Fuzzy Data Mining Technology Based on Time Sequence in Archives Management System

Wei Wang¹*, Xiaohui Hu¹, Mingye Wang¹
¹College of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191, China

*Corresponding author e-mail: wangw5901@buaa.edu.cn

Abstract. The rapid development of data mining technology enables the processing of unstructured data in archives. It can help realize the multiple classification of archives, the accurate retrieval of archive information, the integrated presentation of archive content, and the scientific standardization of archive appraisal. This article mainly introduces the research of fuzzy data mining technology based on time series in archives management system. This paper uses the research on the time series-based fuzzy data mining technology in the archives management system, uses the data mining technology to plan and design the archives management system, and combines the relevant procedures of the archives management work to discuss what should be followed when designing the archives data mining model. It also introduces the basic framework of archive data mining in detail, mainly from the aspects of database construction, data preprocessing, data document establishment, training process design, application process architecture and so on. The experimental results of this paper show that the research on the time series-based fuzzy data mining technology in the file management system has increased the efficiency of file management by 19%. The limitation of the research on the time series-based fuzzy data mining technology in the file management system is the data Mining applications provide good methods and ways to analyze, discuss and summarize, so as to enrich the academic research results.

Keywords: Time Series Application, Data Mining, Fuzzy Data, File Management

1. Introduction

Under such a background, if the archives business cannot keep up with the trend of the times, it will inevitably not be able to manage the mass production of archives in the new era [1-2]. For this reason, this article discusses the research of fuzzy data mining technology based on time series in the file
management system. While expounding related theories, it focuses more on the analysis of practical content, and specifically shows the operating principles of the entire data mining process to show this. It is not a conjecture, but a project that can be actually completed [3-4]. Since data mining technology itself is a comprehensive discipline, coupled with the specific application in archival work, what is described in this article can only be regarded as a basic framework. Some of the details still need to be studied in depth, and the relevant content can be further expanded. I hope to help Everyone has a basic understanding of the application of data mining in archives, and at the same time, it also serves as a starting point for academic and industry research [5-6].

With the advancement of science and technology and the rapid development of the Internet, the volume of archive data is too large to use traditional data analysis tools and techniques to process them. Bhattacharyya B believes that the data set is relatively small, and if the data itself has heterogeneous characteristics, it cannot be processed by traditional methods [7]. In order to solve these problems by Scott JK, researchers from different disciplines began to jointly develop effective tools that can handle different data characteristics [8]. The technology of data mining was born under this background. It is based on the combination of traditional data analysis and big data processing algorithms, which provides the possibility to explore the potential value contained in a large amount of data. In the long-term work, various institutions are A large number of archives have been formed. Most of these archives are used in the form of "vouchers", and the way of using them is relatively traditional and single. The way of manual system retrieval and utilization still occupies the mainstream position [9-10]. However, there are errors in their experimental process, which leads to inaccurate results.

The innovation of this paper is to put forward the research of fuzzy data mining technology based on time series in file management system. The research aims at the research of fuzzy data mining technology in archives management system, and analyzes the effective countermeasures of fuzzy data mining algorithm. This type of algorithm is used in file management. Mainly expounds the concepts and relationships of archive data mining, machine learning, and database management. The value of archive data mining in reality is embodied in the multiple classification of archives, the accurate retrieval of archive information, the integrated presentation of archive content, and the scientific norms of archive appraisal. The purpose of this research is to find a new path suitable for the development of current data mining technology in the file management system.

2. File Management under Fuzzy Data Mining

2.1. File Management Analysis

File organization and other related management activities are mainly done manually by archivists. Once a large number of files appear at the same time, work efficiency will inevitably be greatly reduced. On the whole, my country's archives industry is still in the initial stage of informatization, and there is a certain gap compared with the archives management model of developed countries. However, with the rapid development of the information age and the acceleration of the internationalization process, the paperless era will eventually come slowly. No matter whether the traditional dual-system mode or dual-track mode is replaced, a large number of electronic files will surely be produced. These original documents generated by the computer system during the office process are essentially different from the documents formed after the digitization of the archives. The
former is another manifestation of the archives and the latter is just a copy of the archives. This type of original electronic document load With more information than digital documents, it is also more convenient in processing.

As the final destination of these electronic documents, the archives department will accumulate a large amount of original information. If these documents reflecting the original information can be used efficiently, it will definitely generate value. Therefore, how to organize these complex archive data scientifically has become the most important problem for utilization. Fortunately, data mining technology has been able to process these heterogeneous data, solve problems that may actually be encountered, make file management related activities more intelligent, reduce the burden on archivists, and at the same time allow the general public to get better quality service. An effective algorithm can be designed in file management, with the following forms.

$$K = (U \times V + 1)^{p1}$$  \hspace{1cm} (1)

The calculated value of the file management algorithm is as follows;

$$\text{Precision} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|}$$  \hspace{1cm} (2)

The test shall use the following formula:

$$\begin{ aligned} I_{A1} &= U \setminus \varnothing C_{01} \\ I_{B1} &= 0 \\ I_{C1} &= U \setminus \varnothing C_{01} \end{ aligned}$$  \hspace{1cm} (3)

2.2. Archive Data Mining Application

In the whole process of archival data mining, many algorithms are involved, mainly in the three stages of Chinese word segmentation, feature extraction and classification. Correspondingly, this article mainly uses statistics-based machine learning algorithms, TF-DF algorithms and naive Bayes algorithms to discuss. The reason is that these three algorithms are more commonly used and easier to understand. Therefore, the above content is mainly used for Explain the entire archival data mining framework. In actual applications, these three algorithms do not necessarily have a good classification effect. In addition, the files of each organization are also different. Even the same algorithm may have greater differences in different application environments. For this reason, other commonly used algorithms that can be used for archival data mining are introduced here. Under appropriate conditions, the algorithms used above can be replaced for analysis. With the continuous deepening of the informatization construction of Kunshan Archives Bureau, massive data continues The amount of archive data has formed a certain scale. The society's demand for in-depth knowledge is also becoming stronger. It is no longer sufficient to meet the needs of economic and social development if the use is
limited to personal inquiries. The large amount of archive data formed by many years of archive digitization is also a waste of resources.

3. File Management System Based on Time Series

3.1. Time Sequence Analysis of File Management

The construction of the verification model for the sequential consistency of the archive management workflow. In order to monitor and verify the timing consistency state during the execution of the archive management workflow, this paper constructs corresponding timing consistency verification models from two different perspectives. As far as the archive management workflow is concerned, one is based on throughput and time. The time-series consistency verification model of the propagation effect is constructed. As far as the activities executed on a single queuing system are concerned, a time-series consistency verification model based on the response time of workflow activities is constructed. The proposed time-series consistency model can accurately reflect the sequence status during the operation of the workflow. Field research method: Through field visits, interviews and other forms, we can understand the relevant work done by the Kunshan Archives Bureau on the archives data mining platform, analyze the principles of its operation, and confirm whether there is something worth learning from. Interdisciplinary research method: Because the thesis involves the content of other disciplines, it needs to be structured with the thinking of related disciplines and discussed in combination with the content characteristics of archives. Experimental method: When conditions permit, do simulation experiments on part of the content in the paper to analyze and explain the internal architecture principles.

3.2. Timing System in File Management

According to the time sequence consistency model, the running sequence status of the file management workflow is verified, so as to detect the sequence abnormality. Compared with the existing detection point selection strategy based on fixed observation time interval, this paper proposes a detection point selection strategy based on dynamic observation time interval, which uses the workflow execution of the previous detection time point to determine the next observation time interval the size of. This strategy can not only effectively monitor the occurrence of timing abnormalities, but also reduce the monitoring cost of timing abnormalities.

Data mining is a process of discovering and identifying potentially valuable and understandable information from a large amount of data. It can be seen from this that the first premise of data mining is the amount of data, and the amount of data directly affects whether the potential between the data can be discovered. Contact, the larger the amount of data, the more reliable the results obtained. Secondly, the larger the amount of data, it means that the stability of data including data structure and data content will be reduced, the structure of data will become more and more complex, and the content of data will become more and more messy, making traditional computers Statistical analysis has become impossible to start. It needs to simulate human thinking to analyze and mine. Therefore, machine learning is the realization method of data mining. On the whole, data mining can be considered as the intersection of machine learning and database management. With the support of database management technology, a large amount of data is extracted from the database and analyzed by machine learning technology to mine potentially valuable information. The specific results are
shown in Table 1.

**Table 1.** Comparison of several wireless data transmission methods

|                        | Short Message | Gsm Circuit Connection | Gprs        |
|------------------------|---------------|------------------------|-------------|
| **Way of Communication** | Full Duplex   | Full Duplex            | Full Duplex |
| **Transmission Rate**   | 26s/Piece     | 9.6kbps                | Maximum 171.2kb |
| **Reliability**         | Poor          | Extremely High         | High        |
| **Time Delay**          | Large Fluctuation Range | Small            | Medium      |

4. **File Management Analysis under Fuzzy Data Mining**

4.1. **Archive Information System**

Accurate retrieval of archive information is an important way to use archives. The data formed by digitizing the form and content characteristics of archives is stored in the database, and the result data is returned through the retrieval system, so as to achieve the purpose of archive retrieval. The retrieval algorithm has basically reached a satisfactory level, and the improvement in satisfaction brought about by the improvement of the retrieval algorithm due to the limitation of the retrieval scope has reached the bottleneck period. The real influencing factor is the total amount of relevant information in the system, namely including the original content text data and metadata of the archive, the original content text data of the archive is directly presented to the user and can be seen directly by the human eye. The metadata of the archive is highly generalized when the archive is described, which can express the relevant characteristics of the archive. Data, such as subject headings, etc. In archival information retrieval, the main source of information retrieved is metadata, and some metadata such as subject terms are rarely described. Therefore, the type of metadata retrieved is very limited, and the title is often used as the main. The specific results are shown in Figure 1, the abscissa is the number of file synchronizations N, and the ordinate is the absolute value of the error. The test results show that the increase of the synchronization times makes the error decrease exponentially.
4.2. *Queuing Model Analysis*

Most archival organizations use this type of database. As the process of archival data mining involves more heterogeneous data, it requires the participation of non-relational databases, especially document-based databases. Compared with relational databases, non-relational databases do not need to define related table structure before use, and they also have greater flexibility when used. Specifically, the management of the database should be adjusted in time according to the tasks of the data mining process to meet its system requirements, and this type of information is often not within the predetermined design knot. Therefore, in the process of archival data mining, data is often brought in. Uncertainty of variables such as type, thereby exacerbating the instability of the defined structure. In the archives industry, with the development of the information society and the advancement of the paperless policy, the amount of digital archives is constantly increasing, and most of them are unstructured data, making archive management more difficult, archival data mining is to discover the useful information from a large number of digital archives that is beneficial to archive management activities, improve the current situation of archive managers, and provide the general public with better archive services. The specific results are shown in Table 2.

| Table 2. File service |
|----------------------|
|                     |
| Normal  | Ageing | Malfunction |
| Number of Transformers | 17     | 566         | 82     |
| Total Sample       | 305    | 160         | 160    |
| Training Samples   | 204    | 106         | 106    |
| Validation Sample  | 101    | 54          | 54     |
Queuing theory (or "queuing theory") refers to the mathematical study of the random process of queuing queues and their characteristics, trying to evaluate queuing behavior based on various assumptions. The queuing simulation of the concurrent execution of file management workflow is usually composed of large-scale concurrent workflow instances, and each workflow instance has a strict business logic execution sequence. In the big data environment, a single workflow instance generally has simple execution steps, while the process of concurrent execution of a large-scale workflow is quite complicated. In addition, because business system software systems generally have potential errors and abnormal propagation phenomena. The specific results are shown in Figure 2. When the number of attributes is the same, the efficiency of queuing theory has been significantly improved as the number of sample data increases.

![Figure 2. Random queue and its characteristics](image)

5. Conclusion

Although this article is studying the fuzzy data mining technology based on time series in the archives management system, there are still many shortcomings. Since the development of data mining technology in the 1950s, with the help of the upgrading of hardware equipment, it has given new vitality in today's industrial revolution. However, specific to the archives industry, there are still not many researches on the practical application of data mining in archives management. Most of them are at the macro level, and there is no complete framework system. There are still many in-depth contents worthy of research on the design and research of indoor path planning algorithms for complex environments. There are still many steps in the research of fuzzy data mining technology analysis because of space and personal ability, etc., which are not covered. In addition, the actual application effects of the related experiments of the file management system can only be compared with traditional models from the theoretical and simulation levels.

References

[1] Li S, Huang J X, Tohti T. Fake Plate Vehicle Auditing Based on Composite Constraints in Internet of Things Environment[J]. IOP Conference Series: Materials ence and Engineering, 2018, 322(5):204-205.

[2] Kumar N, Rodrigues J J P C, Chilamkurti N. Bayesian Coalition Game as-a-Service for
Content Distribution in Internet of Vehicles[J]. IEEE Internet of Things Journal, 2017, 1(6):544-555.

[3] Wang X, Han S, Yang L, et al. Parallel Internet of Vehicles: ACP-based System Architecture and Behavioral Modeling[J]. IEEE Internet of Things Journal, 2020, 7(5):3735-3746.

[4] Lee D G. A multi-level behavior network-based dangerous situation recognition method in cloud computing environments[J]. The Journal of Supercomputing, 2017, 73(7):1-16.

[5] Jing M, Jie Y, Shou-Yi L, et al. Application of fuzzy analytic hierarchy process in the risk assessment of dangerous small-sized reservoirs[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(1):113-123.

[6] Chen H, Feng S, Pei X, et al. Dangerous Driving Behavior Recognition and Prevention Using an Autoregressive Time-Series Model[J]. Tsinghua Science and Technology, 2017, 22(006):682-690.

[7] Ruddy J, Meere R, Terence O’Donnell. Low Frequency AC transmission for offshore wind power: A review[J]. Renewable and Sustainable Energy Reviews, 2016:14-15.

[8] Bhattacharyya B, Raj S. Swarm intelligence based algorithms for reactive power planning with Flexible AC transmission system devices[J]. International Journal of Electrical Power & Energy Systems, 2016, 78:158-164.

[9] Scott J K, Laird C D, Liu J, et al. Global Solution Strategies for the Network-Constrained Unit Commitment Problem with AC Transmission Constraints[J]. IEEE Transactions on Power Systems, 2018:1-1.

[10] Liang H, Liu Y, Wan L, et al. Penetrating power characteristics of half-wavelength AC transmission in point-to-grid system[J]. Journal of Modern Power Systems and Clean Energy, 2019:277-299.

[11] Liang H, Liu Y, Wan L, et al. Penetrating power characteristics of half-wavelength AC transmission in point-to-grid system[J]. Journal of Modern Power Systems and Clean Energy, 2018, 7(10):1-8.