Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

Product Recommendation System using Scalable Alternating Least Square Algorithm and Collaborative Filtering using Apache Spark in E-Commerce

Bineet Kumar Jha[1], Sivasankari G.G[2], Venugopal K.R[3]

[1]Research Scholar (AMC Engineering College, VTU), CMR Institute of Technology, Bengaluru, India

[2]Professor, Department of Computer Science and Engineering, AMC Engineering College Bengaluru, India

[3]Vice-Chancellor, Bangalore University, Bengaluru

Email: yoursbineetjha@gmail.com[1] (corresponding author), shivashankarigg@gmail.com[2], venugopalkr@gmail.com[3]

ABSTRACT`

Recommender System is tremendously used in numerous spaces, such as e-commerce and entertainment to enhance businesses by increasing the chance of sales. Earlier researches have focused more on traditional Machine Learning (ML) and Artificial Intelligence (AI)-based approaches. Developing a scalable recommender system has been challenging concerning high availability and fault tolerance. The traditional collaborative filtering approach used with the recommender system also faces challenges due to the absence of explicit product ratings by the customer and the cold start problem. We have proposed a scalable Alternating least square (ALS) and collaborative filtering-based approach for the recommender system. The experimental results of the proposed hybrid approach show improved performance as compared with the traditional approach.

Keywords

Alternating Least Square; Collaborative Filtering; Recommendation System; Spark Framework

Introduction

The product recommender system has got a lot of popularity in recent years. Personalize recommendations save user search time and increase the chance of a sale. It is an emerging area of research with the growth of e-commerce. A product recommender system uses various techniques to recommend products based on user activities such as - user browsing history, frequently bought items, recently viewed items, the customer who bought this also bought other items, bestselling products, product bought in a bundle, highest rated items. The personalized recommendation of individual users varied based on their purchase behaviors. A good recommender system improves e-commerce businesses around the globe. According to VentureBeat the Amazon's recommender system has boosted the sale by 35%. The recommender system is used in various e-commerce and social platforms such as as- Amazon, eBay, Netflix, YouTube, Hulu, etc.

Wang and Zhao [2] have proposed a learning algorithm for a recommendation system, based on OMTCF for a higher accuracy level. Users' temporal information can be used to determine their behavior and it will help do analytics. The problem related to time drift is addressed using TimeSVD++ algorithm proposed by Korenv[3]. A similar approach is proposed by Ling, Yang and King [4], in that SGD-RMF/DA-RMF algorithm is used to determine the dynamic changes of users. A random walk model-based TrustWalker algorithm is proposed by Jamali and Ester [5] to solve the problem of users' interest.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

In this paper, we have proposed a parallel recommendation algorithm on the Apache Spark framework. It includes collaborative filtering based on users' shopping. The spark ALS model is used for collaborative filtering. Before the model is trained the user information is fed inside the model. The Collaborative Filtering (CF) algorithm is implemented in the Spark MLlib. Apache stream is a widely used framework for stream analytics. It has distributed processing power to improve accuracy and speed up the computation. The Spark computational framework uses in-memory computation which gives the product recommendation in quick time.

The Recommender algorithm depends on various approaches such as- content-based filtering, collaborative filtering, Bayesian network, association rules, etc. The collaborative filtering depends on memory that is utilized to find historical data of existing clients and the closest target neighbor. To train and predict the model, the collaborative filtering approach can be used. Figure 1 shows the classification of the Recommender system.

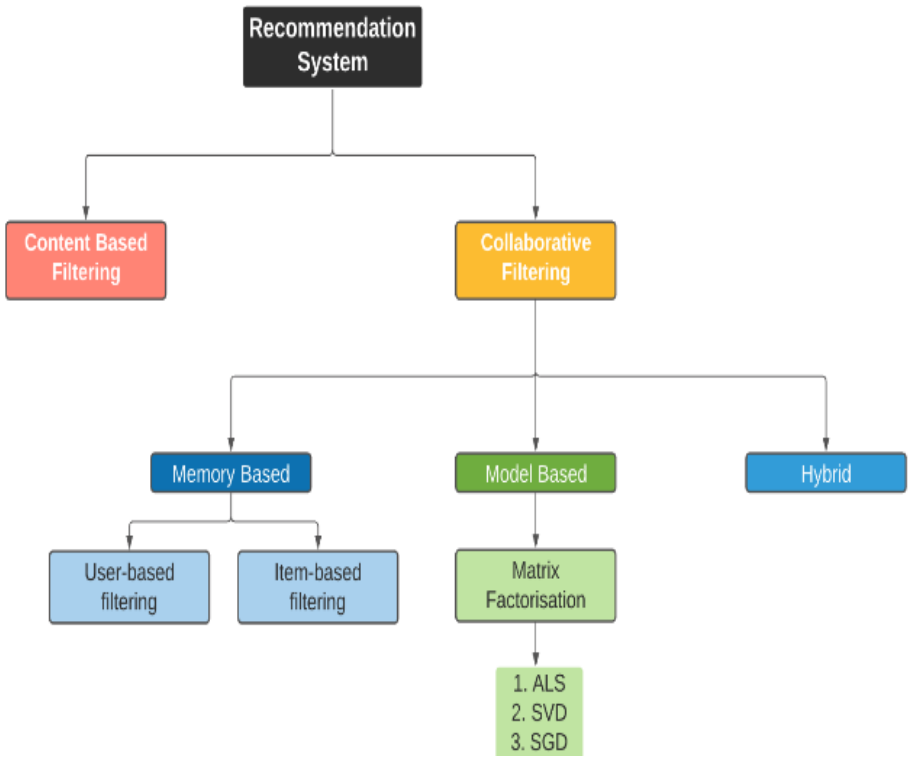


Figure 1. Classification of Recommender System

In recent times the Spark framework has been used to a greater extent as it provides in-memory computation power that is useful for fast iterative calculation. The Spark framework even has the capabilities of handling historical data. It provides fast and universal analysis support. Compared to the Map-Reduce framework it has better performance in terms of scalability and quick response. It provides various computational frameworks such as- SparkSQL, Streaming Spark, MLBase, and GraphX. Streaming Spark provides the stream processing that is useful to handle real-time large streaming data. GraphX is a parallel graph computing framework. The MLib is providing all supports to implement a Machine Learning application is distributed environment. SparkSQL is a

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

SQL query engine to handle structured data. Figure 2 shows a typical Apache Spark platform for data processing and analysis. The framework relies on HDFS for organizing data whereas Hadoop Yarn is used for resource allocation.

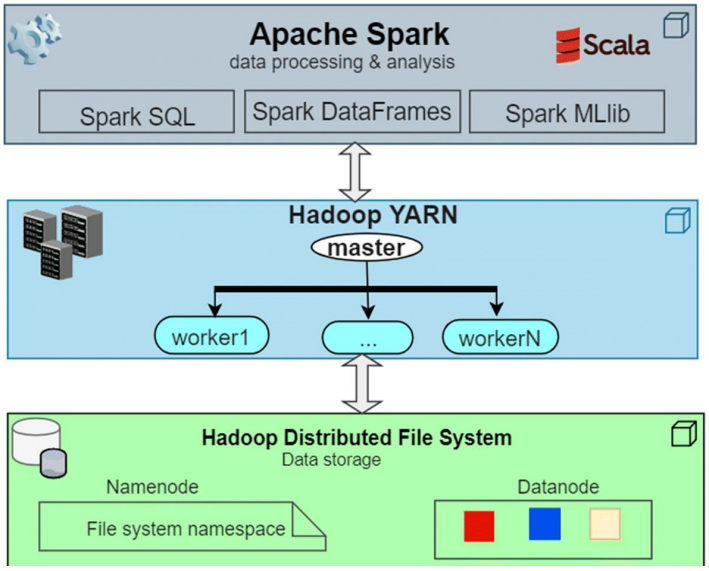


Figure2. Big data frameworks

In this research work, we have proposed a parallel and scalable recommender system based on ALS algorithm and collaborative filtering, runs on the Spark platform in cloud environment. The parallel implementation of the recommender algorithm includes- collaborative filtering based on users, content, and the ALS model.

Related Work

Riyaz P A et al. [6] proposed a product recommendation system to resolve the scalability issue. In this paper Collaborative filtering technique is used in the Hadoop platform. The performance of the system has significantly increased with increased data nodes. An improved ALS-based collaborative filtering approach is used to improve the performance of the recommender system [15]. In this approach, the similarity data between user and product are put together. Mustafa Fatih Cetin and Serkan Ayvaz [7] proposed a hybrid recommender system based on negative similarity. To enhance the accuracy and scalability a cloud platform can be used. The Elastic Map-Reduce services are offered by Amazon's AWS cloud platform [8] that can be used for the product recommendation system. In this proposed approach an advanced version of ALS is used with matrix factorization.

Sachin Gulabrao Walunj and Kishor Sadafale[9] have addressed the scalability issues in the product recommendation system and proposed an Apache Mahout platform to resolve the issue. Developing a real-time application to meet the challenges associated with large data is one of the challenging tasks faced by programmers. Implementing and deploying scalable algorithms needs a proper framework and API support. The machine learning-based PredictionIO is an open-source server that can be used to develop and deploy scalable applications [10]. It has rich APIs that help to deploy and test the learning algorithms effectively and easily. The traditional recommender systems have

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

lower performance in real-time applications. Such applications also have lower efficiency and scalability issues. These challenges can be resolved using Singular Value Decomposition (SVD) methods on Apache Hadoop and Spark platforms. The TrustSVD algorithm has shown a significant improvement in performance in terms of RMSE [11]. In the proposed work the RMSE value reduces a lot with an increase in the dataset size. Apache Spark can do a similar task 100 times faster than the Hadoop Map-Reduce. In another research work, Ruo Huang et al. [12] have proposed a novel recommender system using an attention-based long-term and short-term model (ASLM) based on Recurrent Neural Network (RNN) that has a preference and attention layers.

Proposed Product Recommendation System

The tremendous growth in data on e-commerce platforms leads to scalability issues. A spark-based framework is used to resolve scalability issues.

A) System Architecture

Our approach consists of obtaining the hidden patterns based on the user's activities on an e-commerce site. We have used Amazon's online shopping site to gather customer activities. The gathered data is processed and analyzed in Google™ Colaboratory environment.

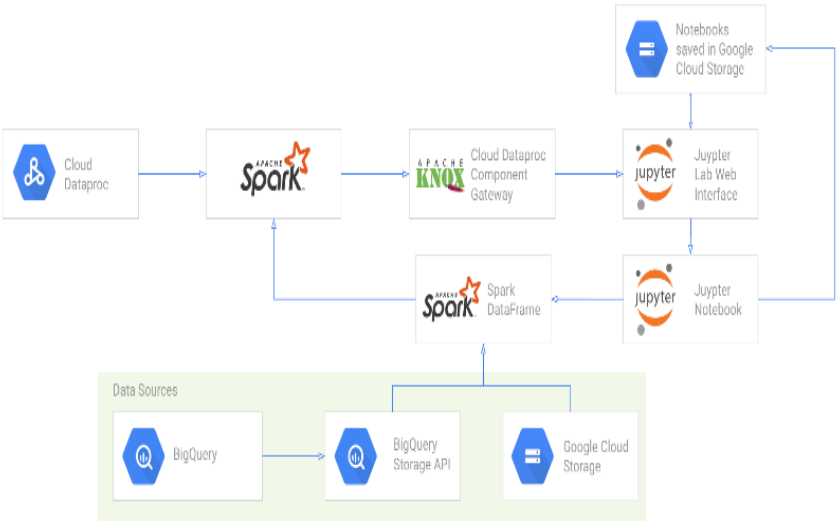


Figure3. Spark Processing Architecture

The Google compute engine in the backend process the task with allocated 6.83GB RAM, and 31.97GB disk space. The number of products analyzed is 491192 which took 25 sec of CPU time. Figure 4 shows the Tensor Processing Unit (TPU) for the Tensor Flow framework.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

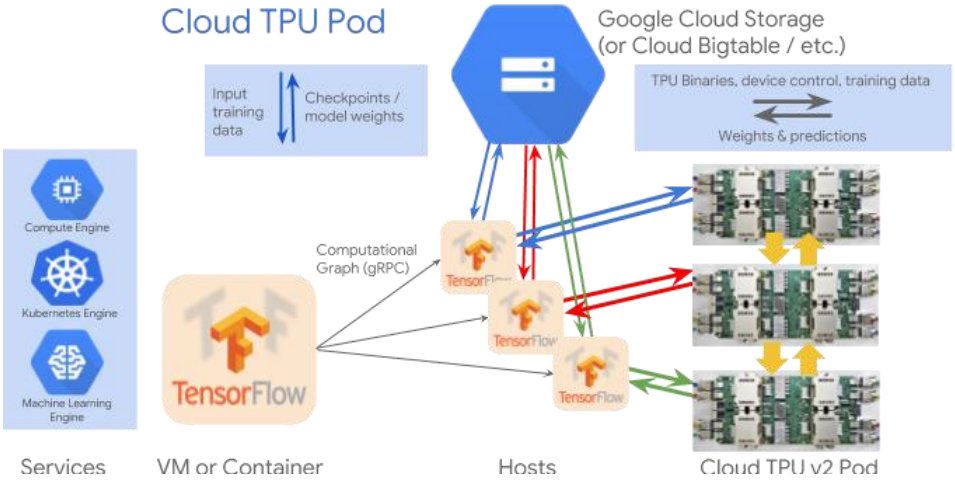


Figure 4. TPU Pods in Google Data Center

B) Spark Computing Framework

Spark context is useful for the Spark functionality that allows the application to access the Spark cluster with the help of the cluster manager. Work node runs an application in the cluster. Figure 5 shows a Spark processing flow chart.

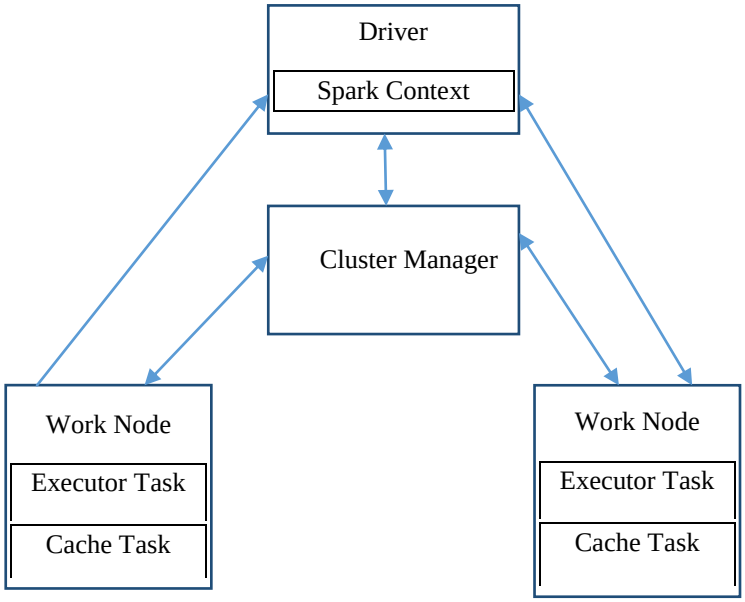


Figure 5. Spark Processing Flow Chart

C) Optimizing Hybrid Recommendation Algorithm

The collaborative Filtering algorithm captures the interaction between users and products. Resnick et al. [14] described Collaborative Filtering for product recommendation based on the nearest neighbor principle. The users have different scores based on their interaction with various products.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

The main algorithm of collaborative filtering used for the recommendation is based on the item recommendation algorithm. It works in the following steps-

a) Gather users purchase behavior related historical data
b) The similarity between the objects is calculated to determine the customer preferences, based on Tanimoto Similarity.

$$sim(x,y) = \frac{\sum_{u=U}^{\cdot} (W_{a,u} - \overline{r_a})(r_{b,u} - \overline{r_b})}{\sqrt{\sum_{u\in U}^{\cdot} (r_{b,u} - \overline{r_b})}\sqrt{\sum_{u\in U}^{\cdot} \left(W_{a,u} - \overline{r_a}\right)}} \quad (1)$$

In that $U = \{u_1, u_2, u_3, ...u_n\}$ is the set of users. And the item set is denoted as $I = \{i_1, i_2, i_3, ...i_n\}$ R is an evaluation matrix of *m x n;* $sim(x,y)$ is the similarity between user x and y.

c) Based on the user's item selection preference obtain the ***K*** item set with high similarity indices. The user ***u*** preferences are the neighbors in item set ***I.***The already recommended items are not recommended again therefore the candidate recommendation item set is removed from the list when user ***u*** prefers such items. The user ***u*** calculates the interestingness of the candidate recommendation set, such as the formula is given below-

$$p(u,i) = k\sum_{v\in S(i,K)\cap N(u)}^{\cdot} W_{j,i} r_{u,i} \quad (2)$$

The similarity calculation of collaborative filtering can be also calculated as

$$sim(u,v) = \frac{\overline{u}.\overline{v}}{|\overline{u}\,||\,\overline{v}\,|} \quad (3)$$

Here $\overline{u}$ and $\overline{v}$ are the sore vector of items u and v **,** respectively. There are many other ways to calculate the similarity based on the Manhatten distance and hash method. These methods have their advantages and disadvantages.

$W_{j,i}$, represents the similarity between the items in item-set I.The R_u represents the preferences of user u's for item *I*, and *K* denotes the normalization factor. In that N_u denotes the set of user preferences. The Top N represents the interest degree obtained in the above formula.

d) For Top N recommendation, the Jaccard formula is used-

$$sim(x,y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|} \quad (4)$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

Where $N(x)$ and $N(y)$the individual user has its Top N recommendation sequences therefore first 2 of all users make a new recommendation sequence.

The weighted factor is set in the weighted recommendation if the new user D solves the cold boot problem. The amount of user preference item in the new sequence is used to determine the weighting factor.

$$\alpha = 1 - \frac{T_u}{N} \tag{5}$$

Where $T_u$represents the number of users in suggested sequences of new preference things,N is the total number of sequences. The old user preference is recommended when the weighting factor is lower to achieve a higher preference score of the items. The Mean Absolute Error (MAE) measures the accuracy of the algorithm is calculated as-

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |P_i - a_i| \tag{6}$$

Where P_i is the ratings predicted by the recommender algorithm and a_i is the actual rating given by the users. The smaller value of MAE indicates the higher accuracy levelof recommendations.

D) Collaborative Filtering using Improved ALS algorithm in the Spark Framework.

The matrix decomposition in collaborative filtering algorithm includes ALS and SVD. Matrix $R = (R_{ij})^{m\times n}$represents the rating of n items by m users.The loss function is given as-

$$L(U,V) = \sum_{ij}^{\cdot} (R_{ij} - X_{ij})^2 \tag{7}$$

In Equation (7) $(R_{ij} - X_{ij})^2$ is the error found commonly in approximation.

To reduce the loss function the Equation (8) can be written as-

$$L(U,V) = \sum_{ij}^{\cdot} (R_{ij} - U_i V_j^{\,T})^2 \tag{8}$$

To avoid the overfitting issue a regularization function is added in Equation (8). The modified function is shown as-

$$L(U,V) = \sum_{ij}^{\cdot} (R_{ij} - U_i V_j^{\,T})^2 + \lambda \left(\|U_i\|^2 + \|V_i\|^2\right) \tag{9}$$

The ridge regression is used to predict each line of U and vice versa. Equation (10) shows the mathematical formulation of that.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

$$U_i = R_i V_j^T (V_{xi}^T V_{xi} + \lambda n_{xi} I)^{-1}, i \in [1, m] \quad (10)$$

In thatR_i denotes the score vector by user i, V_{xi} represents the number of items evaluated by the users. In a similar way derivative of V is taken while fixing U.

$$V_j = R_j^T U_{mj} (U_{mj}^T U_{mj} + \lambda n_{mj} I)^{-1}, i \in [1, n] \quad (11)$$

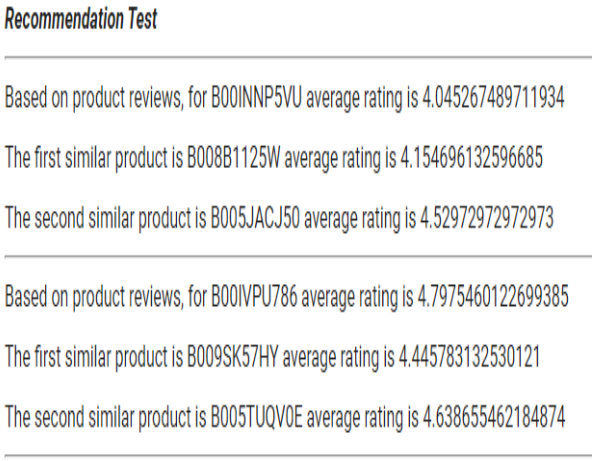
In thatR_j is the score vector of an item by user j andU_{mj} is the characteristics vector of items evaluated by user j. The improved algorithm is shown using Pseudocode-

Algorithm ***ImprovedALS()***
Input: vectors R, V,U, I and W
Output: Similarity, Regression line
for x ← 1 to N – 1 do
for y ← x + 1 to N do
compute sim(x, y) using Equation (1)
end for
end for
for x ←1 to M – 1do
for y ← x + 1 to M
Compute sim(x,y) using Equation (4)
end for
end for
For i ← 1 to m do
Fixed V
Compute U using Equation (10)
Fixed U
Compute V using Equation (11)
end for

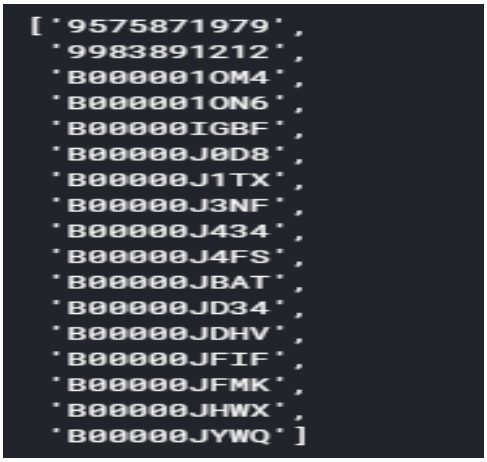
Experimental Result Analysis and Discussion

The quality of the recommendation system can be evaluated based on accuracy and coverage. Accuracy is the ratio between the correct recommendation and overall recommendations. The matrices to evaluate accuracy are classified into two major categories- Statistical Accuracy and Decision support accuracy matrices. The Mean Square Error (MSE), Root Mean Square Error (RMSE) and Correlation are used to evaluate the accuracy of the recommender system. The decision support accuracy matrices can be evaluated based on F-measure, Recall and Precision. The k-nearest neighbor algorithm is used to find the most related product. The Figure 6(a) shows the related products and their average ratings.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.



(a) Most Related Product



(b) Top recommended product IDs

Figure 6. Most related and Recommended Product

The model-based collaborative filtering is used to recommend the top 24 products from all using Truncated SVD.Figure 6(b) shows the top- recommended products.Table 1 shows the efficiency of the proposed model in terms of accuracy score and Mean Square Error(MSE).

Table 1. Performance Evaluation

Algorithm	Accuracy	MSE
Improved ALS (Regression & NLP)	0.95	0.11
Item-based collaborative filtering (K-Nearest Neighbour)	0.83	0.17
Model-based Collaborative Filtering	0.81	0.33

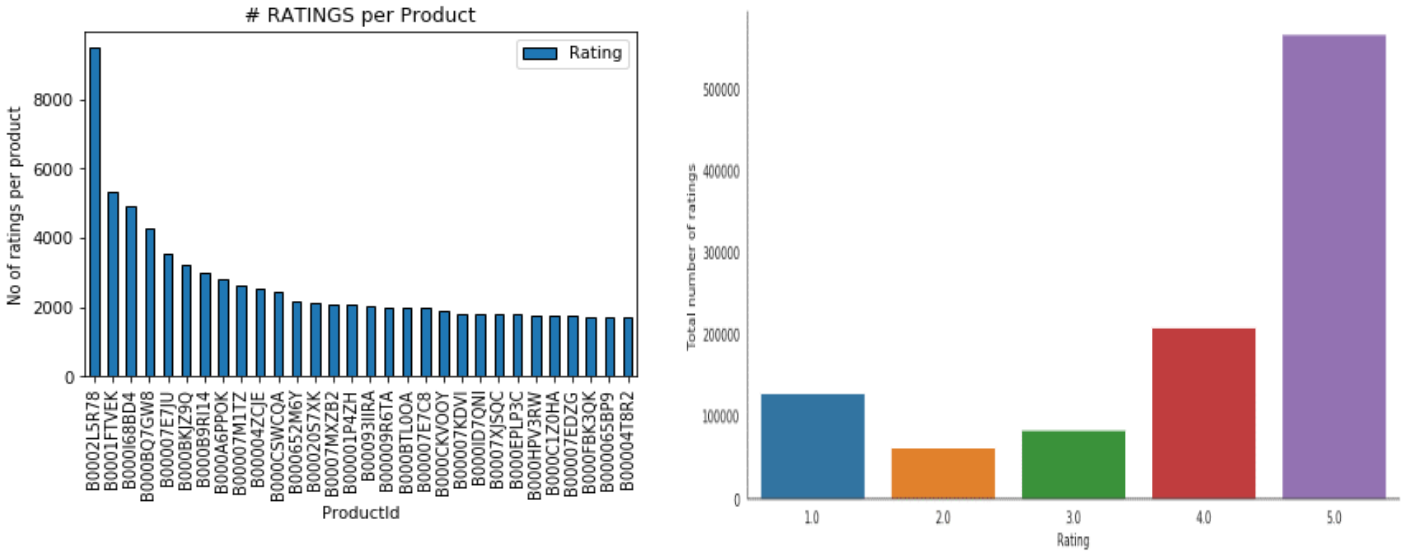
The Item-based collaborative filtering model performance can be evaluated in terms of precision, recall, f1-score and support. Table 2 shows the performance matrix for Item-based collaborative filtering.

Table 2. Performance Matrix

	precision	recall	f1-score	support
3	0.40	0.50	0.45	34
4	0.92	0.88	0.90	208
accuracy			0.83	242
macro avg	0.66	0.69	0.67	242
weighted avg	0.84	0.83	0.83	242

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

The Amazon electronic store dataset is collected and analyzed for the product ratings. Figure 8 shows the ratings of various products.



(a) Product wise ratings

(b) Rating type and number of ratings

Figure 8. Product Ratings

Figure 9 shows the tradeoff between Mean Absolute Error (MAE) and the number of items using collaborative filtering.

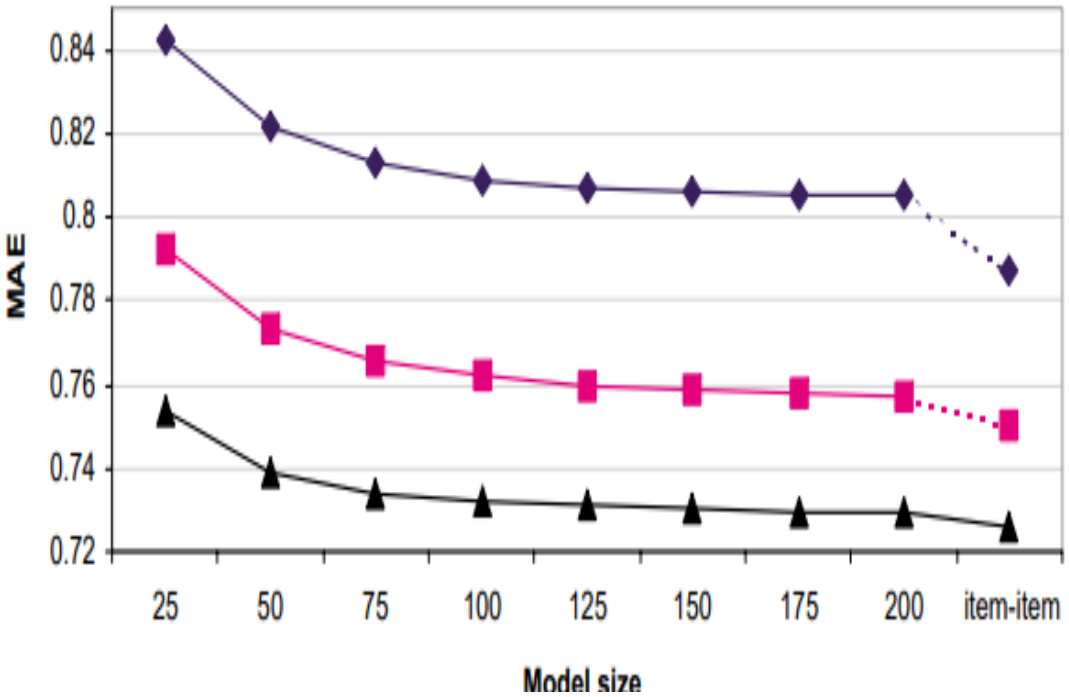


Figure 9. MAE in Collaborative Filtering

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2611 – 2622
Received 05 March 2021; Accepted 01 April 2021.

Conclusion and Future Work

In the proposed research work we have used the scalable ALS algorithm with collaborative filtering for the Amazon product recommendation. We have evaluated the performance of the recommender system using various approaches. In the proposed research work we have used PySpark based ALS library to implement a recommender system using collaborative filtering. To develop the recommender system we have collected a dataset from Amazon electronic store. We have also evaluated the performance of different collaborative filtering and ALS based recommender system in the Google cloud environment. The improved ALS approach has better accuracy as compared to Item-based and model-based collaborative filtering. We have also examined the various performance parameters for the recommender system mathematically. In future work, we can implement a recommender system using REST APIs in cloud environment to collect and analyze real-time review using Spark framework.

Acknowledgement

We are thankful to our research center at the Department of Computer Science and Engineering, AMC Engineering College, Bangalore for providing excellent resources and support to complete the research work.

REFERENCES

[1]. Liu, J. G., T. Z h o u, B. H. Wang. Research Progress of Personalized Recommendation System. *Regress in National Science*, Vol. 19, 2009, No 1, pp. 1-15.

[2]. Wang, J., P. Zhao. Online Multi-Task Collaborative Filtering for On-the-Fly Recommender Systems. *In: Proc. of 7th ACM Conference on Recommender Systems*, New York, ACM, 2013, pp. 237-244.

[3]. Koren, Y. Factorization Meets the Neighborhood: A Multifaceted collaborative Filtering Model. *In: Proc. of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, ACM, 2008, pp. 426-434.

[4]. Ling, G., H. Yang, I. King. Online Learning for Collaborative Filtering. *In: Proc. of 2012 International Joint Conference on Neural Networks*, IEEE, 2012, pp. 1-8.

[5]. Jamali, M., M. Ester. TrustWalker: A Random Walk Model for Combining Trust-Based and Item-Based Recommendation. *In: Proc. of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York*, ACM, 2009, pp. 397-406.

[6]. Riyaz P A, Surekha Mariam Varghese. A Scalable Product Recommendations using Collaborative Filtering in Hadoop for Bigdata. *In:Proc. Of International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST)*, 2015, pp. 1393-1399.

[7]. Mustafa Fatih Cetin, SerkanAyvaz. A Negative Similarity Based Hybrid Recommender System Using Apache Spark. *In:Proc. of the 2019 3rd International Conference on Advances in Artificial Intelligence(ICAAI)*, October 2019, Pages 166–172, https://doi.org/10.1145/3369114.3369152

[8]. L. Chen, R. Li, Y. Liu, R. Zhang and D. M. Woodbridge. Machine learning-based product recommendation using Apache Spark. *In:Proc. of the2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City*

Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, 2017, pp. 1-6, doi: 10.1109/UIC-ATC.2017.8397470.

[9]. Walunj, Sachin & Sadafale, Kishor. (2013). An online recommendation system for e-commerce based on apache mahout framework. *SIGMIS-CPR '13: Proceedings of the 2013 annual conference on Computers and people research*, May 2013, pp. 153–158, https://doi.org/10.1145/2487294.2487328

[10]. Ujwal U J, Dr. Antony P J, Sachin D N. Ecommerce User Data Analysis and Product Recommendation Using PredictionIo. *International Advanced Research Journal in Science, Engineering and Technology (IARJSET)*, Vol. 3, Issue 9, September 2016, pp. 153-157.

[11]. S.N. Patil, S.M. Deshpande, Amol D. Potgantwar. Product Recommendation using Multiple Filtering Mechanisms on Apache Spark. *International Journal of Scientific Research in Network Security and Communication (IJSRNSC)*, Vol.5, Issue.3, pp.76-83, 2017.

[12]. Thom Lake, Sinead A. Williamson, Alexander T. Hawk, Christopher C. Johnson, Benjamin P. Wing. Large-scale Collaborative Filtering with Product Embeddings.*arXiv:1901.04321v1 [cs.IR] 11 Jan 2019*

[13]. Ruo Huang, Shelby McIntyre, Meina Song, Haihong E and Zhonghong Ou. An Attention-Based Recommender System to Predict Contextual Intent Based on Choice Histories across and within Sessions.*Appl. Sci.2018,8, 2426; doi:10.3390/ app8122426*

[14]. Resnick P., Iacovou N., Suchak M., Bergstrom P. and Riedl J. GroupLens: An open architecture for collaborative filtering of net news. *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work*, October 1994 Pages 175–186, https://doi.org/10.1145/192844.192905

[15]. Li Xie, Wenbo Zhou, Yaosen Li. Application of Improved Recommendation System Based on Spark Platform in Big Data Analysis. *Cybernetics and Information Technologies*, Volume 16,2016, No 6, pp. 245-255.