Effective and Scalable Recommendation Model Combining Association Rule Mining and Collaborative Filtering In Big Data

Manikandan R, Ramesh R, Saravanan V

Abstract*: Due to the huge volume of information over the internet, The process of retrieving apt information is becoming more and more challenging. Many researchers have been carried out to sort this issue and the recent ones include Recommender Systems that are intelligent enough to predict the apt information and web pages that an user is anticipating. Collaborative Filtering is the well known method of any Recommendation model but the it has major drawbacks such as scalability and accuracy. The presented work is intended to combine the CF and association rule mining which is generically used for Big data, The aim of the research is to give a Recommendation model that is more scalable and accurate. We have taken the personalized e-book recommendation model that takes the previous users' browsing pattern.

Index Terms – Recommendation Collaborative filtering Association rule mining, Big data

I. INTRODUCTION

The increase in the size of data in internet is exponentially increasing with the number of users in the internet. The first ever prediction of the ddta growth was when the information explosion was felt[1] [4]. More than ninety percent of the available data in the internet is added in a span of just five years [1]. The sources of data production is more and more increasing which includes Social media , repositories, government organizations and many which obviously increase the size of data available in the internet. The inability of the old and novice Relational data base management to handle the growing data is felt.. Data that are normally measured in Terabytes and more are called as big data. When there is no availability of technology to collect and process, the data sets are termed Big data. [2] [3]. The enormous growth in the count of the users and the information makes a single end RSs a failure. These are normally in capable of meeting the huge computational needs which eventually result in more time consumptions. The scalability is also doubted [4] [10]. A variant architecture is needed to handle the case. RSs are videly used in almost every domain to increase the production by acquiring the Business intelligence of the domain and making suitable recommendations to the user in midst of enormous data available.

Revised Manuscript Received on March 25, 2019.
Manikandan R , College of Enginering , Trivandrum , Kerala, India
Ramesh R , Computer Science, Sri Krishna College of Arts and Science
Saravanan V, Computer Applications, Sri Venkateswara College of Computer Applications and Management

The generic RSs have two common apaches. Content Based filtering and Collaborative filtering. The Content Based method do not require all the characteristics of the users and it can recommend based on only the unique characters of an user. This has some serious issues in terms of Scalability, data sparsity and the cold start problem. The users would give ratings for only a few items or only a few users would have given ratings to a particular product or service. Either way affects the performance of the RSs. The personalized recommendation, The dynamically increasing users and items will face Cold start problem as the data would be insufficient for the Collaborative filtering to work efficiently. The scalability refers to the non working of an RS in real time. This may happen due to the increase in number of web users and the data sparseness [6]. The generic RSs have two common apaches. Content Based filtering and Collaborative filtering. The Content Based method do not require all the characteristics of the users and it can recommend based on only the unique characters of an user. This has some serious issues in terms of Scalability, data sparsity and the cold start problem. The users would give ratings for only a few items or only a few users would have given ratings to a particular product or service. Either way affects the performance of the RSs. The personalized recommendation, The dynamically increasing users and items will face Cold start problem as the data would be insufficient for the Collaborative filtering to work efficiently. The scalability refers to the non working of an RS in real time. This may happen due to the increase in number of web users and the data sparseness [6]. The rest of the paper is drafted as follows, Section II gives an overview of the existing methods of CF in hadoop . Section III provides the explanation of the proposed methodology and results and Section IV details about the conclusion and enhancements.

II. EXISTING WORKS

In paper [9] the mapReduce is introduced to achieve scalability. The CF is achieved using
a)Building of item-user matrix where rows are no of users and column are the no of items and each entry is a rating.
b) The similarity among the items and the neighbor that are nearest.
c) The ratings of the items that are not known are predicted. The stated job are performed using 3 maps and 3 reduce.

In paper [10] the problem of time consumption is attended and the authjor has proposed a hybrid method to achieve the scalability of a Recommendation System. The hybrid algorithm is achieved with the three important steps such as user-cluster which is obtained using K means and user-item using the asme and at last the Slope-one algorithm for rating prediction ands recommending the Top N results among the top n items. Mapreduce is used to achieve this.

In paper [11] The scalability and the sparsity is increased in a RS. The method used is the machine learning and a hybrid model is proposed with two modules namely the user module that is present already and the newly added user profiles.

Existing User module:

K, The user-Item matrix is created of the size |N| × |M| is constructed after the selection of features are over.Here N represents the no of users and M represents the number of E. The sparsity is overcome by the ALS (Alternate Square Mean) method[7]. Cluster of similar users are obtained usingf the K-Means algorithm[17]

New User module: Each book is assigned a tag to make the users enable to interpret the tag and a specific storage for the tags are used called as tag storage. The users are expected to select the most liked tag and the corresponding section comes in top of the recommendations. Thus the asimilarity is ascertained.

In paper [12] the author took the problem of route recommendation for tourists. The scalability and stability is good as it was deployed in the cloud platform. The behaviour pattern of the user is captured and the association rules are generated. The A-priori algorithm is used to generate rules. The user through an interface will access the recommended rules.

S.no	Paper	Advantage	Disadvantage
1	CF recommendation in distributed scenario[9]	This model is more robust to data sparsity.	Less scalable
2	Hadoop based hybrid recommendation [10]	Speed and used in large Data	Thye quality of prediction is less
3	Apache Spark based RS in hybrid frameworkA [11]	Good speed and prediction quality	Not scalable to large Data sets
4	Intelligent system for tourist routing [12]	Good Scalability	Poor Prediction Quality

Table 1 : Comparison of Existing methods

III. PROPOSED WORK

The intended research is to acquire a Recommendation System that is scalable and efficient for large data sets that are complex and un capturable. The objective is to arrive at a RSs that is strong and which does not compromise in the quality. The proposed method is segmented into six major activities. The structured data set is obtained by prep processing which gives out the user-item matrix. Step 1: To extract the feature, The User-Item matrix is obtained. If the number of users is denoted by N and there are M number of e-books , Then user-item matrix K is a matrix of size |N| x |M| which has all the ratings. K[l][m]=1 if the book is rated and and K[l][m]=0 if the book is not rated

Step 2: The low dimension latent factor is then obtained using the ALS "Alternating Least Square" Eventually, The objective is to find the matrices R(|N| x K) and S(|M| x K) such that the combination is equal to K

Step 3: The recommendations at various stages are obtained using the FP- growth algorithm [15].

Step 4: The Association rules that are obtained are stores accordingly after sorting them in line with the confidence. The results are immediately stored.

Step 5: The CF method is then applied to calculate the similarity between users and the items' ratings are predicted.

Step 6: Once the CF is applied, The recommendations are sent to the users based on their preferences.

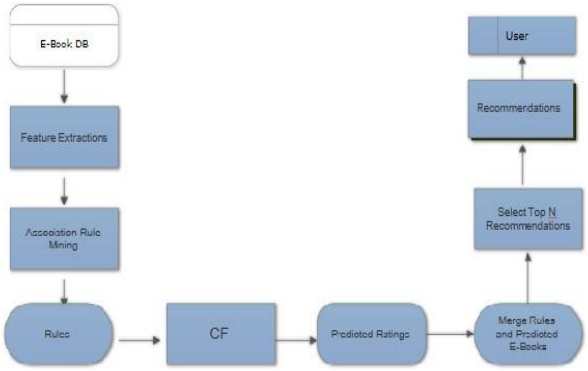

Figure1 - Architecture Of Proposed System

IV. EXPERIMENTAL EVALUATIONS

A. Setting up of Hadoop

The experiment is conducted using Hadoop 2.6.0 and. Latest version og Java is used. Two clusters were used namely the one master- one slave and one-master –two slave architecture. Each has enough disk space.

B .Dataset

The dataset consisted of testing and training data. The Google Books dataset is used that has over 15 million ratings. The ratings fall between 1 and 10 point scale and the corresponding tags are also assigned. The other data set used is also from Google books to test the scalability of the proposed work. All the data were of the .CSV formats with appropriate fields like bookID, UserID and TagID for the appropriate books.

C. Experimental Results

An 80:20 ratio is maintained between the training and testing data sets. As the motive is to find out the optimum value for MSE, The model is trained for values for λ=0.01 and nf = 10, which are the two other important parameters of ALS [7] [8]. The results obtained were as follows Figure.2 shows that there is an obvious decrease in MSE value as the no of iterations are increased.

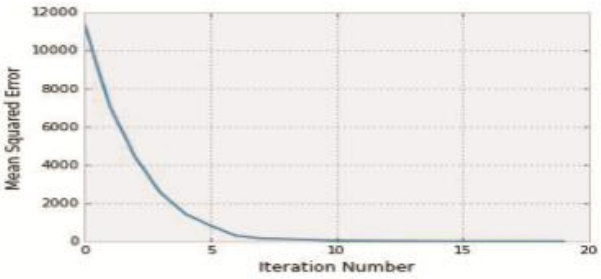

Figure2 – Iteration number Vs MSE

The performance is evaluated by measuring the run time in line with the increasing no of nodes and the volume of data. Figure.3 shows the comparison when running time when the number of nodes are considered. Ans sixe of the data are increased.

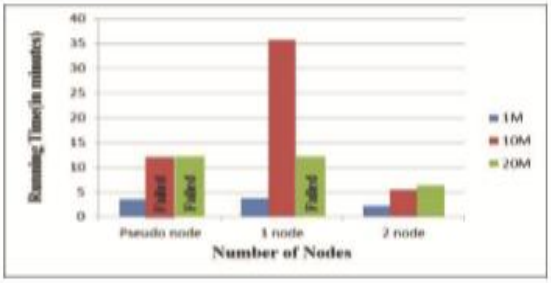

Figure3- Runtime Vs Increased Nodes

Figure 4 shows the comparison between the model taken from literature survey the Base model and the proposed model.

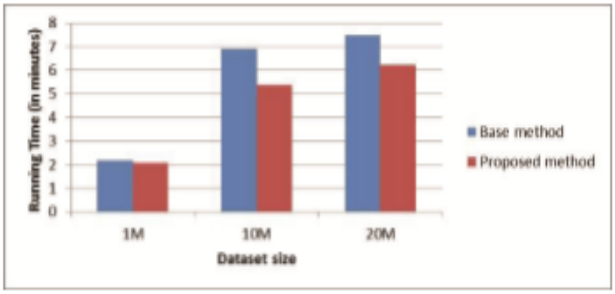

Figure4: Comparison Of Runtime

V. CONCLUSION AND FUTURE SCOPE

The available models in recommendation systems have been studied. The analysis of the existing methods proves to be not scalable. The inception of developing a scalable and accurate RS model is taken The Association rule mining and the CF methods are combinely used to achieve the task. This work also intends to parallel achievement of both CFs and association rule mining that is capable of handling big data. The future work is to introduce sequential association rule mining for the growing database.

REFERENCES

1. S. P. Menon and N. P. Hegde, "A survey of tools and applications in big data," 9th International Conference on, Intelligent Systems and Control (ISCO), Sept2015, pp 1-7.
2. [2] C.L. Philip Chen, Chun-Yang Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Volume 275, 10 August 2014, pp 314-347.
3. Amir Gandomi, Murtaza Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, Volume 35, Issue 2, April 2015, pp 137-144.
4. J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, "Recommender Systems survey, International Conference on Knowledge Based Systems, 2013 pp.109-132.
5. D Park, H. Kim, Y. Choi and J. Kim, "A Literature review and classification of recommender systems research," International Conference on Expert Systems with Applications, 2012, pp. 10059-10072.
6. Aberger, Christopher R. "Recommender: An Analysis of Collaborative Filtering Techniques." Stanfords.edu
7. A Zhou, Yunhong A Wilkinson, Dennis A Schreiber, Robert A Pan, Rong Fleischer, Rudolf E Xu, and Jinhui B, "Large-Scale Parallel Collaborative Filtering for the Netflix Prize", Algorithmic Aspects in Information and Management (AAIM)4th International Conference on, July 2008, pp. 352359.
8. S. K. Joshi and S. Machchhar, "An evolution and evaluation of dimensionality reduction techniques A comparative study," IEEE International Conference on Computational Intelligence and Computing Research, Dec 2014, pp. 1-5.
9. Poonam Ghuli, Atanu Ghosh and Dr. Rajashree Shettar "A Collaborative
10. Filtering Recommendation Engine in a Distributed Environment" 9th International Conference on Computer Science and Education (ICCSE), Nov 2014,pp. 568-574.
11. Kunhui Lin, Jingjin Wang and Meihong Wang "A Hybrid Recommendation Algorithm based on Hadoop" 9th International Conference on Computer Science and Education (ICCSE), Aug 2014, pp. 540-543.
12. Sasmita Panigrahi, Rakesh Ku. Lenka, Ananya Stitipragyan, A Hybrid Distributed Collaborative Filtering Recommender Engine Using Apache Spark, Procedia Computer Science, Volume 83, 2016, pp 1000-1006.
13. X. Chen and L. Zhou, "Design and implementation of an intelligent system for tourist routes recommendation based on Hadoop," Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on, Beijing, 2015, pp. 774-778.
14. J. P. Verma, B. Patel and A. Patel, "Big Data Analysis: Recommendation System with Hadoop Framework," 2015 IEEE International Conference on Computational Intelligence & Communication Technology, Ghaziabad, 2015, pp. 92-97.
15. P. Chandarana and M. Vijayalakshmi, "Big Data analytics frameworks," 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), Mumbai, 2014, pp. 430434.
16. Chang, Hong-Yi, et al. "A Hybrid Algorithm for Frequent Pattern Mining Using MapReduce Framework." First International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA), Dec 2015, pp. 19-22.
17. Sonali Gandhi, Monali Gandhi. "Hybrid Recommendation System with Collaborative Filtering and Association Rule Mining Using Big Data" , 2018 3rd International Conference for Convergence in Technology (I2CT), 2018

AUTHORS PROFILE

Manikandan R , an academician with 7 years of teaching experience and in research . Area of interest includes Data mining and Big Data.

Ramesh R , Is currently associated in the research area of Associative Classification and working as a faculty in Sri Krishna College of Arts and Science in Coimbatore,India

Dr. V Saravanan has a PhD degree in the computer science and has over 15 years of academic experience .His research area includes Data Mining, Artificail Intelligence, data cleaning and Software Agents.