International Journal of Engineering Research and Technology.
ISSN 0974-3154 Volume 11, Number 6 (2018), pp. 881-896

http://www.irphouse.com

Big Data Mining Application in Fasteners Manufacturing Market by using Apache Mahout

Hemlata[1], Preeti Gulia[2]

[1, 2] *Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India*

[1]ORCID: 0000-0002-6105-7399
[2]ORCID 0000-0001-8535-4016

Abstract

In the digital communicating era, data is generated on a very large scale in a fraction of second. This all unstructured data and information collectively is termed as Big Data. Big Data has its application in every field of our life. To extract the meaningful information out of the whole data is really challenging. There exist a number of big data mining techniques which have diverse applications in every field like medicine, e-commerce, social networking etc. In this paper, Mahout – a machine learning algorithm of big data is used for predicting the demand of fastener market. Mahout is a popular tool used in predictive analytics. For building a user based recommendation system, collaborative filtering technique is used. The proposed system provides the recommendation to the user for purchasing fastener items.

Keywords: Collaborative filtering, Eclipse, Mahout, Predictive Analytics, Recommendation System.

I. INTRODUCTION

Industrial fasteners have a wide range of products like bolts, screws, nuts etc. The automobile industry is the largest consumer of fasteners in our country. The market is expected to grow with the rise in automotive industry. The growth in automotive industry will propel the demand for fasteners products [1]. There are several factors which affect the demand of fastener products. They can be growth of fastener application industries, price, quality, service, durability and many more.

In order to forecast or predict the future trend of the fasteners product, predictive analytics can be used. For prediction a huge amount of data is needed which can be Big Data. So, there is a need of big data mining application in fasteners manufacturing

industry. Due to a little work done in this field, data collection and data cleaning was a little difficult. Also Mahout is a good machine learning software, which is used in the present study for predicting the unknown values by taking the known values as basis.

The remaining paper is organized as: Section II provides a summary of the related work. Introduction of the concepts used in the present study are given in Section III. It includes predictive analytics, recommendation system, similarity measures, Apache Mahout, its architecture and the evaluation techniques. Proposed work is presented in Section IV. It includes the experimental set-up, dataset taken, System configuration required, Experimental procedure followed. Section V describes the experimental evaluation and results. Conclusion and future work are given in Section VI.

II. RELATED WORK

Revisiting the work done in the related area it has been observed that the recommendation systems built in Mahout are very accurate and precise. Also Mahout is very efficient in terms of accuracy and time if distributed big data is considered. Owen, S and et.al [2] in their book explains the whole process of installing and using Mahout in non-distributed as well as distributed environment. They explained various machine learning techniques using Mahout with practical implementation. Gupta, Ashish [3] in his book explains the popular open source project Apache Mahout from the scratch. The book explains all the classification algorithm available in Mahout under Hadoop environment. Tiwary, Chandramani [4] in his book explains the machine learning using Mahout in detail. It covers the basic concepts of both machine learning and Mahout. Withanawasam, Jayani [5] in her book explains and implements the machine learning algorithms for classification, clustering and recommendation using Apache Mahout. Giacomelli, Piero [6] in his book explains the theoretical and practical concepts of various machine learning algorithms using Apache Mahout.

F.O. Isinkaye and et.al [7] explains the characteristics and power of different prediction techniques of recommendation systems. They describes the strengths and weaknesses of the traditional recommendation techniques, different machine learning algorithms for generating recommendation models and various evaluation measures for testing the efficiency and quality of the recommendation system. Al-Noukari, Mouhib and et.al [9] proposed a data mining application in car manufacturing industry. The results show that the data mining techniques were used in two types of analysis – Launch analysis and Slow Turning Analysis. Such analysis helped the car market in accurate prediction of future car demand. Kumar, Thangavel Senthil and et.al [10] describes the usage of collaborative filtering technique of recommendation system in Mahout on Hadoop environment. They presented that the results improved by combining user based and item based collaborative filtering. The performance was checked by using efficiency and speedup. Bagchi, Saikat [11] explains the performance and quality analysis of various similarity measures used on collaborative filtering in Mahout. The analysis of various similarity measures was conducted through experiments. The similarity measures used are Euclidean Distance, Pearson

correlation, Tanimoto coefficient, Uncentered Cosine, City Block, Log Likelihood and Spearman Correlation. It was concluded that Euclidian distance gave best performance in every situation. Sondur, Mr. Sridhar Dilip and et.al [12] provide a comparison of various similarity measures of collaborative filtering in recommendation system.

A new and efficient Multi-Criteria Collaborative Filtering (MC-CF) algorithm was proposed by Kumar Bokde, Dheeraj and et.al [13]. The proposed algorithm used dimensionality reduction techniques for the improvement of recommendation quality and prediction accuracy. The two dimensionality reduction technique used in the study are – Singular Value Decomposition and Principal Component Analysis. Walunj, Sachin Gulabrao and et.al [14] provides a study of the process of building a recommendation system for online businesses. It was concluded that Apache Mahout provides the recommendations that enable the online companies to perform better.

III. PRELIMINARIES

In this section a brief outline of the concepts and terms used in the study are presented.

III.I PREDICTIVE ANALYTICS

Predictive Analytics means the process of deriving new insights from the historical data with the help of data mining tools, statistical models and machine learning techniques and to predict the unknown events. It discovers the hidden patterns by applying the observed patterns with the help of data mining techniques. It uses the known observations to develop (train) a model which can be used to predict values for new data[15]. The model provides the data in the form of predictions that represent a probability of target variables. The process of Predictive analytics[16] involve following five steps which are also shown in fig. 1:-

- *Defining Objective*: - The outcomes, deliverables or objectives are identified. Data sets to be used are collected or gathered on the basis of objectives.
- *Data Preparation*: - Here, all the possible data is collected from various sources to make it Big Data. The data is then cleansed in the form which can be used for our objective and prediction.
- *Modelling*: - A model is prepared from the available data for the prediction of unavailable data. It can also be called train model which is trained according to the known observations and can be used to forecast the unavailable data/event.
- *Deployment*: - The dataset taken is implemented in the model proposed to find the unknown values.
- *Model Monitoring*: - The model is monitored and checked regularly so that correct/expected results are prepared.

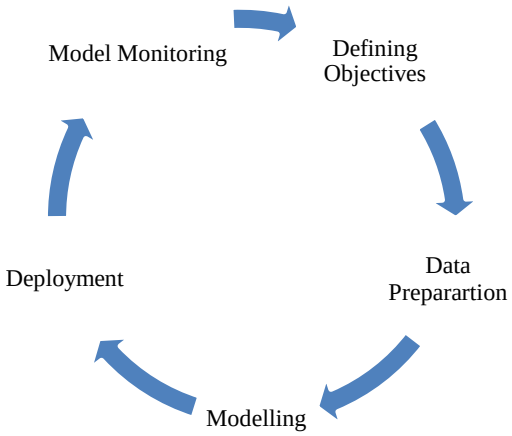

Fig.1 Predictive Analytics Process

Some of the popular freeware predictive analytics softwares are R, Orange, RapidMiner, Weka, KNIME, Actian Analytics platform, Apache Spark MLib, Apache Mahout, Tanagra.

III.II RECOMMENDATION SYSTEM

Recommendation system means the filtering technique, which gives users with information according to his interest. It has the ability to predict user's particular preferences. Recommender systems have improved the decision making process of both the users and service providers[17]. Some of the online recommender systems available are Netflix, amzon.com, Linkedin, Pandora etc. The process of Recommendation [18] is depicted in fig. 2.

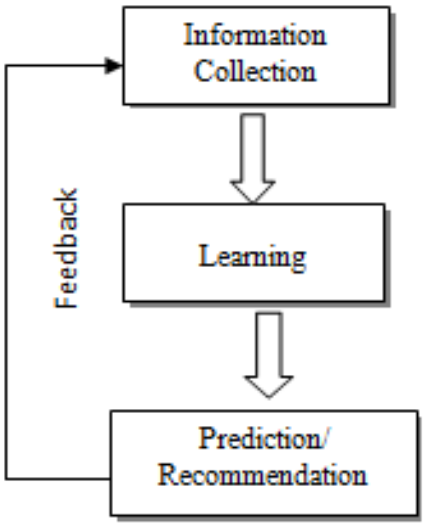

Fig.2 Recommendation Process

Recommendation process starts with the data/information collection which will become the base for generating a model. The collected information is then used to train the data or generate a model by applying a learning algorithm. This process of training or modelling is also termed as learning from the known information [19]. This learning helps in predicting the unknown or future values. Efficient and correct recommendation technique is necessary to give accurate and useful prediction to the user. Various techniques used in recommendation are represented in fig.3.

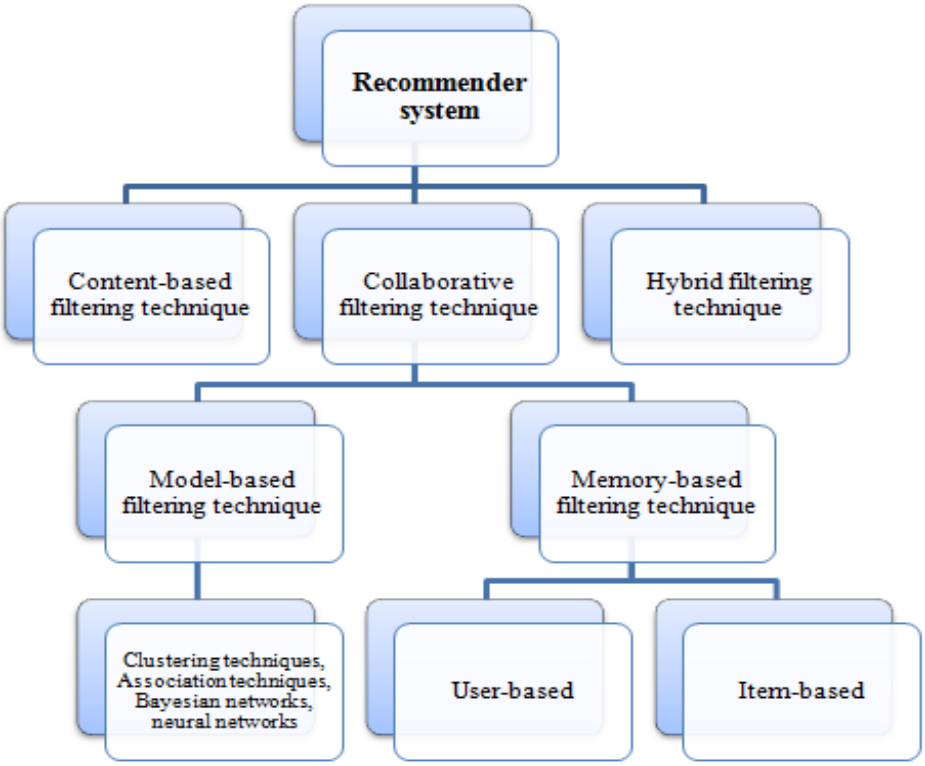

Fig. 3 Recommendation Techniques

- **Content-based Filtering**: - Content-based filtering is a domain-dependent algorithm which analyses the characteristics of items to predict. Recommendation is based on the characteristics extracted from the contents of the items which were evaluated by the users in the past [20]. The model generating techniques used are Vector Space Model or Probabilistic model. This technique does not require the profile of other users. It only uses the historical data of the same user whose recommendation is required. [7] Various examples of content based filtering systems are News Dude, CiteSeer and LIBRA.
- **Collaborative Filtering**: - Collaborative Filtering is a domain-independent recommendation / prediction technique. In this technique user-item matrix of the preferences/likes for items by users is generated [8]. Then similarities between the users are calculated on the basis of their preferences or interests. Similar users are

grouped under the heading neighbourhood. A particular user is provided with recommendation on the items on which he has not given any feedback. These recommendations are according to his neighborhood's preferences [7]. The output given by the collaborative filtering can be recommendation or prediction. Collaborative Filtering technique can be divided into two:-

- *Memory based filtering technique*: - After finding the neighbor of a user various algorithms are implemented to find users preferences or interests [21]. Memory based collaborative filtering involves two techniques: item-based and user-based. User-based collaborative filtering involves the extraction of similarity between users on the basis of their preference of same item whereas Item-based collaborative filtering finds the similarity between the items and not the users [7]. The architecture of User based filtering and Item based filtering [10] is shown in fig. 4 and fig. 5.

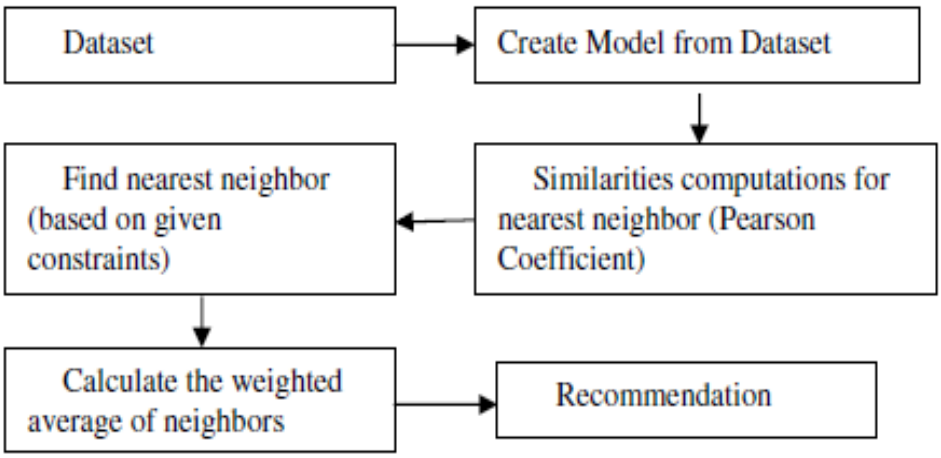

Fig. 4 User based Collaborative Filtering

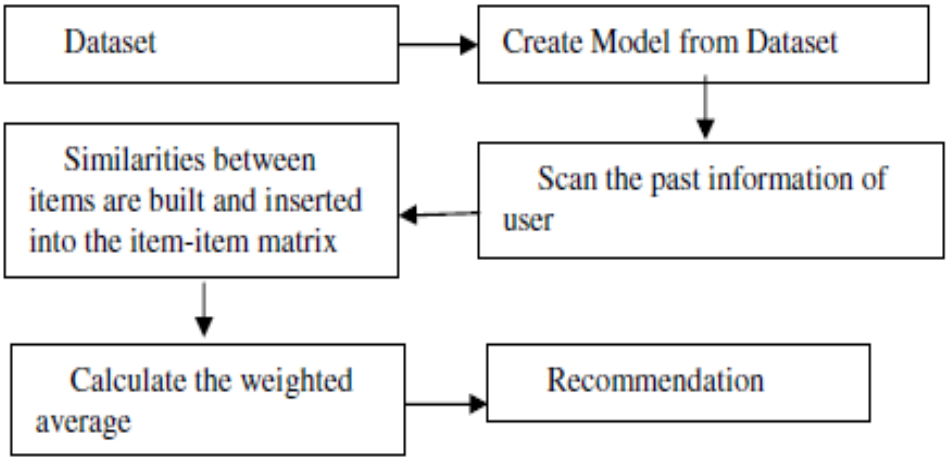

Fig. 5 Item based Collaborative Filtering

- *Model based filtering technique*: - In this technique known values are used to build/learn a model. Model building is done by various machine learning algorithms or data mining techniques [22]. This technique also uses user-item matrix for finding the relation between items. Different learning algorithms used by model-based filtering are Association Rule, Clustering, Decision Tree, Artificial Neural Network, Regression, Bayesian Classifier etc.

- **Hybrid Filtering**: - Hybrid filtering means joining of different recommendation technique to achieve system optimization [7]. In this different combination of algorithms are used for better and effective results as it overcomes the problems of a single algorithm.

III.III SIMILARITY MEASURE

Similarity measure or similarity function can be defined as an operation that computes the similarity between two vectors or objects [11]. It finds a numerical value for similarity between objects. Various similarity measures used are Eucliean Distance, Pearson Correlation Coefficient, Cosine similarity, Jaccard Coefficient, Minkowski distance, Spearman Correlation etc [24]. One of the popular measures, Pearson Correlation Coefficient, used in this study is explained below:

- Pearson Correlation

Pearson Correlation measures the correlation between two variables or object. The measuring value ranges between -1 to +1. Like other similarity measures 1 indicates perfect correlation and -1 indicates no. As the values increases from zero, it indicates the increasing relation between variables or objects under study [12]. Pearson coefficient can be represented as:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where r is the pearson coefficient and x,y are preferences of the two variables.

III.IV APACHE MAHOUT

Apache Mahout is an open-source machine learning library which provides a framework relation for both distributed and non-distributed environment [13]. It is an application developed under the umbrella Apache organization. The word "Mahout" is an Indian hindi word which means the person who drives an elephant. In this context Apache Mahout is a machine learning application which drives on the head of Hadoop (whose official logo is yellow elephant) [25]. Mahout came into existence in 2008 by the Apache Lucene project. Mahout provides the free implementation of the following three machine learning algorithms:

- Collaborative Filtering
- Clustering
- Classification

For Collaborative Filtering Mahout has its own open source project namely "Taste" [14]. Mahout is highly scalable and supports distributed processing (Map-Reduce) of Big data along many clusters of computers using Hadoop. In the present study the non-distributed algorithm is being used. It provides various similarity measures, neighbourhood computations and evaluation techniques for collaborative filtering recommendation method.

III.V EVALUATION METRICES

The recommender system predicts the value or rating of a user for the items which are not available or which are not predicted by the users. The evaluation of the predicted value means how the predicted rating matches with the actual rating. The quality of a recommendation system depends on the accuracy of the prediction. Filtering technique used by the recommendation system specifies the type of evaluation metric to be used [23]. There are two types of accuracy metrics used – Statistical accuracy metric and decision support accuracy metric. Statistical accuracy metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Correlation [26]. Decision support metrics include Reversal rate, Weighted errors, Receiver Operating Characteristics (ROC) and Precision Recall Curve (PRC), Precision, Recall and F-measure.

In the present study, precision and recall decision support metrics are used. These are also called information retrieval metrics. Precision means the fraction of recommended/predicted items that are actually relevant to the user [27]. In other words, precision is how precise or correct the predicted value to the actual one. It can be computed as follows:

$$\text{Precision} = \frac{\text{Correctly recommended items}}{\text{Total recommended items}}$$

Recall means the fraction of relevant items that are the part of recommended items. It can be calculated as under:

$$\text{Recall} = \frac{\text{Correctly recommended items}}{\text{Total useful recommended items}}$$

IV. PROPOSED WORK

The objective of the present study is to predict the demand for fastener products on the basis of ratings given by the customers. A manufacturing industry predicts its demand on the basis of many factors. In the study prediction is done on the basis of recommendation system. Customer preferences for different fastener products along with the selling unit are recorded. User-item matrix is created to find the relation between the user and item. Pearson correlation similarity measure is used to find the

similarity ratio between the users i.e. customers. A new data model is created which can predict the interests of the customers whose preferences for some specific item were not recorded because they have never used the item before. On the basis of recommendation or prediction given by the model the manufacturers can predict the demand of their product.

IV.I EXPERIMENTAL SET-UP

Dataset Taken - Original data is collected from the customer industries about their preferences for different fastener products. To limit the data, only seven customers are taken from whom data was collected. The customers are given ID's from 11 to 17. They were asked to give the rating about 8 fastener products of 5 manufacturers. Product ID is given according to the manufacturer. If one manufacturer produces 8 products, the Item ID given to it is from 1001 to 1008. Similarly, Item ID of second manufacturer is given from 2001 to 2008 and so on up to 5008. Customers gave their preferences only for the items they have used in the past. The preferences are given in the form of ratings scaling form 1 to 5. The customer who likes a particular item of a particular manufacturer rates it as 5 and the item which is not liked at all or by which the customer is not satisfied is given 1 rating. Data of 248 preference rating is collected from all the customers. The dataset imported in Eclipse is shown in fig. 6.

System Configuration - The study is carried out by using following system, software and other configuration.

Table 1. System and Other Configurations

Processor	*Intel Core i3 M350*
RAM	*2.27 GHz*
Operating Syatem	*Ubuntu 15.10*
JVM	*JRE 1.8.0_112*
Mahout	*Apache Mahout 0.9*
Eclipse	*Eclipse 4.7 OXYGEN.2*
Data File	*Fasteners Dataset with 248 records*
User-item preference rating scales	*1 to 5*

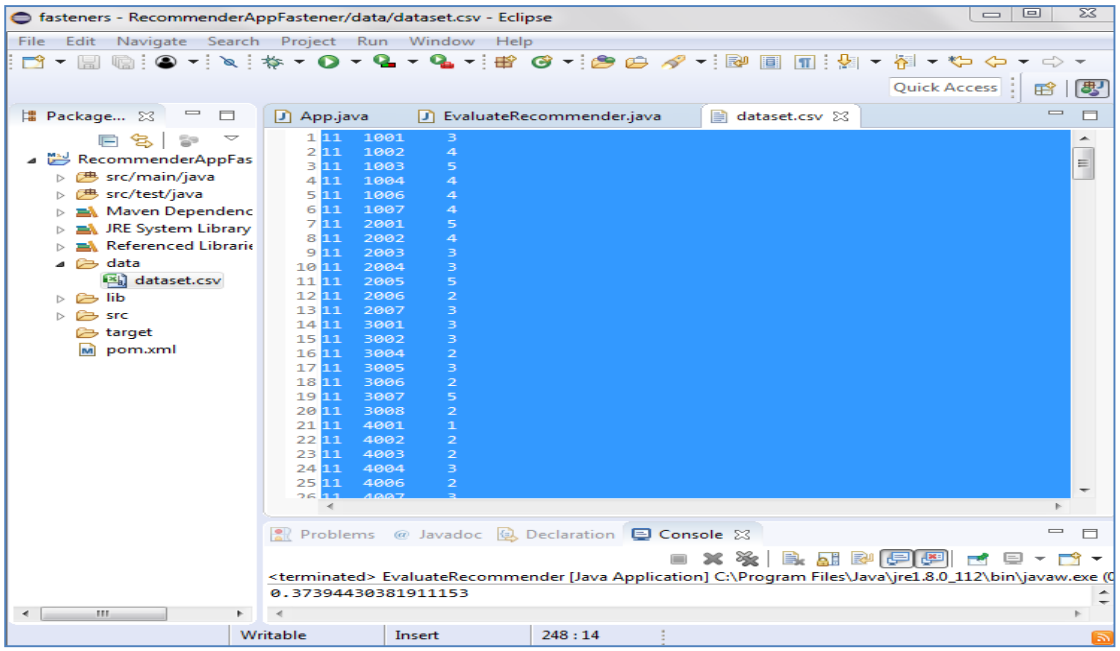

Figure 6. DataSet Taken as inserted in the Eclipse shows 248 rows

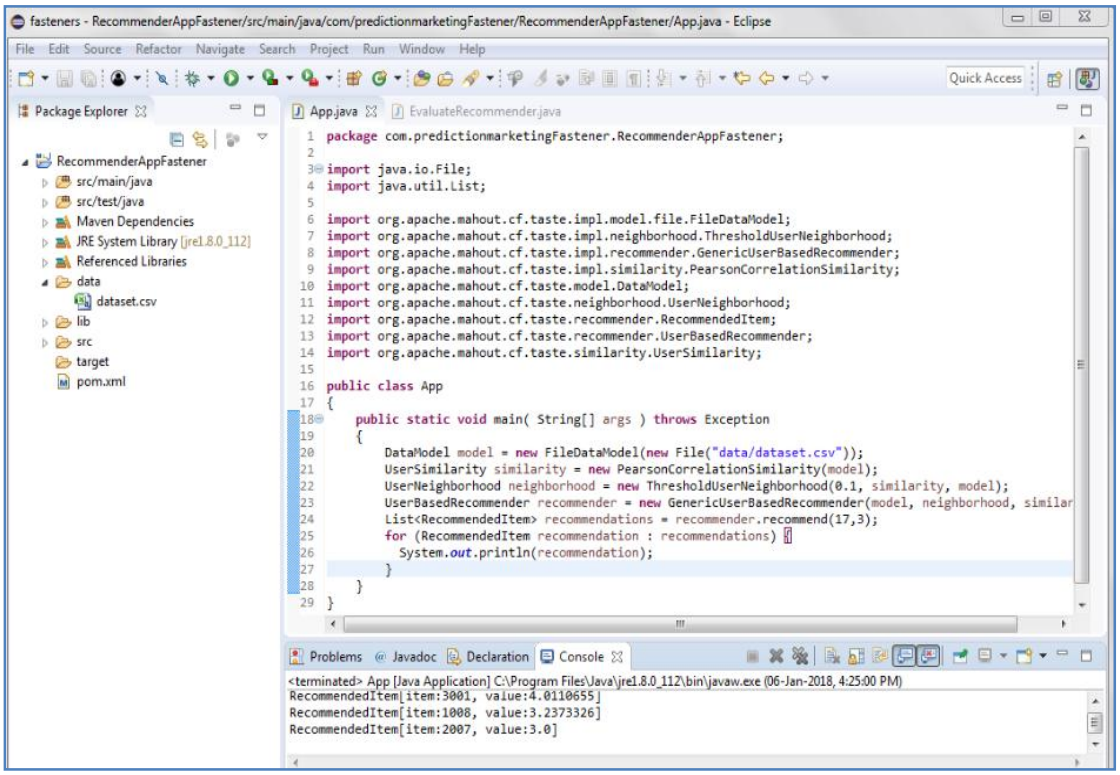

Figure 7. Recommender system built for Fasteners Dataset

IV.II EXPERIMENTAL PROCEDURE

- First of all data is collected for analysis and System is prepared according to the configuration stated above.
- Eclipse is an Integrated Development Environment for implementing various codes. It provides the workspace in which many java projects can be created [10]. It includes a plug-in Maven which inherently calls various Mahout Packages and interfaces. So, eclipse is used to execute various Mahout Classes and methods. Mahout and Eclipse are both written in Java.
- User based recommender system is generated by building a learning model in Mahout. The recommender system built is depicted in fig. 7.
- This recommender predicts the values preferences of the customer about an item which is not at all used by the user.

After creating recommendation system it is evaluated on the basis of two evaluation techniques used by Mahout: - Precision and Recall. The Evaluate Recommender created is shown in fig. 8.

IV.III MAHOUT APIs USED FOR THE MODEL AND EVALUATOR

Following APIs of Mahout are used for building the model and model evaluation:-

Table 2. Mahout APIs Used

Parameter	**Mahout API**
Data Model	*FileDataModel*
Neighbourhood Algorithm	*ThresholdUserNeighbourhood*
Similarity measure	*PearsonCorrelationSimilarity*
Evaluator	*AverageAbsoluteDifferenceRecommender Evaluator*
IR stats evaluator (Precision /Recall)	*GenericRecommenderIRStatsEvaluator*

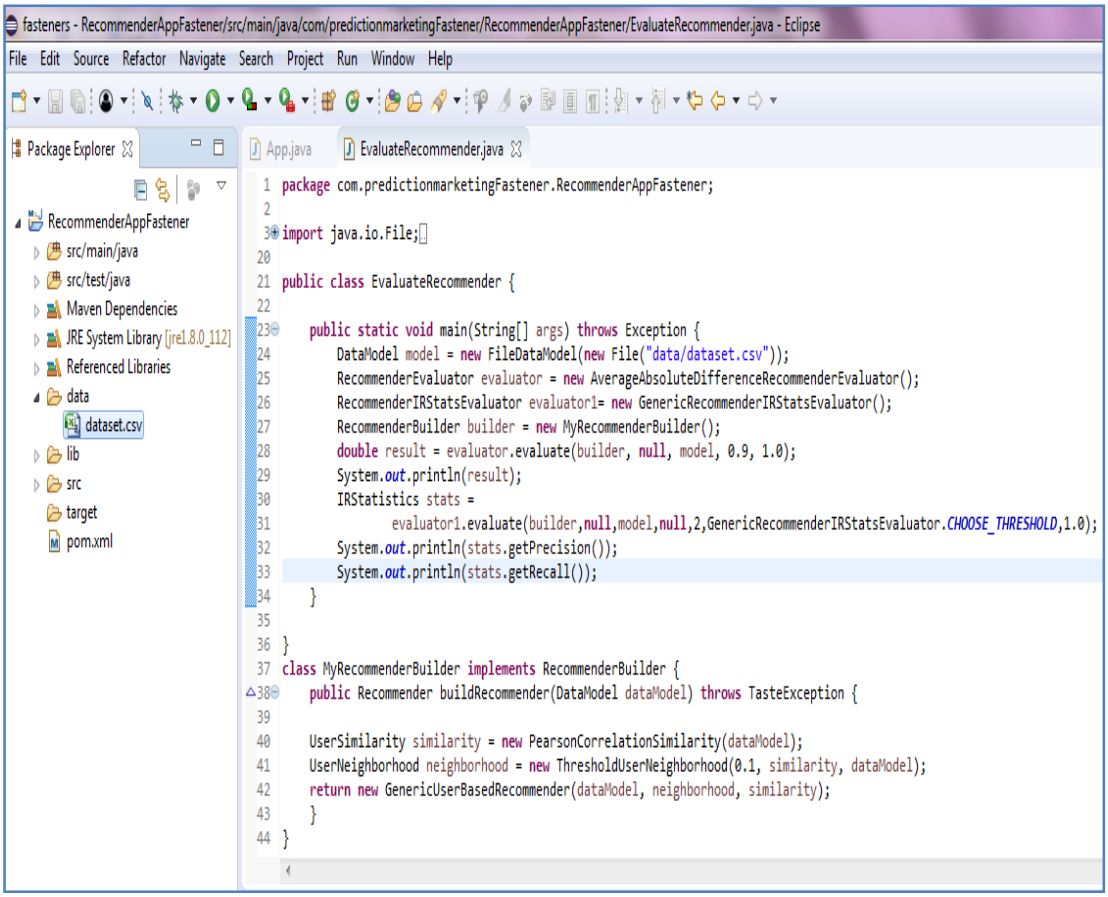

Fig. 8 Evaluate Recommender which uses both precision and recall evaluation techniques

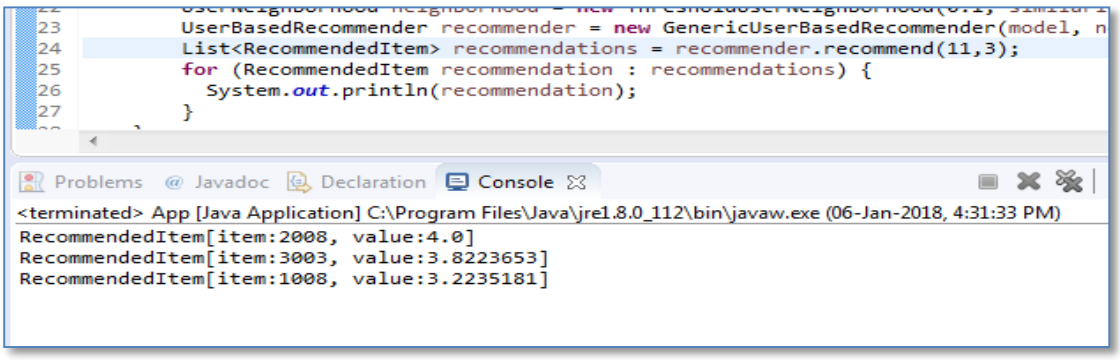

Fig. 9. Recommendation for User ID 11

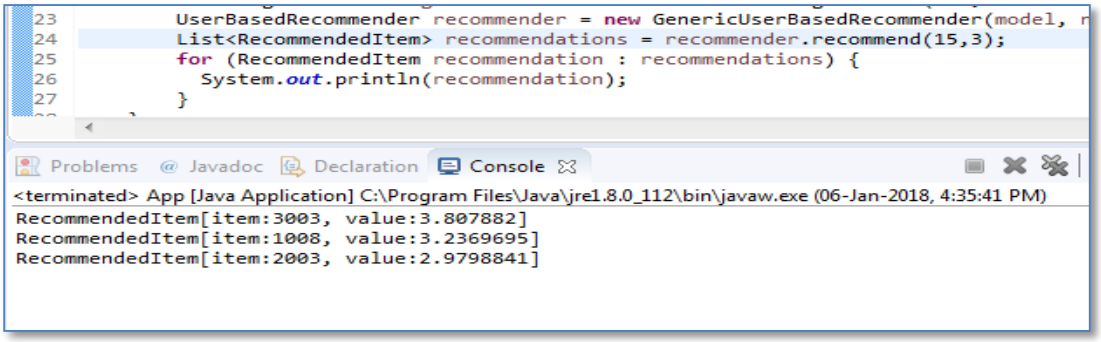

Fig. 10 Recommendation for User ID 15

V. EXPERIMENTAL EVALUATION AND RESULTS

The recommender system gives the recommendations to the users about the items to be purchased according to the user's interests and preferences. In the study, the recommender system gives the output as three recommendations to customer 11. The output to customers 11 and 15 are shown in fig. 9 and fig.10.

The output of the recommendation model suggests user 11 to purchase item no 2008 with recommendation value 4. User 11 i.e. Bajaj Auto Ltd. should purchase nut from Sundaram Fasteners Ltd. The recommendation value 4 is also very high which means that the model strongly recommends this specific item from the manufacturer mentioned. Next two recommendations given are item no. 3003 and 1008 whose recommendation value is 3.9 and 3.2 respectively.

Similarly, user 15 is recommended with items 3003, 1008, 2003 respectively in the order of the preference. If user 15 wants to purchase some item the system recommends it to purchase 3003with recommendation value as 3.8.

By considering all recommendations for all the users the manufacturing unit can predict that the demand for a particular item will increase if the recommender system recommends that particular item to many users. Likewise, if recommender system does not recommend a particular item, it can be predicted that the demand will not increase but it will rather decrease. Hence the proposed recommendation system can be used by the customer industries for purchasing the fastener products. Also, the proposed system can be used by the manufacturing industries to predict the sale, in turn demand, of its products in future.

The proposed recommender system is the evaluated by using AverageAbsoluteDifferenceRecommenderEvaluator and RecommenderIRStatsEvaluator API's of Mahout. AverageAbsoluteDifferenceRecommenderEvaluator class has a method evaluate() which accepts the model and the data partition as parameters. In the study, the data is partitioned in the ratio 90:10. 90% of the data is considered to be training data and 10% data is used as test data. The evaluate method analyses the model and finds the evaluation value which means that how much errors are present in the model. The proposed recommender model has an evaluation value as 0.37 which is

not so significant. It can be said that the proposed system gives almost correct prediction. This value is given by only 248 data records. It will reduce significantly if the data records are increased because Mahout works excellently as the dataset increases.

Two evaluation measures: Precision and Recall are used in the study. An in-built class GenericRecommenderIRStatsEvaluator has a method getPrecision() and getRecall() which finds the value of precision and recall for the proposed recommender system. The precision and recall comes out to be 0.78 which shows that our proposed model gives correct prediction for almost 80% of the time it is executed. Figure 11 shows the output of the evaluation recommender. It gives the value of the evaluation of training data over test data, Precision value and recall value respectively.

VI. CONCLUSION AND FUTURE WORK

Recommender systems play an important role in online-trading or e-commerce. These can be used to predict the unknown values on the basis of known preferences. In this paper, a recommender system is proposed to recommend the user an item to purchase. The dataset of fastener industries is taken for study to predict the future demand of fastener products by the customers. The proposed system recommends the user to purchase the particular item by calculating the similarity between the preferences of users for different items. Here, collaborative filtering technique of Mahout is used to calculate the similarity between users. User-based recommendation system is generated using various algorithms of Mahout in Eclipse API. The proposed system recommends the user all those items which are not used by the user before. On the basis of recommendation various manufacturing organizations can predict their demand for different items in future. The proposed system is evaluated by using two evaluation measures namely precision and recall. The evaluation metrics give precision and recall value 0.78 to the proposed recommender system.

In future, other machine learning algorithms supported by Mahout can be used to build a recommendation system or prediction. Also the dataset can be increased to find the efficiency of proposed system and Mahout.

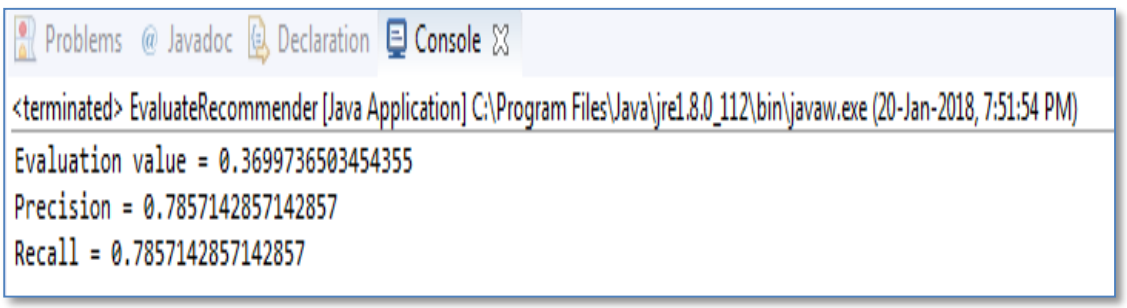

Fig. 11 Output showing evaluation, precision and recall

REFERENCES

[1] Herring, Daniel H. "Fastener Heat Treatment."

[2] Owen S., Anil R., Dunning T. and Friedman E.: "Mahout In Action", 2012. Manning Publications Co. ISBN 978-1-9351-8268-9.

[3] Gupta, Ashish. Learning Apache Mahout Classification. Packt Publishing Ltd, 2015.

[4] Tiwary, Chandramani. Learning apache mahout. Packt Publ., 2015.

[5] Withanawasam, Jayani. Apache Mahout Essentials. Packt Publishing Ltd, 2015.

[6] Giacomelli, Piero. Apache mahout cookbook. Packt Publishing Ltd, 2013.

[7] Isinkaye, F. O., Y. O. Folajimi, and B. A. Ojokoh. "Recommendation systems: Principles, methods and evaluation." Egyptian Informatics Journal 16.3 (2015): 261-273.

[8] Herlocker, Jonathan L., et al. "Evaluating collaborative filtering recommender systems." ACM Transactions on Information Systems (TOIS) 22.1 (2004): 5-53.

[9] Al-Noukari, Mouhib, and Wael Al-Hussan. "Using data mining techniques for predicting future car market demand; DCX case study." Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on. IEEE, 2008.

[10] Kumar, Thangavel Senthil, and Swati Pandey. "Customization of recommendation system using collaborative filtering algorithm on cloud using Mahout." Intelligent Distributed Computing. Springer, Cham, 2015. 1-10.

[11] Bagchi, Saikat. "Performance and quality assessment of similarity measures in collaborative filtering using mahout." Procedia Computer Science 50 (2015): 229-234.

[12] Sondur, Mr Sridhar Dilip, Mr Amit P. Chigadani, and Shantharam Nayak. "Similarity Measures for Recommender Systems: A Comparative Study." Journal for Research| Volume 2.03 (2016).

[13] Kumar Bokde, Dheeraj, Sheetal Girase, and Debajyoti Mukhopadhyay. "An item-based collaborative filtering using dimensionality reduction techniques on Mahout framework." 4th Post Graduate Conference for Information Technology (iPGCon-2015), Sangamner, Published in Spvryan's International Journal of Engineering Science and Technology (SEST). 2015.

[14] Walunj, Sachin Gulabrao, and Kishor Sadafale. "An online recommendation system for e-commerce based on apache mahout framework." Proceedings of the 2013 annual conference on Computers and people research. ACM, 2013.

[15] Apache Mahout, Available at URL: http://mahout.apache.org/

[16] Eclpse Oxygen, Available at URL: https://www.eclipse.org/

[17] https://en.wikipedia.org/wiki/Eclipse_(software)/

[18] Landset, Sara, et al. "A survey of open source tools for machine learning with big data in the Hadoop ecosystem." Journal of Big Data 2.1 (2015): 24.

[19] Singh J. Big Data Analytic and Mining with Machine Learning Algorithm. Int J Inform Comput Technol. 2014;4(1):33–40.

[20] B. Wang and R. Wang, "A Collaborative Filtering Algorithm Fusing User-based, Item-based and Social Networks," IEEE Int. Conf. Big Data (Big Data), pp. 2337–2343, 2015.

[21] S. K. Zhuo Zhang, Paul Cuff, "Iterative Collaborative Filtering for Recommender Syatems with Sparse Data" Princeton University, Princeton , NJ 08544," IEEE Int., pp. 1–6, 2012.

[22] J. Kim and S. Hwang, "Big Data Platform of a System Recommendation in Cloud Environment," Int. J. Softw. Eng. Its Appl., vol. 9, no. 12, pp. 133–142, 2015.

[23] Hemlata, Gulia, P. (2017). Novel Algorithm for PPDM of Vertically Partitioned Data. International Journal of Applied Engineering Research, 12(12), 3090-3096.

[24] Jannach D., Zanker M., Felfernig A., Friedrich G.: "Recommender Systems: An Introduction". (http://www.cambridge.org/uk/catalogue/catalogue.asp?isbn=9780521493369).CUP. ISBN 978-0-521-49336-9.

[25] Alan Said, Alejandro Bellogín: "Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks". In Proc. of RecSys'14, 2014.

[26] https://cwiki.apache.org/MAHOUT/recommender-documentation.html

[27] C. Desrosiers and G. Karypis. "A comprehensive survey of neighborhood-based recommendations methods". In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor,editors, Recommender Systems Handbook. Springer, 2011.