The Research and Design of Big Data Trading Platform for Entity Business

Qian-Ying ZOU[1,a,*] and Lan LUO[1,b]

[1]Chengdu College of University of Electronic Science and Technology, China

[a]zqy_bb@163.com, [b]480706826@qq.com

* Corresponding author

Keywords: Big data trading, Entity business, Decision-making analysis

Abstract. An entity business based big data trading platform will be researched and designed against the defects of the traditional large scale business under the environment of big data. The platform will follow five steps: step one, use hardware equipment to collect basic information of the merchants, commodities and users; step two, store the collected data via distributed database; step three, filter invalid data with data cleaning technology; step four, find the connections among merchants, commodities and users by data mining; step five, visually present the needed results to customers. This brand new trading platform will be mainly applied to some large scale entity business because bringing it online can narrow the current gap between the entity business and the e-commerce platform in such aspects as user experience, users' behavior analysis, business intelligence and so on, thus improving the plight the entity business is confronted with to some extent.

1 Introduction

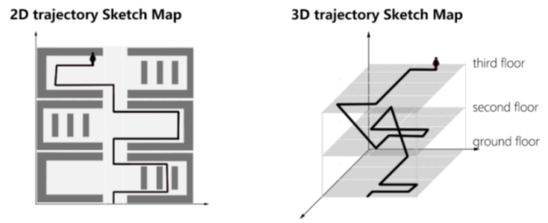
With the tide of "Internet plus" sweeping across the globe, new business platform typically represented by electronic commerce (e-commerce) prevails among people, while the traditional entity business, like department store and so on, suffered a great hit in this wave. Therefore, for the currently existed large entities, actively exploring new business philosophy, taking advantage of advanced big data technology and realizing the transformation of business platform are the essential problems currently confronting all entity business leaders and demanding prompt solutions [1].

This plan proposes a big data trading platform aiming at entity business. The platform will automatically generate the logical strategy of customer needs, carry out the data processing flow of collection, storage, cleaning, and then mining, and finally display the result visually. This platform is mainly used in the traditional commercial quantitative information service, namely, providing support for business decision and prediction, as well as reverse marketing program, enhancing customers' experience, and providing data basis for the ranking of advertising auctions to a certain extent.

2 Data Collection

Data collection module is divided into the following two parts: commodity position information collection and users' coordinates collection.

This platform can obtain the FCE (Food, Clothing and Expenditure) data by collecting the merchants' historical trading data, and provide convenience for the following data mining process by collecting the position information of the commodities and the coordinates of the users. The MAC information of the user's mobile terminal is collected as the unique identification of the user, and combined with the stored timestamp, it is treated as the time and space coordinates information [2]. The user's behavior trail can be constructed by data mining to predict the future trend. Fig. 1 is the diagram of the user's behavior trajectory coordinates, in which Fig. 1a) shows the two dimensional behavior trajectory of the user on the same layer of the business entity and Fig. 1b) shows the three dimensional behavior trajectory of the user in the whole business entity.



a)Two Dimensional Map of the User's Behavior Trajectory
b)Three Dimensional Map of the User's Behavior Trajectory
Fig. 1 Diagram of the User's Behavior Trajectory Coordinates

3 Data Storage

With the increasing amount of data in large-scale business bodies, the necessary data storage environment has greatly changed. The traditional centralized data environment which focused on the relational database is no longer suitable for the current situation. Therefore, it is urgent to solve the problem of data storage in big data environment. There are many drawbacks in the relational database in the big data environment, such as the imperfections of horizontal expansion, poor concurrent literacy of mass data and so on. Therefore, through the analysis of the shopping center's business environment characteristics, this scheme selects the HBase [3] distributed database in NoSQL database as the data storage carrier.

Based on the practically collected data, three module tables are set up, including Store Information Table, Product Information Table, and User Information Table. Merchant Information Table includes Store_Id, SL (Store Location), SN (Store Name), Product_Id, User_Id and so on, shown in Table 1.

Table 1 Store Information Table.

Rowkey	CF_Store
Store_Id	CF_Store: SL CF_Store: SN CF_Store: Product_Id CF_Store: User_Id

Commodities Information Table includes Product_Id, Product Name, PC (Product Category), PQ (Product Quantity), PT (Product Tag), Product Position, User_Id and so on, shown in Table 2.

Table 2 Product Information Table

Rowkey	CF_Product
Product_Id	CF_Product: PN CF_Product: PC CF_Product: PQ CF_Product: PT CF_Product: PP CF_Product: User_Id

User Information Table includes User_Id, US (User_Sex), UA (Use Age), UT (User Tag), UC (User Coordinate), Product_Id and so on, shown in Table 3.

Table 3 User Information Table

Rowkey	CF_User
User_Id	CF_User: US CF_User: UA CF_User: UT CF_User: UC CF_User: Product_Id

Stores and users can receive personalized feedback information through the diverse connections among these tables. The store's own attributes, Product_Id and User_Id are stored in HBase. The management of store attribute information can be optimized by storing Product_Id, and the information type analysis of users who visit the stores can be optimized by storing User_Id, thus providing marketing program and decision making support. According to the Product_Id stored in the table, users' recent preference judgment can be accomplished based on the differential value between the stored time stamp and the present time, thus storing the latest User Tag information.

4 Data Cleaning

In order to improve the data quality, regular data cleaning can clear away the missing, redundant, and abnormal data stored in the database. This scheme includes data cleaning module, which mainly presents the automatic data cleaning framework aiming at the large-scale business body and includes five modules, namely, early-stage preparation, data detection, quality evaluation, data correction, and data output.

During the early-stage preparation, a simple analysis should be made on the data stored Store Information Table, Product Information Table and User Information Table. In order to obtain impeccable data cleaning scheme, the cleaning target and concrete implement method need to be identified by combining corresponding applied system facilities.

Data detection needs to complete the data pre-processing and basic detection first, and then to conduct statistics about the detection results. The data pre-processing mainly aims to eliminate the inconsistent data, empty data, invalid data, etc. When the coordinate value which is collected by the hardware equipment placed around the shopping center exceeds the normal setting parameters value, or when users walk into the blind area of the business district and are covered by no signal, inconsistent data, empty data and valid data are then generated. In addition, data detection also needs to detect redundant data, missing data and abnormal data [4]. For example, when the information of the same user in the same position is collected repeatedly, redundant data is generated. While missing data, null value and abnormal data may probably be caused by the equipment trouble of the hardware or a manmade break down. When user's mobile device abruptly disconnects with Wi-Fi, an independent connection point emerges, causing that only the user's MAC information can be collected, but the attribute information is missed; therefore, an abnormal data is produced. Data detection process can help to obtain data information of a higher quality level, paving the way for the following data mining.

Data quality evaluation judges and evaluates the data quality according to the statistical result of the data quality detection. By integrating the business impact analysis and that of the problem essence, and by referring to the previously prepared data cleaning scheme, the method is improved and a new data cleaning scheme is obtained, i.e., data correction scheme.

Data correction adopts various methods to modify the detected data. Its general functions include eliminating tagged inconsistent data [5], deleting empty data and merging replicate data to reduce redundancy by sorting, merging, rule-based ways and so on. In order to realize this function, the multiple value problem caused by the same positioning information of the user should be solved firstly by using the priority team algorithm, in which each recorded field is considered as a long character string and is ranked twice. The second time ranking adopts the way of anti-tone sequence ranking compared with the first time sorting, and then puts the sequence into the queue again. Finally, the algorithm scans the priority queue and compares it with the stored data. In the function of missing data estimation and filling up, the K-means clustering is taken to reduce the impact of ignoring the missing attribute values on statistical analysis of data. It means that data cleaning tool gives K initial samples, merge the nearby samples into a larger set, and reintegrate the nearby sample points with the new set as the center, until that the last two clustered centers do not change much. Based on this, the sample points are sorted into k sets.

When submitting the data, the previous cleaning scheme should be integrated with the quality evaluation to verify the authenticity and rationality of the cleaned data. If it is true and rational, the data will be then submitted to the data mining module, but if it isn't, this framework can be repeatedly used to improve the data quality.

Data cleaning always runs through the whole life cycle of the data processing, which selects the data source that meets the input requirements, evaluate and correct the mined data, and ensure the correctness of the output data. This data cleaning framework provides high quality data at different stages with the merits of loose coupling, high flexibility, extensibility, and great interactivity.

5 Data Mining

Data mining is fundamental to providing support for decision-making. YARN frame is closely combined with multiple components, and various techniques such as MLlib, spark streaming, offline calculation, statistics and data flow processing are utilized to comprehensively analyze the transaction data of large-scale entity business, and properly make inductive reasoning, from which some potential correlation can be mined out. The basic framework is shown in Fig. 2.

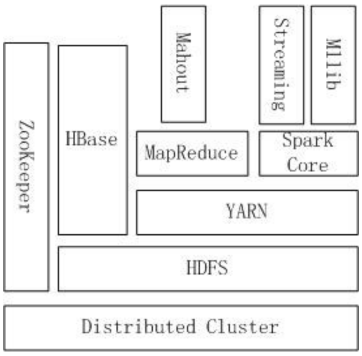


Fig. 2 YARN-based Data Mining Framework Using Offline and Spark Streaming

5.1 Offline Data Processing

Offline data processing mainly deals with the historic transaction data and mines it. The automatic prediction of the new data type can be achieved by firstly training the products sample set with the help of the classification algorithm in the Mahout data mining tool, and then establishing the classification model which includes classification rules.

5.2 Real-time Data Processing

Both the transaction between users and stores and the users' movement trajectory are new data occurred in real-time. In order to achieve real-time push, the MLlib machine learning program in Spark is adopted to quickly explore the association among stores, users and products. The Association rule algorithm is helpful to mine the potential relationship between stores and users, relationship between users and products, the relationship among stores, and the relationship among users, while the Dimensionality reduction algorithm can find the association among the products, users and their preferred products.

6 Data Visualization

In order to present a graphical correlation among the data, the Data visualization module graphically draws the results of data mining by some visualization tools like Baidu's open source project Echarts.

6.1 Product Information Visualization

With the guidance of the classified data of FCE platform, the visualization of product information can give off-line feedback on the products sales conditions, the information of peer competition and users attributes. It also gives real-time feedback on the crowd migration in order to judge the distribution of visitors flow rate, which serves as a reminder that merchants need to cope with the short-time surge phenomenon of visitors flow rate. The platform also provides relevant decision support so as to assist the merchant to carry out reverse marketing strategy.

6.2 User Information Visualization

Users' information can be pushed and spread by means of mobile terminal interface; thereby their information can conduct visual interaction. For instance, users of the same type can be clustered through the behavior trajectory information, and then real time information about group purchase and set meal can be pushed to them. Users' future position can be predicted through the movement direction, and then business information nearby can be sent to them. UT can be combined in order to push some selling platform to them, including their favored goods and discount information. Besides the above aspects, the preference degrees of users can be analyzed according to the consumption weight of FCE different types of commodities, thus the individualized push can be achieved.

6.3 Platform Provider Information Visualization

The visualization of the integral selling information of platform providers can remind merchants to carry out flexible marketing, extend market development, increase market penetration and achieve secondary profit by offering key words. For example, many campaigns can be held periodically to realize the maximization of market potential of hot-sale commodities. Moreover, promotion strategies, such as packing the present product and redemption for some unsalable goods with low prices, can also be adopted to promote selling.

7 Conclusion

The whole framework of the big data trading platform consists of hardware layer, architecture layer, storage layer, processing layer and application layer. The hardware layer includes distributed cluster, data collection unit, various network facilities and transmission facilities. The architecture layer is composed of Yarn framework. The storage layer takes advantage of non-relation distributed database Hbase to store source data, clean mined data and backup data. The processing layer carries out repeat mining and cleaning of the data from the storage layer, and eventually gets qualified information. The framework is shown in Fig. 3.

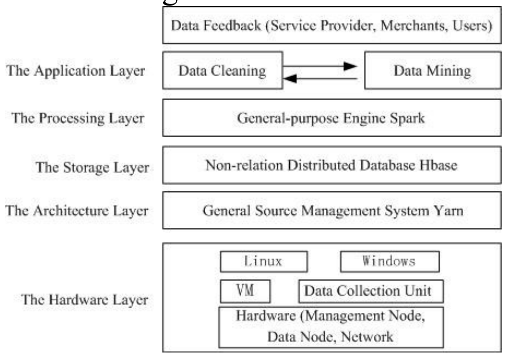

Data Feedback (Service Provider, Merchants, Users)
The Application Layer
Data Cleaning
Data Mining
The Processing Layer
General-purpose Engine Spark
The Storage Layer
Non-relation Distributed Database Hbase
The Architecture Layer
General Source Management System Yarn
Linux
Windows
The Hardware Layer
VM
Data Collection Unit
Hardware (Management Node, Data Node, Network

Fig. 3 Hierarchical Flow Graph

Although this framework of big data trading platform achieves the aims of automatically operating data, and providing decision support and personalized marketing program for decision makers, it still demands to integrate more functional parts, and needs continuous improving. In the future, the framework can be optimized from the following two aspects.

For data security, it requires mature solutions on data security issues to duly handle some confidential information, like users' behavior trajectory and merchants' trading data. For example, deploying related network security management facilities, like UniNAC network access control, UniAccess terminal security management system and so on to detect and find all kinds of abnormal behaviors and security threats in the net and then take relevant security measures.

For public opinion analysis, statistic analysis can help to analyze such information as reprint amount, revisit amount, and users' positive comments, and then research and determine the development trend within the "golden four hours" after the public opinion occurrence. In this way, we can know better about merchants' and users' evaluation on products and services, and then optimize some related services.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (61370073), and the foundation of Educational Department of Sichuan Province (172A0819).

References

[1] JiulaiHong. Salvation, needs feel the pulse accurately ——talk about the life and death of the physical bookstore. [J].Editors Monthly, 2013, (03):21-26.

[2] ShaliLiu, XizhongTan, ZhenhongJia. Research on WiFi indoor positioning system based on compressed sensing signal reconstruction [J].Laser Journal, 2014, (09):82-85.

[3] Cattell R. Scalable SQL and NoSQL data stores [J]. ACM SIGMOD Record, 2011, vol. 39(no.4): 12-27.

[4] JinyuSong, ShuangChen, DapengGuol. Data Quality and Data Cleaning Methods [J]. Command Information System and Technology,2013 (4):63-70.

[5] YuefenWang,ChengzhiZhang,BeibeiZhang,et al. Asurvey of data cleaning [J]. New Technology of Library and Information Service, 2007(12):50-56.