Data Mining and Its Application Problem Analysis in the Era of big Data

Xiangyu Han, Xiaomeng Wei, Zhenhang Wang

School of Computer Science & Cyberspace Security, Hainan University, Haikou, China

Keywords: Big data era, Data, Data mining, Data application.

Abstract: With the development of information technology, the amount of data accumulated by humans has increased dramatically. However, due to the expansion of data volume and the widening of data, the previous data analysis methods are no longer applicable. A large amount of data needs to be analyzed and processed, and valuable data and information are extracted from it, data mining technology is born. This paper analyzes the research status of data mining at home and abroad, analyzes the main methods of data mining technology, introduces the application fields of data mining technology, and summarizes the future development trend of data mining technology in the era of big data.

1. Introduction

With the advent of the era of big data, society's data requirements for “mining” have become more stringent, and each precise result has its own “value”. At this time, the new attribute of the era of big data-”value” It interpreted to be very vivid. Data mining (DM) is an emerging cross-disciplinary discipline that brings together multiple disciplines. This extraordinary process involves unknown, implicit, and potentially valuable information from vast data. The process of extracting. In August 1989, at the symposium of the 11th Joint Conference on Artificial Intelligence held in Detroit, USA, scientists first proposed knowledge discovery in database (KDD). At the same time, knowledge discovery was also called Data mining, but the two are not exactly the same [1]. In 1995, the term KDD was accepted at the first International Conference on Knowledge Discovery and Data Mining in Montreal, Canada. The conference analyzed the entire process of data mining [2]. In essence, data mining is a sub-process of knowledge discovery.

2. The overview of the big data era

Big data refers to a large sample and information integration. Based on some sample problems, statistics are taken to sample and analyze to achieve the required progress, because some problems have higher dimensionality and require dimensionality reduction, compression and decomposition based on statistical principles. In addition, big data can be understood as a multi-angle, multi-domain information integration, can be a natural science, but also can be integrated information such as the humanities, and each discipline can be interspersed with each other. The original statistical method is suiTable for analyzing the information storage of a single computer, but it cannot adapt to big data analysis. The main purpose of current big data is to transform the cumbersome and incomprehensible information into people's easy-to-understand content, to explore the source and mechanism of information and then formulate countermeasures [1]. It takes a long time to consider the conversion of data into knowledge, so the information collected by people will be stored in memory for later application.

Big data was first used in the field of technology. With the advancement of the Internet platform, the information age has emerged, and the daily data processing volume has increased. The previous data systems have been unable to save a large amount of data, and the emerging information system will handle it, data rigidity and more intelligent analysis of data, as shown in Fig. 1, improve the efficiency and ability of people to process information on a daily basis.

  DOI: 10.25236/isrme.2019.072

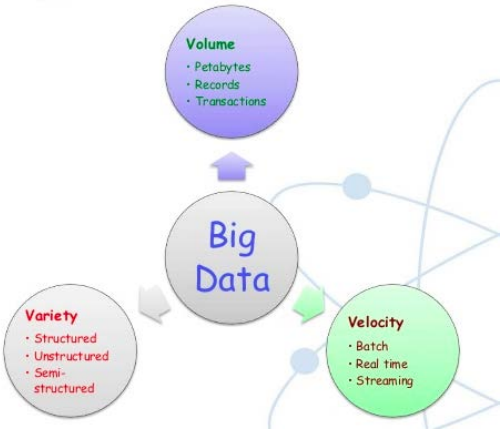

Fig.1 Characteristics of big data

The processing performance of big data for large amounts of data means that big data needs to be matched with cloud computing. The operational capability of a single computer cannot support the processing of information [2]. However, the development of big data technology is not just a simple processing of information. Data processing and processing, this process is a key prerequisite for the unit to obtain profits. There are many special technologies in big data applications that deal with the information that is passed in the short-term. These technologies integrate and calculate data, arrange it for hundreds and thousands of computers, and work together to process data.

3. Research status of data mining technology

Data mining combines theories and technologies of high-performance computing, machine learning, artificial intelligence, pattern recognition, statistics, data visualization, database technology, and expert systems. For data mining, the era of big data is both an opportunity and a challenge. It analyzes big data, establishes an appropriate system, continually optimizes, and improves the accuracy of decision-making, to better grasp and adapt to the multi-faceted changes in the market. In the era of big data, data mining has been recognized as the most commonly used data analysis method in various fields [3]. At present, domestic and foreign scholars mainly study the application of classification, optimization, recognition and prediction in data mining in many fields.

Classification. With the progress of the times and the rapid development of science and technology, as a populous country, China's public data generated in health care and aging society has grown geometrically, and the value of mining data based on big data is attached. The problem needs to be solved urgently. The structure, scale, scope and complexity of health care data are constantly expanding. Traditional calculation methods cannot fully satisfy the analysis of medical data [4]. Data mining technology can be based on some characteristics of medical data: pattern polymorphism, information Loss (missing values in the data due to personal privacy issues), timing, and redundancy classify health care data to provide accurate decision-making for doctors or patients.

Optimization. The traffic conditions of the roads are closely related to people's travel. With the rapid development of the city and the improvement of living standards, the scale of motor vehicles has gradually expanded, bringing problems such as traffic congestion. Data mining technology can effectively solve the optimization problem between traffic roads and logistics networks. Pan et al. proposed a data mining forecasting model, which is used to “real-time predict” short-term traffic conditions and bring drivers caught in traffic jams [3].

Identification. Since the advent of digital images in the 1950s, digital images have become an indispensable “data” in human society. In computer applications, data mining is becoming more and more popular in image recognition applications. Representative applications are face recognition and fingerprint recognition. Face recognition further analyzes and processes reliable and potential data-by-data mining of the obtained information base, and fully prepares the data analysis work and future development work. Wright et al. elaborated on robust face recognition based on sparse representation and gave a detailed theoretical analysis and practical summary [4].

Forecasting. Forecasting is the most studied problem in all fields. Its purpose is to predict future data values or trends through historical data. Most of the historical data is time series data, which means that they are arranged in chronological order and a series of observations, are obtained. Due to the continuous advancement of information technology, time series data has also increased dramatically, such as weather forecasting, oil exploration, and finance [1]. The ultimate goal of time series data mining is to predict the trend of the future and its impact by analyzing the historical data of the time series.

4. Application of data mining in the age of big Data

In the era of big data, data mining has been widely applied to various fields of life, and it has become a hot issue in the development of high-tech today. Whether in software development, medical and health, or in finance, education, etc., you can see the shadow of data mining everywhere [5]. You can use data mining technology to discover the intrinsic great value of big data, Fig. 2 below shows the way data mining in the era of big data.

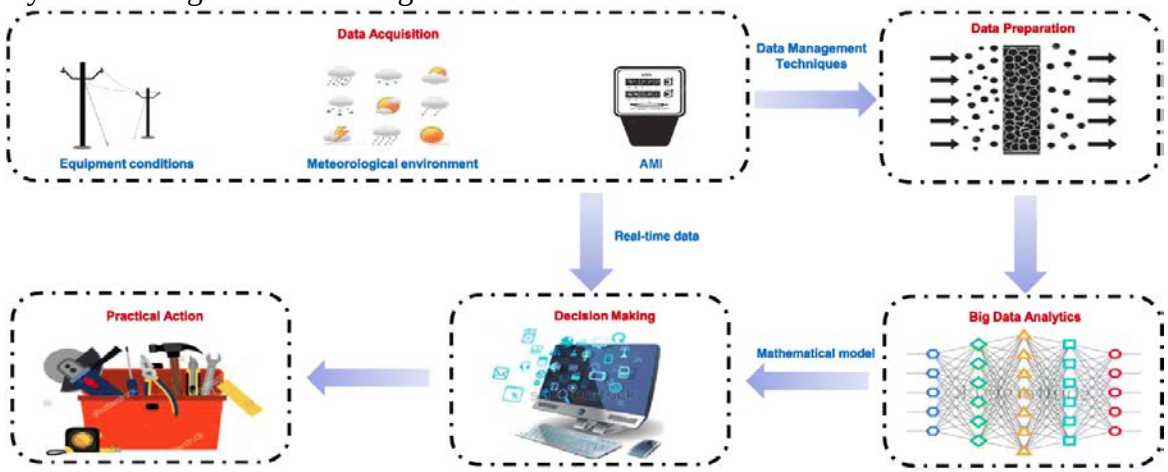

Fig.2 The way data mined under big data

4.1 Smart detection of malware.

In the era of big data, data mining technology has been widely used in malware detection. Malware seriously damages the network and computers. The detection of malware relies on the signature database (SD). Through SD, the files are compared and checked. If the number of bytes is equal, the suspicious files will be identified as malicious files. Some topics based on tagged malware detection are concentrated in a vague environment, and thus, dynamic modification of malware behavior is not possible, and hidden malware cannot be identified. Conversely, behavior-based malware detection can find the true behavior of a malicious file [5]. If a classification method based on data mining technology is adopted, the detection of malware can be detected according to the characteristics and behavior of each malware.

4.2 Wide application in bioinformatics.

Bioinformatics is an interdisciplinary subject that combines life sciences, computer science, information science, and mathematics. With the rapid development of technology, technology improvement and optimization of results, high-tech information technology has been extended to the field of biological research. However, it is far from enough to rely solely on the original computer technology. It is necessary to use computer science as an aid to integrate the interdisciplinary disciplines of life science, information science and mathematics, and to process it through data mining techniques and carefully analyze the biological data. The intrinsic link to mine potential information within biological data [5]. There are many characteristics of biological information data. Sun Qianlong summarizes the characteristics of current biological information data, including large quantity, variety, high dimension, wide form and sequence. The current hotspots of bioinformatics

include: the transition from compositional analysis represented by sequence analysis to functional analysis; the transformation from single bioanalytical research to gene regulation; and the overall analysis of genomic data. Humans' current research in the biological genome project is only the tip of the iceberg. Future research on biological genes such as differential gene expression, cancer gene detection, and encoding of proteins and RNA genes is inseparable from data mining technology. Only better. The use of data mining technology can unearth the extraordinary value of the biological genome.

4.3 Credit card default forecast.

Nowadays, with the rapid development of technology, the amount of information has increased dramatically, and content has become more and more abundant. Credit cards have a status that cannot be ignored in people's lives. As we all know, credit cards are issued by banks. Banks need to verify the applicant's personal information, and then issue credit cards after confirmation. Chen et al. proposed a fuzzy algorithm for credit ratios for commercial banks. Before the credit card is processed, the bank first needs to conduct a detailed investigation of the applicant, and judge whether it has the ability to repay the loan amount according to the actual situation of the applicant [6]. Liu Ming et al. use the gray wolf optimization algorithm to calculate the neural network based on the traditional neural network. The initial weights and thresholds are proposed, and an improved fuzzy neural network algorithm is proposed. By comparing the default prediction model of credit card customers, it compares with other current prediction methods to obtain better prediction results. Further, it is verified. Fuzzy neural network has better robustness, accuracy and efficiency in the prediction of credit card customers.

5. The development trend of data mining in the era of big data

Regardless of the research field or commercial application, data mining is a hot issue. More and more people are paying attention to it. People gradually understand, learn and use it, and the related fields are becoming more and more mature. When using data mining technology to solve and solve practical problems, Wang Guangdong proposed three noteworthy angles: using data mining technology to solve the problem types, solving data mining data preparation work and the theoretical basis of data mining. In the era of big data, the development trend of data mining will be reflected in the following five levels around the value of data mining.

5.1 Multimedia data mining.

In the era of big data, video, audio and images, all belong to the category of multimedia. With the development of the times, the massive data structure has become complicated and dynamic, and the traditional mathematical methods are used to manage real life. In the problems, the results obtained often fail to meet people's expectations [6]. The practical application of drones and unmanned vehicles, the development of public security Skynet projects, and the comprehensive development of smart medical projects require rapid processing of multimedia data. In order to obtain better results, the effects obtained are optimized and need to be developed. In addition, new intelligent algorithms for designing data mining.

5.2 Mining of potential data in the financial sector.

In the credit card business, the data mining of default prediction has the advantages of prophetic, effective and practical. In the process of credit card transactions, there are many types of data mining applications, such as the detection of abnormal credit card behavior, the maintenance of high-end credit customers and credit card risk control [7].

5.3 Improvement and visualization of data mining algorithms.

When data mining algorithms are used to analyze and process massive data, the improvement of the algorithm mainly depends on the accuracy and speed of the algorithm, that is, the accuracy and efficiency of the algorithm. Today, academic research focuses on setting appropriate thresholds

between accuracy and efficiency and visualizing the results of data mining [7]. A new series of deep learning algorithms such as RNN, CNN, DNN, and Capsule for data mining algorithms will become the vane for leading big data research methods.

5.4 Data mining and privacy protection.

When it comes to solving practical problems, it is ineviTable to involve privacy data. For example, when researching the relationship between credit cards and users, it is ineviTable that there will be personal information of users in the data, in the study of cervical cancer (risk factors) and when people's age, number of pregnancies, number of sexual partners, etc., there will be some privacy information that is inconvenient to disclose the outside world [5]. In the process of data mining, without revealing the user's personal privacy issues, desensitization of data will become another important aspect of people's research data mining.

5.5 Data mining technology integration with other systems.

Data mining is a complete process, rather than simply a single algorithm or a few of them can be simply mixed. Applying data mining to the actual exercise process, that is necessary to integrate data mining with other fields and systems in a structured manner [7]. It cannot be understood as a single algorithm to solve a problem, thus maximizing the advantages of data mining.

6. Summary

In the era of big data, when using traditional mathematical methods to encounter difficulties, it is particularly important to skillfully apply data mining techniques. Today, data mining technology can be applied to all areas and industries. Data Mining Technology Data Mining Technology can be used in almost every aspect of people's lives. Not only does it bring great changes and influences to our daily lives, but also it also profoundly changes our way of life.

References

[1] Sh.F. Han and L.Ch. Chen, Data Mining Technology and Application Review, Mechanical Management Development, 2006, vol.2, pp.22-24.

[2] Y.F. Yan and D.X. Zhang, Big Data Research, Computer Technology and Development, 2013, vol.4, pp.168-172

[3] W. Pan, The challenge and development of statistics in the era of big data, Science and Technology, 2018, vol.3, pp. 26-28.

[4] L.T. Liu, Economic Value Evaluation and Over-Excavation Risk Research of Big Data Analysis, Tianjin University of Finance and Economics, 2017, vol.4, pp.55-57.

[5] Y.T. Sun, Discussion on the problems and solutions of data mining in the era of big data, Talent, 2017, vol.5, pp.22-24.

[6] X.D. Liu, Exploring the thinking of data mining in the era of big data, Talent, 2016, vol.35, pp.44-48.

[7] Y.P. Fu, Opportunities, Challenges and Trends in Data Mining in the Big Data Era, China Management Informatization, 2016, vol.14, pp.245-247.