Research on Scientific Research Information Management System Under the Background of Big Data

Zhengqian Feng, Wang Li, Ning Dang(✉), Xikai Ding, and Zhongwei Chen

Shandong Scicom Information and Economy Research Institute Co., Ltd., Jinan, China
{fengzhq,liw,dangn,dingxk,chenzhw}@sdas.org

Abstract. With the rapid development of cloud computing and big data technology, information technology has had a huge impact on the field of education, and scientific research management driven by big data has become a new topic in the development of modern education. Scientific research activities will generate massive data, and how to collect, process, analyze and use the massive data has become an important goal of scientific research managers. This paper focuses on the four stages of scientific research projects, based on data mining, calculation and analysis technologies such as semantic analysis, web crawler and decision support, to demonstrate the availability and efficiency of key technologies in the process of scientific research management, and promote the use of modern information technology in scientific research management. The application in the education system promotes the scientific development of the modern education system.

Keywords: Big data · Scientific research management · IT in education

1 Introduction

With the continuous mining of massive data, big data technology has become another disruptive technological change after the Internet of Things and cloud computing. In the field of education, big data technology is increasingly deepened in teaching management, scientific research management, education reform, etc., and the innovative development of big data fusion education modernization is also an inevitable requirement of the times. Based on the perspective of scientific research management, this paper introduces data mining, data calculation and analysis, decision support technology, builds and improves the scientific research management system, and promotes the high-level development of higher education modernization.

2 Application Exploration of Big Data Method in Research Management

The whole process of Science and technology projects from start to finish follows the law of the life cycle, reflecting the phase and periodicity of the project. In modern

V. Balakrishnan et al. (Eds.): ICMEIM 2022, AHSSEH 6, pp. 740–746, 2022.
https://doi.org/10.2991/978-94-6463-044-2_93

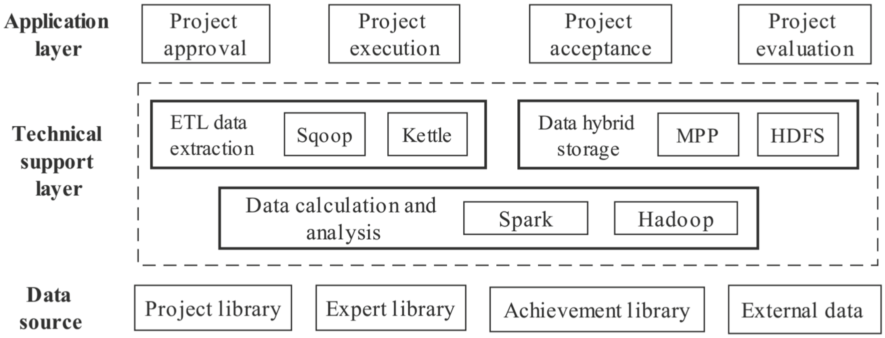

Fig. 1. Big data application framework for Science and technology project management

management theory, the life cycle of Science and technology project management is usually divided into four stages: project application, project execution, final acceptance and performance evaluation [1]. This paper will discuss the application of big data technology in Science and technology project management in combination with the project life cycle.

2.1 Research on Key Technologies in the Process of Project Data Collection and Integration

The collection of data information is the first link of Science and technology project management. The sources of data collection include business system data, Internet data, social network data, etc., and the types of data obtained are also different, including structured, semi-structured and unstructured.

First of all, for business systems or office application systems, traditional relational databases (such as Oracle, MySQL, SQL Server, etc.) are usually used to store data, and ETL tools such as Sqoop and Kettle can be used to extract and transmit massive project data from ordinary databases to distributed databases [2]. Secondly, for the collection of various academic social networks, Internet websites, the content of this article can be extracted from huge heterogeneous web pages with the help of web crawler technology. Due to the different sources of various academic data, the overall data is in a heterogeneous state, so it is necessary to integrate and standardize the data, and clean and transform the original data. Apache Hive is a commonly used tool in big data cleaning.

In this paper, a combination of semantic analysis technology and web crawler technology is used to design an algorithm system that can be used for data analysis of web pages, and effectively filter information in web pages that is not related to the content of Science and technology projects. The system uses the vector space model (VSM) in the semantic analysis technology to screen the collected web page data to ensure the quality of the data captured by the crawler technology [3]. The core of the VSM algorithm is to map web page information into a vector space and convert it into vector operations. The process of VSM algorithm processing data includes:

·Information preprocessing, segmenting and filtering web page data collected by web crawlers, and filtering out wrong data.

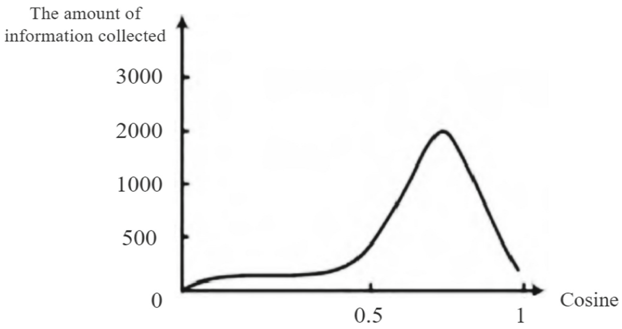

Fig. 2. Distribution diagram of the relationship between the amount of collected information and the cosine value of the VSM

·The word segmentation results obtained in the above steps are counted and weighted, and the corresponding frequency is calculated.

·The web page data is processed as a web page vector with n components, the weight of the keyword in the web page is each component, and the weight depends on the frequency of the keyword. The calculation expression required to calculate the vector similarity is:

$$Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k-1}^{n} W_{1k} \times W_{2k}}{\sqrt{(\sum_{k-1}^{n} W_{1k}^2)(\sum_{k-1}^{n} W_{2k}^2)}} \tag{1}$$

Sim (D1, D2) represents the similarity of two web pages, the calculation result is between 0 and 1, and W1k represents the weight of the kth keyword in the web page. The higher the correlation degree of web page data, the closer the calculation result is to 1, and the calculation and filtering of web page data crawling results can be realized by this indicator (Fig. 2).

In this paper, the crawler web page information crawling efficiency test is carried out with a single project, and the time taken by the crawler to crawling 500, 1000, 2000 items of related information and the average similarity between the collected information and the project theme are tested. The results are shown in Fig. 1 Show. It can be seen from the results that the average collection speed of the system's public opinion information is about 12 ms per piece, indicating that the system has a high collection efficiency, and the collected web page information has a high degree of relevance to the project (Table 1).

2.2 Research on Key Technologies in the Process of Project Data Storage Management

The project data collected in the whole process from Science and technology project declaration and execution to final acceptance and evaluation come from different channels, the data types and characteristics are also different. It needs to be reasonably stored

Table 1. Crawler web crawling efficiency test results

The amount of information collected	Acquisition time/s	Average time/ms
500	5.5	11
1000	12	12
2000	23	11.5

and managed so that it can be efficiently used by the upper algorithm. Currently, technologies commonly introduced in various scenarios include: Data Warehouse, HDFS, HBase, etc. Due to the diverse structure and low value density of massive scientific and technological data, the storage technology of big data cannot rely on a single file system, and needs to be classified and stored according to different data types. This paper introduces four types of mixed data storage technologies in the management of Science and technology projects [4].

First, for large-scale structured data, a new MPP-based data cluster is used, which can support the analysis and storage of PB-level structured data. Second, for semi-structured and unstructured data, Hadoop-based technologies can be used. Hadoop is suitable for the storage and management of unstructured data, and its easily scalable properties are more conducive to the application of big data analysis. Third, for the mixture of structured and unstructured data, it is necessary to combine the advantages of MPP and Hadoop for hybrid storage, which is also the current trend of processing big data. Fourth, for those data that require high real-time performance, due to its high requirements on the efficiency of database reading and writing, traditional databases can no longer meet it. InfluxDB is the best choice for such data storage.

2.3 Research on Key Technologies in the Process of Project Data Calculation and Analysis

Data calculation and analysis is the final application stage of massive Science and technology project data. Through mathematical modeling, predictive analysis, machine learning and other technologies, the correlation between data is discovered and the potential value of data is found.

First of all, in the stage of project approval review, rational use of big data technology to achieve scientific topic selection. The innovation and timeliness of Science and technology projects are an important basis for project approval review. By combing the existing literature, it can be found that the idea of checking Science and technology projects basically starts from the project declaration form, mining useful information from massive data, and performing data processing such as word segmentation on it, extracting feature vectors, and the innovation of projects is determined by the similarity of feature vectors. There are two factors that determine whether the duplication check of a Science and technology project can be successfully implemented: one is the establishment of the duplication checking database and the maintenance of the original data, and the other is the selection of key fields in the duplication checking model.

Table 2. Generate rules based on user hierarchy and data attributes

Data attribute		**User level**				
		Access attribute 1	Access attribute 2	Access attribute 3	Access attribute 4	Access attribute 5
National	User A	√	√	√	√	√
Provincial	User B	√	√	√		√
Municipal	User C		√	√		√
Personal	User D					√

Secondly, in the stage of acceptance and project evaluation, it is necessary to focus on expert retrieval and matching based on big data technology, and the evaluation of project management benefits. The expert recommendation method in traditional Science and technology project management is to formulate some simple screening and avoidance rules, and use the manual selection method to match the expert database with the project, but this method is inefficient and difficult to achieve complete objectivity and fairness. Based on the research situation at home and abroad, at present, expert retrieval and recommendation are mainly completed by building matching models through big data [5]. For example, TF-IDF and other algorithms are used to process project text, extract key fields, and calculate semantic similarity. In this way, the semantic similarity between items and experts is calculated, so as to realize expert recommendation.

Finally, the intelligent decision support system—ADSS (AI + DSS) is applied to the management of Science and technology projects, through the comprehensive analysis of the data from the expert database, knowledge database, method database, semantic system, problem processing system, etc., it is presented in the interface of human-computer interaction to provide supporting information for high-quality decision-making.

2.4 Research on Key Technologies in Project Data Security Management

The diversification and wide distribution of big data information make Science and technology project management face huge challenges in data security management. The sustainability and value of big data analysis can only be achieved by ensuring that data information is used safely and reasonably. First, in terms of data security, classify and store data from different sources, set different security levels and access passwords according to the importance of data information, and build a data access control authority mechanism combined with user attributes (Table 2).

In addition, on the premise of not affecting the applicability of the data, the associated data information is desensitized to reduce the risk of data leakage [6]. Secondly, in the process of project management, strengthen the "trace" management of the whole process, and comprehensively use information systems, video recordings and other means to ensure that the whole process of the project is "searchable and traceable". Based on the big data computing and analysis platform, establish risk early warning models and risk information bases for research technology routes, major issues research, and fund use management, carry out real-time supervision of the entire process of Science and technology projects, and timely prevent and resolve risks in project management.

3 Conclusions

Big data technology has given new features and paths to educational information management, and put forward higher requirements for scientific research management. This paper conducts research on key technologies in data collection, integration, storage, computational analysis, privacy and security, etc., to improve the scientific nature of scientific research project process management, auxiliary decision-making, expert recommendation, and evaluation of scientific and technological achievements, in order to improve the quality of modern education. Social and economic benefits.

Acknowledgment. This research was supported by the Key R&D Program of Shandong Province (Soft Science Project) (2021RZA01023).

References

1. Liang Dai. Research on science and technology project management based on life cycle management. East China Science and Technology, no. 7, 2021, pp. 68-71.
2. Junwu Ren, Bin Xu, Jiayao Liu. Design of project management system for full-cycle government affairs informatization. China Construction Informatization, no. 21, 2021, pp. 64-66.
3. A. R. Mohd, H. Zainuddin. A Theoretical Framework of Critical Success Factors on Information Technology Project Management During Project Planning. International Journal of Engineering & Technology, vol. 7, 2018, pp. 650-650.
4. Paton S, Andrew B. The role of the Project Management Office in product lifecycle management: A case study in the defence industry. International Journal of Production Economics, vol. 208,2019, pp. 43-52.
5. Hao Zhang. Practical Analysis of Information Management of Engineering Projects. Electronic Technology, vol. 51, no. 2, 2022, pp. 250-251.
6. Wei Gao. Analysis on the Optimization and Innovation Path of Science and Technology Project Management in the Era of Big Data. Heilongjiang Science, vol. 12, no. 8, 2021, pp. 120-121.