ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on anomaly detection and financial statement auditing based on data mining

Jiangling Huang[1,*]

[1] School of Business, Xiamen Institute of Technology, Xiamen, Fujian, 361021, China

SUMMARY: *With the arrival of the digital era, the scale of data is getting bigger and bigger, and the traditional auditing methods, including sampling audit, circular audit and other methods, can no longer meet the current needs. In this regard, this paper uses association rule algorithm to complete the data mining work on the basis of laws and regulations database, financial and business database, obtains a total of 30 financial and business association rules, and verifies their usability by calculating evaluation indexes. From the definition of anomaly rules and the principle of outlier detection, the anomaly detection model is designed to facilitate the understanding, judgment and audit of business personnel. Finally, under the theoretical guidance of EA-LDA combination algorithm, the hidden relationships between financial statement auditing entities are obtained, which are added to the original database, and finally the financial statement auditing knowledge graph is constructed and analyzed by in-depth exploration. After analyzing, the difference degree of financial statement auditing entities is less than 0.1, and the probability value is less than 0.5, and at the same time, some examples of financial statement auditing knowledge graph based on the EA-LDA combination algorithm are shown, which fully verifies the practical research value of this paper. This paper has certain reference value for enterprise financial statement auditing, which in turn improves its audit quality and efficiency, with a view to ensuring the reliability of its financial statements.*

KEYWORDS: *association rule algorithm; data mining; EA-LDA combination algorithm; anomaly detection model; financial statement auditing; knowledge map*

1 Introduction

China's Securities Law and the Securities Regulatory Commission (SRC) stipulate that listed companies must regularly disclose their operating conditions to the public in accordance with the principles of “truthfulness, accuracy, completeness, timeliness, and fairness”, and provide corresponding financial statements to assist in the explanation [1]. With the financial statements have been plagued with the public is an important issue of financial anomalies, the capital market to bring the return of low investment and high profits to tempt the company to use improper means to carry out financial counterfeiting. Financial counterfeiting refers to the enterprise in violation of the laws and regulations issued by the state, the internal accounts using non-compliant fraudulent means of forgery and concealment, in order to cover up the real operating conditions of the enterprise [2, 3]. The occurrence of financial fraud is inevitably accompanied by the emergence of financial data anomalies, the current effective financial fraud identification model is based on almost all from the enterprise's financial data [4]. The normal operation of the capital market cannot be separated from the openness and transparency of

*huangjiangling1989@163.com
https://doi.org/10.65102/is2026235
Published: 30 April, 2026

enterprise data, and it has become the focus of market supervisory organizations, investors and relevant experts to determine whether the financial data of enterprises are abnormal from the open financial data of enterprises [5, 6]. However, the existing method of analyzing publicly disclosed financial data of listed companies by relying solely on experts from regulatory agencies often suffers from such shortcomings as low data credibility, excessive time lag in data disclosure, and non-uniformity of identification data indicators. The continuous development of data mining technology provides a new solution to the problem.

The application of data mining is extensive and deeply integrated into numerous fields of production and life, with strong coverage and a constructive role. Certified public accountants should resolutely shoulder the responsibility of safeguarding public interests and enhancing the efficiency of capital market allocation on the path of digital economic development. In auditing work, they should also shift towards a direction centered on "improving quality and efficiency through the application of data mining technology" [7-9]. Big data-based financial statement auditing refers to the auditor to big data platform, technology as the support, mining the source of complex, diverse format, the number of huge scale of data information, cross-project team, cross-functional departments, cross-audit geographic area of the deep analysis, and on this basis, to improve the accuracy of the anchoring of the audit problem, analysis of the audit risk of the comprehensiveness of the [10, 11]. Enhancing the ability to use data mining technology to check risks, evaluate judgments, and comprehensively analyze in auditing work marks China's data mining technology auditing with data resources as a new orientation toward a new development era, in which all audit subjects are facing unknown changes [12, 13]. Data mining technology is the core driving force of enterprise digital transformation, and at the same time will bring many opportunities and challenges to financial statement auditing.

How to accurately identify corporate financial anomalies has been the focus of academic attention, and the development of data mining technology in recent years has pointed to a new research direction. Literature [14] points out that anomaly detection is an important data analysis task on the fields of finance, computer network, behavioral analysis, etc. The study launched an in-depth investigation on its application in the field of finance, and compared the effect of multiple clustering algorithms in anomaly detection from different perspectives. Literature [15] investigates the effectiveness of the combination of blockchain technology and data mining technology in the assessment of financial anomalies, the article overviews the blockchain technology and data mining technology, and confirms the potential value of the combination of the two technologies in financial anomaly identification through a practical case. Literature [16] used nearest neighbor, clustering and statistical methods to detect anomalies using the trading data of the Australian Stock Exchange from 2009-2013, and showed better anomaly detection performance by testing the local outlier factor and the clustering-based multivariate Gaussian outlier score. Literature [17] designed a data mining based anomaly detection method that can identify anomalies in trading at the business level and operational level, by determining the trust factor at the business level and the credibility factor at the operational level to identify whether the current trading is anomalous or not.

Machine learning is one of the core techniques of data mining and it is also widely used in financial anomaly detection tasks. Literature [18] conducted a comprehensive review on the topic of fraud detection in the financial field and found that data mining techniques are the most widely used techniques for financial fraud detection, where machine learning methods such as logistic models, neural networks, Bayesian belief networks, and decision trees play an important role in solving specific fraud problems. Literature [19] compared seven supervised machine learning techniques and two unsupervised machine learning techniques for anomaly detection in financial auditing, different forms of machine learning methods have their own strengths and weaknesses, and all of them show great potential in extracting financial data to

identify anomalous entries. Literature [20] investigated the financial fraud of Brazilian export enterprises, applied deep learning-based anomaly detection model to analyze the product export data of relevant Brazilian enterprises in 2014, and the study accurately detected the data anomalies of more than 20 enterprises, effectively preventing the money laundering efforts of enterprises. Literature [21] used the method of Meta learning to identify 9006 financial samples in the United States, of which 9191 are non-fraudulent samples and 815 are fraudulent samples, and the final predicted recall rate of the fraudulent samples is 81.5%. Literature [22] used decision tree, random forest and other algorithms to detect anomalies in financial data, in which the random forest algorithm assesses the degree of data anomalies based on the similarity of the samples, and the algorithm effectively improves the computational practice and accuracy of the anomaly detection model, which is more advantageous than other methods. Literature [23] designed a financial fraud detection framework called CoDetect, which analyzes network information and financial characteristic information to identify fraudulent behavior, and tests on the dataset verified the feasibility of the framework, which seriously combats financial criminal activities. The machine learning based continuous fraud detection system proposed in literature [24] enables real-time monitoring of financial information while tracking financial transaction patterns, user activities and audit logs, and the performance of the system's case study is recognized by the crowd, with an anomaly detection accuracy and recall rate of more than 0.9.

Big data and data mining complement each other, the relationship between big data and financial statement auditing research, for the use of data mining technology to guide the audit work to provide a solid theoretical foundation. Literature [25] points out that many industries and enterprises for the application of data to the direction of big data continues to evolve, pay more attention to the data itself, the process of data generation and analysis, not only in the simple presentation of the data, these evolutions for the auditing industry to bring new challenges and opportunities for reform. Literature [26] investigated the impact of the interaction patterns of big data technology features in the user auditing process on the relationship with the company's internal relations, and that big data technology optimizes the financial statement auditing process, automates audits on a large scale, and helps auditors to communicate more efficiently. Literature [27] suggests that the importance of Big Data and Artificial Intelligence in accounting and auditing is unquestionable, however the impact and application of these technologies to these areas has not been fully explored to date, and Big Data auditing still has a large scope for development. Literature [28] emphasizes the importance of big data analytics in financial statement auditing, where information technology, data categorization, and diverse skills leverage audit data, in addition to demonstrating strong advantages in audit efficiency improvement and assisted decision making.

Currently, the application of data mining techniques in financial statement auditing is mostly focused on the field of fraud detection research. Literature [29] screened and comprehensively analyzed the articles related to financial statement auditing research and determined that the biggest problem in financial statements is fraud detection, for which it explored the solutions based on machine learning and data mining techniques, and looked forward to the future development of this field. Literature [30] builds a classification system to study the application of data mining technology for fraud detection in financial statement auditing, which can accurately match the data mining technology and fraud detection scheme, and greatly improves the detection efficiency in response to different fraudulent behaviors. Literature [31] in order to improve the auditor in the implementation of financial statement audit task is the identification of abnormal patterns and potential risks, using data mining technology to design a financial audit system, the system's audit quality and efficiency has been greatly improved, and can be visualized to identify the error nodes of the audit operation.

Literature [32] applied data mining and generative adversarial network model in fraud detection in financial statement auditing, which overcomes the high-dimensional nature of the data feature space, and has a good performance in monitoring fraud in homemade datasets. Literature [33] compares the differences of several data mining techniques in the identification of fraud factors in financial statement auditing, where logistic regression, decision trees and artificial neural networks have correct classification rates of 88.5%, 90.3%, and 92.8%, respectively, in the test sample. Literature [34] investigated the application of data mining techniques in the classification of financial statement tampering, and the classification methods used include decision tree, logistic regression and artificial neural network, in which artificial neural network has the best classification performance, and the fraud in the company can be effectively detected through classification. Literature [35], on the other hand, conducted a systematic analysis of financial statement audit data, the study randomly selected 12 attributes in the data set to build 10 stochastic decision tree model, the classification accuracy of this model on the data compared to the comparison of the model to improve the accuracy of nearly 10%, and with the increase in the number of samples the accuracy of the model will continue to improve.

In this paper, using the association rule algorithm, data mining is carried out on the laws and regulations database, financial and business databases, and 30 financial and business association rule sets are collected, and their validity is verified based on the degree of support, confidence, enhancement, leverage, and certainty. On this basis, the anomaly detection model is constructed by combining the definition of anomaly rules and outlier detection method, in which the laws and regulations database, financial and business database are divided into three data sets, which are data set Q, data set S, and data set T. The feasibility of anomaly detection is illustrated based on the anomaly detection time overhead. Subsequently, the combined EA-LDA algorithm is used to construct a knowledge graph for financial statement auditing, which aims to reveal the business logic relationships between financial auditing entities, helps to improve the efficiency of auditors' retrieval and correlation comparison, and is of great significance in promoting the digitization and intelligence of financial statement auditing.

2 Linkage mining for finance and business

In the face of audit projects in different industry sectors, auditors often have to learn from scratch, and it is difficult to control the core and important business of the audited unit, which makes the audit work inefficient and the audit risk difficult to control. The association mining in data mining technology can reasonably associate business processes and financial accounting processes together, discovering the close association and role between financial information and business information, which can reveal the degree of association inherent in financial and business indicators, and help to improve the effectiveness of financial auditing based on business indicators.

2.1 Relevance mining dataset

2.1.1 Database of laws and regulations

In the collection of laws and regulations, it is necessary to crawl data from the Internet, and the database of laws and regulations is shown in Figure 1. Based on the perspective of the audited unit, the use of python program on the Internet to crawl a series of accounting laws and regulations such as accounting laws, accounting systems, financial systems, accounting standards, accounting regulations, a total of more than 500 laws and regulations, these laws and

regulations will become the main basis for the construction of the correlation of the financial and operational indicators between the corresponding relationship.

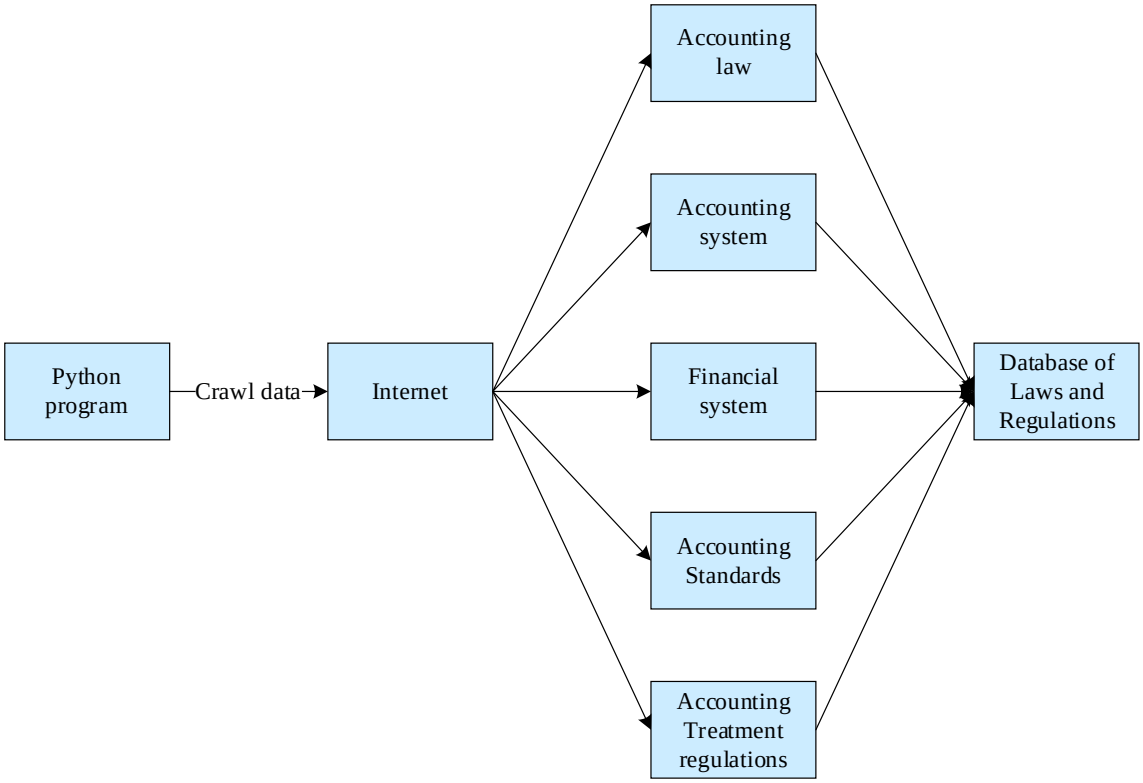


Figure 1: Database of Laws and Regulations

2.1.2 Financial and operational databases

In addition to the financial and business correspondences embedded in laws and regulations, there may also be relevant correspondences between financial and business indicators in various media reports on the Internet about the audited entity, hot news, and current events. Auditors in the process of understanding the audited unit and its environment, often through the external environment to determine whether the audited unit's financial data and financial data changes are reasonable. The financial and business database, as shown in Figure 2, is established through the establishment of a database of news reports and information on the external environment such as the market and competition status of the industry in which the audited unit operates, relevant production technology, energy supply and cost, and statistical data.

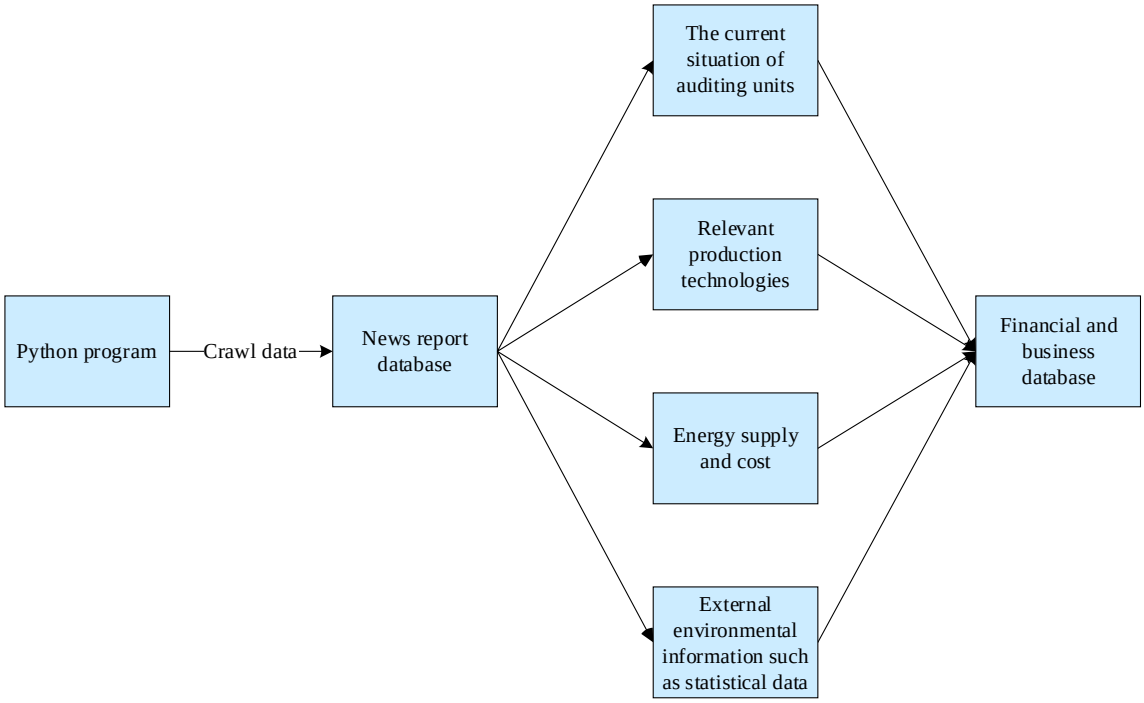


Figure 2: Financial and business database

2.2 Evaluation metrics for correlation mining results

In the training sample stage, as relationship extraction becomes an automated computer program, however, it is not the case that if a financial information and a business information appear in the same text, it means that the two are related, which only means that the two may be related. Without human labeling, we need to establish a set of criteria to determine the strength of the relationship between financial and business. The strength of the relationship between a financial information and a business information can be judged by five indicators: support, confidence, enhancement, leverage and certainty.

2.2.1 Level of support

The degree of support indicates the frequency of occurrence of the former item and the latter item in a dataset. Assuming that X is a certain financial information and Y is a certain business information, the support degree indicates the proportion of the amount of text in which both X and Y appear to the total amount of text. The formula is as follows:

$$Support(X,Y) = \frac{P(X,Y)}{P(All)}, \text{ scope: } [0,1] \tag{1}$$

where $P(X,Y)$ denotes the number of texts in which X and Y appear simultaneously in the same text.

2.2.2 Confidence level

The confidence level indicates how often the latter item occurs in the same dataset given the occurrence of the antecedent of the antecedent. Again assuming that X is a certain financial information and Y is a certain business information, then the confidence level of X over Y indicates the proportion of the amount of text in which both X and Y appear to the amount of

text in which all X appears. The formula is as follows:

$$Confidence(X \rightarrow Y) = P(Y / X) = \frac{P(X,Y)}{P(X)},\ \text{scope: } [0,1] \tag{2}$$

As can be seen from the formula, since the confidence level is calculated with the inclusion of the prior, the larger the value, the higher the strength of the association between X and Y is indicated.

2.2.3 Degree of elevation

Enhancement is the ratio of "the frequency of occurrence of the latter item in the same dataset, provided that the precondition, the former item, occurs" to "the frequency of occurrence of the latter item". Again assuming that X is a particular piece of financial information and Y is a particular piece of business information, the degree of enhancement of X to Y reflects how much the occurrence of X changes the frequency of Y's occurrence. The formula is as follows:

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{P(Y)},\ \text{scope: } [0,\infty] \tag{3}$$

If $Lift(X \rightarrow Y) = 1$, it means that there is no relationship between X and Y. If $Lift(X \rightarrow Y) < 1$, it means that X and Y are exclusive of each other. If $Lift(X \rightarrow Y) > 1$, it means that there is an association between X and Y. And the larger the value of $Lift(X \rightarrow Y)$, the stronger the association between X and Y.

2.2.4 Leverage value

The value of leverage represents the difference between the calculated observed frequency of co-occurrence of the preceding and following items and the frequency of independent occurrence of the preceding and following items. Again assuming that X is a particular piece of financial information and Y is a particular piece of business information, the leverage between X and Y reflects the degree of correlation between X and Y. The leverage between X and Y is calculated by using the following formula. The formula is as follows:

$$Levarage(X,Y) = P(X,Y) - P(X) * P(Y),\ \text{scope: } [-1,1] \tag{4}$$

If $Levarage(X \rightarrow Y) < 0$, then X and Y are negatively correlated. If $Levarage(X \rightarrow Y) = 0$, then X and Y are independent of each other. If $Levarage(X \rightarrow Y) > 0$, then X and Y are positively correlated, and the closer the value of $Levarage(X \rightarrow Y)$ is to 1, the closer the relationship is.

2.2.5 Conviction

Confidence denotes the probability that the former term occurs and the latter term does not, again assuming that X is a certain financial information and Y is a certain business information, then the X to Y confidence denotes the probability that X occurs and Y does not. The formula is as follows:

$$Conviction(X \rightarrow Y) = \frac{1 - support(Y)}{1 - Confidence(X \rightarrow Y)} = \frac{P(X)P(!Y)}{P(X,(Y)},\ \text{scope: } [0,\infty] \tag{5}$$

where $P(!X)$ denotes the number of texts in which X does not occur, the denominator becomes 0 at confidence level $Confidence(X \to Y) = 1$, when $Conviction(X \to Y)$ is defined as "inf". If $Conviction(X \to Y) = 1$, then X and Y are independent of each other. If $Conviction(X \to Y) < 1$, it means that X and Y are exclusive. If $Conviction(X \to Y) > 1$, it means that X and Y are related, and the larger the value of $Conviction(X \to Y)$, the stronger the strength of the association between X and Y is.

3 Analysis of mining results for financial and business linkages

3.1 Association mining results

When applying data mining techniques for rule extraction, there are several analysis methods to choose from, such as logistic regression, support vector machine, decision tree, and association rule mining. According to the understanding of several mainstream machine learning models, association rules are preferred for data mining and rule extraction because they do not require parameter assumptions and have the advantages of strong interpretability and easy regularization, which facilitates the subsequent extraction of entity relationships. On the basis of legal and regulatory databases, financial and business databases, the association mining in data mining technology is utilized to conduct text word frequency mining and obtain the key concepts of financial statement auditing, and the results of text word frequency statistics are shown in Table 1. Based on the data size in the table, it can be seen that X1~X10 are current assets, inventory, intangible assets, liabilities, debt payable, owner's equity, revenue from main business, cost, financial expenses, operating profit in financial information X, respectively, and the corresponding number of text word frequency is 54, 56, 36, 49, 42, 52, 52, 44, 59, 58, whereas Y1~Y3 are the business information Y of financing, operation, and investment, their text word frequency quantities are 171, 173, and 158, which provide data support for the following support, confidence, enhancement, leverage, and certainty calculations. Overall, relying on manual operations to analyze audit data and discover audit suspicions is difficult to meet the complex audit needs, applying the above research on association rule technology to the field of financial statements in an attempt to enrich the audit knowledge base with the mined rules and provide a reasonable theoretical basis for audit data quality detection.

Table 1: The statistical results of text word frequency

X	N	Y	N
X1	54	Y1	171
X2	56	Y2	173
X3	36	Y3	158
X4	49		
X5	42		
X6	52		
X7	52		
X8	44		
X9	59		
X10	58		

3.2 Evaluation of indicator results

It is known that there are 10 items of financial information X, while there are 3 items of business information Y. According to the permutations and combinations, we can get 30 items of financial and business correlation rules, and under the role of the above formula, we can find out the degree of support, confidence, enhancement, leverage value, and certainty, and the values of the evaluation indexes are as shown in Table 2. After calculation, the value domains of support, confidence, enhancement, leverage, and certainty are 0.0113~0.0203, 0.0244~0.0437, 1.586~2.964, 0.04~0.987, and 2.071~4.826 respectively, which show the strength of the association between the financial information and the business in full, and to a certain extent validate the method of this paper to obtain the financial and business association rules obtained by the method of this paper. This can explain the financial and business logic relationship, but also insight into the data interaction relationship of the financial statement audit, thus helping the auditors to grasp the actual situation of the enterprise financial data, while confirming the matching of the data of the form and the data flow of each asset, to identify the financial statement data risks and vulnerabilities to provide data support, and then enhance the efficiency of the audit of the enterprise financial statements, so as to better adapt to current The requirements of enterprise financial auditing.

Table 2: Evaluation index value

N	Support level	Confidence level	Enhancement degree	Leverage value	Certainty
1	0.0183	0.0396	1.333	0.536	3.326
2	0.0190	0.0410	1.347	0.221	4.41
3	0.0122	0.0264	2.692	0.146	3.141
4	0.0166	0.0359	1.3	0.148	3.053
5	0.0142	0.0308	1.594	0.251	2.468
6	0.0176	0.0381	2.236	0.124	4.757
7	0.0176	0.0381	2.151	0.55	3.799
8	0.0149	0.0322	2.387	0.101	2.865
9	0.0200	0.0432	2.048	0.916	3.797
10	0.0197	0.0425	1.058	0.883	4.807
11	0.0185	0.0400	2.573	0.472	3.596
12	0.0192	0.0415	1.409	0.377	4.826
13	0.0124	0.0267	1.138	0.971	2.623
14	0.0168	0.0363	1.406	0.987	3.049
15	0.0144	0.0311	1.438	0.752	2.399
16	0.0178	0.0385	1.055	0.313	4.648
17	0.0178	0.0385	1.621	0.04	3.865
18	0.0151	0.0326	1.084	0.369	3.23
19	0.0203	0.0437	2.964	0.737	4.489
20	0.0195	0.0430	2.241	0.904	3.027
21	0.0169	0.0366	1.59	0.524	3.788
22	0.0176	0.0379	1.879	0.126	3.657
23	0.0113	0.0244	1.586	0.231	2.921
24	0.0154	0.0332	1.162	0.957	3.926
25	0.0132	0.0284	1.626	0.508	4.502
26	0.0163	0.0352	1.9	0.547	2.764
27	0.0163	0.0352	2.337	0.462	2.559
28	0.0138	0.0298	1.273	0.387	4.772
29	0.0185	0.0399	1.377	0.088	2.071
30	0.0182	0.0393	2.112	0.985	4.224

4 Anomaly Detection Model

In this paper, for the actual financial data auditing work facing high cost, rapid changes and experience transmission difficulties, the use of association rule mining and outlier detection methods, the above association mining data for anomaly detection model construction, to realize the financial data anomaly identification rules of automatic mining, the generated rules itself with explanatory, easy for business personnel to understand, judgment and audit.

4.1 Exception Rule Definitions

The main goal of anomaly rule mining is to learn the general laws of general behaviors from a large amount of historical data, and if a specific behavior contradicts the identified laws, it is considered as anomalous behavior. The anomaly identification rule adopts the representation of generative rule, i.e. R=X→Y, where X is the antecedent of the rule and Y is the consequent of the rule. Using this method is closer to the human way of thinking, which can be understood as Y because of X. It is suitable for expressing causality, and the rules can be directly transformed into textual descriptions with explanatory properties.

4.2 Outlier detection

Outliers are data points that deviate significantly from other observations, and outlier detection is the process of extracting distinctive data objects through statistical or modeling methods. Outliers are not always abnormal data points, but can also be caused by errors or data variability. Outlier detection can be realized based on model, based on clustering and based on statistical methods, the detection rules can be interpreted for reasons of consideration, this paper chose the association rule method, and the Z-score as a metric for outlier detection, the Z-score is a parameter anomaly detection method in one-dimensional or low-dimensional feature space. The technique assumes that the data is Gaussian distributed and the outliers are data points in the tails of the distribution and therefore far from the mean of the data. The distance depends on the interset value Z_i of the normalized data points Z_{thr} calculated using the formula, the Z_i score is expressed as:

$$Z_i = \frac{x_i - \mu}{\delta} \tag{6}$$

where x_i is the data point to be detected, μ is the mean of all points x_i, and δ is the standard deviation of all points x_i.

4.3 Specific processes

The raw data is processed by difference processing for date type fields and semantic parsing for text type fields to obtain labels, and finally all the raw data information is transformed into numeric and enumerated fields to obtain the processed data point set $D = \{x_1, x_2, \cdots, x_n\}$ where $x_i = \{fn_1, fn_2, \cdots, fn_j, fe_1, fe_2, \cdots, fe_k\}$, fn, and fe represent numeric field values and enumerated field values, respectively.

In order to improve the computational efficiency of the algorithm, the antecedent of the rule uses the set of all enumerated labels obtained from the frequent item set computation, and the frequent item set is mined using the association rule algorithm.

The backend of the rule targets different field types: enumerated fields use association rule

learning for rule generation and numeric fields use discrete point detection for rule generation. The idea of the algorithm is to iterate over each frequent itemset to generate different non-empty subsets as rule antecedents. On the one hand, the difference set is used as the rule posterior to compute the support, confidence and lift metrics. On the one hand, for each numeric field, the corresponding value of the z-score threshold is computed to generate the rule posterior. Finally, an explanatory description is generated for each rule and rules with the same type and the same antecedent are merged.

5 Anomaly Detection Model Testing

5.1 Test experiment setup

The region of anomalous association determination is shown in Fig. 3.In this experiment, the minimum support is set to 0.0113, i.e., the set of items with support lower than 0.0113 is considered to be an infrequent item set. In addition the maximum confidence level is set at 0.04 level so as to find out as many anomalous association problems as possible without missing any outliers. Overall the parameter setting of the anomaly detection model requires some experience with the data and depends on the size of the number of samples in the experiment. Combined with the definition given before, in order to simplify the means of observation, in the experiment, if the thresholds of the infrequent itemsets to be mined and the anomalous association rules are set to the same value, using the value of the degree of support as the horizontal coordinate, and the value of the degree of confidence as the vertical coordinate, to establish a planar right-angled coordinate system, we can get the following region of the judgment of the anomalous association. The red region is the infrequent item set judgment region, and the blue region is the abnormal association rule judgment region. By observing the scatter distribution composed of support and confidence, it helps to study the law of abnormal association in the data set. In the experiment, the data in the legal and regulatory databases, and the financial and business databases are divided into the full set of data “containing suspected abnormal data” with a total of 39,114 samples (hereinafter abbreviated as set Q). The “does not contain suspected anomalous data” dataset, with a total of 330,251 items (hereinafter abbreviated as set S). The “contains only suspected anomalous data” dataset, with 63,377 entries (hereafter referred to as set T). The “Does not contain suspected anomalous data” and “Contains only suspected anomalous data” are subsets of the “Contains suspected anomalous data” data set. That is, $Q \bigcup ST$ and $Q \bigcap S \neq \phi$. In addition, suspected abnormal data refers to the last status information of the financial statements of an enterprise that has been modified or canceled before the modification is retained, but since there is no further information to indicate that it must be an abnormal record, so it is called “suspected abnormal data” here.

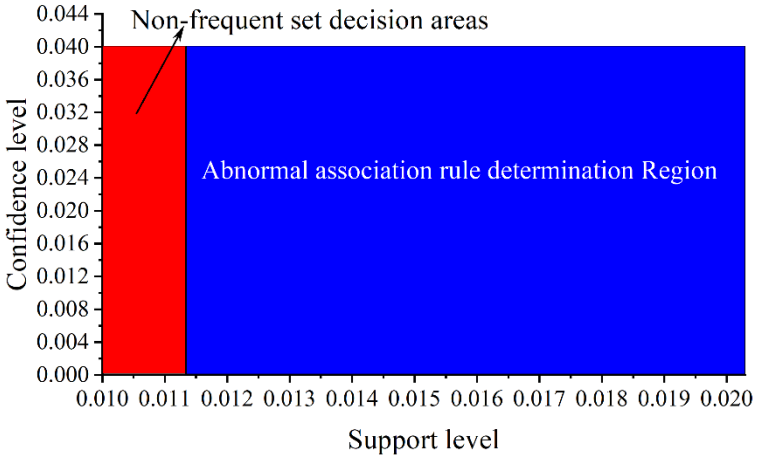


Figure 3: Abnormal association determination area

5.2 Data analysis

The experiment compares the distributions of the scatter points of the three datasets to roughly derive the data patterns dedicated to financial statement auditing, aiming to realize the abnormal data detection, the support confidence scatters containing the suspected data are shown in Fig. 4, the support confidence scatters without the suspected data are shown in Fig. 5, and the support confidence scatters containing only the suspected data are shown in Fig. 6. Comprehensively comparing the three parts of the experimental dataset in Fig. 4, Fig. 5, and Fig. 6, some interesting features can be found.

Feature 1, there are scatters with confidence equal to 0.04 in all three figures. By definition, the existence of an association rule with a confidence level equal to 0.04 indicates that when a set of items that meets some specific conditions occurs, a set of items containing specific values must be introduced. In fact, this characteristic reflects the fact that the fields within the former and latter itemsets are highly likely to be predefined values and there is no possibility of human modification.

Characteristic two, even though the number of records in the set S is quantitatively different from the set Q and the set T, if one observes the interval of support equal to 0 to 0.018, one can find that the scatters of the three sets in this interval are denser. This indicates that even if the scatter falls in the infrequent item set determination region, it does not mean that the item set where the scatter is located is necessarily an infrequent item set, because the fields involved in the modification or cancellation on the financial statement are not necessarily the fields examined in the experiment.

Feature three, similar to feature two, even though the number of records in set Q, set S, and set T are not identical, there is a clear clustering of the distribution of the scatters at the confidence level, with support and confidence at 0.018 or 0.04.

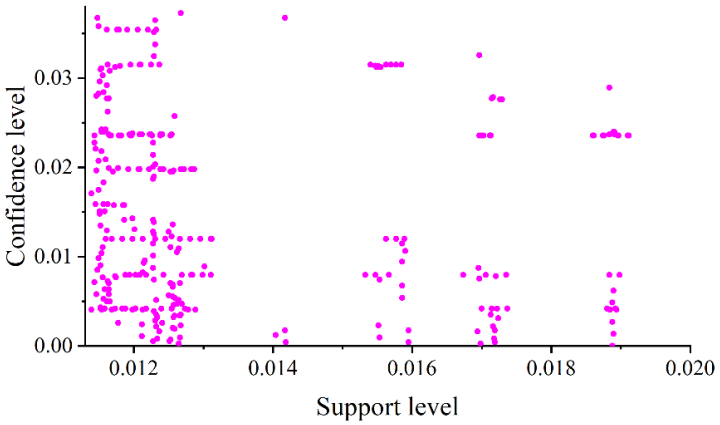


Figure 4: Support and confidence scatter points containing suspected data

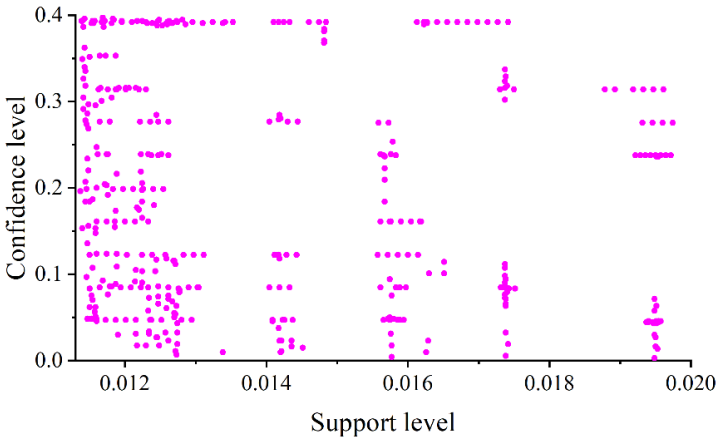


Figure 5: Support and confidence scatter points that only contain suspected data

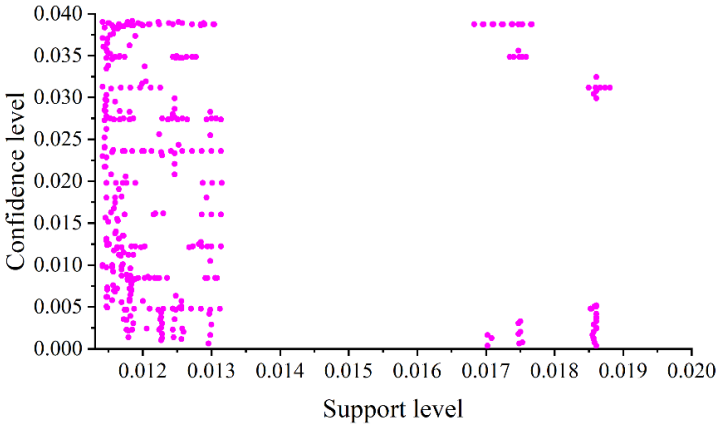
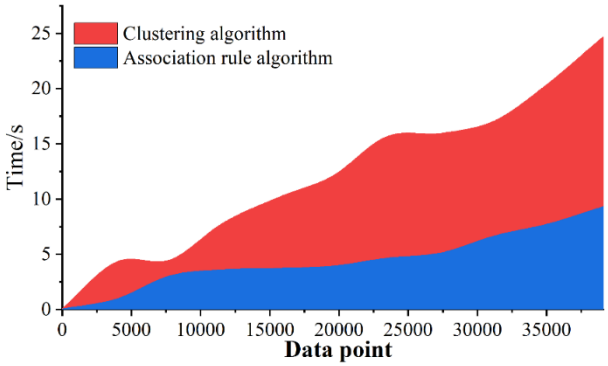


Figure 6: Support and confidence scatter points that only contain suspected data

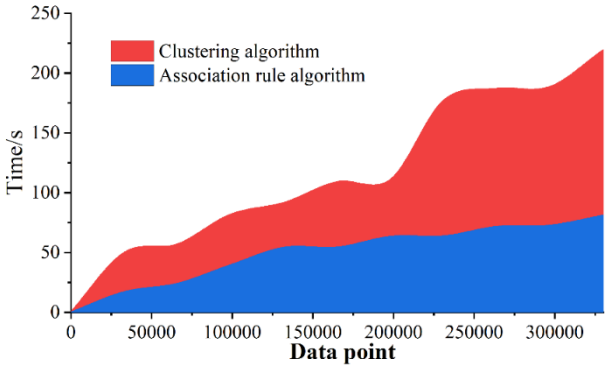
Combined with the above characterization, the infrequent itemsets and normal result sets of abnormal association rules are determined based on the abnormal association determination region, which improves the efficiency and quality of detecting anomalies, so further processing is needed to eliminate the common parts of the dataset. Through practical experiments, it is found that a more feasible approach is to take a given time interval as a reference benchmark to obtain its itemsets and association rules. Then the data in the next time interval is used as the experimental test dataset to obtain its itemsets and association rules, and then the common parts of the two are eliminated. The addition of this experimental step makes the results of anomalous

association detection more practical.

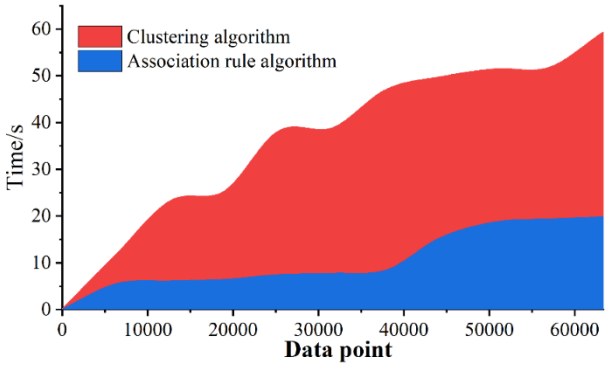
In order to better verify the priority of the association rule algorithm in financial statement anomaly data detection, the clustering algorithm is used as a control algorithm, and the time overhead comparison results under the same conditions are shown in Fig. 7, in which (a) ~ (c) are the dataset Q, dataset S, and dataset T, respectively. The comprehensive performance of the data in the figure shows that, compared with the clustering algorithm, the association rule algorithm in the detection of financial statement anomalies has a The time overhead is smaller, and its time is 9.272s, 80.919s, 19.728s respectively, i.e., the association rule algorithm in this paper is able to detect the abnormal data in financial statements quickly.



(a)Dataset Q



(b)Dataset S



(c)Dataset T

Figure 7: The comparison result of time cost

6 Financial Statement Audit Knowledge Mapping Construction

With the development of informatization and business intelligence, the amount of information that needs to be processed by today's financial departments has proliferated, and it has become a difficult task to mine valuable and relevant information from among the data that needs to be processed by multiple departments, to discover the hidden relationships among the departments, and then to improve the financial statement audit knowledge graph. To address this problem, this paper proposes a method based on inter-entity association analysis and topic analysis, i.e., EA-LDA algorithm to mine the relationships between entities, firstly, the association rules are used to mine the association relationships between entities, and then the LDA topic extraction method is used to analyze the relationships between topics in the data related to the entities, and then get the hidden relationships between the entities of financial statement auditing, and then the newly discovered relationships are added to the original database, and finally complete the financial statement audit knowledge graph design work. The construction of a knowledge map for financial statement auditing can reveal the business logic relationships between entities such as financial accounting entities, organizational entities and audit objects, which helps to improve the efficiency of search and correlation comparison by auditors.

6.1 Association Rule Based Entity Relationship Acquisition

Since the data related to entities contain a large amount of potential association information, this paper applies the association rule algorithm to calculate the frequent keyword set based on the data related to entities in a continuous iterative manner, and forms the entity association rule set based on the n frequent keyword sets obtained, and obtains the hidden relationships between entities through further analysis of the association rule set.

(1) Organize the data associated with different entities to obtain multiple entity-associated datasets $D_1, D_2 \ldots D_n$, preprocess each dataset to obtain multiple corresponding keyword sets $W_1, W \ldots W_n$, and compute the support degree of each word in each keyword set $\sup(i)$. The formula is shown in equation (7):

$$\sup_m(i) = P_m(i) = num_m(i) / num_m(All) \tag{7}$$

where $\sup_m(i)$ denotes the support of the i th keyword in the mth keyword set W_m; $P_m(i)$ denotes the probability of the i th keyword in the keyword set W_m appearing in the current keyword set; $num_m(i)$ denotes the number of occurrences of the keyword set W_m the number of times the i keyword appears in the current keyword set; $num_m(All)$ denotes the number of data records related to the current keyword in the data set D_m.

(2) Based on the support results of each keyword in the obtained keyword set, the keywords with values of support greater than or equal to the threshold α are retained to obtain a 1-item frequent keyword set L_1^m for each entity-associated data set.

(3) For each entity-associated dataset, iteratively use the $(j-1)$ th frequent keyword set, calculate the support of each keyword in the frequent keyword set, and retain the keywords with support greater than or equal to the threshold value α to obtain a new candidate k -item frequent keyword set L_k^m for the m th entity-associated dataset , until no new frequent item set is generated and the algorithm ends.

6.2 Entity Relationship Acquisition Based on LDA Topic Discovery

Based on each topic probability distribution, in order to get the relationship between topic probability distributions, this paper defines a way to calculate the degree of association of the topic probability distributions, for the given two topic probability distributions A and B, first screen to get the common topic words in them, and sum the probability distributions corresponding to the common topic words in the topic probability distributions A and B, respectively, and then calculate them to get the degree of association of the two topic probability distributions. The degree of association of the probability distributions, by comparing the size of the degree of association of the probability distributions of each two topics, we get the entity topic document with a larger degree of association, and retain the common topic words in the entity topic document as the relationship between the entities. The formula for calculating the degree of association of topic probability distributions is shown in equation (8):

$$D = \sum \frac{1}{n^2} \log_{10} \frac{P(x)}{Q(x)} \tag{8}$$

where D is the difference of the topic distribution, $P(x)$, $Q(x)$ denote the probability of the same topic word in the topic distribution P and the topic distribution Q, respectively, and n is the number of the same topic words in the two topic distributions, if the difference degree of the two topic distributions computed, D, is smaller, then the association between the corresponding two-entity topic documents the higher the degree of association between the corresponding two entity topic documents.

Since each entity association data contains multiple topics, and there are potential relationships between multiple topic distributions, this chapter takes each entity association data as an entity topic document, applies the LDA topic extraction algorithm to get the topic distribution of each entity topic document, and analyzes the degree of association between the topic documents to get the hidden relationships between the entities by calculating and analyzing the degree of association between the topic documents. Each word in an article is selected with a certain probability to a certain topic, and a certain word is selected with a certain probability from this topic to generate a document, and the probability of each word appearing in the document is calculated as shown in equation (9). That is:

$$p(word \mid doc) = \sum_{topic} p(word \mid doc) * p(topic \mid doc) \tag{9}$$

where $p(word \mid doc)$ denotes the word frequency, i.e., the probability of occurrence, of each word in each document; $p(word \mid doc)$ denotes the probability of occurrence of each word in each topic; and $p(topic \mid doc)$ denotes the probability of occurrence of each topic in each document. The relative distribution of document-topic, topic-word item, document-word item is obtained by iteratively calculating the topic distribution of each document.

6.3 Potential relationship discovery between entities

In this paper, the algorithm is based on the association data of financial and business entities, and applies the association rule algorithm to mine the entity association rule set of each entity association data, and analyzes the association relationship between the association rules to get the potential relationship between auditing entities. And apply the LDA topic extraction method to obtain the topic probability distribution of each part of entity association data, and further analyze the correlation relationship between each topic probability distribution to obtain the

hidden relationship between audit entities. The newly obtained entity relationships are compared with the entity relationship triples in the original knowledge base, and the new entity relationship triples are added to the database, so as to realize the construction of the financial statement audit knowledge graph, and the overall flow of the EA-LDA algorithm is shown in Figure 8.The specific flow of the EA-LDA algorithm is as follows:

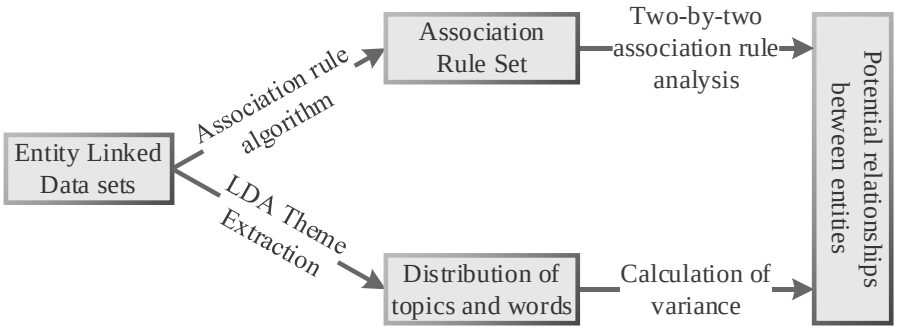


Figure 8: EA-LDA algorithm process

(1) Firstly, we organize the laws and regulations database and the financial and business database to obtain multiple entity-associated datasets $D_1, D_2 \ldots D_n$, and carry out preprocessing operations such as data segmentation and deactivation of words for each entity-associated dataset.

(2) Iteratively calculate the support degree of each frequent keyword set in each entity-associated dataset, retain the k frequent keyword sets that satisfy the support degree threshold of α, perform multiple combinations of keywords for the k frequent keyword sets to obtain the candidate association rule sets, and retain the association rules with the confidence interval of β to obtain the association rule sets of each entity-associated dataset $C_1, C_2 \ldots C_n$.

(3) Analyze the two sets of association rules and retain the common association rules between the two sets of association rules, and the association relationship between the common association rules is the relationship between the corresponding entities.

(4) For the preprocessed entity association data set, Gibbs sampling formula is applied to iteratively update the topic number corresponding to each word, and the distribution of topics θ_d is obtained by counting the topics corresponding to each word, and the distribution of topics and words β_k is obtained by counting the distribution of each topic word in the data set.

(5) Calculate the difference degree between the topic distributions corresponding to each entity's association data two by two, keep the topic distributions whose difference degree D is less than 0.1, and analyze the association relationship between the topic distributions to get the inter-entity relationship.

(6) The relationships between entities obtained by association rule analysis and topic probability distribution analysis are de-emphasized and added to the database in the form of <entity, relationship, entity> to realize the construction of knowledge graph.

7 Financial Statement Audit Knowledge Mapping Analysis

7.1 Entity Relationship Acquisition Analysis

Also on the laws and regulations database, financial and business database, using the EA-LDA algorithm to obtain the financial statement audit entity relationship, this paper sets the number

of topics in the topic model K = 30, with the above set of associated items corresponding to the entity relationship to obtain the results of the analysis shown in Table 3. Through the EA-LDA algorithm for financial statement auditing entity difference degree and probability calculation, it can be seen that the difference degree of its 30 topics are less than 0.1, the value of the domain of 0.01 to 0.096, can be obtained from the financial statement auditing entity relationship, the corresponding probability of the value of the domain of 0.05 to 0.485, which means that the introduction of the LDA algorithm on the basis of association rule algorithms, which makes the financial statement auditing entity more close to the reality, which ensures the application effectiveness of financial statement auditing knowledge graph.

Table 3: Obtain the analysis results of entity relationships

Topic	Degree of difference	Probability	Topic	Degree of difference	Probability
1	0.084	0.485	16	0.01	0.076
2	0.03	0.285	17	0.059	0.337
3	0.016	0.246	18	0.086	0.121
4	0.083	0.229	19	0.068	0.184
5	0.052	0.313	20	0.085	0.477
6	0.059	0.066	21	0.062	0.212
7	0.084	0.274	22	0.072	0.468
8	0.062	0.005	23	0.024	0.198
9	0.051	0.006	24	0.038	0.156
10	0.034	0.047	25	0.089	0.239
11	0.07	0.162	26	0.052	0.356
12	0.094	0.478	27	0.095	0.416
13	0.022	0.207	28	0.096	0.43
14	0.056	0.236	29	0.062	0.287
15	0.049	0.301	30	0.046	0.316

7.2 Application Analysis of Knowledge Mapping

Under the guidance of EA-LDA algorithm theory, the financial statement audit knowledge map is obtained, and an example of financial statement audit knowledge map is shown in Figure 9. Auditors can directly find the corresponding legal provisions by searching for audit doubts, avoiding the mistakes and errors that may arise from human search, and solving the problems of “inappropriate citation of legal provisions, not cited” and so on. The existence of a large number of non-performing accounts receivable leads to poor capital turnover of the unit, limited funds are occupied, resulting in the loss of state-owned assets, which is a serious violation of the “Financial Rules for Institutions” of the unified accounting, unified management of the relevant provisions. In accordance with the provisions of the performance pay system for institutional staff, each unit shall not introduce or change the payroll policy without authorization, and the issuance of other labor expenditures outside the performance payroll is contrary to the unity and seriousness of the payroll policy. In accordance with the provisions of the annual audit of the financial budget of the institution, in order to ensure that the accounts are consistent with the actual situation, it is necessary to adjust the inventory surplus and deficit in a timely manner, regular or irregular inventory inventory count, if the audited unit does not exist inventory count records, it shows that it does not implement the provisions. Administrative institutions need to follow the principle of comprehensiveness in the preparation of financial budgets, the annual budget to include all budget management income and expenditure, reflecting the requirements of the integrated budget, if there is incomplete budgeting in the

financial audit, indicating that the audited object violates the principle of financial budgeting. Comprehensive several situations can be seen, thanks to the financial statement audit knowledge mapping, so that the audit of qualitative work has been significantly simplified, the auditor to audit the problem of searching for legal and regulatory provisions can be realized quickly locate, financial statement audit efficiency has been significantly improved.

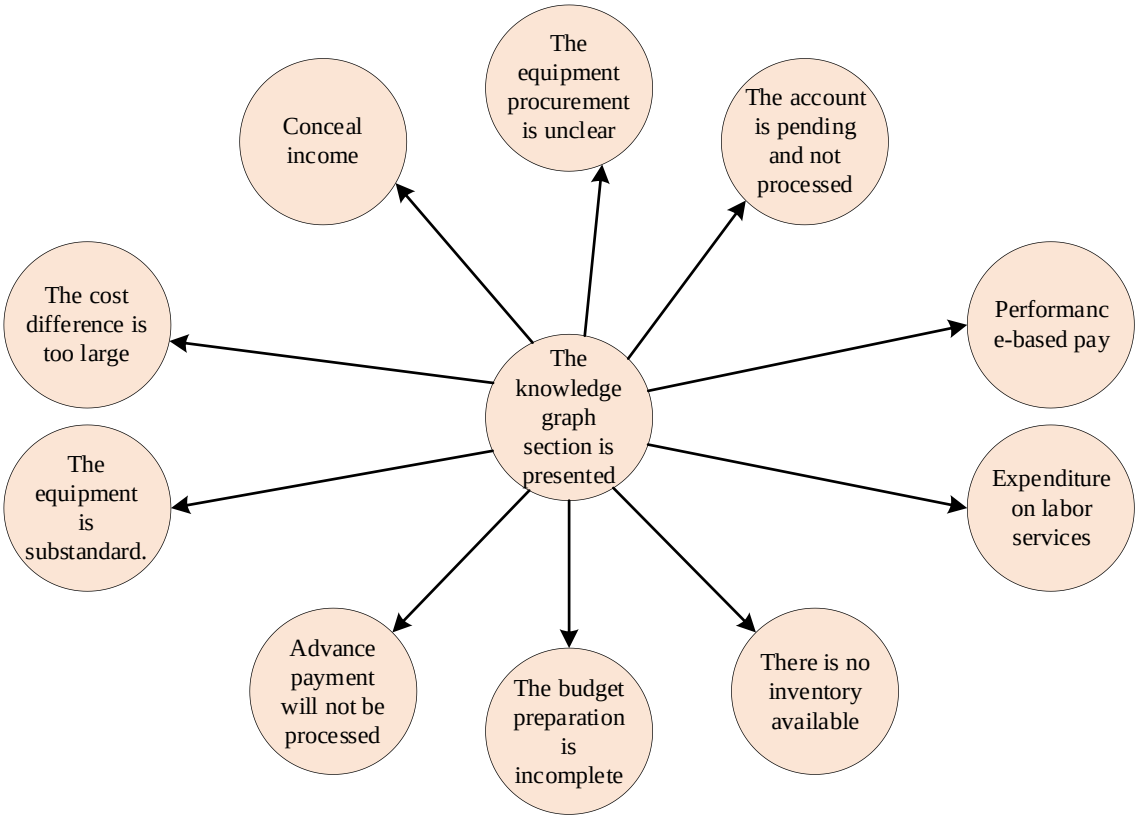


Figure 9: An example of the knowledge graph for financial statement auditing

8 Conclusion

With the development of digital technology, the complexity of the enterprise financial information system has been increasing, and the traditional auditing method has been unable to meet the needs of the current stage of enterprise financial statement auditing. In this regard, this paper utilizes the association rule algorithm and LDA algorithm to construct a financial statement anomaly detection model and knowledge map respectively, and carries out in-depth verification and analysis.

(1) After the association mining of finance and business, we get current assets, inventory, intangible assets, liabilities, debt payable, owner's equity, revenue from main business, cost, financial expenses, operating profit, corresponding to the number of text word frequency is 54, 56, 36, 49, 42, 52, 52, 44, 59, 58, and financing, operation, investment, the number of text word frequency is 171, 173, 158, in addition to calculating the support, confidence, enhancement, leverage, and certainty of each association rule, with specific value domains of 0.0113~0.0203, 0.0244~0.0437, 1.586~2.964, 0.04~0.987, and 2.071~4.826, which demonstrates the strength of the association between the financial information and the business in the auditing work.

(2) Setting the minimum support and maximum confidence level of 0.0113 and 0.04,

respectively, the data from the laws and regulations database, financial and business database can be divided into three data sets, referred to as data set Q, data set S, and data set T, which are 39,114, 3,305,25, and 6,337, and it is found that most of the data points have a support level of 0~0.018, and the corresponding confidence level of 0~0.04, which are normal data, otherwise they are abnormal data. In addition, compared with the clustering algorithm, this paper's algorithm has a smaller time overhead for anomaly detection, which is 9.272s, 80.919s and 19.728s, which comprehensively verifies the anomaly detection model constructed by this paper's algorithm.

(3) Through the EA-LDA algorithm for financial statement auditing entity discrepancy degree and probability calculation, the value domain of discrepancy degree and probability is 0.01~0.096 and 0.05~0.485 respectively, which makes the constructed financial statement auditing knowledge graph more suitable for the actual situation, and also shows some examples of financial statement auditing knowledge graph and summarizes the application value of this graph.

References

[1] Yang, T., Chen, Y., Zhang, S., Qiao, V., Wang, Z., & Zheng, S. (2021). Highlights of the new PRC securities law. Journal of Investment Compliance, 22(1), 20-28.

[2] Akomea-Frimpong, I., & Andoh, C. (2020). Understanding and controlling financial fraud in the drug industry. Journal of Financial Crime, 27(2), 337-354.

[3] Nemati, Z., Mohammadi, A., Bayat, A., & Mirzaei, A. (2025). The impact of financial ratio reduction on supervised methods' ability to detect financial statement fraud. Karafan Journal, 22(Special Issue), e232554.

[4] Naz, I., & Khan, S. N. (2025). Impact of forensic accounting on fraud detection and prevention: a case of firms in Pakistan. Journal of Financial Crime, 32(1), 192-206.

[5] Han, N., Liu, P., Zhong, F., & Zhao, D. (2025). Does public data access improve fiscal transparency?--On a quasi-natural experiment from government data platform access. Socio-Economic Planning Sciences, 98, 102184.

[6] Li, J. (2022). Modern enterprise financial accounting abnormal statistical data in the biopharma industry. Journal of Commercial Biotechnology, 27(1), 98-107.

[7] Kuo, Y. H., & Kusiak, A. (2019). From data to big data in production research: the past and future trends. International Journal of Production Research, 57(15-16), 4828-4853.

[8] Mazloum, H., Mazloum, N., & Saleh, A. M. (2025). FORENSIC AUDIT: ENHANCING CERTIFIED PUBLIC ACCOUNTANT USING ARTIFICIAL INTELLIGENCE TECHNIQUES. BAU Journal-Science and Technology, 6(2), 10.

[9] Werner, M., Wiese, M., & Maas, A. (2021). Embedding process mining into financial statement audits. International Journal of Accounting Information Systems, 41, 100514.

[10] Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. Accounting horizons, 29(2), 423-429.

[11] Musunuru, K. (2025). Big data analytics for financial auditing practices: Identification of conceptual patterns, implications and challenges using text mining. Contaduría y administración, 70(2), 1-36.

[12] Fu, M. (2024). Study on Audit Risk Model Based on Data Mining Algorithm. International Journal of High Speed Electronics and Systems, 2540014.

[13] Saeedi, A. (2021). Audit opinion prediction: A comparison of data mining techniques. Journal of Emerging Technologies in Accounting, 18(2), 125-147.

[14] Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems, 55, 278-288.

[15] Tien, H. T., Tran-Trung, K., & Hoang, V. T. (2024). Blockchain-data mining fusion for financial anomaly detection: A brief review. Procedia Computer Science, 235, 478-483.

[16] Ahmed, M., Choudhury, N., & Uddin, S. (2017, July). Anomaly detection on big data in financial markets. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 998-1001).

[17] Song, F. (2024, September). Financial Data Anomaly Detection Based on Data Mining Technology. In 2024 International Conference on Intelligent Computing and Data Analytics (ICDA) (pp. 103-106). IEEE.

[18] Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.

[19] Bakumenko, A., & Elragal, A. (2022). Detecting anomalies in financial data using machine learning algorithms. Systems, 10(5), 130.

[20] Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagao, T. (2016, December). Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In 2016 15th ieee international conference on machine learning and applications (icmla) (pp. 954-960). IEEE.

[21] Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. Mis Quarterly, 1293-1327.

[22] Zhang, Q. (2022). Financial data anomaly detection method based on decision tree and random forest algorithm. Journal of Mathematics, 2022(1), 9135117.

[23] Huang, D., Mu, D., Yang, L., & Cai, X. (2018). CoDetect: Financial fraud detection with anomaly feature detection. Ieee Access, 6, 19161-19174.

[24] Muntala, P. S. R. P. (2022). Detecting and Preventing Fraud in Oracle Cloud ERP Financials with Machine Learning. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(4), 57-67.

[25] Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. Accounting Horizons, 29(2), 381-396.

[26] Salijeni, G., Samsonova-Taddei, A., & Turley, S. (2021). Understanding how big data technologies reconfigure the nature and organization of financial statement audits: A sociomaterial analysis. European Accounting Review, 30(3), 531-555.

[27] Agustí, M. A., & Orta-Pérez, M. (2023). Big data and artificial intelligence in the fields of accounting and auditing: a bibliometric analysis. Spanish Journal of Finance and Accounting/Revista Española de Financiación y Contabilidad, 52(3), 412-438.

[28] Mohammed Ismail, I. H., & Abdul Hamid, F. Z. (2024). A systematic literature review of the role of big data analysis in financial auditing. Management & Accounting Review (MAR), 23(2), 321-350.

[29] Ashtiani, M. N., & Raahemi, B. (2021). Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review. Ieee Access, 10, 72504-72525.

[30] Gray, G. L., & Debreceny, R. S. (2014). A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. International Journal of Accounting Information Systems, 15(4), 357-380.

[31] Yu, L., & Wang, J. (2025). Research on the design of a data mining-based financial audit model for financial multi-type data processing and audit trail discovery. Systems and Soft Computing, 200295.

[32] Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and GAN models. Expert Systems with Applications, 227, 120144.

[33] Lin, C. C., Chiu, A. A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. Knowledge-Based Systems, 89, 459-470.

[34] Ozdagoglu, G., Ozdagoglu, A., Gumus, Y., & Kurt Gumus, G. (2017). The application of data mining techniques in manipulated financial statement classification: the case of Turkey. Journal of AI and Data Mining, 5(1), 67-77.

[35] Shan, R., Xiao, X., Che, J., Du, J., & Li, Y. (2022). Data mining optimization software and its application in financial audit data analysis. Mobile Information Systems, 2022(1), 6851616.