A Study on E-commerce Recommender System Based on Big Data

Xuesong Zhao
Oxbridge College, Kunming University of Science and Technology
Kunming, China
e-mail: 156613792@qq.com

***Abstract*—Recommender system algorithms are widely used in e-commerce to provide personalized and more accurate recommendations to online users and enhance the sales and user stickiness of e-commerce. This paper discusses several recommendation algorithms and the challenge of tradition recommender system in big data situation, and then proposes a framework of distributed and scalable recommender system based on Hadoop. The recommender system based on Hadoop, combining the advantage of computational ability and scalability of MapReduce and hybrid recommendation algorithms, brings a solution to information overload problem in big e-commerce.**

Keywords-recommender system; big data; algorithms; Hadoop

I. INTRODUCTION

Recommender systems provide products information and related suggestion to e-commerce buyers by imitating intelligent salesman and help e-commerce users to make decision and finish online purchases. The development of e-commerce provides personalized shopping experience to online users and improves the sales and stickiness of users of e-commerce website by accurately predicting users' preference and potential demands [1]. For example, Amazon.com, one of the e-commerce giants, has constructed a personalized online store with 29 million customers and several million items by fully applying recommender system. Every web user sees different items on Amazon websites. According to Microsoft Asia Research Academy, about 30% web page browsing of Amazon comes from recommender system. Similarly, Netflix widely uses recommender system and about 80% of the video watched were introduced by recommender system. The value of the Netflix recommender system reached one billion US dollar each year according to the Chief Product Officer of Netflix [2].

E-commerce recommender systems build users' models reflecting users' attributes and behaviors by collecting users' information as much as possible. Users' information can be acquired either by explicit feedback, such as purchasing and rating, which expresses obvious preference or implicit feedback, such as navigation history and links followed, which indirectly infer the preference of users [3]. After acquiring the information of users, the recommender systems filter and mine the attributes of the users by several learning algorithms to predict or recommend products that users may like to purchase. The recommendation process can be generally expressed by Figure 1.

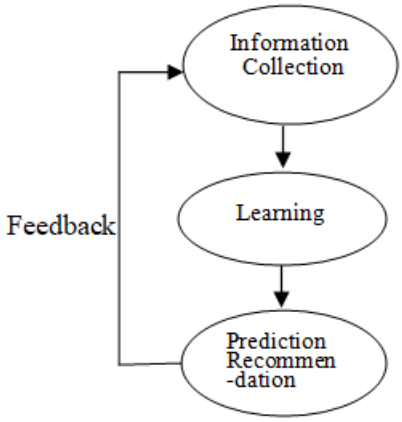

Figure 1. Recommendation process

II. RECOMMENDATION ALGORITHMS

In e-commerce recommender systems, the algorithms either focus on the users, finding the nearest neighbors of a target user and making recommendations to the target user with his neighbors' purchases or preferences, or focus on the products, recommending items that are similar to the items already purchased by the users [4]. Some algorithms provide personalized recommendations to target user while others make general recommendations. The commonly used algorithms are Collaborative Filtering, Content-based Filtering and User Clustering Models.

A. Collaborative Filtering

1) Collaborative filtering algorithm

Collaborative Filtering is widely used in e-commerce and is one of the most successful recommender systems in e-commerce. Collaborative Filtering focus on finding the nearest neighbors of target user who either purchased the same precuts or rated similarly on the same products with target user [5]. To find the nearest neighbors, Collaborative Filtering calculates and compares the similarities of a user with target user. The mostly common used algorithms to calculate the similarity of different uses is Cosine similarity. Suppose there are two users A and B, the Cosine Similarity of A and B can be expressed by (1).

$$\text{similarity}(\vec{A},\vec{B}) = \cos(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\| * \|\vec{B}\|} \qquad (1)$$

Since the nearest neighbor is found, Collaborative Filtering can recommend products that have purchased by the neighbor but have not purchased by the target user to the target user. The algorithm is bases on the following hypothesis: First, web users tend to have relatively stable preference; second, e-commerce can predict the future behavior of web user by his past purchasing behavior [6].

2) The disadvantage of collaborative filtering

While widely accepted, the effectiveness of Collaborative Filtering is limited by several factors [7]. The first one is data sparsity. Since most users tend not to rate or comment on what they buy, the user-item matrix will become data sparse which will lead to the difficulty of calculating and comparing the similarity between users and negatively affect the accuracy of recommendation.

The second factor limiting the effectiveness is Cold start which means when a new user first time comes to e-commerce website, the recommender system will recommend nothing to him since recommendation is provided with a user's past behavior.

The third factor is scalability of the recommender system. With the increase of the user and products, the computation ability needs to grow linearly. For some big e-commerce, Collaborative Filtering will possibly not generate satisfactory recommendation results if the data grows too fast. Scalability is a crucial problem for modern e-commerce with huge amount of users and products and transactions since the computation is very resource consuming.

We can partially solve the scalability problem by dimension reduction. For example, we can randomly sample in M users or just discarding those users buying only several items or discard the most popular and the most unpopular products in N items. But these data reduction measures will lead to poor recommendation quality meanwhile.

B. Content-based Filtering

Content-based Filtering mainly generates recommendation by analyzing products' attributes. In Content-based Filtering, products related to the positively rated or commented will be recommended to the user without knowing other users behavior or preference. Furthermore, if a user's preference change, Content-based Filtering will adjust recommendation in a short time. The main disadvantage of Content-based Filtering is that its effectiveness greatly depends on the extensive and in-depth description of products' attributes.. Another problem is that Content-based Filtering only recommend items similar to the products described in users' profile with the problem of content over-specialization.

The commonly used algorithm is Content-based Filtering is Association Rule Mining, which is a very useful method to discover the relationship hidden in large datasets. The uncovered relationships can be represented in the form of association rules or frequent items. The method aims to find the rules which can satisfy the preset minimum support and minimum confidence thresholds among the given data. The process can be divided into two steps. The first step is to find the frequent items set. Assume *minsup=s*, the frequent items are those items whose support probability is bigger than *s*. The second step is to find the association rule in which the confidences of frequent items are bigger than the minimum confidence. Assume *minconfi=c*, X and Y are frequent item sets which satisfy $X \cap Y = \emptyset$, and then when $P(Y \mid X) > c$, $X \rightarrow Y$ is the association rules.

C. CLuster Models

Cluster Models find the similar user set with the target user by dividing the users into many small segments. The goal of Cluster Models is to assign target user into the segment containing most similar users and make recommendation to target user by comparing purchased or rated items of similar users. Clustering or other unsupervised learning algorithms can help create the segment. A good clustering algorithm can produce high quality segment with high intra-segment similarity and low inter-segment similarity.

K-means is one of the commonly used algorithms of clustering. The first step is to select k data-objects from database as initial cluster center of k clusters. The second step is to assign every unselected data-object to one of the clusters by calculating the Euclidean value of this data-object from cluster center. Then calculate the clustering rule function as below.

$$E = \sum_{i=1}^{k} \sum_{0 \in C_i} \|O - M_i\|^2 \quad (2)$$

In (2), O stands for the selected data-object, M stands for the original cluster center, and E stands for the totaling of all the Euclidean values of all the data-objects in the database. The third step is to recalculate the cluster center by mean of all the data-objects in each cluster, if the new cluster center is the same as original one, the calculation finishes. Otherwise, repeat the second and the third step till E becomes convergent.

Compared with Collaborative Filtering, Clustering Models have better online scalability since they only compare the target user with limited number of clusters rather than all the users in the data base. What's more, large amount of computation in Clustering is finished offline. But the recommendation quality is low meanwhile.

III. The Challenges for E-Commerce Recommender System Based on Big Data

A. Big Data in E-Commerce

Big data generally refers to complex data that has 4V features, the technology to store, process, and analyze these data, and even the talents and organizations that can obtain practical meanings by analyzing these data. The 4V features of big data are: big volume, high velocity of data production and updating, rich variety and high value of data. The technology to storage, process and analyze the data refers to Hadoop: a distributed processing framework for large-scale data, NoSQL database with good scalability, machine learning and statistical analysis [8].

In modern e-commerce, big volume heterogeneous data are produced every second. The data set includes not only

the structured data such as uses' data, products' data and transaction data, but also non-structured data such as ratings, comments, thumbs-up, re-twitters and so on. It also includes the real time data source such as online click stream. The big data brings new challenges for recommender system in data process and data framework.

B. *The Challenges for Recommender System Based on Big Data*

In big data circumstances, recommender systems are different from those in traditional situation in many ways. The differences can be expressed by Table I.

TABLE I. THE DIFFERENCE BETWEEN RECOMENDER STSTEM BASED ON BIG DATA WITH TRADITIONAL RECOMMENDATION

	Recommendation based on big data	**Traditional recommendation**
Data input	Big volume, heterogeneous data	Small volume, structured data
Data type	Mainly implicit feedback data	Mainly explicit feedback data
Data update	Fast and augmenting update	Periodic and accurate computation update
Recommendation result	High accuracy	Low accuracy requirement
Recommendation Real time	High real time requirement	Normal real time requirement

We can see from the comparison that big data brings the following challenges for e-commerce recommender system.

- The recommender system with centralized framework has only limited computational ability and stand-alone algorithms are limited in data processing volume and effectiveness.
- Big e-commerce owns huge amount of users and products with many attributes, therefore, it's very challenging to build accurate and effective models for high-dimension users and products.
- The recommender system based on fixed models and parameters are hardly adaptive to the dynamic changes of e-commerce marketing especially when the users change their focus in different scenarios.

IV. THE DESIGN AND REALIZATION OF RECOMMENDER SYSTEM BASED ON BIG DATA

A. *Design Concepts*

E-commerce recommender system based on big data is composed of distributed file system managed data framework which is realized by cluster technology. The big data recommender system can satisfy the needs of effective, real time processing scalable huge volume data.

A typical clustering distributed computational framework is Hadoop MapReduce. MapReduce divides the data process into 2 functions: Map and Reduce. Map is responsible to segment the main task into many small tasks, and Reduce is responsible for integrating the results of many small distributed tasks. On the other hand, Spark, a rising big data processing engine, can satisfy the need of real time recommendation in big data circumstances by storing the intermediate computational results of users' real time click stream into memory and combining with traditional offline recommendation [9].

It's suitable for modern big e-commerce to use parallel algorithms other than sequential algorithms to generate recommendation since parallel algorithms are good at processing big volume and heterogeneous data effectively.

B. *The Realization of Recoomender System Based on Hadoop*

1) *The recommendation process*

We first extract the users' information, products information and preference information from the heterogeneous, multi-sources and noisy data to build users' model and products model by data ETL. And then construct several independent recommendation engines by utilizing different algorithms. After that combine all the recommendation generated by the engines into initial recommendation set. Sometimes, the e-commerce needs to recommend some specific items to the user, we can construct a candidate recommendation set including these items and combine with initial recommendation set to generate final recommendation list to the user after filtering. The process can be expressed by Figure 2.

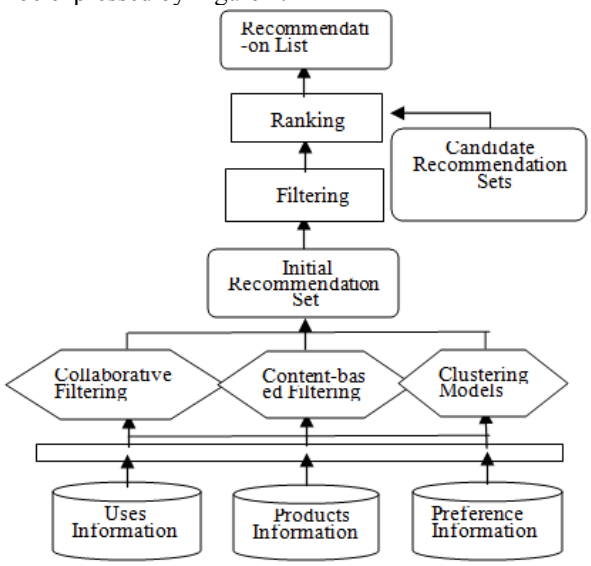

Figure 2. Process of recommender system based on Hadoops

2) *The framework of recommender system based on Hadoop*

The recommender system based on Hadoop is realized by layered architecture with different layers interacting by interface.

- Data access layer. This layer is responsible for integrating the heterogeneous, multi-structured and multi-sources data and extracting users' attributes and products' attributes which can help a lot to improve the effectiveness of recommender system. Meanwhile, this layer is also responsible for integrating the results of Hadoop clustering large scale data analysis.

- Data Model layer. Data Model layer is composed of users' model and products' model. Users' model includes users' profile, users' behaviors and preferences while products' model includes products' data, ratings and other feedbacks. Users' feature vectors and products feature vector can be acquired by users' attribute sets and products attribute sets extracted from original data.
- Algorithms layer. This layer includes several data mining algorithms and machine learning algorithms, such as Association Rules mining, Clustering and Collaborative Filtering. The algorithms layer is independent from recommendation layer for the purpose of integrating more algorithms.
- Recommender system layer. This layer is the core of the system. It packs algorithms into independent recommendation engines by calling the algorithms to complete the similarity calculation, association analysis and clustering and so on. The layer is functional scalable for more engines.
- Application layer. This layer is the interface of users with system which including configuration, management, interaction and exhibition. It allows configuring parameters and measures of hybrid recommendation models to assure the flexibility of the system.

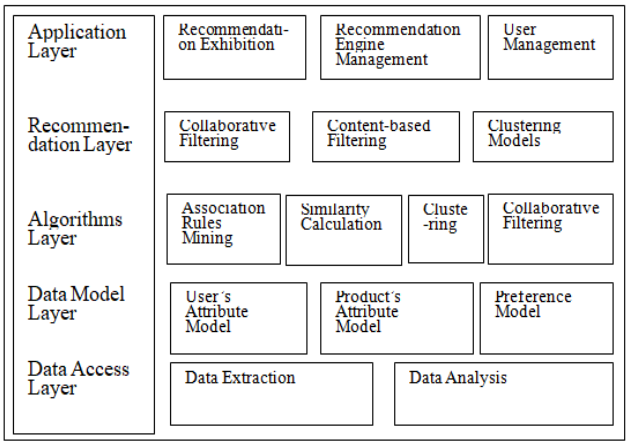

Figure 3. Framework of recommender system based on Hadoop

C. *Online and Offline Framework of Recommender System*

1) *Online Part*

There are 2 core modules in the online architecture. The first one is the service gateway which is responsible for the validity check of recommendation request and configuration of request response results. The second part is recommendation engine which is the essence of recommendation system and includes processing such as online logic, recall, filtering, feature computation, ranking and diversification and so on.

The data goes to 2 directions. Firstly, user's request is transmitted to the service gateway through traffic allocation module for validity check, and then it is transmitted to recommendation engine. Secondly, user behavior's data goes from gateway to Flume, which collects data from various web servers and stores them in HDFS, HBase and other memories, then to Kafka, providing real-time data processing for Spark. Meanwhile, it also provides data for offline part.

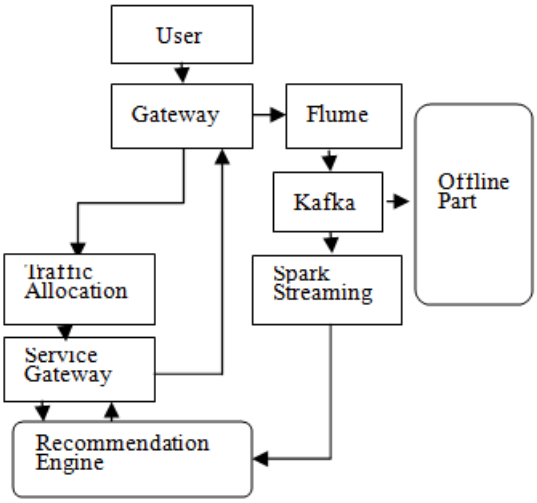

Figure 4. Online Part of recommender system

2) *Offline Part*

Offline part of the recommender system can be divided into data layer, recall layer and ranking layer.

Data layer contains data generation and data storage. It mainly uses various data processing tools to clean the original data, processes it into formatted data, and stores it in different types of storage systems for the use of algorithms and models.

Recall layer mainly uses various triggering strategies to generate recommendation candidate sets from the perspectives of user's historical behavior and real-time behavior, filters and combines the candidate sets according to product rules. Since the online system cannot rank such large number of candidate sets responding a single online request, rough ranking will be carried out in the recall layer first.

Ranking layer mainly uses machine learning model to rank candidate sets selected by recall layer.

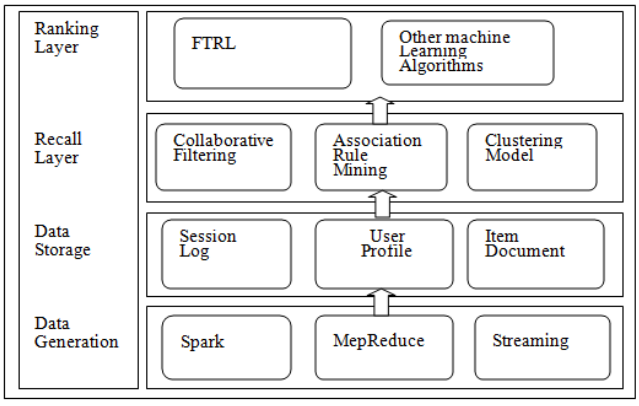

Figure 5. Offline Part of recommender system

V. CONCLUSION

In this paper, we discussed several commonly user e-commerce recommendation techniques and algorithms. With the challenge of big data, a Hadoop based mixed recommender system was proposed to be a reference for the design and realization of big data-based scalable, distributed and effective recommender system. With the combined advantages of Hadoop distributed computation ability and mixed recommendation, the scalable, flexible and diversified recommender system can obviously bring a solution to information overload problem in big e-commerce and provide sustainable competitive advantage for e-commerce with personalized marketing. The recommender system based on big data can be composed of online part and offline part to generate the optimized recommendations according to the real-time requirements and big volume of data in big data environment.

ACKNOWLEDGMENT

This paper is supported by scientific research fund (2015Z195) of Yunnan Provincial Department of Education.

REFERENCES

[1] Li Wenhai, Xu Shuren. "Design and implementation of recommendation system for e-commerce on Hadoop". Computer Engineering and Design .2014.1(35),pp:130-136;143.

[2] Brent Smith,Greg Linden. ”Two Decades of Recommender Systems at Amazon.com”. IEEE INTERNET COMPUTING.2017.vol 21,pp:12-18.

[3] Isinkaye FO et al. "Recommendation systems: Principles, methods and evaluation". Egyptian Informatics (2015).

[4] Greg Linden,Brent Smith,and Jeremy York."Amazon.com Recommendations Item-to-Item Collaborative Filtering". IEEE INTERNET COMPUTING.2003.Jan,pp:76-80.

[5] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. "Analysis of Recommendation Algorithms for E-Commerce". Proceedings of the 2nd ACM conference on Electronic commerce - EC '00 (2000).

[6] ZhiDan Zhao,Ming-Sheng Shang."User-based Collaborative-Filtering Recommendation Algorithms on Hadoop". 2010 Third International Conference on Knowledge Discovery and Data Mining. IEEE Computer Society,pp:478-481.

[7] Dong Liu. "A Study on Collaborative Filtering Recommendation Algorithms". 2018 IEEE 4th International Conference on Computer and Communications.,pp:2256-2261..

[8] Xuesong Zhao. "A Study on the Application of Big Data Mining in E-commerce". 2018 IEEE 4th International Conference on Computer and Communications,pp:1867-1871.

[9] CEN Kai-lun,YU Hong-yan,YANG Teng-xiao."Design and Implement of E-Commerce Real-Time Recommender System with Spark Based on Big Data". Modern Computer. 2016.8,pp:61-69.