Research on Improved Model of Electronic Commerce Data Mining Based on Big Data Technology

Hongsheng Xu[1,2 a *], Ganglong Fan[1,2] and Ke Li[1,2]

[1]Luoyang Normal University, Luoyang, 471934, China

[2]Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang, 471934, China

[a]85660190@qq.com

Keywords: Big data; Data mining; Hadoop; MapReduce; Electronic commerce

Abstract. Big data collection is the use of multiple databases to receive data from the client, and users can use these databases for simple queries and processing. Big data is pointing to massive, comprehensive, highly correlated complex data forms. At present, most Internet companies use Hadoop's HDFS distributed file system to store data and analyze them using MapReduce. Statistics and analysis mainly use distributed database or distributed computing cluster to analyze and classify the mass data stored in it. The paper presents research on improved model of electronic commerce data mining based on big data technology.

Introduction

Big data refers to data that exceeds the processing power of traditional database systems. Its data size and speed of transfer requirements are high, or its structure is not suitable for the original database system. In order to obtain the value of large data, we must choose another way to deal with it. Valuable patterns and information are hidden in the data, and it takes a considerable amount of time and cost in the past to extract the information [1]. Leading companies like WAL-MART or Google have to pay high prices to tap information from big data. Today's resources, such as hardware, Cloud Architecture, and open source software, make it easier and cheaper to process large data. Even companies that start up in the garage can rent cloud services at a lower price.

With the development of social economy and the increase of personal income, people's individual demands begin to be highlighted. And enterprises need a lot of data support to meet these personalized demands efficiently. The emergence of the Internet and the development of related technologies have made it possible to collect and analyze mass data. The characteristics of the Internet also cause these data to be transmitted at a high speed and capacity. The Internet introduces patterns that generate data from users. This model is characterized by multi source, low cost, and timelier. Of course, the authenticity and reliability of these data need to be certified.

Big data boom did not fade signs, on the contrary, including aviation, finance, electricity providers, government, telecommunications, electricity, and even F1 racing and other industries in the Nuggets are big data [2]. As can be seen, in promoting large data enterprise applications really see big data potential business value of enterprises than big data technology vendors have to worry about. For example, IT manager network once reported that WAL-MART big data lab directly involved in the development of large data tools and open source work.

From the category of data, "big data" refers to information that cannot be processed or analyzed using traditional processes or tools. It defines data sets that go beyond the normal processing range and size and force users to adopt non-traditional processing methods. Amazon Web Service (AWS), big data scientist, refers to a simple definition: big data is any amount of data that exceeds the processing power of a computer. R & D groups define big data: big data is the biggest publicity technology, is the most fashionable technology, when this phenomenon occurs, the definition becomes very confusing.

Volume refers to the amount of data, the amount of data, and the integrity of its scale. The storage of data is extended to ZB TB. This is closely related to the development of data storage and network

technology. With the improvement of data processing and processing technology, the increase of network bandwidth and the rapid development of social networking technology, the amount of data and storage have increased exponentially. In essence, to some extent, the magnitude of the data is not important, and the important thing is that the data is integrity. The application of data scale are reflected, such as the daily 12 TB tweets analysis, to understand the people's psychological state, can be used in the research and development of emotional products; analysis of Face book based on tens of thousands of information, can help people to deal with the reality of the relationship between the interests of the circle of friends.

Big data collection refers to the use of multiple databases to receive data from the client, and users can use these databases for simple queries and processing. For example, the electricity supplier will use traditional relational databases such as MySQL and Oracle to store every transaction data. In addition, NoSQL databases such as Redis and MongoDB are also commonly used for data collection. The paper presents research on improved model of electronic commerce data mining based on big data technology.

Overview of Big Data Processing and Analysis Techniques

The type of big data is used to produce or deal with complex data type is relatively single, most of the data is structured. Now, social networking, networking, mobile computing, online advertising and other channels and new technologies continue to emerge, resulting in a large number of semi-structured or non structured data, such as XML, e-mail, blog, instant the message, leading to new data types increase. Enterprises need to integrate and analyze complex from traditional and non-traditional sources of data, including internal and external data [3]. Along with the explosive growth of sensors, smart devices and social collaboration technology, the type of data to count, including: text, micro-blog, sensor data, audio, video, click stream, log file.

Large amounts of unstructured data are also waiting to be collected and analyzed. The future of the financial industry will be driven more by the science and technology innovation, more and more inclined to retail marketing: for the financial industry, big data means huge business opportunities, can enhance the customer experience, and improve customer loyalty. The development of big data technology brings about the transformation of enterprise management decision model, drives the industry change, and brings forth new business opportunities and development opportunities. The ability to manage large data has been confirmed as the core competitiveness of enterprises in the lead, this ability can help enterprises to break the boundary of the data, draw a panoramic view of business operations to make business decisions, and the optimal development strategy.

Mass level refers to the amount of data that has been completely ineffective or cost too high for databases and BI products. There are a lot of massive data level excellent enterprise products, but the cost of hardware and software based, most Internet companies using the HDFS Hadoop distributed file system to store data, and analyzed using MapReduce. Later in this article, we will mainly introduce a MapReduce based multidimensional data analysis platform on Hadoop, as is shown by equation (1) [4].

$$x_{i+1}^{2}(t+1)=\left(1-d_{i}^{2}(t)\right)x_{i}^{2}(t)+\left(\frac{N^{2}(t)}{N^{3}(t)+N^{2}(t)}\right)s_{i}\alpha N^{1}(t) \tag{1}$$

The uncertainty of large data is represented by high dimension, variability and strong randomness. Stock trading data streams are a classic example of large, uncertain data. Big data has stimulated a lot of research. Individual performance, unstructured and semi-structured data, the general characteristic and basic principle is not clear, these are required by multi disciplines including mathematics, economics, sociology, computer science and management science, to study and discuss.

Cloud computing and big data is both sides of a coin, cloud computing is the basis of big data IT, and big data is a killer application of cloud computing. Cloud computing is the driving force for growth of large data, because the data more and more, more and more complex, more and more real-time, which requires more cloud computing to deal with, so the two are complementary. In essence, the relationship between cloud computing and big data is static and dynamic. Cloud computing emphasizes computing,

which is the concept of moving; and data is the object of calculation, is the concept of static. If it is combined with the actual application, and it is the former emphasizes the computing ability, and it is storage capacity or value, and it is but that does not mean that the two concepts so quite distinct from each other.

The data is expanding rapidly and become larger, which determines the future development of enterprises, although the enterprise may not be aware of problems of the explosive growth of data hidden, but as time goes on, people will be more aware of the importance of enterprise data [5]. The era of big data poses a new challenge to the ability of human data control, and provides unprecedented space and potential for people to gain more profound and comprehensive insight.

In the process of collecting data, the main characteristics and challenges is the high number of concurrent, because at the same time there may be tens of thousands of users to access and operate, such as train ticketing website, visit their concurrent at the peak reached millions, as is shown by equation (2), so in the end need to support the deployment of a large number of data acquisition. And how to load and distribute between these databases really needs deep thinking and design.

$$f'(\xi)=\frac{1}{\xi}=\frac{\ln a-\ln b}{a-b} \tag{2}$$

It has been an important research object of traditional data analysis. At present, the mainstream structured data management tools, such as relational databases, all provide data analysis functions. Analysis of commercial and scientific research fields will produce a large amount of structured data and structured data management and analysis of these depend on the database, data warehouse, OLAP and business process management maturity of commercial technology. Thanks to the development of relational database technology, the method of structured data analysis is more mature. Most of them are based on data mining and statistical analysis.

Big data preprocessing technology: the main completion of the received data analysis, extraction, cleaning and other operations.

1) For data extraction: may have different structures and types, data extraction process can help us put these complex data into a single or a convenient disposal configuration, in order to achieve the purpose of rapid processing.

2) cleaning: for big data, not all the value, some data is not our concern, while other data is interference completely wrong, as is shown by equation(3), so it is necessary to filter the data by "denoising" to extract the effective data [6].

$$p_0(t)=\frac{u}{\lambda+u}+\frac{\lambda}{\lambda+u}e^{-(\lambda+u)t} \tag{3}$$

Data quality and data management) data quality and data management are some of the best practices in management. Data processing through standardized processes and tools ensures a well-defined, high-quality analysis result. Visual analysis, whether for data analysis experts or ordinary users, data visualization is the most basic requirement of data analysis tools. Visualization can visually display data, let the data speak, and let the audience hear the results [7].

Big data analysis theory is the core algorithm of data mining, data mining algorithm of data types and formats can be different based on more scientific data shows itself has, it is precisely because of these various statistical methods is the world recognized by the statistician (known as truth) to in-depth data mining. Accepted is value. Another aspect is because these data mining algorithms can deal with large data faster, and if an algorithm takes several years to draw conclusions, the value of the big data is out of the question.

Innovation Analysis of the Era of Big Data Business Model

Big data can development, continuous generation of innovation and innovation in thinking. Direct analysis of PB data of big data, no longer rely on the random sampling; data processing is not excessive

pursuit of accuracy of individual data, the prediction has become the focus of large data processing; no longer pay much attention to causality, pay more attention to the correlation data set.

Big data is bound to subvert many traditions [8]. In the past, the commonly used "sampling survey" in social science research was once regarded as a solid foundation for the establishment of social civilization, and it was widely used. In fact, it is a helpless solution to a particular problem at a specific time when the technology is limited. Now, the information that has not been collected in the past can be collected, so the sample is equal to all". Moreover, it is much more accurate to draw conclusions than using sampling methods. Big data has filter and prediction capabilities. The so-called filter is to seize the key. The so-called prediction is to obtain valuable information ahead of time, as is shown by equation (4).

$$L = \sum_i \sum_j np_{ij} = 0p_{00} + 1p_{01} + 1p_{10} + 2p_{11} + 2p_{b1} = \frac{4\rho + 5\rho^2}{H} \tag{4}$$

The impact of big data has increased the need for information management experts. In fact, the impact of big data is not limited to the information and communication industry, but is "swallowed" and reconstruction of many traditional industries, extensive use of company data analysis operation management and optimization of its essence is a data company [9]. The flagship stores such as McDonald's, KFC and Apple Corp are all located on the basis of data analysis. While in the retail industry, technology and method of data analysis is widely used, traditional enterprises such as WAL-MART through data mining to reshape and optimize the supply chain, the rise of new electricity providers such as Amazon, Taobao, through the understanding and analysis of massive data, provide more professional and personalized service for users.

Big data can be divided into large data technology, big data engineering, big data science and big data applications and other fields. At present, the most talked about is big data technology and big data applications. Engineering and scientific issues have yet to be taken seriously. Big data refers to the system of engineering project planning and construction management of large data; data discovery and scientific attention to validate the relationship between big data rule and natural and social activities of the big data network development and operation process.

Experiments and Analysis

Data sets are often very large and difficult to handle with traditional database management tools. As of 2012, data sets consisted of tens of megabytes to several gigabytes of data [10]. These include accessing web pages, landing, and online trading, and so on. At present, the size of the data set is increasing. Enterprises should use appropriate tools to compress and screen data, showing only data related to specific content. At present, some enterprises have implemented big data strategy; some enterprises are developing or intend to develop big data.

Big data analysis of various industries will usher in more applications. More and more enterprises will not be satisfied with large data management capabilities and seek outside experts. Mobile analytics increased significantly. Mobile push analysis will change consumer spending information and consumer habits. The emergence of more intelligent devices and appliances, a large degree of embedded analysis. More emphasis on real-time analysis, although I am not optimistic that it will make great progress this year, unable to deal with a large number of data, variety or speed of product analysis company will be eliminated. Hadoop's challenge will begin to emerge. Users will reach a point of frustration with performance limitations, versioning, confusion, and various standards and interfaces.

Statistics and analysis of the main use of the distributed database, or distributed computing analysis and classification of common summary of mass data storage within the cluster, in order to meet the demand analysis of the most common, in this regard, some real-time requirements will be used EMC GreenPlum, Oracle Exadata, and MySQL based storage Infobright so, some of the batch, or based on semi-structured data needs can use Hadoop. The main features and challenges of statistics and analysis

are that the amount of data involved in the analysis is large, and the system resources, especially I/O, will be greatly occupied.

Summary

The paper presents research on improved model of electronic commerce data mining based on big data technology. Big data is a kind of artificial nature has the hidden law, searched for scientific mode of big data will bring a general method to study the beauty of big data, although this exploration is very difficult, but if we find the unstructured and semi-structured data conversion method of structured data, data mining methods known will become a major tool for data mining.

Acknowledgements

This paper is supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, and also supported by the science and technology research major project of Henan province Education Department (13B520155, 17B520026).

References

[1] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: The next frontier for innovation, competition, and productivity. May 2011. MacKinsey Global Institute, 2011.
[2] Melnik S, Gubarev A, Long J J, Romer G, Shivakumar S, Tolton M, Vassilakis T. Dremel: interactive analysis of web-scale datasets. Proceedings of the VLDB Endowment, 2013, 3(1-2): 330-339.
[3] Zheng QL, Fang M, Wang S, Wang XQ, Wu XW, Wang H. Scientific Parallel Computing Based on MapReduce Model. Micro Electronics & Computer, 2015, 26(8):13-17.
[4] Hongxin Wan, Yun Peng, Clustering and Evaluation on Electronic Commerce Customers Based on Fuzzy Set, IJACT, Vol. 5, No. 3, pp. 199 - 206, 2013.
[5] Thusoo A, Sarma J S, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R. Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment, 2014, 1(2): 626-629.
[6] Tao XJ, Hu XF, Liu Y. Overview of Big Data Research. Journal of System Simulation, 2013, 25S: 142-146.
[7] Stonebraker M, Çetintemel U, Zdonik S. The 8 requirements of real-time stream processing. ACM SIGMOD Record, 2016, 34(4): 42-47.
[8] Dawei Sun, Ge Fu, "Using Big Data Technology for Information Management in Hybrid Learning System", RNIS, Volume 12, pp. 179 ~ 182, 2015.
[9] Wei Li, Hongtu Zhang, "The Research of E-Commerce Recommendation System Based on Cloud Computing", IJACT, Vol. 5, No. 20, pp. 256 ~ 263, 2014.
[10] Li YL, Dong J. Study and Improvement of MapReduce based on Hadoop. Computer Engineering and Design. 2016, 3(8):3110-3116.