Research on Library Big Data Cleaning System based on Big Data Decision Analysis Needs

Jianfeng Liao [1, 2, *], Jianping You [2] and Qun Zhang [2]

[1] Department of Computer Science Wenhua College, Wuhan, China.

[2] College of Computer Science, Hubei University of Technology, Wuhan, China.

* jazz981012@163.com

Abstract. In the era of big data, university library information management services must be based on actual conditions, using high-quality data to improve big data management. However, high-quality big data is useful data that needs to be filtered and classified. Big data cleansing is an effective way to improve data quality. To this end, the paper proposes to integrate the data resources of efficient libraries, analyze the source and type of useless data, and design a hierarchical management model of data. The model includes management operation level, data cleaning and filtering level, data integration level and big data. At the resource utilization level, after attempting to filter invalid data through data cleaning, the complexity of big data decision analysis is reduced, library big data integration is promoted, big data decision-making is realized, and the possibility of library big data integration and sharing is improved.

Keywords: University library; Big data decision analysis; Data cleaning and filtering; Data integration application.

1. Introduction

In the process of modern sensor, cloud computing, high-speed wireless transmission and integrated manufacturing technology, it is widely used in the library operation process, and the daily operation and management of the library have undergone comprehensive changes [1]. In this context, the library data center computing system scale, storage system capacity, service load and network structure complexity are constantly improving, the library big data scale is expanding, and the traditional data processing methods are not suitable for current needs. Therefore, comprehensive collection, real-time analysis, short-term processing and accurate decision-making for modern library security big data. The library industry has begun to think about how to eliminate the disordered and low-value information through the integration of big data information, thereby increasing the value of information resources, thereby reducing the cost of big data operation and maintenance.

2. Security Issues and Challenges Facing Libraries in the Era of Big Data

Library big data has 4V features such as Volume, Variety, Value (low value density) and Velocity (fast processing speed), which makes the data environment more complex and changeable, and increases the security management of the library. Difficulty, uncertainty and uncontrollability. The original big data collected and stored in the library is mixed with incomplete, erroneous and repetitive "unclean" data, resulting in inconsistencies, incompleteness, low value density, uncontrollable and unusable characteristics of library big data [2]. If the library emphasizes the improvement of big data processing performance of IT infrastructure, the scientific nature of data analysis methods, and the data literacy of data analysts, without improving the quality and usability of data through big data cleaning, it will lead to the library. The yield and data decisions of data applications are scientifically declining.

3. Library Big Data Integration Platform Design

The big data integration system is intended to adopt a layered system structure. Hierarchical design is already a mature platform resource integration design model. After continuous verification in the IT market, it does have strong scalability and loose coupling[3]. The application layered design facilitates the addition, deletion and modification of any functional module without reducing the overall functionality, operability, and security controllability of the entire big data integration system platform. The design of the library big data resource integration platform system architecture is shown in Figure 1.

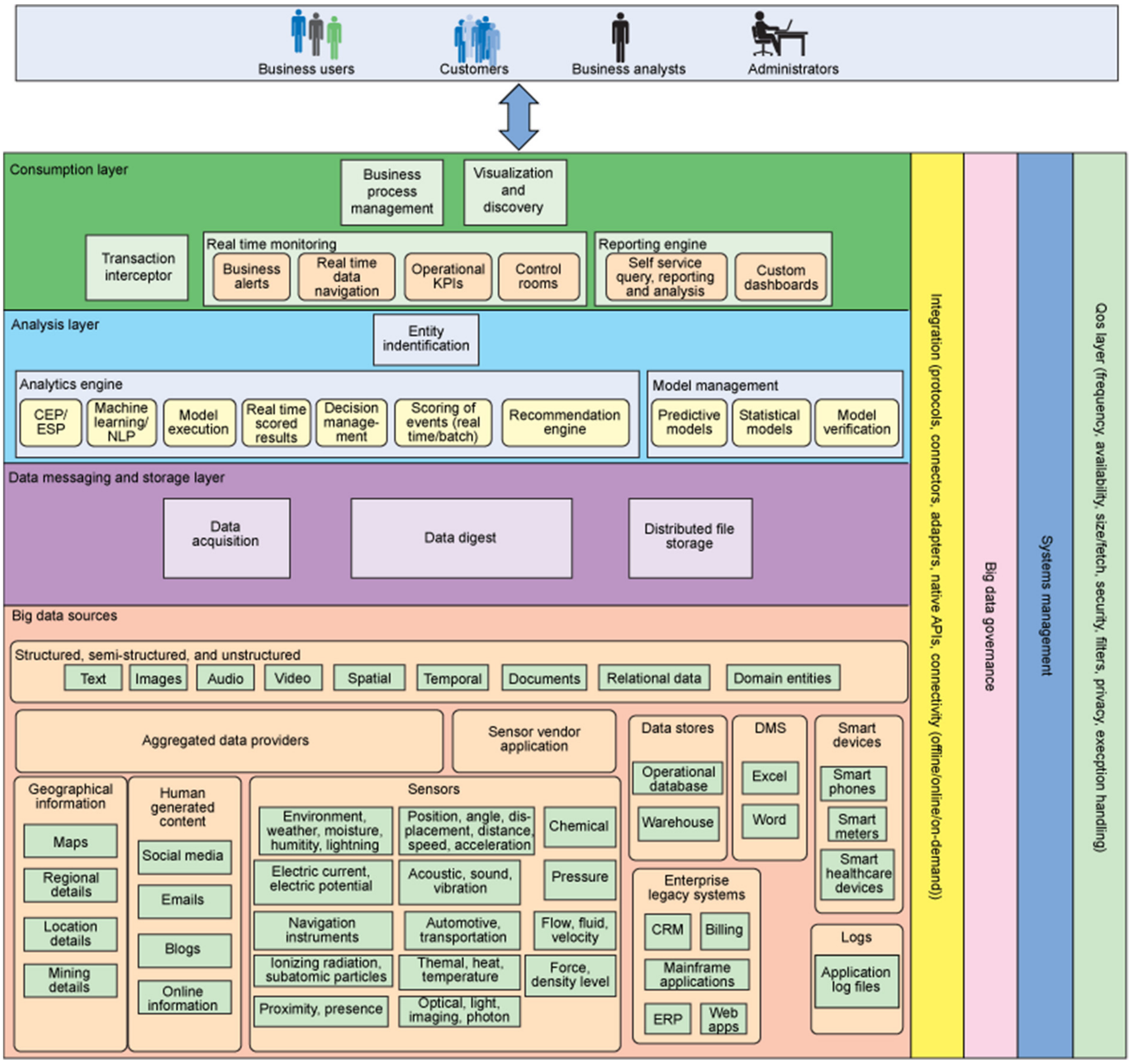

Fig. 1 Library big data resource integration platform system architecture diagram

According to the layered design principle, the library big data integration platform system is divided into the following layers from top to bottom: The top is the management operation layer, the data pre-cleaning and filtering layer, followed by the data integration layer. The lowest end is the big data resource layer.

3.1 Management Operation Layer

The design integrates data storage, system parameter setting, query and management modules, mainly provides the administrator with a unified interface for daily maintenance and operation management of the platform.

3.2 Data Pre-Cleaning and Filter Layer

Through pre-set cleaning filter rules, data analysis is performed on one pair of big data collected, usability check, filtering out unordered and unrelated data, retaining valuable operational data, and integrating data for the next layer. Provide prerequisite preparation[4].

3.3 Data Integration Layer

Think of data as a service that uses data and services to construct new storage patterns. Data services are first of all the collection and storage of data. In addition to traffic data and customer data from the organization's website, it also includes national policies, economic data, and so on. These data are typically stored in a database for data mining and user information queries. Through the integrated universal data interface and exchange mechanism, these data resources are extracted, converted, loaded (ETL), and repackaged and stored for system analysis. By reading the pre-cleaned and filtered data temporarily stored in the external data repository, the data is converted by the conversion controller under the integrated control system, and then the platform base layer accesses these parsed and converted processes through the integrated runtime engine. Data, and finally enter the primary database storage[5]. Thereby completing the integration process of library big data resources.

3.4 Big Data Resource Layer

The big data decision application layer is based on the support of the lower platform layer to complete the library development strategy decision, functional department work and service decision, reader reading needs analysis and forecast, library QOS evaluation, intelligent service report, and other changes with the library, Develop big data decisions related to reader services, provide reliable big data decision support for library strategic decision-making, system management and operation, readers' QOS assurance, and sustainable development of service productivity. Mainly include big data resources, temporary databases and main database and application system databases. Perhaps the same big data is limited by the integration process[6]. Different databases at different time points will have different inconsistencies. Therefore, it is urgent to introduce a data synchronization mechanism to ensure the consistency of different database data at the same time. Guarantee the security and controllability of data support.

4. Research on Data Cleaning Strategy for Real-Time Demand Decision-Making of Big Data

Library big data decisions can be divided into real-time decision-making and offline decision-making. Real-time decision-making is mainly applied to real-time management of library security and services, real-time judgment of online service demand of readers, real-time service policy formulation and service push, real-time evaluation and optimization of service systems, etc., requiring big data application platform to obtain in a short period of time. Clean, analyze, and analyze real-time data to provide scientific data support for real-time, dynamic decision making of big data. Real-time decision-making requires high timeliness for big data acquisition, transmission, cleaning, analysis, and decision-making processes[7]. Real-time big data with small data volume, low cleanliness, and limited value may affect the scientific and usability of library real-time decision making. Offline decision-making is mainly applied to the library's macro-strategy formulation, user service model change, service effectiveness assessment and service market competition environment analysis. Although these offline decisions reduce the timeliness of decision-making due to the complex analysis of massive big data, However, the decision-making results are highly scientific, accurate and reliable. Therefore, libraries must develop relevant big data cleaning strategies based on the real-time needs of big data decisions.

4.1 Several Important Dimensions of Library Big Data Cleaning Quality Assessment Standards

The scientific nature of library quality assessment of unclean data cleaning is not only the key to the value assurance of library data and the availability of big data decision, but also an important basis for the library to control, optimize and improve the data cleaning system. Therefore, it must choose the scientific, comprehensive, complete and operational big data cleaning quality assessment dimension ensures that the big data cleaning process is efficient, high quality, fast, economical and controllable. The construction of library big data cleaning quality evaluation system should adhere to the principle of wide coverage of evaluation indicators, reasonable distribution of index factors, open evaluation system and easy operation [8]. The dimensions of library big data cleaning quality evaluation are shown in Table 1.

Tab. 1 Library big data cleaning quality assessment dimension table

Numbering	Dimension content	Description of the big data cleaning evaluation dimension
1	Normative	Data existence, quality and storage standards
2	Integrity and accuracy	Data structure integrity, accuracy and availability
3	Repeatability	Data is repeated outside of fields, recorded content, or data sets
4	Consistency and synchronization	Consistency and synchronization across different databases, applications and systems
5	Timeliness and availability	Non-real-time data cleaning and value usability measurement
6	Identifiability and relevance	Data understandable, value measurable and compatible
7	Ease of use and maintainability	The extent to which data can be accessed, used, updated, maintained, and managed
8	Data value coverage	The value, object and content coverage of data in decision making

4.2 Design of the Library Big Data Cleaning Process

The scientific nature of the big data cleaning process, the controllability of data flow and the validity of the evaluation criteria are important issues related to the safe, efficient, fast and economical process of the library's big data cleaning process. Therefore, in the design of the library big data cleaning process, this paper insists on improving the value density, availability, real-time decision-making and reducing the cost of big data application under the premise of keeping the total value of big data unchanged. It effectively guarantees the scientific, reliable, real-time and economical decision-making of big data. The library big data cleaning process is shown in Figure 2.

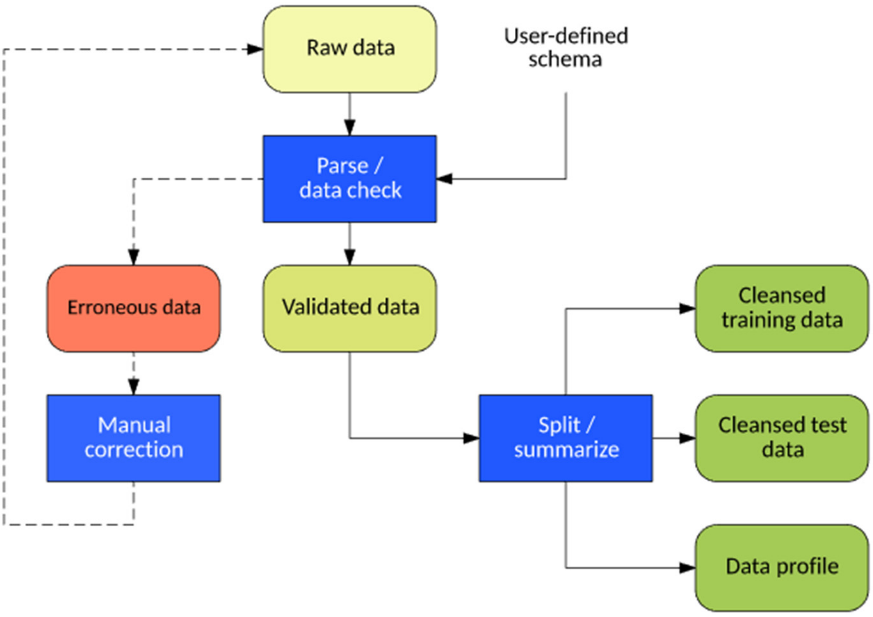

Fig. 2 Library big data cleaning flow chart

The library big data cleaning system first preprocesses the received big data, and imports the big data to be cleaned into the data cleaning system to check whether the metadata such as field interpretation, data source, code table, etc. are correct, preliminary statistics Cleaning data types, structures, real-time requirements and presence patterns [9]. Secondly, in the cleaning of big data, according to the complexity of big data cleaning, the total amount of big data and the process correlation, according to the removal and repair of missing big data, removal of duplicate and logical error big data, detection of abnormal big data and Processes, non-requirement big data cleaning, big data correlation verification, etc., finally evaluate the quality and availability of cleaned big data, and transmit unclean data that does not meet the big data decision requirements to the data cleaning input interface. Secondary cleaning.

5. Conclusion

At present, the era of big data in libraries has arrived. As an important part of production materials, big data has become an important decision-making basis for libraries to improve service productivity, management and service model transformation, gain competitive advantage and build smart libraries [10]. With the growth of reader service requirements and the transformation of library service models, the library's big data environment presents 5, "Volume", "Variety", "Velocity", "Value"and "Veracity". V characteristics, the rapid increase of the total amount of unstructured data and the proportion of large data, will lead to more complex and variable library big data environment, the rapid growth of large data collection, transmission, storage, management and processing, and the scientific, cycle controllability, and return on investment of unstructured big data analysis are affected.

In order to ensure that the analysis and decision-making process of unstructured big data in the library is scientific and efficient, the library should focus on readers' reading needs and service productivity improvement, and insist on "data-driven" as the basis for unstructured big data analysis and decision-making. Strengthen the availability and controllability of unstructured big data in acquisition, noise filtering, value extraction and storage, and strive to improve the real-time analysis, human-computer interaction, scientific evaluation and feedback optimization level of unstructured data. The scientific and usable level of big data decision making in the library can provide reliable big data decision support for personalized reader reading activities.

Acknowledgments

This work was financially supported by Hubei Natural Science Foundation No. 2013CFC113.

References

[1]. Dong Wei. Design of Big Data Analysis System for Readers' Demand Information in University Libraries. Computer Programming Techniques & Maintenance,Vol.5 (2017) No.20, p.57-59.

[2]. He Li. Discussion on Librarian Data Literacy Based on Big Data Vision. Library and Information Service,Vol.6 (2016) No.28, p.54-58.

[3]. Bian Qian. Research on Public Library Service Work under the View of Big Data. Library and Information Service, Vol.20 (2015) No.25, p.25-27.

[4]. Liang Junrong. Research on Security Analysis and Management of Library Information System Based on Big Data Decision. Library Theory and Practice,Vol.3 (2017) No.28, p.93-98.

[5]. Cui Li. Challenges and Opportunities of Library Public Opinion Information Service under the Perspective of Big Data. Jintu Journal, Vol.5 (2015) No.32, p.20-24.

[6]. Ma Xiaoting. Research on the Construction of Library Big Data Analysis Platform Based on Personalized Service Demand. New Century Library, Vol.4 (2014) No.6, p.20-23.

[7]. Yan Na. Research on Information Management and Information System Specialty Construction from the Perspective of Big Data. Library Science Research,Vol.11 (2013) No.14, p.9-12.

[8]. Liu Guifeng, Lu Zhangping, Hua Hui. Research on Library Big Data Knowledge Service Ecosystem and Its Dynamic Mechanism. National Library Journal, Vol.3 (2016) No.25, p.52-60.

[9]. Zhang Kai, Guo Jianqi. Library Big Data Survey and Forward-looking Concept——Based on Baidu Index Analysis. Journal of Library Science in China, Vol.6. (2016) No.42, p.51-65.

[10]. Qin Shuai. Research on Library Big Data Value Analysis and Service Quality Assurance Based on User Service Value. Journal of Jiamusi Vocational College, Vol.1 (2017) No.14, p. 103-105.