the preferences of a particular user which they can very well cluster and use for further promotions.

In the online multimedia platform, movie recommendation is one of the most popular and widely used, which predicts and recommends the movies to the users based on their previous history or their taste for a particular attribute. Main two techniques used for recommendation systems are content based filtering and collaborative filtering. Content-based filtering is also known as cognitive filtering and works by establishing correlation between the user and his/her preferences. [2] Here only the user’s previous history is focused as to what movies the user has watched. But the limitation of content-based filtering is the limited data present or analyzed. Collaborative filtering can solve this limitation. Here similarity between user-user and item-item will be considered. [3] This technique makes use of a user-item matrix which is build considering the preferences of every other user. It basically builds a database of user-item matrix of preferences for items given by users after which the users with similar tastes are matched by calculating the similarities and finally recommending them choices relevant to their preferences. Unlike content-based filtering the user gets recommendation of a movie even though he/she have not rated the movie. This recommendation is based on a prediction value which is numerical, so the highest predicted scores for a movie as given as recommendations to the user.

DOI 10.17148/IJARCCE.2019.8216

This approach uses collaborative filtering to achieve the goal. It is an efficient technique which makes automatic predictions for users. The limited content analysis is the drawback of content-based filtering which will be overcome by collaborative filtering. The alternating least squares (ALS) algorithm is an algorithm for collaborative filtering. ALS algorithm is used for implementing collaborative filtering and dealing with missing values. Based on the similarity score it will recommend the movie to the user.

II. LITERATURE SURVEY

This paper tends to implement different collaborative algorithms and thereby combining them and using a hybrid filtering approach. Also stating some of the limitations for the collaborative filtering technique and the advantages of hybrid filtering technique. Further Spark is used for stripping down time interval. Main algorithms analyzed and used are Tanimoto, Pearsons, Slope and SVD Algorithms. [4]. [5] Similarity between two users is found out by using the Euclidean distance and the prediction value is calculated for those users. Pearson correlation suggests how close two sets of data are similar. The limitations are that Pearson correlation and Euclidean distance are not suitable or rather efficient when we consider large dataset. This paper also implements K-Nearest Neighbor approach as KNN can mostly store the historical data and thereby helping the correct recommendation.

[6] This paper focuses on the preprocessing area for the data for improving effectiveness and accuracy of collaborative filtering. The concept of vector space model is used to differentiate the weights for different items for each customer who purchase similar products regularly. The for large number of items this wont really make a difference and can be the drawback. [7] A hybrid approach is proposed in this paper. This approach is combination of content-based and collaborative filtering techniques. [8] Also propose a hybrid system known as content-boosted collaborative filtering system. Content-based technique is used to populate the prediction values in the sparse user-item matrix built by the collaborative filtering technique. Advantages and disadvantages of both content-based and collaborative filtering techniques are described in this paper.

[9] Item based collaborative filtering is proposed in this paper. Firstly, the user-item matrix is built and then the relationships between different items are identified and then the recommendation is provided to the user. This method is more efficient when the number of users is far more than the number of items otherwise the limitations are data sparsity, cold start and shriller attacks for the new users. [10] A Probabilistic relational model is proposed which works on the problem of data sparsity for the training data matrix. A predicted value can be filled in the sparse matrix, this predicted value can be found out by using one of the following namely Mean Square Differences algorithm, Vector Similarity algorithm, etc. [11] This paper makes use of natural language processing to convert the dataset into proper format. This method focuses on a particular feature and then calculate the prediction score based on that. For users who have watched less movies, feature selection can increase the number of recommendations successfully. [12] The feature selection specified in this paper are rating the genre. Recommending the movies based on genre will increase the accuracy of a recommendation system. The drawback here is that it will become one dimensional prediction.

[13] This paper applies inductive learning algorithms for the recommendation systems. A decision tree is constructed to represent the user preference. Collaborative filtering is not transparent, and this approach solves that problem. The algorithms used are C4.5 and CART. The drawback here is preprocessing time is increased for a new item added. [14] Primary objective of this approach is to suggest a recommendation through computer intelligence and clustering the data, hence the data is partitioned on similarities and then intra-cluster similarities can be used for prediction. The drawback here is if the initial partition is weak then the efficiency is vastly decreased. [15] This paper proposes the use of machine learning to be incorporated in the recommendation systems. Various machine learning algorithms are discussed along with various content-based algorithms.

[16] This paper briefs the matrix factorization techniques that can be used in the recommendation systems. The advantages of these techniques are discussed over the classic nearest-neighbor techniques. [17] Pictorial data like frames and posters can be used for the prediction purpose. The attractiveness of the pictorial representation can be handy for the recommendation systems. [18] This paper puts forth the analysis of ALS algorithm and the advantages of matrix factorization algorithms for generating implicit feedback. The main advantage of using ALS is the accuracy and speed of training time over the tradition collaborative filtering algorithms. [19] SVD algorithm is analyzed in this paper and how SVD splits user item rating matrix into two matrices with lower ranks and how the overfitting of data is avoided and also the sparsity problem can be solved.

DOI 10.17148/IJARCCE.2019.8216

III. PROPOSED SYSTEM

The proposed system is built once with using the Alternative Least Square algorithm and the other time with the Singular Vector Decomposition algorithm. In this approach the dataset is firstly loaded into Apache Spark. Spark being a library for machine learning helps to train the dataset more efficiently thereby reducing the processing time of the queries. [20] Spark.ml currently supports collaborative filtering, in which users and items are described by a small set of latent factors that can be used to predict missing entries. Spark gives the power of quick computation for the large dataset. The dataset then is cleaned the pre-processed accordingly using the library functions. We are using elastic search for retrieving the results quickly. The movie dataset loaded into spark and then after cleaning the dataset using it is loaded into elastic search using machine learning libraries. Elastic search is the distributed search and analytic engine, it stores the data centrally based on the index. The index can be any attribute of a dataset. It helps in computation of personalized user choices and similar movie recommendation with search, indexing and filtering.

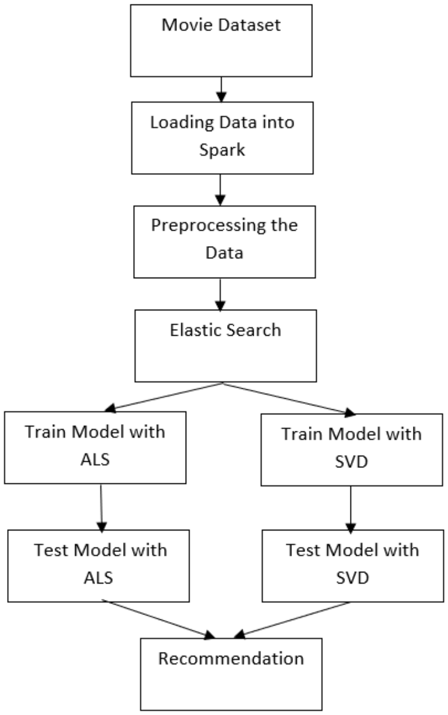


Fig.1: Movie Recommendation Engine Framework

The dataset is then trained using Alternative Least Square algorithm as well as Singular Vector Decomposition algorithm respectively. Spark.ml contains the ALS algorithm and it is one of the commonly used algorithms. ALS works on the latent factors and regularization parameters.

$$f(\mathrm{U,M})=\sum_{(i,j)} w_{i,j}\,(r_{i,j}-u_i\times m_j)^2+\lambda\left(\sum_i n_{u_i}||\,u_i||^2+\sum_j n_{m_j}\,||m_j||^2\right)$$

The first summation represents the completion term followed by the cost function included in the first bracket. Regularization term is the one followed by the lambada. Here n represents the rating given by users with respect to a latent factor. The various parameters like how much user likes a particular latent factor, how much each movie scores according to some latent factor, weight matrix, number of ratings available for a user are also defined. For SVD a given matrix is divided into the best lower rank approx., the given matrix consider R is divided into two unitary matrices U and V and one diagonal matrix Sigma.

$$R=U\sum V^T$$

IJARCCE

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

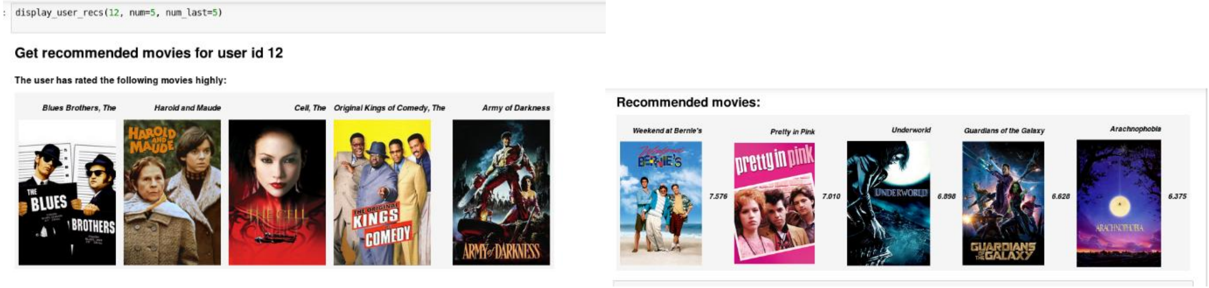
International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 2, February 2019

Where R is the user rating matrix, U is the user features matrix, V is also the user features matrix and Sigma is the diagonal matrix of singular values. Here U and V are orthogonal matrices and the meaning of user features matrix is how relevant each feature is to each movie. The dataset is trained and tested using the SVD algorithm. Finally, we compare both the algorithms based on the accuracy as the dataset is very huge. This data set is 10M MovieLens dataset which contains approx. 10000054 ratings and 95580 tags applied to 10681 movies by 71567 users of the online movie recommender service MovieLens. [21]

IV. RESULTS and DISCUSSIONS

Datasets and tools used for this experimentation are as follows, to perform recommendation on movies MovieLens dataset is used, it contains anonymous rating of movies. MovieLens 10M dataset contains that much amount of records. [21] The dataset is made available on GroupLens Research webpage. Jupyter notebook is the open source web application. Using Jupyter notebook we can create the code. Jupyter notebook is used for data visualization and machine learning and performing the code. Python is programming language using python we can integrate our system. Apache spark is used for the large data processing while HDFS used for storing large amount of data that data is processed using the spark. Cluster management is done in spark.



```
display_user_recs(12, num=5, num_last=5)
```

Get recommended movies for user id 12

The user has rated the following movies highly:

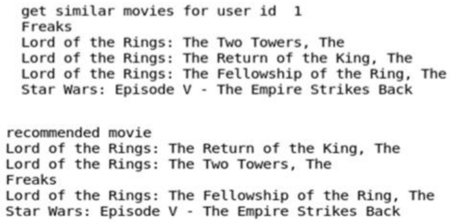
Blues Brothers, The   Harold and Maude   Cell, The   Original Kings of Comedy, The   Army of Darkness

Recommended movies:

Weekend at Bernie's 7.576   Pretty in Pink 7.010   Underworld 6.898   Guardians of the Galaxy 6.628   Arachnophobia 6.375

Fig.2: Movie Recommendation using ALS algorithm

Here given the userId the system will predict the user with movies based on his/her preferences.



```
  get similar movies for user id  1
  Freaks
  Lord of the Rings: The Two Towers, The
  Lord of the Rings: The Return of the King, The
  Lord of the Rings: The Fellowship of the Ring, The
  Star Wars: Episode V - The Empire Strikes Back

recommended movie
Lord of the Rings: The Return of the King, The
Lord of the Rings: The Two Towers, The
Freaks
Lord of the Rings: The Fellowship of the Ring, The
Star Wars: Episode V - The Empire Strikes Back
```

Fig.3: Movie Recommendation using SVD algorithm

Here given the userId the system will predict the user with movies based on his/her preferences.

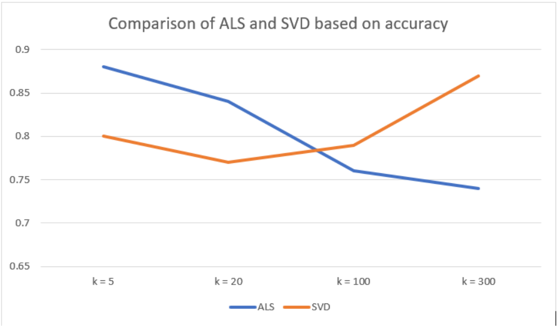


Fig.3: Comparison of ALS and SVD algorithms based on accuracy for different training data

DOI 10.17148/IJARCCE.2019.8216

Here k is the number of training set, the graph shows that for smaller values of k ALS works more accurately but as the size of k increases relatively the ALS's accuracy comes down compared to SVD, whereas SVD works more efficiently on large datasets.

V. CONCLUSION AND FUTURE WORK

This paper mainly focuses on effective calculations in python and using spark libraries, utilizing Jupyter and elastic search saving data and index automatically for faster computations. ALS algorithm is used for implementing collaborative filtering and dealing with missing values. SVD algorithm is used to manipulate the matrix and obtain the best results by predicting the sparse matrix values. Altogether Apache Spark and Elasticsearch are used for better computation, processing and retrieving the huge amount of data. The futuristic work is avoiding cold start problem, because cold start problem is the limitation of collaborative filtering approaches.

REFERENCES

[1]. F.O. Isinkaye, Y.O. Folajimi – "Recommendation systems: Principles, methods and evaluation", Egyptian Informatics Journal – 2015
[2]. Robin van Meteren, Maarten van Someren – "Using Content-Based Filtering for Recommendation", NetlinQ Group- 2000
[3]. Muh Hanafi, N. Suryana – "Paper survey and example of collaborative filtering implementation in recommender system", Journal of Theoretical and Applied Information Technology – 2017
[4]. Howal S., Mote A., Vanjari R., Desai V. - "Movie Recommender Engine Using Collaborative Filtering". 46th ISTE National Convention and National Conference Journal of Advance Research Innovation – 2017
[5]. Mishra D. P., Mukharjee S., Mahapatra S., Mehta A. – "Analysis of Movie Recommender System using Collaborative Filtering", International Journal of Recent Trends in Engineering and Research - 2017
[6]. Jonghoon Chun, Sang-goo Lee – "A Preprocessing Method for Improving Effectiveness of Collaborative Filtering", ResearchGate - 2003
[7]. George L., Caravelas P. – "A hybrid approach for movie recommendation", Springer Science + Business Media, LLC - 2006
[8]. Hande R., Gutti A., Shah K., Gandhi J., Kamtikar V. –Moviemender-A MovieRecommender System", International Journal of Engineering Sciences and Research Technology – 2016
[9]. Lakshmi T. P., Sreenivasa D. P., Siva N. N., Srikanth Y. - "Movie Recommender system using item based collaborative filtering technique", IEEE – 2016
[10]. Patel H. M., Shah J. B. – "Collaborative Filtering Approaches for Movie Recommendation System Using Probabilistic Relational Model", International Journal of Advance research and Development - 2015
[11]. Zehra C., Mahiye U.– "Feature selection for movie recommendation", Turkish Journal of Electrical Engineering & Computer Sciences- 2016
[12]. Ashwani Kumar Singh, P. Beaulah Soundarabai – "International Journal of Advanced Research in Computer and Communication Engineering- 2017
[13]. Peng, L., Yamada S. – "A Movie Recommender System Based on Inductive Learning", Conference on Cybernetics and Intelligent Systems Singapore, IEEE – 2004
[14]. Verma O.P., Katarya R. - "An Effective Collaborative Movie Recommender System with Cuckoo Search", Egyptian Informatics Journal - 2017
[15]. Bhumika Bhatt, Premal J Patel. – "A Review Paper on Machine Learning Based Recommendation System", International Journal of Engineering Development and Research - 2014
[16]. Robert Bell, Yehuda Koren – "Matrix Factorization Techniques for Recommender Systems", Cover Feature- 2009
[17]. Lili Zhao, Zhongqi Lu – "Matrix Factorization+ for Movie Recommendation", International Joint Conference on Artificial Intelligence - 2016
[18]. Balazs Hidasi, Domonkos Tikk – "Speeding up ALS learning via approximate methods for context-aware recommendations", Springer- 2016
[19]. Youchun Ji, Wenxing Hong – "Regularized singular value decomposition in news recommendation system", 2016 11th International Conference on Computer Science & Education (ICCSE) - 2016
[20]. Spark Programming Guide - Spark 1.6.0 Documentation, http://spark.apache.org/docs/latest/programming-guide.html
[21]. http://grouplens.org/datasets/movielens/10m/

DOI 10.17148/IJARCCE.2019.8216