A Review Document on Apache Spark in E-Commerce Industry

Nikita Das
M. Sc. Big Data Analytics, Jai Hind College, Mumbai, Maharashtra, India

Abstract

E-Commerce is a largely growing industry day by day with millions of suppliers and customers carrying out transactions remotely and in a trusted environment. A product recommendation system in E commerce will be a great chance in increasing sales and improve business. Not only on the sellers side it would be beneficial, but customers will also get amazing recommendations of products at great deals and of different brands to choose for the best. Another way to use spark in E-commerce systems is Product quality risk assessment systems to avoid the risk of poor-quality product being sold in the name of online selling. Artificial Intelligent systems and Machine Learning systems work well on such case scenarios, but when it comes to E-Commerce, the volume of data and constant change in volume as well as the variety of data becomes a challenge. Scalability becomes a challenge and thus Big Data Technologies are the key to achieve such efficient systems with enormous amount of data. the recommendation system developed in this research is implemented on Apache Spark. Also, the matrix factorization using Alternating Least Squares (ALS) algorithm which is a type of collaborative filtering is used to solve overfitting issues in sparse data and increases prediction accuracy. This paper analyzes Apache Spark and highlights the role of Apache Spark (and eco-system) in the working of a modern E-commerce platform. Apache Spark, the trendy big data processing engine that offers faster solutions for any failures compared to Hadoop, can be effectively utilized in finding patterns of relevance useful for the common man from these sites.

Keywords: Recommendation System · Alternating Least Square (ALS)· Collaborative Filtering · Spark mllib · E-commerce · Risk Assessment · Performance Monitoring

1. Introduction

Product Recommendation Systems are E-Commerce strategies to build better business models and increase sales. A product recommendation is basically a filtering system that seeks to predict and show the items that a user would like to purchase. It may not be entirely accurate, but if it shows you what you like then it is doing its job right. The Recommendation System employs a range of technologies to filter the best result and to provide users with information they need. Recommendation System is divided into three broad categories: first one is collaborative filtering system, next one is content based system and the last one is hybrid recommendation system. Collaborative filtering is based on other user's similar choices and feedbacks and recommending to those users who have something in common. Content based filtering is related to keywords or the basic functionality of the products etc. They both can be combined to provide a better model which is the hybrid approach.

On the other hand, online transactions comes with a huge risk of receiving poor quality or defected products which creates negative imprints on public for e-commerce industries. E-commerce platforms are also concerned with monitoring its performance on everyday basis. Promoting products, providing cashback offers, festive deals and suggesting products based on customer behaviour also enhances the quality of the platform. The motive is to find out real-time problems and fix it as soon as possible before any major failure occurs and results in loss to businesses. To monitor performance of the Ecommerce model there

must be a model which constantly analyses the customer behaviour based on their browsing information and transactions made through the platform. These various models use apache spark so as to overcome the issues or real time large volume of data and perform distributed processing to achieve real time processing of data.

Spark being the second-generation big data engine comes with extended features. It provides machine learning library to perform ML algorithms for classification, clustering and association. It integrates functions of big data with machine learning models offering distributive processing of large data performing various functions.

2. Related Work

Previously Big Data framework Hadoop was used for data processing. Although it has been a major change in the field of data processing, Apache Spark provides better performance with extended features.
For a Product Recommendation System, the authors of " Product Recommendation System using Scalable Alternating Least Square Algorithm and Collaborative Filtering using Apache Spark in E-Commerce", they 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.

In another paper, the authors of "Recommendation System for E-commerce using
Alternating Least Squares (ALS) on Apache Spark" for developing a recommendation system a computationally effective algorithm for matrix factorization ALS (Alternating Least Square) is proposed. Here, the matrix factorization is compared with the collaborative filtering method.

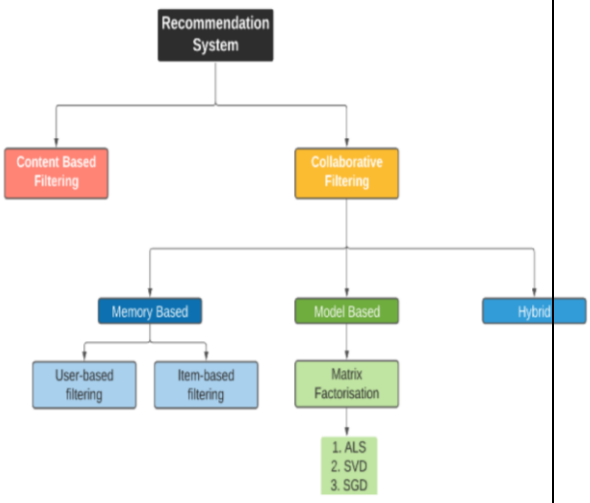
For the Risk assessment model, the creators of "The product quality risk assessment of e-commerce by machine learning algorithmon spark in big data environment" they have made a logistic model considering commodity environmental factors, product quality factors and customer service factors. The risk assessment indicators for the model are customer reviews, rating stars, commodity brand, material etc.

For performance monitoring, the authors of " Big Data Application Performance Monitoring in Retail E-Commerce using Spark" Spark Streaming is used to keep track of real-time data and processing according to it. Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analytics.

3. Apache Spark in E-Commerce Industry

3.1 Product Recommendation System

The collaborative Filtering algorithm captures the interaction between users and products. Resnick et algorithm described Collaborative Filtering for product recommendation based on the nearest neighbour principle. The users have different scores based on their interaction with various products. The algorithm gaters the users purchase related historical data.



In Hybrid Recommendation Algorithm the similarity between the objects is calculated to determine the customer preferences,

based on Tanimoto Similarity. Top N recommendations are calculated based on the Jaccard formula. The Mean Absolute Error (MAE) measures the accuracy of the algorithm, Where it considers the ratings predicted by the recommender algorithm the actual rating given by the users. The smaller value of MAE indicates the higher accuracy level of recommendations.

Collaborative filtering is done using the improved ALS algorithm in which the matrix decomposition in collaborative filtering algorithm includes ALS and SVD. It reduces the loss function and solves the overfitting problem by adding a regularisation function and Ridge regression is used to predict the outcome. The matrices to evaluate accuracy are classified into two major categories- Statistical Accuracy and Decision support accuracy matrices. ALS gives an accuracy of 0.95 and MSE of 0.11 whereas, Model based Collaborative filtering gives an accuracy of 0.81 and MSE of 0.33

E-commerce companies encounter challenges in personalized product recommendation like Amazon and Netflix. In these personalized settings users rate the items and the ratings data are used to predict to find ratings for other items. The rating data can be represented as an m x n matrix R where n = users and m = items. The R matrix is sparse matrix as items do not receive ratings from many users. Therefore, the R matrix has the most missing values. Matrix factorization is the solution of this sparse matrix problem. There are two k dimensional vectors which are referred to as “factors". Popular predictive accuracy metrics such as mean absolute error (MAE), root mean squared error (RMSE) and mean user gain (MUG) are used to measure the accuracy of the prediction made by the recommendation system.

3.2 Product Quality Risk Assessment

The paper proposes Naïve Bayes algorithm in Spark to measure product quality risk. Bayesian function calculate the prior probability and conditional probability values of various categories for the classified items, which the classified item belongs to the category with the greatest probability. Spark uses parallel processing as it has advantage of Hadoop and Map Reduce but it can be stored in memory without reading and writing in HDFS which gives the advantage of fast computing.
The, Naive Bayesian algorithm has some shortcomings, such as assuming that attributes are independent from each other and needing to know the prior probability, which will affect the accuracy of classification. The cost sensitive learning (CSL) algorithm can used to deal with imbalance data classification problems. The feature attribute is considered e.g. reviews, ratings etc., then train the prior probability of each sample and calculate conditional probabilities for each feature attribute. In the prediction phase, calculate the gailv of sample of each category and selecting categories with maximum probability of selected feature attribute. The model then evaluates the accuracy based on confusion matrix.

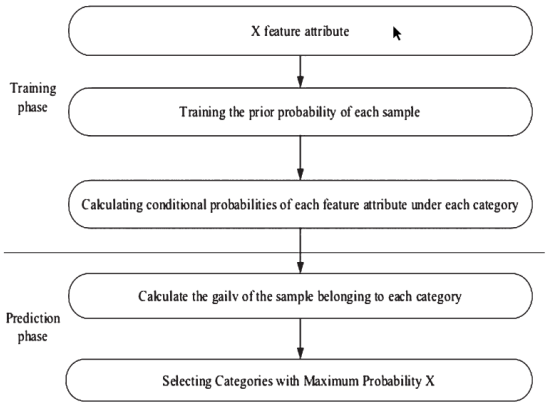


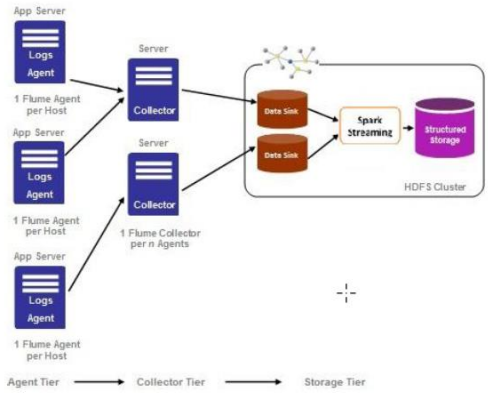
Fig. 6. The flow of Naive Bayesian algorithm.

3.3 Performance Monitoring in Retail E-Commerce

Apache Spark provides programmers with an application programming interface centred on a data structure called the resilient distributed dataset (RDD), Spark added a metrics system to allow reporting and monitoring of various internal and custom Spark application metrics. Install and configure the SPM monitor on each node in the Spark cluster using the standard package manager. Once all nodes are restarted, you should start seeing metrics appearing in the SPM dashboard within a few minutes. The main dashboard

provides a useful overview of what's going on in the cluster. The detail tabs on the side allow you to drill down into more detailed metrics for the Master / Driver, and Workers / Executors, and, of course, all key JVM and server metrics.

Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analytics. It ingests data in mini-batches and performs RDD transformations on those mini-batches of data. the streaming applications receives the retail stream and group them into batches with a suitable selection of batch interval. transactions are filtered at real time from the streaming based on browsing history. Using windowing function, all the relevant advertisements collected over a chosen interval of time is written to a text file. This intermediate result with listed transactions itself can serve as a source of information to the retailers.



4. Conclusion

This paper reviews various papers based on applications of Apache Spark in E-Commerce industry and narrowed it to three main models viz. Product Recommendation System, Product Quality risk assessment, Performance Monitoring of e-commerce platform. Each of the model is overall efficient when implemented on apache spark than just implementing a simple ML model as the E-Commerce data is huge and real-time, growing every minute and Spark provides faster computational power along with its ML library which makes it an ideal option.

5. References:

Product Recommendation System using Scalable Alternating Least Square Algorithm and Collaborative Filtering using Apache Spark in E-Commerce Bineet Kumar Jha1, Sivasankari G.G2, Venugopal K.R3

Recommendation System for E-commerce using Alternating Least Squares (ALS) on Apache Spark Subasish Gosh1, Nazmun Nahar2, Mohammad Abdul Wahab3, Munmun Biswas4, Mohammad Shahadat Hossain5 and Karl Andersson6

Challenges in Storing and Processing Big Data Using Hadoop and Spark
SHAIK ABDUL KHALANDAR BASHA, MTECH •
SYED MUZAMIL BASHA, MTECH • DURAI RAJ VINCENT, PHD •
DHARMENDRA SINGH RAJPUT, PHD

A Reference Architecture and Road map for Enabling E-commerce on Apache Spark
Mohit Sewak ,Sachchidanand Singh

A Review Document on Apache Spark for Big Data AnalyticswithCase Studies
Vivek Francis Pinto [1], Sampath Kini [2], Igneta Mcluren Dsouza [3]

Big Data Application Performance Monitoring in Retail E-Commerce using Spark
1Lavanya Marasa, 2Kalyani Kunchum

The product quality risk assessment of e-commerce by machine learning algorithm on spark in big data environment.
Yi Liua,b,∗, Jiahuan Lua, Feng Maoa and Kaidi Tonga