Building a Scalable Eservice Recommender System

S.J. Savitha [1], D. Betteena Sheryl Fernando [2], K. Saranya [3]
Assistant professor, Department of Computer Science and Engineering,
Sri Ramakrishna Institute of technology
Tamilnadu, India

***Abstract-** E-commerce Recommender system is proposed to solve Big Data problem due to huge amount of data, prevailing in many of the service recommender systems in the market. And to build scalable, efficient and precise service comparison and recommender system is highly needed. This system enables the shoppers to deeply analyses on what product to choose in various services. This system recommends the user to purchase the product and grab the data from various web services, loads to hadoop file system and clustered and classified the product using mapreduce framework. This recommender system will recommend the product based on the Case Based Collaborative Filtering (CBCF). CBCF is to filter the product information from huge amount of data for product comparison. Model based method is used to predict the item. Pearson Correlation Coefficient is used to measure the similarity value of the items. This proposed system avoids the scalability problem of existing recommender system. It reduces the overall time required by the user to analyses the services on the e-commerce environment and the users can effectively retrieve and identify the suitable product from the e-commerce system*

Keywords- *Hadoop,Mapreduce,Fuzzy-KmeansClustering ,Naïv Base Classification, Recommendation System, Case based Collaborative Filtering, Similarity Measure.*

I. INTRODUCTION

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years. Recommender System (RS) has been successfully exploited to solve information overload. Big Data applications are to explore the large volumes of data and extract useful information or knowledge for future actions. This will reduce the complexity and ambiguity of user to analysis the services provided by the application. Case Based Collaborative Filter (CBCF) Recommendation is used recommend the product based on the case that detail description of product are matched with the user submitted target query. Case Based Recommender system to make fine grained judgements about the similarities between items and queries for providing high quality suggestions to user.

The proposed system Case Based Collaborative Filter performs as well as the best among the alternative algorithms when the data is sparse or static. Pearson Correlation Coefficient finds the probability of item similarities between user query and detailed description of the individual item. It reduces the overall time required by the user to analysis the data and services on the ecommerce environment, and effectively user can retrieve and identify the data from the environment. Discovering useful knowledge from the Big Data in this system uses the hadoop frame work. Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources to combine and recommend the product.

II. RELATED WORK

This section provides a brief survey of the Recommendation works that are done earlier

Collaborative filtering and content – based filtering are the two prominent approaches for product recommendations. Resnick [5] proposed a content-based recommendation technique. Provide recommendations by considering mapreduce the description of products.

Resnick [5] proposed content based filtering for hybrid recommendation. These systems consider both the rating of the user and item features to recommend the item to the user the features of limited amount of data can be analyzed with the existing data analysis tools. Considering large amount of dataset Terabytes, the big data analysis tool hadoop is used. MapReduce paradigm to perform distributed processing over clusters of computers to reduce the time involved in analyzing the item's feature. Existing recommendation system recommends books to the user based on the book name and the ratings. Given by that user to the book or based on the number of views for that book.

Prem Melville [3] proposed neighborhood-based techniques. In this method users are chosen based on their similarity and a weighted combination of their ratings is used to produce predictions for these users, weight all users with respect to similarity. Users are selected based on highest similarity. Similarity is weighted combination of the selected neighbors 'ratings. *Item-to-item* collaborative filtering where rather than matching similar users, they match a user's rated items to similar items. These neighbors based on a small number of overlapping items tend to be bad predictors. This type of recommender system would give worst prediction.

Yunwei Zhao [4] proposed clustering performance index (CPI), that takes into consideration of homogeneity, relative population, and number of clusters aggregated and propose a new hierarchical clustering algorithm by adopting homogeneity as its key similarity. Experimental results show that proposed clustering algorithm can achieve a good balance among CPI, the number of clusters aggregated, and the time cost of the algorithm reduced. The decision makers usually have difficulties to define the number of clusters to be formed. Because they do not have prior knowledge about the quality of cluster and what happens in the unsupervised clustering process.

Xueming Qian[1] proposed a user interested item based recommendation which is effective with limited data size. This approach uses the data mining techniques K-means clustering and Naïve Bayes classification to provide recommendation. One challenging task for data mining algorithms is how to handle the large volume of data. Mapreduce fuzzy K-means clustering is used, which can adaptively achieve large volume of data.it will able to solve the scalability problem of the data mining techniques.the existing recommenders system used the collaborative filtering method to filter the user behavior and interest to recommend the product.

Xinhua [2] proposed a novel method Sequence- (CSGM) algorithm to generate closed sequences and sequence generators for non -redundant sequential rule mining. By applying this method on web logs, to extract sequential associations among products which reflect users preference on products. Extract sequential associations among products which are viewed or visited by users when they are navigating the web site for products. These associations reflect user's preferences towards products. In their system they do not accurately find user's interests because if the user searching irrelevant product it will not provide the best match and thus generate very less accurate recommendations. Accurate item recommendation is achieved in this user selected item based recommendation system.

III.SYSTEM ARCHITECTURE

The system has been proposed to the proposed system has modeled the item recommendation in an online e-commerce system. The proposed system will vary from existing systems in a number of ways. The proposed system used multiple resources to product recommendation. In order to accomplish this, Hadoop framework is implemented. In this framework have two parts, 1. HDFS. 2. Mapreduce. HDFS will handle the large amount products.

In Figure 3.1 shows the entire process of recommender system. Mapreduce is a distributed processing to cluster product nodes. Mapreduce Classification is used to classify the product features using its <product id, list [price information >. Each mapper takes chunk of files and spilt to nodes for processing. The iteration will continue until the intermediate result. The output of mapper is given to the reducer phase to provide reduced classified output.

Case Based Collaborative Filtering is performed in this project to recommend the product to the user. This collaborative filtering filter the item from user selected from different services.

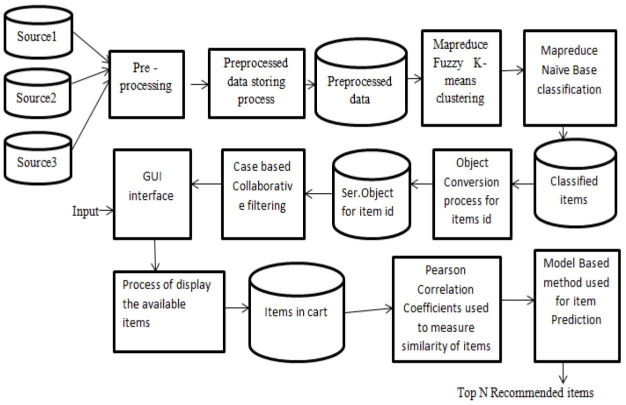

Source1
Source2
Source3
Pre - processing
Preprocessed data storing process
Preprocessed data
Mapreduce Fuzzy K-means clustering
Mapreduce Naïve Base classification
Classified items
Object Conversion process for items id
Ser.Object for item id
Case based Collaborativ e filtering
GUI interface
Input
Process of display the available items
Items in cart
Pearson Correlation Coefficients used to measure similarity of items
Model Based method used for item Prediction
Top N Recommended items

Figure 3.1 System Architecture

The shoppers will be provided with clean indexes of various products with its spec ,cost and also Service Ratings which is done in a statistical way .Our System crabs the data's from various web application and loads in its datasets collaboratively and process, so as to categories classify and to Index the data's in a distributed and Parallel processing Manner.

Shoppers can analyze, get recommendations and can pick products and add to cart irrespective of the service provider. Hence our application stands unique as it does not rely on the single service provider. The cart can be reviewed at any time and can be processed whenever the shopper wants the product. All the information will be securely and precisely stored in the user's session.

The Purchase phase look up for the web services of the products service provider and can make the online payment with the banks from service provider. Once it got over process gets back to our gateway bringing out the track id from product service provider.

A. Product dataset gathering

The product dataset is collected from the amazon.com, Flipcart, eBay. Dataset is in CSV files which are converted into TSV files to avoid line space problem. Dataset contains the more number of product attributes. So need to extract the relevant features from dataset. This dataset preprocessed and load to the HDFS system. It contains more than 10 lakhs products features which are processed in this Recommender system. This collect the thousands of feature variables, not all the feature variables inherent in this application are useful for sophisticated data analysis. So in this method remove the features such as transaction id, date of transaction, jpg, product dimensions, weight of product, and based on inconsistency words. After the preprocessing the output is stored in the HDFS files.

Inconsistency features=No of inconsistency feature-No of instances .... (3.1)

B. Mapreduce fuzzy k-means clustering

Mapreduce cluster the products. Mapreduce fuzzy k-means clustering is used to cluster product. Large sets of book, cd, can be distributed among the nodes of a cluster and processed parallel. There are two types of node such as master node and slave node. Master node 18 allocates the tasks to the slave and slave nodes carries out the task assigned to it. Master node then collects the results.

This model has two main steps which are 1) Map phase distribute the task among the slaves with the < product id, list(price information)> and it calculate the centroid for it form the cluster and similarity is measure based on the features, it calculate the Euclidean distance. 2) Reduce phase it will reduce the minimum number of features such that<product id, list (price information)> pairs as output.

C. Naïve baye's classifier

Input: Cluster Files items

Output: Classes item titles

Calculate probabilities product types

$P(c\,j / d) = p\,(d / c\,j * p(c\,j) / p(d)$.

$P\ (c\,j \mid d)$ = probability of total product being in class c j.

$p\ (d \mid c\,j)$ = probability of generating product feature vector d given class c j, we can imagine that being in class c j.

$p\ (c\,j)$ = probability of occurrence of product features in class cj

This just the frequent of product features the class c j

p (d) = probability of similar product features vector

The Naïve Bayes classifier is used to classify the product dataset which is taking the training data and test data to classify the features. In the classes which can be classified based the feature occurrences in the test cases and the total number of feature occurrences in the training data and which can be classify the class as price , product id, offers in the Classes .Classified item is distributed to master node and store into the HDFS.

Naïve Bayes classification has an assumption that attribute probabilities P $(x \mid j)$ are independent given the class j, where x is ith attribute of the data set. This assumption reduces the complexity of the problem to practical and can be solved easily. Despite the simplification of the problem, the Naïve Bayes classifier still gives us a high degree of accuracy. In this project the product id and offers and price are classified.

D. Case Based Collaborative Filtering

Input: Item Name.
Output: Similar items.
Algorithm: Case Based Collaborating filter.

```
For each item j
    Compute k most similar feature of item
Where k<n   number of items
    Generate prediction for each item i
 Target item, t is compared to each items
    Select the k most similar features
  Item contained within these selected features s
     Relevance to the target item r
```

Collaborative Filtering used to select the relevant items for target item. The large amount of product where reducing the irrelevant item occurrence. The Case Based Collaborative Filtering technique is used to fetch the relevant item of given input.

E. Similarity measure

Input: Cart Items
Output: Similarity value of items
Algorithm: Pearson Correlation Coefficient

Calculate the Sales price and offer price <x_i,x> of input item X

Combine the x / x_i for particular item.

Calculate the sales price and offer price of target item Y

Find the similarity between input item and target item using the Pearson Correlation Coefficient.

$P_{x,y}$ = Pearson Correlation Coefficient,

Where x_i sales price and x offer price take as numerical values of input item X and target item Y.

The Pearson score of items is generated as numerical value.

$$P_{X,Y} = \frac{\sum_{i=1}^{n}((x_i - x)(y_i - y))}{\sqrt{\sum_{i=1}^{n}(x_i - x) * \sum_{i=1}^{n}(y_i - y)}} \ldots$$

…….(3.1)

IV EXPERIMENTAL RESULT

The system can be implemented in java language using Eclipse ide. Database used is hdfs and MySQL database. Isoefficiency and speedup are two useful scalability metrics are used in this project. To measure the Speedup Amdahl's law are used to measures the performance of speed of the processor.

1. Amdahl's law

Sp =T1/T p

……..(4.1)

The single processor time T1, the total amount of time with processor 1 and processor n is Tp. T1 is the amount of sequential execution time with single processer. Tp is the amount of parallel execution time with p processor. If the algorithm is scalable, the speedup has a linear relation with the numbers of nodes with the data size fixed

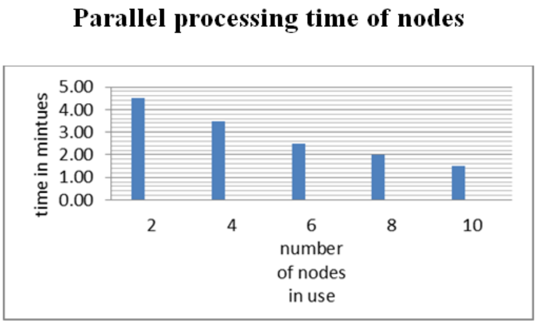
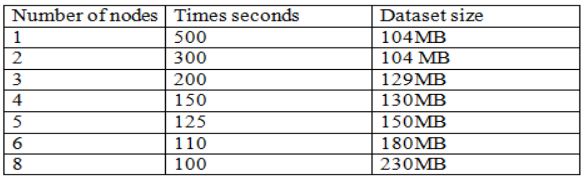


Figure 4.1 Graph for parallel processing time of nodes



Number of nodes	Times seconds	Dataset size
1	500	104MB
2	300	104 MB
3	200	129MB
4	150	130MB
5	125	150MB
6	110	180MB
8	100	230MB

Table 4.1 Parallel processing time of nodes

Various nodes running time for different dataset size counted the number of nodes time for various nodes in seconds which is running for different dataset is shown in table 4.1.

2.Performance of classifier

	Data types	Dataset size	Precision	Recall	F measure
Book	Tsv files	10203	0.85	0.7912	0.8125
Car	Tsv files	32051	0.8361	0.7619	0.8012
Nursery	Tsv files	35560	0.8125	0.7346	0.7981

Table 4.2 Performance of classifier

In table 4.2 show in three different dataset precision , recall and f measure for book , car and nursery datasets are calculated.

3.Precision

Precision is computed from a 2 × 2 table. The item set must be separated into two classes relevant or not relevant. That is, if the recommended items Ns in the recommended items are relevant to the user is Nrs based on that the Evaluation is measured. Equation 4.1 used for calculate the precision of relevant items

P= N_{rs}/N_s

….….. (4.2)

Nrs=Number of relevant items recommended.

Ns=Number of items Recommended.

4.Recall: Recall represents probability of that a relevant item will be selected.

R= N_{rs/N_r}

……… (4.3)

Nrs=Number of relevant items selected.

Nr=Number of relevant items.

	Precision	Recall
Relevant items	0. 8333(5/6)	0. 7142(5/7)
Irrelevant items	0.1666(1/6)	0.2222(2/7)

Table 4.3 Evaluation of Recommendation

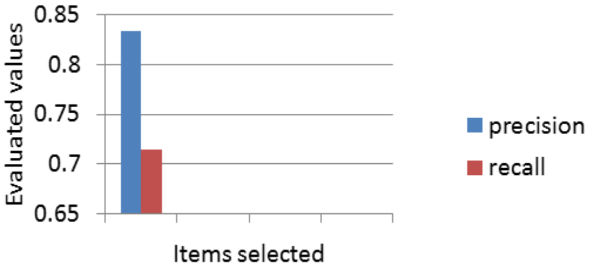


Figure 4.2 Graph for Performance of recommendation

V CONCLUSION

In this research, a Mapreduce framework for product recommendation was implemented. The main optimization of this method was the usage of multiple resources is taken for product recommendation framework where Fuzzy K-means clustering and Naïve Bayes classification were used as the corresponding functions. Case Based Collaborative filtering used to filter the relevant item, and Pearson correlation coefficient used to measure the similarity value items and Model Based Method are used to predict the relevant item to user input. The application of multiple resources is to have huge amount of data because a single source cannot effectively provide the quality of item recommendation.

REFERENCES

[1] S. Papadimitriou and J. Sun, "Disco: Distributed Co-Clustering with Map-Reduce", *8th Conference in Data Mining*, pp. 512-521, June 2008.

[2] D. Luo, C. Ding, and H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining",*12th International conference on Data Mining*, pp. 489-498, Nov 2012.

[3] C. Ranger, R. Raghu Raman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating Mapreduce for Multi-Core and Multiprocessor Systems", *13th Conference on Computer Architecture*, pp. 13-24, June 2007.

[4] C. Ranger, R. Raghu Raman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating Mapreduce for Multi-Core and Multiprocessor Systems", *13th Conference on Computer Architecture*, pp. 13-24, June 2007.

[5] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks", *IEEE Transaction on e-commerce recommender systems*, vol.no.1, pp.337-341, Dec.2012.

[6] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks", *15th conference on Knowledge and Information Systems*, pp. 603-630, Aug 2013.

[7] P. Resnick and H. R. Varian, "Recommender systems," *IEEE Transaction on Communications of the ACM*, vol. no. 3, pp. 56–58, June 2013.

[8] M. J. Pazzani and D. Billsus, "Content-based recommendation Systems," *IEEE Transactions in The adaptive web*, vol. no.5, pp. 325–341, Dec 2012.

[9] T. Zhang and V. S. Iyengar, "Recommender systems are using linear classifiers", *The Journal of Machine Learning Research*, vol. 2, pp. 313–334, May 2002.

[10] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources", *IEEE Transaction on Knowledge and Data Engineering vol.* No. 2, pp. 353-367, Mar/Apr 2003.