Recommendation Engine Using Apache Mahaout and MongoDB

Abhishek Girme, Akshay Ashtekar, Aishwarya Dhamale, Devayani Deshmukh, Mayura Kulkarni
Maharashtra Academy of Engineering, Alandi, Pune 412105, Maharashtra, India

Abstract

With so many e-Commerce website in place ,it become a need that you understand your customer in the best possible way and keep them surprised by providing them what they desire under their eyes. This can be done by building a powerful recommendation engine backed up by a non-traditional way of displaying the popularity through visual aid like graph and charts. Existing systems only intend do give recommendations only on historical user experience data. There is no such a innovation involved into it which may keep the user thinking. We proposed to built a strong Recommendation Engine with database support from MongoDB Replicated Engine and Warehouse support from Hadoop. MongoDB will take care of all transactional data generated on the website and will also work on input obtained from Hadoop.Hadoop will hold historical data and also data from differect API's like news and twitter.

KEYWORDS- Recommendation, Prediction, Hadoop, Apache Mahout, Big-Data.

I. INTRODUCTION

The information becomes an essential part of our daily life operations, but its huge quantity makes it difficult and time consuming to retrieve relevant information according to user preferences. An important drawback of the existing system is that its not flexible to users need and interest.

On the Internet today, an overabundance of information is created and used by users which makes it difficult for them to choose appropriate options, is termed as information overload. Recommender Systems (RS) are most commonly used software tools to help users in decision making process by applying information filtering, data mining and prediction algorithms. Therefore, RS is the solution to the problem of information overload and plays vital role in e-commerce

RS have the effect of guiding the user in a personalized way to choose interesting or useful objects in a large space of possible options. RS can be mainly classified into: Content-Based (CB), Collaborative Filtering (CF) and Hybrid RS. The CB method classify the user-item metadata and gives recommendation according to classification results. CF pre-dicts the overall rating for an item based on past ratings regarding both item individual and overall criteria, finally recommend an items to the user with best overall score. Recommendation systems are designed to analyze available user data to recommend items such as movies, music, or other goods to consumers, and have become an increasingly important part of most successful online businesses.

With big data analytics we are able to rapidly analyze both structured and unstructured data sets aiming to improve customer satisfaction.

For large and complex dataset MC-CF frequently gives better performance as well as accurate and high quality recommendations for users considering multiple aspects for items, which serves a win-win strategy to both users and ecommerce industry .In recent years, latent factors models i.e., dimensionality reduction techniques like: Matrix Factor-ization (MF) and Tensor Factorization (TF) have proved to be promising solution to the problem of designing efficient MC-CF algorithms in the Big Data Era .

For solving the problem of making recommendations cus-tomized for each use, recommendation systems apply knowl-edge discovery techniques. The scale of recommendation poses new challenges for recommendation systems due to high amount of web traffic. These systems have to face the dual challenge producing high quality recommendations as well as calculating personalized recommendations for millions of users.

Recommendation Engine can be built by a large scale dis-tributed batch processing infrastructure known as 'HADOOP'. The distributed processing of large data sets called big data across clusters of computers using Map Reduce and HDFS can be done using Hadoop. It makes it possible to store unstructured data by making use of HDFS and other projects that work with Hadoop like HBase, Hive, and Pig and many more.

The input to algorithm consists of user, item and rating to build recommendations using any of the following ways: A. User based recommendations are computed based on users with identical characteristics. B. Item based recommendations
are computed based on similar items. C. Slope-one: In this recommender, similarity metric is not considered as standard component. It is fast and simple approach for item recommen-dation.

II. EXISTING RECOMMENDATION PROBLEM

There are number of recommendation engine already devel-oped on large scale. The basic problem is that they only intent to give recommendations only on historical user experience data. There is no such innovation involved into it which may keep user thinking. Walking into a store and shopping for items is beginning to become a less attractive option of many people. The information becomes an essential part of our daily life operations, but its huge quantity makes it difficult and time consuming to retrieve relevant information according to users preferences. An important drawback of the existing system is that its not flexible to users need and interest. Online Shopping on the Internet has exploded registering leaps of growth by leaps and bounds year after year.

Users register and shop on popular shopping site like Ama-zon.com spending billions of dollars. Amazon collects basic profile information of the person and provides great service on recommending books, electronics and other products based on user profile, previous shopping history and relationships between the items categories derived from

purchases of all the users on the site. e.g. if you are viewing a book by the author Chetan Bhagat, it can recommend what books other people have bought either by the same author or of books in a similar category or a category that the shopper is likely to be associated with.

The mining of user's profile greatly enhances a person's shop-ping experience on Amazon as the person is able to receive information of the items relevant to the person's interests/needs

. Traditional in-store shopping has not been able to provide great service at-all, as you are a stranger every time you walk into a store to buy an item. Instead of a personalized service, you are left with a bunch of leaflets announcing discounts and offers from salt and sugar to High-Definition stereo that is on sale.

III. LITERATURE REVIEW

Manda Winlaw et al., [1] proposed an approach which is applicable to a broad class of optimization methods and collaborative filtering models. For definiteness, they choose a specific latent factor model, the matrix factorization model from and, and a specific optimization method, ALS. Given the data we use, the model is presented in terms of users and movies instead of the more generic users and items framework.

Weider D. Yu et al.,[2] Proposed an approach of Naive Bayes Model For Classification, NB classifier is composed of structures and parameters. The classifier has simple star-like structures and is unnecessary to learn, so the parameter estimation is the core of learning NB classifier with complete data. predictive data mining algorithm is also used for classification.

Anoj Kumar Mohd. et al.,[3] Praposed an novel approach for efficient personalized web search paper defines a an approach which tries to learn the behavior of user search to make search result relevant to the user . In specific, the following has been achieved: An automatically construction and updates user profiles has been defined. Clustering grouped the web pages into specific category it belongs.

Dheeraj kumar Bokde et al.,[4] Praposed an novel approach for The theoretical foundations of the methods used in this study is Multi-Criteria Collaborative Filtering, Dimensionality Reduction Techniques.aim of the author is not only to generate college recommendations for students, but to create awareness and providing exposure among them based on their interest.

Uma Sahu, et al.,[5] praposed an approach for Book recommendation system has been developed rapidly due to Web technology, which provides a new way to fulfill the users demands. The recommendation pages will contain all the essential book information for users to refer to. Users can rate the book of their choice, and the star rating data from different users with similar patterns will be analyzed by the recommendation system to make scientific recommendation decisions.

Lavannya Bhatia, et al.,[6]approach to locate frequent fea-tures across items as well as to find association among those features .They use a novel incremental algorithm to extract features from online item descriptions. They utilize association rule mining and the k-nearest neighbor method to create feature recommendations in the process of domain analysis.

IV. SCOPE OF THE RESEARCH WORK

We proposed to build a strong recommendation engine with data base support from MongoDB replicated environment and warehouse from hadoop. MongoDB will take care of all the transactional data generated on website and also work on the input obtain from hadoop. This will help us to play with the historical recommendations and mix it up with daily trends and design a secondary recommendation based on trends. Hadoop will hold all the historical data and also data from different APIs like News and Twitter.

The text based search will be running continuously on this data which is continuously define the popularity of different products as well as different categories. This sentiment analysis will help us n designing graphs using different visualization techniques and programming languages that will depicts the trends across all the categories as a comparative study plat-form. Apache Mahout will aid us to derive recommendations from the past transactions. This would then be dumped in to the Mongo DB server where further analysis will happen.

V. PROPOSED SYSTEM ARCHITECTURE

We proposed a system in which query from customer is fetch from website to query engine query planner interfaces with the MongoDB Replicated Cluster. data is transfer to flume through master and slaves. External Source Servers connected to API.

The proposed system to provide recommendations is divided into the following modules:

1) Data collection

2) Data pre-processing

3) Recommendation data building(Mahout)

4) Loading the final data to Serving Layer (MongoDB)

The data collection is the first stage for the recommendation system which consists of user profile records and user transactional records. Data pre-processing, which is the next step, describes any type of processing performed on raw data to prepare it for another processing procedure. Hadoop performs pre-processing by automatically clustering the whole data into clusters depending upon the kind of data. A recommendation algorithm

on item based collaborative filtering is run on Mahout. The results of this analyses undergoes post-processing which means that it is re-ordered before it is displayed on the recommendation page.

The model of recommendation system is made up of NoSQL databases, selenium scrapers, user interface, recommendation module and user profiles. Firstly, social networking sites like Twitter are used to obtain dummy user records from Twitter followers. These records are obtained using Selenium, a web scraper that captures data from a browser in the form of videos or images. The user records are extracted from these images and are stored in C.S.V. files which is then stored in NoSql databases like MongoDB and HIVE. In HIVE, a Map Reduce program is used to generate dummy transaction records for the users. Recommendations are then generated using a user based recommendation in Mahout. The recommendation module finds the appropriate recommendations based on user transaction records. These recommendations are sent to MongoDB where the user profile data is stored following which the recommendations page is generated. Recommendations are made on the basis of the transaction records of the users as well as the activity of similar users.So the recommendation system uses the combination of both content base and collaborative filtering to have better understanding of users interest. Recommendations are also generated based on the cost after considering the recommendations generated on the basis of the users interest. It compares the cost of the items from n different websites and computes the best available price for an item.

VI. RESULTS

kinds of recommendations. The data on the website is imported via the MongoDB console. Users can view all the recommendations based on history, transactions and cost. This is a user-based recommendation engine where every user is given a separate unique ID. If we use Mahout alone it provides implementation framework for non-distributed collaborative filtering. Our experiments which run in series using Mahout and Hadoop together enable us to start working with rec-ommender and begin evaluating their accuracy. To consider million preferences we require to implement recommender algorithm using distributed computing approach from Mahout based on map reduce paradigm and Apache Hadoop.

Our implementation framework can handle a data set of million ratings on single node. It is also efficient in computing the recommendations in real time. Our implementation, utilizes open source platform of Apache Mahout with Hadoop. This implementation is platform independent and performs distributed map reduce computation.

Desired result for Recommend Item Based can be described as follows: Users A, B, and C all liked items X, Y, and Z. The recommendation engine would determine that for an item X, similar items are Y and Z based on items that others with similar preferences have liked.Desired result for Recommend Cost Based can be described as follows: A user B wants to buy item X. X is available at store S and T. The recommendation engine would determine that X is available at a smaller cost at store S when compared store T.

VII. CONCLUSION

This involve converting the available data in the MongoDB server into a machine learning pattern with 3 fields product-id, customer-id and rating. Once in hadoop this would be provided to the mahout library where the recommendation will be generated. These generated recommendations can be then dumped back to the MongoDB server for loading on the website. For every user once he logs in we would be showing a graph with popular items based on his/her interest levels.The utilization of recommendation engine on a website is necessary to users or customers .

A website that contains recommendation and cost effective offers is a good way to attract user traffic. In conclusion, we can say that user information, right algorithms for the recommendation, permission from the user to make his trans-action and history available and the internet is effective in making a website grow. Relevance of a particular item to meet user need can be estimated by a recommender system using data mining algorithms. Though several methodologies for the implementation of recommender have been reported but none of these are efficient to compute generic recommendations using unstructured big data.

VIII. REFERENCES

Results of our implemented experiments show that Mahout makes it simple to explore machine learning algorithms and data mining concepts, and we can work with existing data models and test different components to generate different

[1] Manda Winlaw, Michael B Hynes, Anthony Caterini,Hans De Stercks,1 Algorithmic Acceleration of Parallel ALS for Collaborative Filtering:Speeding up Distributed Big Data Recommendation in Spark University of Waterloo, Canada IEEE 39th Annual International Computers, Software Applications Conference2015.

[2] Weider D. Yu, Choudhury Pratiksha, Sawant Swati, Sreenath Akhil, Medarametla Sarath A Modeling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment San Jose State University, One Washington Square, San Jose(Silicon Valley), California, USA 95192-0180 Weider.

[3] Anoj Kumar Mohd. Ashraf Efficient Technique for personalized web search using users browsing historyGautam Buddha University,Uttar Pradesh, India International Conference on Computing, Communication and Automation (ICCCA2015).

[4] Dheeraj kumar Bokde, Sheetal Girase, Debajyoti Mukhopadhyay An Approach to A University Recommendation by Multi-Criteria Collaborative Filtering and Dimensionality Reduction Techniques 2015 IEEE International Symposium on Nanoelectronic and Information Systems.

[5] Uma Sahu, Amiya Kumar Tripathy, Apurva Chitnis, Karen Aubrey Corda, Sharon Rodrigues Personalized Recommendation Engine Using HADOOP Department of Computer Engineering, Don Bosco Institute of Technology, Mumbai, India.

[6]Lavannya Bhatia,S.S.Prasad Building a Distributed Generic Recommender Using Scalable Data Mining Library Department of CSE, JSS Academy of Technical Education,2015 IEEE International Conference on Computational Intelligence n Communication Technology.