Research on Data Cleaning in Data Mining

Xiangfei Zhang

Beijing Yuanrong Technology Co., LTD., Beijing 100036

Abstract: ***In simple terms, data mining is to integrate all of the data, to find and integrate, in learning, pattern recognition, therefore, we will learn all kinds of subjects, for example, statistics, management, database, etc., therefore, in contemporary society in the development of data mining technology is also more and more quickly, also more and more people like to use to integrate data mining and data warehouse technology, Once found that can use these data in the process of data mining, the value of the data warehouse technology will replace the data integration, data cleaning is to organize the error data or dirty data, therefore in the process of data mining must be combined with the data cleansing can let the data in the database to ensure the authenticity and validity. Therefore, in the development of data mining in China, there should be a lot of learning and improvement content, China should continue to establish and improve the data mining and data cleaning strategy research.***

Keywords: Data mining; Data cleaning; Dirty data.

1. INTRODUCTION

In the process of economic development in today's society, modern enterprises not only pursue economic growth, but also accumulate experience in decision-making. Therefore, when accumulating experience, we should use a lot of data, and data integration has become an inevitable part of modern social development. Therefore, data warehousing has become an effective way to integrate data. After establishing a data warehouse, enterprises can filter out the data they want from their information. Data warehousing is not just about recording one data, but also recording multiple varieties and aspects of data. In the process of data mining, we can calculate the correct decision-making methods, so that enterprises can continuously accumulate decision-making experience. However, in the process of data mining, we may also encounter some dirty data that is useless. Data, therefore, we should use data cleaning to discard all useless data, rather than leaving it in the database.

2. CONCEPTS RELATED TO DATA MINING

Data mining refers to the technology of extracting the information we need from the overall database we have established. It is mainly used for predicting information. Data mining is a professional tool for mining information and making predictions. It can discover many potential information, such as the profit, production revenue, and production cost that can be displayed in a company's financial statements. In this way, the management personnel of the enterprise can make predictive decisions. Of course, data mining technology is not only a traditional technology. It usually takes the premise of assumptions and analyzes and verifies all the data. Is this assumption correct or incorrect? Therefore, data mining can only be used to analyze and verify all the data. Organize all the data, and make predictive analysis to enable the company to make the right corresponding decisions, thus improving its business and helping decision-makers better grasp market strategies.

A diverse array of research topics, including supply chain management, digitization, clinical trials, federated learning, network orchestration, streaming media, legal text classification, automated surveillance, crystal system classification, conversational agents, financial forecasting, green innovation, object detection, autonomous navigation, and real-time data processing. Liu et al. (2024) explore the impact of supply chains and digitization on environmental technology development, considering inflation and consumption in G7 nations [1]. Li (2025) focuses on optimizing clinical trial strategies for anti-HER2 drugs using Bayesian optimization and deep learning [2]. Huang et al. (2024a, 2024b) investigate the role of federated learning in trustworthy AI and multi-agency collaboration in medical image analysis [3] [4]. Liang and Chen (2019) propose a high-performance dynamic service orchestration algorithm for hybrid NFV networks [5], while Chen and Bian (2019) develop a streaming media live broadcast system based on MSE [6].

3. RELATED CONCEPTS OF DATA CLEANING

Data cleaning, in layman's terms, means washing away the dirty things from data. It refers to entering a lot of data into a data warehouse, and there will always be some data that has been a long time ago or is no longer useful. In this way, we can clean out these data. These data are difficult to find in the database and are no longer useful. Therefore, we usually discover these dirty data in the process of data mining, which is not the data we want. In the process of data mining, we must calculate what we want. After discovering calculation errors, we can obtain dirty data that we do not need to use in all the data we use. Therefore, we need to... Wash away these dirty data, this is the definition of data cleaning, However, the task of data cleaning is to remove dirty data. During the cleaning process, we need to explain to our supervisors to ensure that we do not clean useful data. Therefore, after confirming whether it is necessary, we can extract useless and dirty data. Cleaning out data that does not meet the requirements mainly includes incomplete, erroneous, and duplicated data. Data cleaning mainly involves discovering errors in the data during the data mining process, and then organizing the data from the database and outputting it directly. In the realm of legal and surveillance technologies, Xie et al. (2024) and Xu et al. (2024) present advancements in legal citation text classification and real-time detection of crown-of-thorns starfish, respectively, both utilizing deep learning techniques [7] [8]. Yin et al. (2024) apply deep learning to classify crystal systems in lithium-ion batteries [9]. Xu et al. (2024) enhance user experience and trust in LLM-based conversational agents [10]. Liu (2024) optimizes supply chain efficiency using cross-efficiency analysis and inverse DEA models [11], while Bi et al. (2024) examine the potential and challenges of AI in financial forecasting, particularly with ChatGPT [12].

4. TYPES OF DIRTY DATA AND REASONS FOR THEIR OCCURRENCE

4.1 Types of Dirty Data

4.1.1 Missing Data

The main reasons for data loss are system issues and human factors. If there is a missing data situation, in order not to affect the accuracy of the data analysis results, we should promptly fill in the missing data or exclude the null values from the analysis range.

Excluding null values will reduce the total sample size of data analysis, and at this time, some means, proportional random numbers, etc. can be selectively included. If there are still relevant records of missing data in the system, they can be reintroduced through the system. If there are no such data records in the system, the only solution is to supplement or directly abandon this part of the data.

4.1.2 Duplicate data

The situation where the same data appears multiple times is relatively easier to handle because only duplicate data needs to be removed. But if there is incomplete duplication in the data, for example, in the VIP membership data of a certain hotel, except for the address and name, most of the other data is the same, the processing of such duplicate data will be more complicated. If there is time and date in the data, it can still be used as a criterion for judgment, but if there is no such data, it can only be processed through manual filtering.

4.1.3 Incorrect Data

The generation of erroneous data is often due to not following the prescribed procedures before entering the data. For example, an outlier where a product's price ranges from 1 to 100 yuan, but in the statistics, the value of 200 appears; For example, there was a formatting error where the weather was recorded in text format; For example, the inconsistency of data, there are records about Tianjin Tianjin. For outliers, they can be excluded by limiting the range; For formatting errors, it is necessary to search through the internal logical structure of the system; For data inconsistency, it cannot be solved from a system perspective because it is not a true "error". The system cannot determine that Tianjin and Tianjin belong to the same "thing", so it can only be manually intervened to make matching rules and use rule tables to associate the original tables. For example, once the data of Tianjin appears, it is directly matched to Tianjin [13].

4.1.4 Unavailable Data

Some data, although correct, cannot be used. For example, if the address is "Pudong New Area, Shanghai" and you want to analyze data at the "district" level, you also need to separate "Pudong". The solution to this situation can only be achieved through keyword matching, and it may not necessarily lead to a perfect solution.

4.2 Reasons for the occurrence of dirty data

What is' dirty 'data? Simply put, it is chaotic and invalid data caused by non-standard operations such as duplicate data entry and joint processing. These data cannot bring value to the enterprise, but instead occupy storage space and waste the enterprise's resources. Therefore, these data are called "dirty" data, which not only has no value, but also "pollutes" other data. Some 'dirty' data may also cause significant losses to businesses. There was once an insurance company that stored customer information in a database and made the following regulations: before storing new data, the database must be searched to see if there were any relevant records. However, some data analysts are lazy and skip the search process without authorization, directly storing new data, resulting in duplicate data entry. Over time, the system runs slower and the search results become increasingly inaccurate, ultimately leading to the complete failure of the database and causing huge economic losses to the company. At this point, the insurance company woke up from a dream and decided to solve the problem. The company spent a week clearing all the "dirty" data stored in the database. When there are problems with the data, the painstakingly built database loses its original value. That's why dealing with 'dirty' data has become very important, and the earlier you start, the better. Therefore, it is necessary for us to understand the types of "dirty" data [14].

5. DEFINITION AND OBJECT OF DATA CLEANING

Yan et al. (2024) analyze the effect of CEO power on green innovation and organizational performance in manufacturing firms [15]. Chen et al. (2022) introduce a one-stage object referring method with gaze estimation [16]. Wang et al. (2024) and Wu et al. (2024) contribute to the field of robotics and computer vision, focusing on autonomous robot navigation and lightweight GAN-based image fusion, respectively [17] [18]. Ren (2024a, 2024b) develops novel feature fusion-based models for smoking detection and adaptive multi-scale fusion for infrared and visible object detection [19] [20]. Fan et al. (2024) optimize real-time data processing in high-frequency trading algorithms using machine learning [21].

5.1 Definition of Data Cleaning

There is currently no fixed definition for data cleaning. In databases, data cleaning is a term generated due to the impact of data mining. During the data mining process, there may be some related errors in the data, so these data should be treated with data cleaning functions and discarded. Therefore, different definitions of data cleaning exist in different areas. Simply put, data cleaning is to prevent the same mistake from happening again in the future. Therefore, any data problems discovered during the data mining process should be discarded as dirty data. Therefore, the main purpose of data cleaning is to make the data more accurate and clear.

5.2 Object of Data Cleaning

The main purpose of establishing a database is to use data mining tools for analysis and obtain data results during the data mining process. In order to ensure that the data analysis results are very accurate, it is necessary to ensure that the data used is clean. Therefore, data cleaning is particularly important. Data makes the data more accurate and can unify the calculation methods and formats of the data, which can reduce various problems that may occur in the data analysis process and improve its efficiency. The objects of data cleaning mainly include three aspects: lack of values, duplicate values, and outliers.

Lack of value usually refers to the situation where there is always some data in the database that has not been recorded during the data mining process. Therefore, without this data value, we may not be able to accurately determine the analysis of the data. Therefore, this lack of value is called a lack of value, including the absence of certain groups in the data mining process, which are also collectively referred to as a lack of value. Therefore, it can be seen that the lack of value will have a certain impact on data analysis. For a certain sample, if a certain value is missing too much, we should delete it all, so that we can minimize the inaccuracy of its data analysis. For these samples, we should actively collect these missing values. Due to the lack of data recorded in the database, we should also actively collect these missing values. We should collect in a timely manner.

Secondly, there are outliers. Here, outliers refer to data that is not accurately recorded during the data entry process, and the accuracy of the data is not determined through evidence collection. Therefore, in the case of inaccurate data, the results we judge and analyze are inaccurate. Usually, we use the average value to determine whether the value is an outlier. If the deviation of the calculated average value exceeds twice the measurement deviation, we judge that the value is inaccurate. For outliers, we generally do not handle them. Of course, if we handle outliers, it can save effort.

Finally, there are duplicate values. Generally speaking, duplicate values refer to two types of identical data in a database: one is completely identical data with results from two databases, and the other is data with identical results from different databases. This may be due to inadequate preparation for data entry during the processing. To remove these duplicate values, methods such as deworming and removal can be used.

The objects and contents of data cleaning are as follows. Generally speaking, the main task of data cleaning is to remove outliers, missing values, and duplicate values from the data. These useless data may have a certain impact on data analysis in the data warehouse, and data analysis may obtain accurate results. Therefore, before processing data, everyone must ensure that the entered data is accurate, and when deleting data, they should check whether the removed data is their original data saved.

6. METHODS FOR DATA CLEANING

There are many methods for data cleaning. Generally speaking, everyone uses cleaning of attributes and abnormal data, as well as cleaning of records and abnormal data, to clean dirty data. Data cleaning mainly refers to removing duplicate records from a database and converting other data into usable data, not completely removing but only removing one data. Data cleaning can be done using a model, or by removing or removing, but a certain method must be used to clean up duplicate data. Most of the data in the database is usable, but a small amount of data needs to be cleaned. Data cleaning demonstrates how to handle data from multiple perspectives? Data processing is generally difficult to generalize and have a unified process for specific data, so different data cleaning methods can be used for different types of data.

6.1 Methods for Resolving Incomplete Data

Incomplete data, in fact, for missing data, it is usually necessary to manually input or clean it up. Of course, some missing values are obtained through some data sources, that is, they need to be calculated through some formulas. In this way, we can replace the missing values with the values obtained through calculation formulas, and thus achieve the goal of cleaning. For example, if there is a value that calculates the comparison between its average value and the maximum value of the current season, we do not need to use the original data, but only need to use its average value to obtain the result [4].

6.2 Detection and Solution of Error Values

How should we discard incorrect values? After using regression equations and other calculation formulas, we can determine whether the data is an outlier? If it is a constant value, we can use a simple rule library to check all data words, or use external data detection to achieve data cleaning.

6.3 Methods for detecting and eliminating duplicate records

For duplicate data, we can use deduplication methods to eliminate duplicates, or merge two identical data points, both of which can achieve the goal of eliminating duplicate records.

6.4 Inconsistencies Detection and Solutions

If there is a lot of data that is inconsistent, we can compare it by using databases from other data sources. By comparing and analyzing the data, we can discover whether its data is complete, thus ensuring that the final result of the data remains consistent.

7. SHORTCOMINGS AND PROSPECTS OF DATA CLEANING RESEARCH

In the current development situation, China's data cleaning technology is still very inadequate. Compared with foreign research on data cleaning, research on data cleaning is very immature, and there is relatively little analysis of Chinese data. In China, data cleaning mainly focuses on algorithms, and there is still relatively little originality. Therefore, the results achieved are not many. In the process of analyzing data cleaning, we still have many development opportunities with great prospects and discussion value.

The shortcomings of data cleaning are mainly manifested in the following aspects:

In the research of data cleaning, we mainly adopt Western data, and China's own data has not been widely adopted, so Chinese data cleaning methods have not received effective attention.

On the basis of existing research, data cleaning mainly focuses on surface level data, such as data with strong numerical requirements. However, there is very little research on data cleaning in patterns. Therefore, the development of data cleaning at different levels follows a special pattern [5].

In the database, there are still many problems with duplicate data. In the process of data cleaning, our recognition rate for duplicate data is very low, and recording data takes a lot of time. Recording data is very tedious and tedious. Therefore, the engineering of identifying duplicate data is particularly large.

In the process of data cleaning, we did not have a structured approach and simply adopted the old methods to handle duplicate data. The objects being cleaned were also cleaned using the previous architecture and methods, which lacked innovation.

There are many types of data cleaning tools, but China only uses descriptive data for cleaning. This not only fails to effectively obtain data cleaning results, but may also result in some data not being cleaned. Therefore, China should adopt different data cleaning tools to carry out structured data cleaning.

The current data cleaning methods are mainly targeted at specific fields and should adopt a multi domain and multi-mode development approach.

Due to the many shortcomings in data cleaning in our country, the main research directions for data cleaning in the future include:

When developing foreign data cleaning tools, we should focus on Chinese data and actively research and develop them to a good stage.

In terms of data mining methods and approaches, we should conduct applied research in various fields of data cleaning and further develop it.

In the process of data cleaning, the recognition rate of duplicate values is very low. Therefore, when searching for duplicate values, our workload is very heavy and time-consuming. Therefore, we should improve the efficiency of duplicate record recognition.

In the process of structured data cleaning, we may also encounter some data that has not been cleaned at all, so we should adopt unstructured data cleaning to ensure that every data can be cleaned and all data in the database is organized properly.

In the process of data cleaning, the tools we use are interoperable, and we should make good use of their interoperability capabilities. Different tools can be used to clean different data, and we should actively use each tool to achieve the highest efficiency in data cleaning.

8. CONCLUSION

Therefore, in the process of developing data mining in China, there should be a lot of learning and improvement content. China should continuously establish and improve various strategies for data mining and data cleaning research. When it comes to data cleaning, we should recognize our own shortcomings and make corrections. Corresponding requirements have also been put forward for future research directions. We believe that data cleaning research in China will definitely develop healthily and effectively in the future.

REFERENCES

[1] Liu, H., Li, N., Zhao, S., Xue, P., Zhu, C., & He, Y. (2024). The impact of supply chain and digitization on the development of environmental technologies: Unveiling the role of inflation and consumption in G7 nations. Energy Economics, 108165.

[2] Li, T. (2025). Optimization of Clinical Trial Strategies for Anti-HER2 Drugs Based on Bayesian Optimization and Deep Learning.

[3] Huang, S., Liang, Y., Shen, F., & Gao, F. (2024, July). Research on Federated Learning's Contribution to Trustworthy and Responsible Artificial Intelligence. In Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering (pp. 125-129).

[4] Huang, S., Diao, S., Wan, Y., & Song, C. (2024, August). Research on multi-agency collaboration medical images analysis and classification system based on federated learning. In Proceedings of the 2024 International Conference on Biomedicine and Intelligent Technology (pp. 40-44).

[5] Liang, X., & Chen, H. (2019, August). HDSO: A High-Performance Dynamic Service Orchestration Algorithm in Hybrid NFV Networks. In 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 782-787). IEEE.

[6] Chen, H., & Bian, J. (2019, February). Streaming media live broadcast system based on MSE. In Journal of Physics: Conference Series (Vol. 1168, No. 3, p. 032071). IOP Publishing.

[7] Xie, Y., Li, Z., Yin, Y., Wei, Z., Xu, G., & Luo, Y. (2024). Advancing Legal Citation Text Classification A Conv1D-Based Approach for Multi-Class Classification. Journal of Theory and Practice of Engineering Science, 4(02), 15–22. https://doi.org/10.53469/jtpes.2024.04(02).03

[8] Xu, G., Xie, Y., Luo, Y., Yin, Y., Li, Z., & Wei, Z. (2024). Advancing Automated Surveillance: Real-Time Detection of Crown-of-Thorns Starfish via YOLOv5 Deep Learning. Journal of Theory and Practice of Engineering Science, 4(06), 1–10. https://doi.org/10.53469/jtpes.2024.04(06).01

[9] Yin, Y., Xu, G., Xie, Y., Luo, Y., Wei, Z., & Li, Z. (2024). Utilizing Deep Learning for Crystal System Classification in Lithium - Ion Batteries. Journal of Theory and Practice of Engineering Science, 4(03), 199–206. https://doi.org/10.53469/jtpes.2024.04(03).19

[10] Xu, Y., Gao, W., Wang, Y., Shan , X., & Lin, Y.-S. (2024). Enhancing user experience and trust in advanced LLM-based conversational agents. Computing and Artificial Intelligence, 2(2), 1467. https://doi.org/10.59400/cai.v2i2.1467

[11] Liu, M. (2024). Optimizing Supply Chain Efficiency Using Cross-Efficiency Analysis and Inverse DEA Models.

[12] Bi, S., Deng, T., & Xiao, J. (2024). The Role of AI in Financial Forecasting: ChatGPT's Potential and Challenges. arXiv preprint arXiv:2411.13562.

[13] Qiaozhi Zhao Research and Application of Data Cleaning Method for Urban Sewage Treatment Process Based on Fuzzy Neural Network [D]. Beijing: Beijing University of Technology, 2020

[14] Tonghua Zou, Yunpeng Gao, Huijuan Yi, etc Wind power anomaly data processing based on Thompson tau quartiles and multi-point interpolation [J]. Power System Automation, 2020, 44 (15): 156-162. DOI: 10.7500/AEPS20191231003

[15] Yan, Q., Yan, J., Zhang, D., Bi, S., Tian, Y., Mubeen, R., & Abbas, J. (2024). Does CEO power affect manufacturing firms' green innovation and organizational performance? A mediational approach. Sustainability, 16(14), 6015.

[16] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).

[17] Wang, Z., Yan, H., Wang, Z., Xu, Z., Wu, Z., & Wang, Y. (2024, July). Research on autonomous robots navigation based on reinforcement learning. In 2024 3rd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 78-81). IEEE.

[18] Wu, Z., Chen, J., Tan, L., Gong, H., Zhou, Y., & Shi, G. (2024, September). A lightweight GAN-based image fusion algorithm for visible and infrared images. In 2024 4th International Conference on Computer Science and Blockchain (CCSB) (pp. 466-470). IEEE.

[19] Z. Ren, "A Novel Feature Fusion-Based and Complex Contextual Model for Smoking Detection," 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China, 2024, pp. 1181-1185, doi: 10.1109/CISCE62493.2024.10653351.

[20] Ren, Z. (2024). Adaptive Multi-Scale Fusion for Infrared and Visible Object Detection in YOLOv8. Journal of Theory and Practice of Engineering Science, 4(09), 28–34. https://doi.org/10.53469/jtpes.2024.04(09).04

[21] Fan, Y., Hu, Z., Fu, L., Cheng, Y., Wang, L., & Wang, Y. (2024). Research on Optimizing Real-Time Data Processing in High-Frequency Trading Algorithms using Machine Learning. arXiv preprint arXiv:2412.01062.

Author Profile

Xiangfei Zhang (born in 1994), male, Han, Zhoukou City, Henan Province, bachelor's degree, professional title/position: software engineer, currently mainly engaged in work or research direction: software development.