Research on Unstructured Data Processing Technology in Executing Audit Based on Big Data Budget

Benzhong Hou¹, Yongqiang Zhang¹, Ying Shang², Xin Liang³, Tiantian Liu⁴ and Jiangwen Su⁵

¹STATE GRID Corporation of China, Beijing 100000, China
²State Grid Beijing Electric Power Company, Beijing 100000, China
³State Grid Fujian Electric Power Co., Ltd., Fujian Fuzhou 350000
⁴Big Data Center of State Grid Corporation of China, Beijing 100000, China
⁵Fujian Yirong Information Technology Co., Ltd, Fujian Fuzhou 350003

Abstract. The wave of big data has promoted the transformation of auditing technology, which has brought about tremendous changes to auditing models and auditing methods. The traditional audit data analysis methods cannot analyze semi-structured and unstructured data, nor can they meet the requirements of the development of audit informatization in the context of big data. New ideas and methods for audit data analysis are urgently needed. In this context, the article proposes an audit data analysis framework based on text mining, and describes the detailed audit data analysis process of collecting and storing, mining and analysing, summarizing and publishing. By using text mining technology to mine the unstructured original audit data collected, different text mining models are established according to the clear audit requirements, the audit data is analysed, and then audit doubts are found, and ultimately formable audit evidence and an audit trail. The construction of the framework aims to provide new ideas for big data audits to reduce the risk of big data audits and improve audit quality.

Keywords: Budget execution audit, big data technology, text mining, audit data analysis.

1. Introduction

With the advent of the era of big data, auditing, as a comprehensive economic supervision department, is bound to face the huge challenges it brings. Massive unstructured data has been generated in enterprise systems. How to analyze these unstructured data is an important part of promoting big data audits. With the widespread promotion and application of big data in the audit field, the importance of text mining technology for audit data analysis has gradually become more prominent. It no longer takes structured audit data as the analysis object, but can deeply analyze large amounts of unstructured data. Perform mining analysis and utilization. Taking the audit of enterprise contract data as an example, combined with the characteristics of text mining technology, this article discusses the auditing method based on text mining analysis technology, and explores the application of text mining analysis technology in auditing.
2. Analysis of audit data and unstructured data

2.1. Traditional audit data analysis
How auditors convert the collected raw data into audit evidence will directly affect the realization of audit objectives. From the process of collecting to obtaining evidence, the most important question that auditors should pay attention to is whether they can mine valuable data for audit data analysis, which plays an important role in the quality of audit projects and the embodiment of audit results. Therefore, the most critical step in the audit work is to conduct audit data analysis.

At present, the audit data analysis methods and computer-aided audit tools (CAATS) frequently used by auditors in audit work include account analysis, economic index ratio analysis, trend analysis, statistical analysis, Excel data analysis, Access, SQL, AO audit software, etc. Excel data analysis and audit software for accounting statements are used by firms; SQL statement query, AO audit, audit data collection and analysis and other audit software are often used in internal audit work of government departments and institutions; For internal audit, large enterprises use a dedicated audit platform or embed internal audit modules in ERP. SMEs rely on Excel and Access for audit data analysis. However, with the advent of the big data era, massive and diverse heterogeneous data has greatly expanded the scope of big data audits. Traditional audit methods and auxiliary audit tools have become incapable of collecting and analysing unstructured data [1].

2.2. Unstructured data analysis
Unstructured data contains complex content and has different structural characteristics. Traditional relational databases cannot respond to the requirements of unstructured data management in terms of description capabilities or management data scale. Therefore, it is necessary to study and establish a new management technology and platform specifically for the characteristics of unstructured data. The key technologies for building such unstructured data management platform for big data include:

1. Unified data model, which describes various heterogeneous unstructured data in a unified manner, and can realize the integrated description of various modal information (such as semantic features, underlying features, etc.), thereby effectively supporting the big data Retrieval and correlation analysis.

2. Distributed storage and parallel processing models and architectures. This architecture should be highly parallelized and extensible to ensure the processing efficiency of big data.

3. Query language and data visualization. Query language and visual data display are the interface for users to access and analyze big data. The query language should be able to accurately express the user's willingness to access data and support flexible and rich data operations. Figure 1 shows the difference between unstructured data and structured data.

![Figure 1. The difference between unstructured data and structured data.](image-url)
3. Audit unstructured data extraction algorithm

3.1. Data analysis
There is a large amount of contract data in the enterprise contract management system. The enterprise contract audit is mainly for the execution of the contract, especially the audit of the contract payment, including the total price of the contract, the first payment time, the amount, and the second payment time, Amount, etc. Enterprise contract data belongs to unstructured document data. The audit of this type of data currently requires the manual extraction of key contract data (contract amount, payment time, etc.), which requires a lot of labour costs. How to quickly extract the information we need from these unstructured data is the key to solving the rapid audit of enterprise contracts.

3.2. Key information extraction based on hidden Markov model
The purpose of information extraction is to extract specified information. There are many different techniques for information extraction, which can be divided into three categories according to the model used: dictionary-based extraction, rule-based extraction, and hidden Markov model (HMM)-based Extract. Dictionary-based information extraction needs to construct an extraction pattern dictionary, and then use the pattern dictionary to extract the required information from unlabelled text. Dictionary-based information extraction requires the definition of concept nodes, the establishment of related models, the knowledge workload is relatively large, the generation of the best dictionary is also more troublesome, the field correlation is too strong, and the application is not very extensive. Rule-based information extraction needs to construct an extraction rule set first, and then use these general rules to extract information from the text. Rule-based information extraction is commonly used, and in many cases the accuracy is very satisfactory. The disadvantage is that the rule structure is more complicated and less adaptable. Text information extraction using Hidden Markov Model (HMM) is an extraction method based on statistical learning [2].

3.2.1. Hidden Markov Model. A one-dimensional HMM can be generalized, giving it a two-dimensional appearance by allowing each state to become an HMM in a one-dimensional HMM. In this way, the hidden Markov model can be composed of a set of super states and a set of embedded states. Then the super state can model the two-dimensional data along one direction, and the embedded HMM models the data along the other direction. This model is different from a real two-dimensional hidden Markov model, because the transition between different super states is not allowed. Therefore, the model is called embedded HMM. Elements of embedded HMM:

1. \( N_0 \), super state number, \( S_0 \) means super state, \( S_0 = \{ S_{0,1}, I \leq i \leq N_0 \} \).

2. \( \Pi_0 \), the initial probability distribution of the super state, \( \Pi_0 = \{ \pi_{0,i} \} \), where \( \pi_{0,i} \) represents the probability that the time 0 is in the super state I.

3. \( A_{ij} \), the state transition probability between super states, \( A_{ij} = \{ \alpha_{0,i,j} \} \), where \( \alpha_{0,i,j} \) represents the transition probability from super state I to super state j. For a continuous embedded HMM, the observation is characterized by a continuous probability density function, the same dimension HMM is similar, you can take a mixed Gaussian probability density function:

\[
b^k(0_{t_0,t_1}) = \sum_{\pi=1}^{N_0} c^k_{\pi a} N\left(O_{t_0,t_1}, \mu^k_{\pi a}, U^k_{\pi a}\right)
\]  

(1)

Where \( 1 \leq i \leq N_f \), \( c^k_{\pi a} \) is the coefficient of the m-th component of state I in super state k, and B is a Gaussian density function containing the mean vector \( N(O_{t_0,t_1}, \mu^k_{\pi a}, U^k_{\pi a}) \) and the variance vector \( U^k_{\pi a} \). We use \( \Lambda^k = (\Pi^k, A^k, B^k) \) to denote the k super state parameter set, then the parameter set of the embedded HMM can be abbreviated as:

3
\[ \lambda = \left( \Pi_0, A_0, \Lambda \right) \]  

(2)

Here \( \Lambda = \left\{ \Lambda^1, \Lambda^2, \ldots, \Lambda^n \right\} \). Although more complex than one-dimensional HMM, the embedded hidden Markov model is more suitable for two-dimensional data, and its complexity is proportional to the square sum of the number of states, \( \sum_{i=1}^{n} k_i^2 \).

3.2.2. Tetrahedron model. In order to process different types of unstructured data, it is necessary to describe these unstructured data uniformly, and implement the management and operation of the unstructured data content based on descriptive information. The description of unstructured data is divided into three categories, which are keyword-based semantic description, low-level feature-based description and concept-based semantic description. Therefore, unstructured data can be composed of four parts: basic attributes, semantic features, underlying features, and original data, and there are various connections between these four parts of data, called the tetrahedral data model of unstructured data, as shown in picture 2. The audit cloud platform extracts metadata such as semantic features, basic attributes, and underlying features of unstructured data by building a tetrahedral model for storage.

1) Semantic features: the unique semantic attributes of unstructured data expressed in words. For data such as meeting minutes and rules and regulations, semantic features generally refer to content summaries and subject descriptions. This attribute is not required.

2) Basic attributes: Refers to the general attributes of unstructured data, these attributes do not involve the semantics of the data. Including name, type, creator, creation time, etc.

3) Low-level features: Unstructured data characteristics obtained through various special processing technologies (such as image, voice, and video processing technologies). For example, related technical attributes such as the format, duration requirements, and bit rate of telephone recordings.

4) Original data: the original ecological file of unstructured data.

![Figure 2. Tetrahedral data model of unstructured data.](image)

3.2.3. Construction standards for middle-tier models. According to the above-mentioned tetrahedral model, perform audits on various types of unstructured data in the budget to build a unified data management model standard. The unified storage and processing of unstructured data of different banks and different types. The intermediate layer model of unstructured data is shown in Table 1. Under this standard, the characteristics of different types of unstructured data are extracted to realize the storage and processing of all unstructured data in a unified data mode. The characteristics of the tetrahedral model are the integration of semantic features and underlying features for integrated expression; the unity of expression of multiple heterogeneous data such as images, text, video, audio, etc.; support for the scalability of semantic features and dynamic changes of underlying features.
Simplicity. The unstructured data management model based on the tetrahedron model can realize the unified storage and association operation of heterogeneous data, thereby better supporting the deep processing of big data [3].

3.2.4. Key text information in budget execution audit. 1) The total contract prices. In the enterprise contract, the total price of the contract will be identified by a clear keyword, such as the total service fee, after the keyword is located to the paragraph to which it belongs, the currency regular expression is used to extract the amount data. 2) Payment time. In the contract, if the payment time is clearly stated, the data can be extracted directly according to the regular expression of time. The extraction expression of the time displayed in different formats is different. The simplest date format is YYYY/MM/DD. 3) Payment amount. The payment amount and payment time will appear in the same paragraph. After the payment time is withdrawn, the payment amount is withdrawn in the same way as with the total contract price. 4) Payer and payee information. The information of the payer is mainly the name of the paying unit, and the information of the payee includes the bank, account number and account name of the account, such information is generally clearly stated in the contract, and the information can be extracted according to the specific keyword positioning. The extracted payer and payee information is used to obtain data corresponding to the contract from the fund flow data returned by the bank, so as to check the contract payment situation. The key information extracted from the enterprise contract is stored in the data table built in advance by the database to achieve structured storage, which is convenient for auditing the contract [4].

| Field Name | Field Type | Field description |
|------------|------------|-------------------|
| HTBH       | Character  | Contract No        |
| HTMC       | Character  | contract title     |
| FKDW       | Character  | Payment Unit       |
| KHYH       | Character  | Bank account       |
| ZH         | Character  | account number     |
| KHM        | Character  | Account Name       |
| HTZJ       | Numeric    | The total contract price |
| FKSJ1      | Time type  | First payment time |
| FKJE1      | Numeric    | First payment amount |
| FKSJ2      | Time type  | Second payment time |
| FKJE2      | Numeric    | Second payment amount |

4. Design of unstructured processing platform in budget execution audit

4.1. Overall structure
Text mining technology is mainly for mining unstructured knowledge, which is an indispensable part of big data audit technology. Especially with the extensive promotion and application of big data in the audit field, the importance of text mining technology for audit data analysis has gradually become prominent. At present, text mining technology is mainly used to analyze the text description contained in documents and web pages. When mining data such as videos, pictures, and voices, the main content is also extracted from it and replaced with easy-to-understand text descriptions. Therefore, this article will focus on text mining technology to analyze audit data, and build an audit data analysis framework based on text mining, as shown in Figure 3.
In order to ensure the integrity and authenticity of the audit data, it is necessary to establish a strict and standardized system for scientific and safe management of the collected unstructured data. By building the HDFS file storage system of the Hadoop distributed framework, centralized storage of unstructured audit data of business systems. On this basis, it is also necessary to build a database suitable for storing unstructured data-HBase. HBase can make up for the shortcomings of HDFS's lack of random read and write operations. All its internally managed files are stored in HDFS. Construct a Hadoop ecosystem based on Hadoop-based distributed file system HDFS, distributed database HBase, and distributed computing framework MapReduce. As shown in Figure 4, unified management of unstructured data. This management model reduces the risk of audit data management, enables data sharing on various platforms, and breaks the embarrassing situation of information isolation [5].

4.2. Processing flow

4.2.1. Data collection. The data collection contents of the audit management personnel executing a certain audit project in a budget include the financial data, part of business data and part of external data of the audited administrative institution. Among them: financial data and business data are mainly carried out through three methods: intranet transmission, reporting by the audited administrative institution, and on-site collection by the on-site audit team; external data mainly refers to the public budget of the audited administrative institution, and the collection method is mainly Internet collection.

4.2.2. Data cleaning. Due to the large number of audited administrative institutions performing audit projects on a budget by audit managers, the financial and business information systems used by them
are different, and the related data formats are not uniform, which is not conducive to subsequent data analysis work, so data cleaning is required. The off-site audit team uses SQL Server 2008 R2 and the on-site audit system (AO) as the audit work platform. When data cleaning is performed, the data conversion and loading in the financial information system database and some business information system databases of each audited unit are extracted, so that it has the same data format and fields and is stored in the database of SQL Server 2008 R2 and on-site audit system (AO) for management. After the data cleaning is completed, the data of the cleaning needs to be verified, mainly through the on-site audit system (AO) to check the amount in the account balance table to ensure that no data errors or omissions are caused during the data cleaning process.

4.2.3. **Data analysis.** In the off-site audit, the off-site audit team will also be divided into multiple groups according to the specific tasks and analyze the data at the same time. During the analysis, SQL statement modelling is mainly used for query. The query results require professional judgment by the auditors, and when they are determined as audit suspects, they are included in the audit doubt summary table [6].

4.2.4. **Summarize knowledge.** According to the audit doubt summary table and the final audit report, you can create an audit knowledge base and summarize knowledge. Audit knowledge can help off-site auditors to improve and design audit indicators and build audit models, which can effectively help auditors improve the accuracy of identifying audit doubt points. Form a data closed loop for budget execution audit. As shown in Figure 5.

![Figure 5. Unstructured data processing flow in data audit.](image)

5. **Conclusion**
In the context of big data, in order to provide new ideas and new methods for audit research, this paper proposes a framework for audit data analysis based on text mining from the perspective of the actual needs of audit data analysis. The framework is based on the Hadoop ecosystem, combined with text mining technology, to establish an audit data analysis framework that integrates audit data collection, storage, analysis processing, and result visualization. The research in this paper aims to provide a reference for the analysis of audit data in a big data environment, but no empirical analysis has yet been conducted. What is certain is that the use of text mining technology can make up for the shortcomings of traditional audit technology methods. How to implement and verify the application of text mining technology in audit work will be the focus of subsequent research.

**Acknowledgments**
This project is funded by the research and technology project of the intelligent related technology of the digital audit platform of the State Grid Corporation headquarters and writes a paper.

**References**
[1] Song, Y., Nam, S., & Kwon, O. Wellnesswordnet: a word net for unconstrained subjective well-being monitoring based on unstructured data and contextual polarity. Journal of Intelligence & Information Systems, 22 (3) (2016)1-21.
[2] Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. Big data in accounting: an overview. Accounting
Horizons, 29 (2) (2015) 381-396.

[3] Mojgan Khademi, Easa Mohammadi, & Zohreh Vanaki. Resources-tasks imbalance: experiences of nurses from factors influencing workload to increase. Iranian journal of nursing and midwifery research, 20 (4) (2015) 476-483.

[4] Thompson, G., & Mockler, N. Principals of audit: testing, data and 'implicated advocacy' | nova. The university of Newcastle’s digital repository. Journal of educational administration & history, 48 (1) (2016) 1-18.

[5] Rao, C., Zhang, H., Gao, H., Zhao, Y., & Zheng, Z. The Chinese cardiac surgery registry: design and data audit. Annals of Thoracic Surgery, 101 (4) (2015) 1514-1520.

[6] Pramod Gupta, Vipin Koushal, Chandradeo Narayan, & Akshay Anand. Building genetic database at medical institutes: implement patient cost audit and improve biomedical research. Annals of neurosciences, 24 (1) (2017) 3-4.