Application of Data Mining Technology Based on Data Center

Longjun Zhang1,*, Kun Liu1, Ilyar Ilham1, Jiaxin Fan1

1State Grid Xinjiang Information & Telecommunication Company, Urumqi, Xinjiang, 830000, China

*Corresponding author e-mail: zhnaglongjun@sgcc.com.cn

Abstract. Data mining technology refers to the use of mathematics, statistics, computer science and other methods to process a large amount of information to obtain useful conclusions and provide valuable decisions for people. With the rapid development and popularization of the Internet era and the more and more extensive application of computers in various fields, data mining technology has become a hot research field in today's society. Based on the data center, this paper studies the data mining technology. Firstly, this paper expounds the definition of data mining, and studies the process of data mining and the steps of processing data. Then, this paper also designs and studies the framework of data mining, and tests the performance of the algorithm. Finally, the test results show that data mining technology can well meet the target requirements.

Key words: Data Center, Data Mining, Mining Technology, Technology Application

1. Introduction

With the advent of the Internet era, computer networks have gradually entered our lives. People understand the information exchange between countries and regions around them through various channels. In this rapidly developing information society, data mining technology is a very important and meaningful technology for human beings. It has strong learning ability, high intelligence and high precision, and is widely used in all walks of life [1-2]. At the same time, with the advent of the Internet era, the computer network has been rapidly popularized and updated, so it will have a great impact on our life in the future.

Many scholars have studied data mining technology. Some scholars help the management to find the best solution by analyzing the data according to the established objectives, and help enterprises understand the market changes and make predictions [3-4]. With the development of the Internet and the continuous improvement of people's living standards, data mining is widely used in various fields, especially in information retrieval. Domestic researchers have combined data mining technology with ERP system idea to develop a set of procurement management system using data mining system to improve company decision-making [5-6]. These studies have laid a foundation for the research and application of data mining technology in this paper.

Data mining technology is to process a large amount of information in a complex, multi-level and diversified database, extract useful and valuable information and transform it into usable or usable tools. This paper mainly introduces the methods of data mining and modeling, as well as the algorithm...
research based on model classification and regression analysis. Firstly, a brief understanding of the existing relevant literature at home and abroad. Secondly, it expounds in detail that the data contains a lot of information and needs to be processed. Finally, an example is used to verify and summarize some common model types and application prospects to further explore how to apply these tools to real life, so as to better provide users with better services.

2. Discussion on Data Mining Technology Based on Data Center

2.1. Introduction to Data Mining

The data center is the collection of activities and data of various business units in the process of digital transformation of government and enterprises, and constructs the data construction, management and use system, including data technology, data management and data use. The data center is the core of the new framework information application system [7-8]. The data center is mainly used for the collection, analysis and extraction of big data from government agencies and enterprises. Data mining is an important technology to support data processing. From a technical point of view, data mining is a process of extracting knowledge with practical application value for industry or business needs from big data by using a series of algorithms and related technologies. These valuable knowledge and potential information are hidden in big data. They didn't know each other in advance. The extracted knowledge has many forms of representation, such as concepts, rules and forms [9-10]. Compared with traditional statistical analysis, data mining has the following characteristics:

(1) Strong big data processing ability, and can use data mining tools without excessive professional statistical training;

(2) Data mining tools are more in line with business needs in terms of use and demand;

(3) From different perspectives of data mining and statistical analysis, the ultimate purpose of data mining is to facilitate the use of enterprise end users, rather than for the detection of statisticians [11-12]. In data mining, we must first obtain data sources according to the actual needs. The obtained data sources may be of different structural types, such as structured, semi-structured, unstructured and so on. When mining data, there are some duplicate, wrong or missing data in the original data set, which should be cleaned up and deleted, and transformed into a data format matching the computer processing format. The algorithm processing of data set involves many technical fields such as data visualization, artificial intelligence and mathematical statistics. The potential information mined is conducive to decision-makers' management and decision-making.

2.2. Data Mining Process

Data mining is an iterative process. It examines the special relationships between data from a large number of data sets and finds patterns, rules and output values. The general execution process applicable to data mining includes the following steps:

(1) Problem definition and hypothesis clarification: according to the actual needs and the purpose of exploration, first put forward assumptions about the unknown correlation, select the appropriate algorithm, and formulate a detailed data mining plan.

(2) Data collection: there are basically two different methods: one is to collect data in a planned way under the control of demand personnel. On the other hand, demanding employees cannot influence the data generation process. Generally, the data is generated randomly. Customer transaction records, company databases, web crawlers, etc. are all data sources. Select a dataset for the mining task according to the target operation object.

(3) Data preprocessing: unavailable and abnormal data that cannot be directly processed by the computer shall be preprocessed before extraction. Delete bad data, clean up worthless records, process sensitive personal data, add missing data, convert data format, etc. Preprocessing can improve the standardization and consistency of data and the efficiency of data mining algorithms.

(4) Data mining: select appropriate data mining methods according to mining objectives, and select appropriate data mining algorithms according to data sources and user needs.
(5) Analytical model: select relevant angles to check the accuracy of mining results according to the initial mining objectives. Valuable information is presented to users visually.

(6) Application steps: apply the excavated information to the actual site.

2.3. Data Mining Framework
The purpose of data mining is to mine valuable hidden information from a large number of irregular data. According to different exploration objectives, appropriate algorithms can be selected to analyze, reconstruct, predict, interpret and evaluate the data, and finally get the results of data mining. It is delivered to professionals in relevant fields for analysis and printing, and the results are visualized through data visualization technology, so that users can draw the required conclusions from the data. The whole data mining process can be divided into the following basic steps: identifying tasks, preparing data, cleaning data, technical features, training and prediction models, analyzing and displaying results. The exploration framework is shown in Figure 1. First, determine the problem to which the goal of the exploration task belongs, such as classification, regression, time series, clustering learning, etc. Once the target is determined, the raw data can be collected in the scene. Data preprocessing cleans and formats the collected original data. This step is an important step in the whole data processing process, because the quality of data directly affects the prediction results, which directly ensures the effectiveness and consistency of data. The preprocessing stage mainly completes and corrects inconsistent data, misplaced data, missing data, abnormal data, etc. The preprocessing stage also needs to ensure that the data information is not lost and the original metadata is not tampered with.

![Data mining flow chart](image)

**Figure 1.** Data mining flow chart

In the functional engineering step, the sample after data preprocessing shall generate the functions related to the target variables according to the business logic, or deduce other functions according to the statistical principle, so as to better describe the data. This step is the most complex and time-consuming task in all data mining, and directly sets the upper bound of the expected result. Once the feature engineering training samples are prepared, the appropriate algorithm can be selected for training prediction. Data mining training is a spiral process. In order to improve the data mining model, we must continuously evaluate and adjust the parameters with the established evaluation function.
2.4. Association Rules

Association rule mining is to explore the interesting association or relationship between transactions and find the rules related to one group of events or data items and another group of events or data items. Frequent itemsets (frequent itemsets) refer to the high-frequency sub items extracted from the data set. The identified frequent itemsets are used to construct the association relationship between transactions. There is an itemset \( I = \{1, 2, \ldots, n\} \), where each element \( I \) is an item. Given a dataset \( D = \{t_1, \ldots, T_n\} \), transaction \( T_i \) has a unique identifier TID and is a subset of itemset \( I \), that is, \( t_i \subseteq I \).

1) \( X = Y \) support: there are itemsets \( x, y, X \in D, Y \in D \) and \( X \cap Y = \emptyset \). The expression of association rules is \( x \geq y \), which means that when itemset \( x \) appears in dataset \( D \), it can be inferred that itemset \( Y \) will also appear. The support of itemset \( X \cup Y \) is called association rule \( X = Y \) support, which is recorded as \( \text{Sup}(X \Rightarrow Y) \).

\[
\text{Sup}(X \Rightarrow Y) = \text{Sup}(X U Y) = P(D)
\]

Where \( P(D) \) represents the number of itemsets containing itemset \( L \) in dataset \( D \), and \( X = Y \) support represents the percentage of itemset \( X \) and itemset \( y \) in dataset \( D \) at the same time.

2) \( X \Rightarrow Y \) confidence: indicates the probability of the occurrence of itemsets \( X \) and \( Y \) when itemset \( x \) appears. The confidence of \( x = y \) is recorded as \( \text{Conf}(X \Rightarrow Y) \).

\[
\text{Conf}(X \Rightarrow Y) = \frac{\text{Sup}(X U Y)}{\text{Sup}(x)} = P(Y | X) = \frac{P(X \cup Y)P(X)}{P(X)}
\]

For itemset \( I \) and dataset \( D \), the association rules that meet the minimum support (min SUP) and minimum confidence (min CONF) are called strong association rules, and vice versa. In general, the user sets the minimum support min_ Sup and minimum confidence min_ Conf. The threshold value of conf, and the relationship satisfying the threshold value of these two parameters is the association rule sought by the user. Itemsets that are greater than the minimum support threshold are called frequent itemsets (i.e. frequency sets). Generating strong association rules using frequent itemsets is to calculate all possible association rules and their corresponding support and confidence one by one, which is divided into two steps:

1) For any frequent itemset \( X \) in dataset \( D \), all its non-empty subsets are generated.

2) For each non-empty subset \( x \subseteq X \), if confidence \( \text{Conf}(X = Y) \geq \text{min_ Con} \), then rule \( x = (x-x) \) is a strong association rule.

3. Experiment

3.1. Steps of Association Rule Algorithm

Generally, the association rule mining process is divided into two steps, as shown in Figure 2:

- **DS**: Data Set
- **Phase 1**: Phase 1
- **Rule**: Rule
- **Minimum support**: Minimum support
- **Maximum confidence**: Maximum confidence
- **User**: User

**Figure 2.** Associating rule mining steps

1) Identify common project sets. User set minimum support threshold min_ Sup and compare the size of each group of support elements with the minimum support threshold. All greater than or equal to min_ Sup. The element set of sup is a common element set. Association rules cannot be generated without frequent itemsets.

2) Generate association rules. User set minimum confidence threshold min_ Conf, search for
association rules with confidence greater than or equal to the minimum confidence threshold in each group of maximum frequency elements.

3.2 Data Acquisition
Data collection is to obtain targeted data from the external network environment and provide materials and resources for further data mining. During network exploration, data can be collected from server, client, agent or website database. Web server log file is an important information source of Web data mining. The data recorded in h-log reflects the user's access to information and the process of requesting information on the network. In addition to the information used, the server can also provide meta information about content, structure and website (such as file size, last change time, etc.). The web server relies on some tools (such as CGI scripts) to process the data information returned by the user's browser. The server identifies the uniformly requested file information according to the CGI standard, performs syntax analysis, and determines whether it is application specific information. After the CGI program runs, the server returns the requested custom information to the client's browser. Commercial websites often generate and collect a large amount of data in their daily operations. Most of this information is usually generated by the network server and collected through the server log. By storing cookies and CGI request parameters, the behavior of different users can be described. However, due to different caches in the network environment, this record may not be completely accurate, and the parameters passed by the post method are not recorded. The selected data elements include client address, cookie, date and time of access request, requested page, bytes, reference page, status, browser agent, and operating system. Another way to collect usage data from server logs is through packet monitoring technology. This technology monitors all incoming data on the server, extracts HTTP request information from it, and calls hidden CGI variables. But the scalability of this technology is very low.

4. Discussion

4.1 Data Mining Algorithm Performance Test
Table 1 is the data to test the performance of data mining algorithm.

| The field name       | Field meaning                      | Type | Length | Storage size (KB) |
|----------------------|------------------------------------|------|--------|-------------------|
| Sjwj goodscode       | Article number                     | N    | 10     | 144               |
| Xlyb_samplecode      | Sample serial number               | N    | 15     | 251               |
| Xlyb_BpNet1          | The network inputs the first node  | N    | 8      | 156               |
| Xlyb_BpNet2          | The network enters a second section point | N  | 8      | 285               |
| Xlyb_BpNet3          | The network enters a third node    | N    | 8      | 263               |
| Xlyb_BpNet4          | The network enters the fourth section point | N  | 8      | 248               |
| Xlyb_BpNet5          | The network enters a section 5 point | N  | 8      | 269               |
| Xlyb_BpNet6          | The network enter a section 6 point | N    | 8      | 230               |
| Xlyb_BpNetout        | Network output                     | N    | 8      | 287               |

As can be seen from Table 1, that when the system is running, the memory volume basically fluctuates at about 250, and the data length is basically very stable, indicating that the whole operation performance is stable and low consumption.
4.2. Mining the Data Performance Comparison

![Diagram showing storage size comparison](image)

**Figure 3.** Algorithm performance contrast

As can be seen in Figure 3, in the design, according to the input parameters, filter, establish the training sample data, select the sample data to be tested, and then integrate the sample data. Through the repeated learning process, finally obtain the learning experience and store the network knowledge. Input the actual inventory data of the enterprise into the learned and experienced network, get the experimental results, and express the conclusion of data mining in a form easy for users to understand.

5. Conclusion

With the continuous development of the Internet, data mining technology has been widely used, but it has not been well popularized and applied. This paper introduces an algorithm based on data analysis and prediction. This method is based on machine learning for information processing to realize the research on user behavior pattern recognition and decision support vector machine, and design the model for real life. Then an example is used to verify its feasibility and practicability. Finally, the experimental results show that this technology has obvious effect in the actual use process, and has good economic benefits and social value.

Reference

[1] Huang Q, Wang L, Li G. Research and Application of Video Encryption Technology Based on Chaotic Synchronization Theory[J]. IEEE Computer Society, 2018:444-447.

[2] Liu L, Zhai C. Research and Application of Data Mining in College Students' Employment Based on Association Rules[J]. Computer and information science, 2017, 10(3):54-59.

[3] DU Bin, TONG Chaoyang, Liu Zhiwei, et al. Research and Application of Biosensing Technology Based on Aptamer-Surface Plasmon Resonance[J]. Chinese Journal of Luminescence, 2017, 038(008):1039-1046.

[4] Li Xin, YU Xin, JIANG Qichen, et al. Research and Application of Dynamic Management System of Construction Materials Based on BIM Technology[J]. Civil Engineering Information Technology, 2019, 011(002):54-58.

[5] Gao X Y, Xiang W, Fang J L. Research and Application of Distributed Data Mining Method for Improving Rural Power Grid Enterprises in Production and Operation Status Evaluation[J]. Journal of Northeast Agricultural University, 2019, 026(002):87-96.

[6] Wu S, Wang M, Zou Y. Research on internet information mining based on agent algorithm[J]. Future Generation Computer Systems, 2018, 86(Sep.):598-602.

[7] Talebian S H, Najafabadi M E, Fahimifar A. Application of improved data-driven diagnostics
workflow based on coupled hydro-mechanical model for stimulation candidate selection in a complex carbonate reservoir[J]. SN Applied Sciences, 2021, 3(7):1-13.

[8] Ma X L, Zhai X, Liu J W, et al. Study on the Biological Basis of Hypertension and Syndrome with Liver-Fire Hyperactivity Based on Data Mining Technology[J]. World Journal of Traditional Chinese Medicine, 2018, 004(004):176-180.

[9] Xie Q, Chen X, Li L, et al. Image Fusion Based on Kernel Estimation and Data Envelopment Analysis[J]. International Journal of Information Technology &amp; Decision Making, 2019, 18(02):487-515.

[10] She Xiaodong. Research and application of optical fiber communication technology based on coal mine[J]. China New Communications, 2017, 019(010):99-100.

[11] WANG Feng, WANG Tao, XUE Xiaojuan, et al. Research and Application of Key Technology of UAV Supervision Based on Beidou/Mobile Communication Technology[J]. Beijing Surveying and Mapping, 2018, 032(012):1418-1423.

[12] Sun Yuhua, Wu Yonghuan, Liao Jiawei. Research and Implementation of Mobile Migration Technology of Application System Based on DAAS[J]. Computer and Digital Engineering, 2019, 047(004):906-910.