Data Mining Technology in Business Data Analysis

Haohua Qing¹,*, Jiali Zhang², Die Fu³
¹Department of Economic And Trade, Guangzhou College of Technology and Business, Guangzhou, 510850, Guangdong, China
²Department of Computer Science and Engineering, Guangzhou College of Technology and Business, Guangzhou, 510850, Guangdong, China
³Sichuan University of Arts and Science, Dazhou, 635002, Sichuan, China

*Corresponding author e-mail: 13986536@asau.edu.cn

Abstract. With the vigorous development of e-commerce, people pay more and more attention to data mining technology. This research mainly discusses the application of data mining technology in business data analysis. This study uses the clustering algorithm of buyers' purchasing behavior as an example to illustrate the definition of VMML in the clustering analysis of purchasing behavior. In this piece of XML, through the integration of data, an e-commerce research data warehouse is constructed, which is also used as the source data of EBizDAS. A certain amount of research has been done on buyers, sellers, commodities and transaction data, and the results obtained by EBizDAS analysis are displayed on the C2C e-commerce data analysis web platform for C2C users to browse and reference online. In this study, the 10 queries in CoHadoop took a total of 849 seconds, while the distributed strategy took a total of 748 seconds, with a performance improvement of about 12%, and the overall performance improvement was still considerable. This research has positive significance for the development of e-commerce.

Keywords: Data Mining Technology, Business Data, Data Analysis, Clustering Algorithm

1. Introduction
The explosive development of e-commerce has led to the massive and explosive growth of its related data. These data seem to be messy but contain great value. All e-commerce companies are eager to mine their own data and related data to gain a competitive advantage, thus Web data Digging is developing vigorously in China.

With the rapid increase of information, the public's perception of data has changed significantly [1-2]. Data has gradually become the main capital of the current business field and the choice of economic capital injection. All the above changes are due to the progress of science and technology, the reduction of computer network storage costs, and the huge changes in data communication
methods [3-4]. The recording function of data has changed from recording the past to a tool that uses past data as the main basis to predict the future development direction of things [5-6]. Throughout the process, the hidden value of data has been paid attention to by us [7-8]. Any e-commerce company can get tens of thousands of lines or even more access logs from the front-end Web server every day. These seemingly disorganized data records in detail every action of every user on the website that day. If these data can be presented in an understandable way, the value of these data will be difficult to estimate. Using the usage record mining in Web data mining, it is possible to realize this kind of "presentation". The data management of e-commerce companies has several similar characteristics: the data is massive and has an explosive growth trend; the competition to introduce Web data mining technology but the effect is not significant [9-10].

The data mining library is built on the Microsoft SQL Server2000 database management system. By parsing the DB Connection string, the program can obtain various information needed to connect to the database. After connecting to the database, the SQL language is used to select, convert, and load various analytical data sources (ETL). Store the data from each data source in the table corresponding to Table Name as the total data source of the analysis process. Each task in the process makes corresponding modifications to the source table according to the task type.

2. Data Mining Technology in Business Data Analysis

2.1 Data Mining

The amount of information in a data warehouse is often very large, and queries may involve multiple complex join and aggregation operations at the same time. To improve the performance of these queries, database administrators often use indexes. However, selecting the best index set is a difficult task because exponential candidate attributes can be used in the selection process. A data mining pruning method based on the largest frequent item set can solve this problem, which represents the candidate attributes of the index selection process. Compared with the existing methods, the main feature of the pruning method is that it uses other parameters except frequency constraints, and respects monotonicity and anti-monotonicity. These indexes minimize query processing costs and meet storage constraints. Classification and finding associations are the two main steps in the field of data mining.

The information entropy is:

$$H(X) = -\sum_{i=1}^{n} p(X_i) \log_2 p(X_i)$$  \hspace{1cm} (1)

Among them, \(i\) is the number of possible symbols for the source \(X\). Then:

$$H(X/Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(X_i/Y_j) \log_2 p(X_i/Y_j)$$  \hspace{1cm} (2)

Let \(|T|\) be the sample size of data set \(T\):

$$Info(T) = -\sum_{i=1}^{t} \left( \frac{freq(C_i,T)}{|T|} \right) * \log_2 \left( \frac{freq(C_i,T)}{|T|} \right)$$  \hspace{1cm} (3)

2.2 Business Data Analysis

The key to data analysis lies in the quality of data, so improving data quality is an important means to improve data analysis and performance. The actual database often has a lot of noise data, missing data (missing value, missing value) and inconsistent data, etc., and the data required for the mining task
comes from different data sources, all of which need to be processed. This task corresponds to a setup wizard on the client. The user can select the data sampling method. You can select the first N records, select one every N records, and select by percentage to sample the data. When the setup wizard is completed, the <Special Param> part of the XML will create a specific task parameter string, including the data sampling method, the number of records in the sampling method, etc. The server parses these task parameters, performs corresponding operations in the database, and finally gives prompt information for task execution. The updated analysis table of the database will be used as the input data for the next task.

3. E-Commerce Data Mining Experiment

3.1 Mining Algorithm
Each mining algorithm has a specific interface, and the parameters of each algorithm are different. These algorithm parameters are specifically defined in <Algorithm Param>, and are allocated to specific algorithm analysis and execution after being parsed by the EJB on the Server side. Here we take the clustering algorithm of buyer purchase behavior as an example to illustrate the definition of VMML in the cluster analysis of purchase behavior. In this XML, we define the algorithm task, the name of the model, the generated result table, the number of divided clusters, and the weight of the field.

3.2 E-Commerce Research Data Warehouse
Through the integration of data, an e-commerce research data warehouse is constructed, which also serves as the source data of EBizDAS. We have done certain research on buyers, sellers, commodities and transaction data, and used the results of EBizDAS analysis to display on the C2C e-commerce data analysis Web platform for C2C users to browse and reference online. At the same time, both buyers and sellers can also customize reports online to serve them in a targeted manner. The detailed information of the seller is shown in Table 1.

| Product category | Sales | Rank | Amount         | Total number of sellers |
|------------------|-------|------|----------------|------------------------|
| Mobile Communications Equipment Card | 1581 | 10044 | 2177290.00 | 295143 |
| Mobile phone communication equipment card >> Mobile phone accessories | 1581 | 578 | 2177290.00 | 42065 |
| Mobile phone accompanying equipment card >> Mobile | 1820 | 1820 | 3280.00 | 9671 |

4. E-Commerce Data Mining Analysis

4.1 Data Analysis
Fuzzy C-means clustering algorithm is used to cluster the captured data sets, that is, to classify customers. Given a threshold ε, judge whether the euclidean distance between the cluster centers obtained by two adjacent iterations is less than ε. If it is less, the iteration stops and the cluster centers are output as the final result. Otherwise, the iteration continues until two The cluster centers obtained in the next adjacent iteration meet the convergence condition. Fuzzy C-means theory is relatively complete. It is not only supported by profound and rigorous mathematical theories, but also has been successfully applied in many production fields and has become one of the most popular algorithms.
First, the customers are clustered into 3, 4, 5, and 6. From when the number of clusters \(\text{cata}=3\), the data characteristics are quite different, which satisfies the rough classification of the data sample; when the number of clusters \(\text{cata}=5, 6\), the trends of some of the central classes are the same and can basically be attributed. It is the same category, so 5 and 6 are not the best choice for the number of clusters; when the number of clusters \(\text{cata}=4\), the data classification display is intuitive, and the characteristics of each category are obvious, so choose \(\text{cata}=4\) as the optimal number of clusters. Fuzzy mean clustering algorithm divides customers into 4 categories according to customer transaction data according to consumption patterns. The attribution categories of customer data are shown in Table 2. Type I customer groups prefer black types and care about commodity prices. They usually choose middle-to-higher price commodities, which have a certain degree of churn. Customers have low trust in the website and have a certain value, but the merchants earn less profit and belong to potential customers. The second type of customer group prefers to choose the silver type, and the price of the product is in the middle range. The customer churn is small, and it belongs to the small customer group. Merchants can transform them into regular customer groups through reasonable marketing methods, and maintaining this type of customer groups is conducive to the development of e-commerce. Type II belongs to the regular customer group, preferring silver products, low prices, keen on discounts and promotions, and has a certain degree of recognition for the products on the website, which can bring higher profits to e-commerce. Category IV belongs to the high-quality customer group. This group mainly buys other forms of commodities such as packages and pays attention to high prices. There are a large number of customer groups, and they have the highest loyalty to website products. They are the key customer groups of e-commerce and can bring the highest profits to enterprises.

### Table 2. Attribution categories corresponding to customer data

| Factor Category | Color category | Unit price | Quantity | Total consumption amount |
|-----------------|----------------|------------|----------|--------------------------|
| I               | 11             | 1473.39    | 1        | 1473.39                  |
| II              | 33             | 1015.17    | 2        | 2030.34                  |
| III             | 33             | 251.49     | 1        | 2510.49                  |
| IV              | 44             | 2261.39    | 1        | 2261.39                  |

#### 4.2 Business Data Application Analysis

From the real business environment of a financial institution, we randomly grab 10 query sentences at business peak points through database logs, and named them Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10. In the experimental environment, it runs under three data distribution strategies and records the running time. In general, the correlation distribution has improved performance compared with the original HDFS distribution and CoHadoop distribution. Although the effect of the improvement over CoHadoop is limited, in this study 10 queries took a total of 849 seconds under CoHadoop, and a total of 748 seconds under the distributed strategy. Seconds, about 12% performance improvement, but considering that e-commerce and other mobile Internet application scenarios may have to deal with more than tens of millions of query scenarios every day, the overall performance improvement is still very considerable.
Figure 1. Comparison of query performance under different data distributions

4.3 Deposits and Loans of the Whole Bank

Figure 2 shows the deposits and loans of customers across the bank. In order to demonstrate the actual effect of the data correlation distribution strategy in query-oriented applications, we built a cloud computing environment. Because it is in the local area network environment, the default block size of the distributed system is increased to 256M. In the test, we did not use the data set generated by commercial data testing tools such as TPC-H. Instead, we selected a test data set containing 100 data tables from the application system of a financial institution. The purpose was to simulate the Internet such as e-commerce as much as possible. The scenario actually faces random query behaviors of a large number of online customers in the query business. This query returns all customer deposits and loan balances of a certain branch of the bank, involving connections between multiple tables. In the multi-table query, because the query results are aggregated, all records in each table must be polled, and keywords cannot be used to improve query efficiency. The main factor affecting query statements is the amount of data transmission. It can be seen that CoHadoop and CovStrategy have a greater effect on query efficiency in this scenario. The main principle is still in the shuffle stage of the Hive task. Since the related files are distributed in the same node space, the data transmission time is reduced.

Figure 2. Deposits and loans of customers across the bank
5. Conclusion
This research builds an e-commerce research data warehouse by integrating data, which also serves as the source data of EBizDAS. We have done certain research on buyers, sellers, commodities and transaction data, and used the results of EBizDAS analysis to display on the C2C e-commerce data analysis Web platform for C2C users to browse and reference online.

Compared with the existing methods, the main feature of the pruning method is that it uses other parameters except frequency constraints, and respects monotonicity and anti-monotonicity. Future research should focus on the method of transforming the existing traditional data warehouse into a flexible, low-cost, bottom-up data warehouse; transform the existing data mining model to make full use of the characteristics of the new data warehouse; Propose more new data mining models based on demand.

References
[1] Tao, Liling, Long. Imaging Examination and Quantitative Detection and Analysis of Gastrointestinal Diseases Based on Data Mining Technology.[J]. Journal of medical systems, 2019, 44(1):31-31.
[2] Mavaluru D , Arishi H . Internet of Things (IoT) Technology for Smart E-learning Using Data Mining: Review and Challenges[J]. Journal of Advanced Research in Dynamical and Control Systems, 2018, 10(13):626-632.
[3] Li L H . Design of College English Process Evaluation System Based on Data Mining Technology and Internet of Things[J]. International Journal of Data Warehousing and Mining, 2020, 16(2):18-33.
[4] Cao F , Gao J , Zhao J . Application of data mining technology in spectral data processing of chemical compounds in pulp[J]. Paper Asia, 2019, 2(3):154-158.
[5] A J X , B G D , A R L , et al. Analysis of medication data of women with uterine fibroids based on data mining technology[J]. Journal of Infection and Public Health, 2020, 13(10):1513-1516.
[6] Shen L , Xing Y , Lu Q , et al. Exploration of the meridian differentiation law in polycystic ovarian syndrome of hirsutism based on data mining technology[J]. Chinese Acupuncture & Moxibustion, 2018, 38(2):165-173.
[7] Jing L , Wang L . Research on data mining technology of intelligent substation information integration platform[J]. IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association, 2018, 38(2):466-475.
[8] Lammatha K K . Data Mining on 5G Technology IOT[J]. International Journal of entific Research & Management Studies, 2019, 8(5):24655-24660.
[9] Carnicer M B . The Use of Educational Data Mining Technique in Technology Assimilation Evaluation: A Response Assessment[J]. International Journal of Advanced Trends in Computer ence and Engineering, 2019, 8(3):654-657.
[10] Liu Y , Hong Y , Hu C . Research on Big Data Mining Technology of Electric Vehicle Charging Behaviour[J]. Elektronika ir Elektrotechnika, 2019, 25(6):55-61.