ARTICLE

Text Analysis and Visualization Research on the *Hetu Dangse* During the Qing Dynasty of China

Zhiyu Wang, Jingyu Wu, Guang Yu, and Zhiping Song

ABSTRACT

In traditional historical research, interpreting historical documents subjectively and manually causes problems such as one-sided understanding, selective analysis, and one-way knowledge connection. In this study, we aim to use machine learning to automatically analyze and explore historical documents from a text analysis and visualization perspective. This technology solves the problem of large-scale historical data analysis that is difficult for humans to read and intuitively understand. In this study, we use the historical documents of the Qing Dynasty Hetu Dangse*, preserved in the Archives of Liaoning Province, as data analysis samples. China's* Hetu Dangse *is the largest Qing Dynasty thematic archive with Manchu and Chinese characters in the world. Through word frequency analysis, correlation analysis, co-word clustering, word2vec model, and SVM (Support Vector Machines) algorithms, we visualize historical documents, reveal the relationships between functions of the government departments in the Shengjing area of the Qing Dynasty, achieve the automatic classification of historical archives, improve the efficient use of historical materials as well as build connections between historical knowledge. Through this, archivists can be guided practically in historical materials' management and compilation.*

INTRODUCTION

China has a long history documented in numerous archives. At present, various local archive departments preserve large numbers of historical documents from different periods. Owing to the development of China's archive digitization, archive management departments at all levels have established digital archive abstracts, catalogs, and subject indexes of historical documents in their collections realizing online retrieval of historical archives. With in-depth research on Chinese history, simple catalog retrieval cannot satisfy researchers' demand for related knowledge in historical archives. Owing to the limitations of the catalog retrieval system, complex catalog data still need to be read manually. However, it is difficult to view the overall picture of the recorded content and impossible to easily distinguish important information in historical materials; this leads to various difficulties, such as the compilation of historical materials for Chinese historical researchers. Thus, in this study, we aim to use text analysis and visualization methods in machine learning to conduct data mining analysis of historical document data. These methods will help us discover the logical relationships of historical records and their purposes, accomplish visual presentations of historical entities and knowledge discovered in historiography, improve knowledge representation and automatic classification of historical data, and provide valuable information for historical archive researchers.

Zhiyu Wang (mikemike248@gmail.com) is PhD Candidate, School of Management, Harbin Institute of Technology and Associate Professor, School of History, Liaoning University. **Jingyu Wu** (734665532@qq.com) is graduate student, School of History, Liaoning University. **Guang Yu** (yug@hit.edu.cn) is Professor, School of Management, Harbin Institute of Technology. **Zhiping Song** (1367123893@qq.com) is graduate student, School of History, Liaoning University. © 2021.

https://doi.org/10.6017/ital.v40i3.13279

During the process of analyzing traditional manual methods for interpreting historical documents, we find the following phenomena: macro description, single angle, selective analysis, and one-way knowledge connection, among others. For example, the *Hetu Dangse* preserved in the Liaoning Archives contains a total of 1,149 volumes and 127,000 pages, making it difficult to fully grasp and understand the overall content of such documents. Relying on manual reading and analysis of entire archives is an unrealistic task. Therefore, this paper proposes using machine learning, natural language processing (NLP), and other technologies to address various problems from traditional manual reading. First, information from historical documents can be revealed from different angles, and this allows the content of the documents to be displayed more comprehensively and scientifically through visual charts. Second, use of objective quantitative analysis methods, such as text analysis and NLP, prevents subjective interpretations of the same content. Third, NLP and other technologies can solve the problem of calculating massive text training data sets while forming systematic knowledge that avoids the omission and one-sided understanding of knowledge in the historical archive.

The application of machine learning in historical data analysis has attracted the attention of researchers in management, history, and computer science. Tao used the Latent Dirichlet Allocation (LDA) topic modeling algorithm to analyze the themes of documents from 1700 to 1800 included in the *German Archives*, providing a more three-dimensional interpretation and explanation of the spiritual world of Germany during the eighteenth century.[1] Chinese scholars Kaixu et al. proposed a method of automatic sentence punctuation based on conditional random fields in ancient Chinese.[2] This method was proved to better solve the problem of automatic punctuation processing compared with the single-layer conditional random field strategy in ancient Chinese as tested on the two corpora of *The Analects* and *Records of the Grand Historian*. Swiss and South African scholars Stauffer, Fischer, and Riesen, and Chinese scholars Wu, Wang, and Ma used the KWS technology and deep reinforcement learning to automatically recognize handwritten pictures in historical documents.[3] Solar and Radovan used the National and University Library of Slovenia's historical pictures and maps as research data. Using GIS technology, they created a novel display method, and interdisciplinary data resource web application to access and research the data.[4] Chinese scholars Dong et al. and Polish scholars Kuna and Kowalski used the WebGIS technology to conduct efficient management and visualization research on historical data of natural disasters in ancient China and Russia.[5] Meanwhile, Latvian scholars Ivanovs and Varfolomeyev and Dutch scholars Schreiber et al. used web technology to develop a web service platform and explored the intelligent environment of cultural heritage service utilization.[6] Korean scholars Kim et al. used machine learning technology to determine the complex relationships between tasks of various classes in a specific historical period through the network of historical figures.[7] Judging from results in related fields, the semantic analysis and visualization of historical archives in an intelligent way are gradually moving from statistical description to knowledge mining. These results provide theoretical feasibility and practical technical experience for this study.

At present, research on historical documents mainly focuses on the retrieval and utilization of historical material databases. Since the words, semantics, grammar, and sentence patterns recorded in historical materials differ from modern texts, using data mining technologies such as machine learning and NLP to intelligently identify historical documents and organize historical data will help us more than traditional methods. This requires the cooperation of artificial intelligence and historical researchers to establish an effective method of historical big data analysis to achieve the transformation from traditional manual historical document analysis to automatic artificial intelligence analysis methods. In this paper, we use machine learning and data visualization as a tool to identify differently the content of the historical documents from traditional literature reading, reveal valuable information in the content of historical documents, and promote more systematic, efficient, and detailed understanding of the literature.

RELATED TECHNOLOGY DEFINITION

To perform text analysis and visualization of the *Hetu Dangse*, we use machine learning technology such as word vector processing, the SVM (Support Vector Machines) model and network analysis.

Word vector is a numerical vector representation of a word's literal and implicit meaning.[8] We segmented the *Hetu Dangse*'s catalog data and used the word2vec model to transform the segmented data's word vector form into a set of 50-dimensional numerical vectors representing a catalog's vector data set. To accurately visualize historical document records' relationship features, we reduced the vector data set's dimensionality. Dimensionality reduction, or dimension reduction, is data's transformation from a high- into a low-dimensional space so that the representation retains some of the original data's meaningful properties, ideally close to its intrinsic dimension.[9] After dimensionality reduction, each catalog data in the vector data set is reduced from 50 to 2 dimensions to facilitate flat display.

We used the SVM model and network analysis technology to analyze the vector data set. The SVM model is a set of supervised learning methods used for classification, regression, and outlier detection.[10] It is given a vector data set as training to represent historical document records as points in space, and learns independently through the kernel algorithm. Using the algorithm, it maps the separated new records to the same space, and predicts their category based on which side of the interval they fall. Network analysis techniques derive from network theory, a computer science system demonstrating social networks' powerful influences. Network analysis technology's characteristics determine that it is suitable for books and historical archives' visualization in the library and information science field, because the visualization technique involves mapping entities' relationships based on the symmetry or asymmetry of their relative proximity.[11] Thus, it helps to discover historical documents' knowledge relevance. For example, citation network analysis can identify emerging relationships in healthcare domain journals.[12]

SAMPLE DATA PREPROCESSING AND CLASSIFICATION

This study uses the catalog of the Qing Dynasty historical archives from the *Hetu Dangse* collected by the Liaoning Archives as the research sample to conduct text analysis and visualization research. China's *Hetu Dangse* is the largest Qing Dynasty thematic archive with Manchu and Chinese characters both in domestic and international. The *Hetu Dangse* is the official document of communication between *Shengjing General Yamen,* the *Wubu of Shengjing* and *Fengtian Office*, and the document communicated between *the Beijing Internal Affairs Office in Charge* and *the Liubu of Beijing* during the Qing Dynasty. The *Hetu Dangse* was published from 2015 to 2018, including *the Hetu Dangse·Kangxi period* (56 volumes), *Hetu Dangse·Yongzheng period* (30 volumes), *Hetu Dangse·Qianlong period* (24 volumes), *Hetu Dangse·Qianlong period* (17 volumes), *Hetu Dangse·Daoguang period* (52 volumes), *Hetu Dangse·Jiaqing period* (58 volumes), *Hetu Dangse·Qianlong period* Official Documents (46 volumes), *Hetu Dangse·Qianlong period* Official Documents (46 volumes), and *Hetu Dangse·general list* (16 volumes).[13] The *Hetu Dangse* is an important document for studying the history of the Qing Dynasty. Owing to the special status of Shengjing in the Qing Dynasty, it has a unique historical significance as the companion capital of Beijing and the hometown of the Qing royal family. This provides original evidence from this time for studying politics, economy, culture, history, and natural ecology in Northeast China.

In this study, we preprocess the catalog data of the *Hetu Dangse* by performing text segmentation, creating a corpus, and labeling data before using text analysis and visualization technology to analyze the catalog data of *Hetu Dangse*. First, we use word frequency analysis and statistics to study the functions of institutions. Second, we use the co-word clustering algorithm to quantify and visualize the institutional relationships. Finally, we use the SVM model to automatically classify and explore the catalog data of the *Hetu Dangse*. Figure 1 illustrates this process.

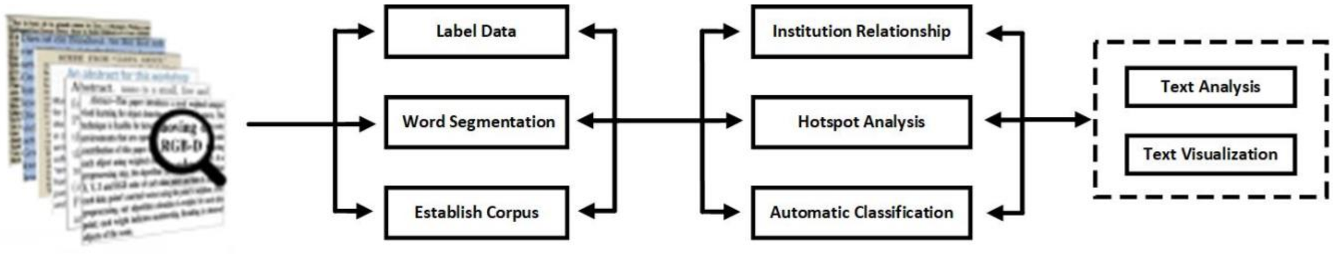


Figure 1. Text analysis flowchart.

Data Preparation and Preprocessing

We collected 95,680 catalog data items in the *Hetu Dangse* of the Liaoning Archives, including 25,148 items from the *Kangxi period*; 1,096 items from the *Yongzheng period*; 23,819 items from the *Qianlong period*; 20,730 items from the *Jiaqing period;* and 15,887 items from the *Daoguang period*. The content of each catalog data includes three parts: title information, time of publication (Chinese lunar calendar), and responsible agency. The proportion for each period was not evenly distributed in the catalog data of the *Hetu Dangse* with the *Kangxi Period* catalog data having the highest proportion (26.2%). Through the catalog data information, we can perform an in-depth analysis of the content of the *Hetu Dangse* from the three perspectives: institutional functions, institutional relationships, and topic classification.

Data Cleaning

As the text recorded in the archives of the *Hetu Dangse* are Manchu and ancient Chinese, using Chinese word segmentation tools (jieba, SnowNLP, THULAC, etc.) based on modern Chinese will cause errors. Therefore, it is necessary to construct a special text corpus for word segmentation. First, we construct a stop vocabulary list to remove words with little impact on semantics in the *Hetu Dangse*, such as *for (为), please (请)* and *of (之)*. Second, we use the word segmentation tools mentioned above for preliminary word segmentation and then perform part-of-speech tagging and word segmentation corrections based on the word segmentation results. The title part of the catalog data of the *Hetu Dangse* mainly contains three dimensions of information: the record title of the catalog, issuing institution, and receiving institution. Accordingly, we set a total of four types of tags in the text corpus: issuing institution, receiving institution, record type, and keywords. The receiving institution and the issuing institution correspond to the institutions at the beginning and the end of the catalog, respectively, such as the words *Shengjing Zhangguan Fang Zuoling*, and *Shengjing Ministry of Justice.* The record type is the front word of the receiving institution, such as *counseling (咨)* and *please (请).* The keywords are words that can represent the overall semantics in the record title of the catalog, such as *arrest (缉拿)* and *advance (进送)*. Table 1 presents the corpus we developed.

Table 1. *Hetu Dangse* corpus

Num	Word	Property1	Property 2
1	盛京掌关防佐领	Organization	Noun
2	为	Stop_words	Preposition
3	缉拿	Keywords	Verb
4	逃人	Keywords	Noun
5	舒廷	Name	Noun
6	官事	Stop_words	Noun
7	咨	Keywords	Verb
8	盛京刑部	Organization	Noun
9	正白旗佐领	Organization	Noun
10	兆麟	Name	Noun
11	呈	Stop_words	Preposition
12	为	Stop_words	Preposition
13	交纳	Keywords	Verb
14	壮丁	Keywords	Noun
15	银两事	Keywords	Noun
┆	┆	┆	┆
61047	收讫事	Keywords	Noun
61048	盛京佐领	Organization	Noun

Label Data

To improve the utilization efficiency of the *Hetu Dangse* and show the document content information from multiple angles, we use a supervised machine learning method to automatically classify the catalog data of the *Hetu Dangse*. Therefore, the original catalog data set must be labeled. We determine the classification and label of the *Hetu Dangse* catalog according to the Chinese Archives Classification Law, Chapter 12. Table 2 presents the 11 categories of the catalog. With this, we complete the *Hetu Dangse* catalog sampling classification and labeling laying the foundation for automatic catalog classification.

The *Hetu Dangse* has a total of 95,680 catalog records involving five periods: Kangxi, Yongzheng, Qianlong, Jiaqing, and Daoguang. We randomly select 500 records from each period and manually label these 2,500 records as the sample data set. The data classification after manual labeling is shown in figure 2. The overall distribution is relatively even, making it suitable for machine learning processing.

Table 2. Data labels

Num	Category
1	Type of Official Documents （政务种类）
2	Palace, Royal Family and Eight Banners Affairs（宫廷、皇族及八旗事务）
3	Bureaucracy, Officials（职官、吏役）
4	Military（军事）
5	Politics and Law（政法）
6	Sino-foreign Relations（中外关系）
7	Culture, Education, Health and Scientific Cultural study（文化、教育、卫生及科学文化研究）
8	Finance（财政）
9	Agriculture, Water Conservancy, Animal Husbandry（农业、水利、畜牧业）
10	Building（建筑）
11	Transportation, Post and Telecommunication（交通、邮电）

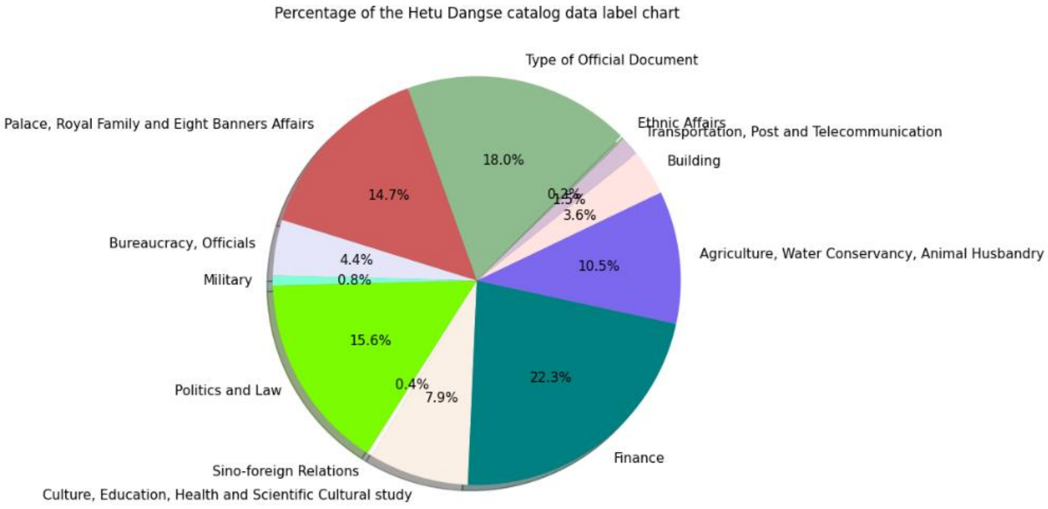


Figure 2. Percentage of the *Hetu Dangse* catalog data label chart.

RESULTS

In this study, we used the catalog data of the *Hetu Dangse* as a sample to analyze and reveal the *Hetu Dangse* catalog data from three perspectives: institutional function, institutional relationship, and automatic classification. This will improve usage efficiency of the *Hetu Dangse*, thus improving researchers' mastery of relevant information about the document. To achieve the functional requirements of text analysis, we adopted four methods: word vector conversion, word frequency analysis, co-word clustering, and the SVM model.

Word Vector Conversion of Text Catalog Data

The automatic classification of machine-learning technology is based on vector data sets. Thus, the *Hetu Dangse* text catalog data set must be vectorized before automatic classification. Currently, word vector conversion technology mainly includes methods such as one-hot, Word2vec, and GloVe. *Hetu Dangse* records the history of the Qing Dynasty for more than 200 years. There are inevitable relationships among the contents recorded in the documents, indicating that they are not isolated from each other. The word2vec model provides an efficient implementation of CBOW and skip-gram architectures for computing vector representations of words, both of which are simple neural network models with one hidden layer. The word2vec model produces word vectors as outputs from inputting the text corpus. This method generates a vocabulary from the input words and then learns the word vectors via backpropagation and stochastic gradient descent.[14] This makes the word2vec model more suitable for catalog data from *Hetu Dangse*. word2vec includes the CBOW model and the skip-gram model, which can enrich the semantic relevance depending on the context, and it is more suitable for the semantic relevance of historical documents such as the *Hetu Dangse*. Therefore, we adopt the skip-gram model to analyze the catalog data of *Hetu Dangse*. We extracted the features of word vectors in catalog data from the corpus, input them into the word2vec model, imported the Gensim library in Python, trained the vector embeddings, and obtained the htd.model.bin vector file and htd.text.model model file. The correlation between each word in the *Hetu Dangse* catalog can be found by implementing the model. For example, if the word *Bannerman (旗人)* is input into the model, the most relevant words are *Minren (民人*, with 0.84726 relevance*)*, *accused (被控*, with 0.812017*)*, and *robbery (抢劫*, with 0.795359*)*.

To visualize the ethnic relationships recorded in the *Hetu Dangse* catalog, we input the first 300 words of the word vector into the trained word2vec model and performed dimensionality reduction to realize a planar graph. To understand the structure of the data intuitively, we used the t-SNE algorithm to reduce the dimensions of the word vector. The t-SNE is a type of nonlinear dimensionality reduction used to ensure that similar data points in high-dimensional space are as close as possible in low-dimensional space. We set the embedded space dimension parameter of t-SNE to 2 and the initialization parameter as pca. This makes it more globally stable than random initialization. The maximum number of optimization iterations is 5,000. Figure 3 presents the results.

In figure 3, the terms *Sanling*, *Yongling*, *Zhaoling*, *Prime Minister,* and *Fuling* form clusters. In Shengjing, the Qing set up the *Sanling Prime Minister's Office*, and the *Prime Minister's Mausoleum Affairs Minister* was appointed concurrently by General Shengjing. Near *Fujinmen*, the *Sanling Prime Minister's Office* was established. In the 30th year of *Guangxu*, the government office was changed to the *Prime Minister's Office* of Shengjing mausoleum affairs, and the governor of the three provinces concurrently served. Under the *Sanling Prime Minister's office*, the *Sanling office* was set up to undertake the sacrifice and repair affairs of the three tombs (*Xinbin Yongling, Shenyang Fuling,* and *Zhaoling*).[15] Therefore, the clustering in figure 3 verifies the close relationship between the *Sanling Prime Minister's Office* and the tombs.

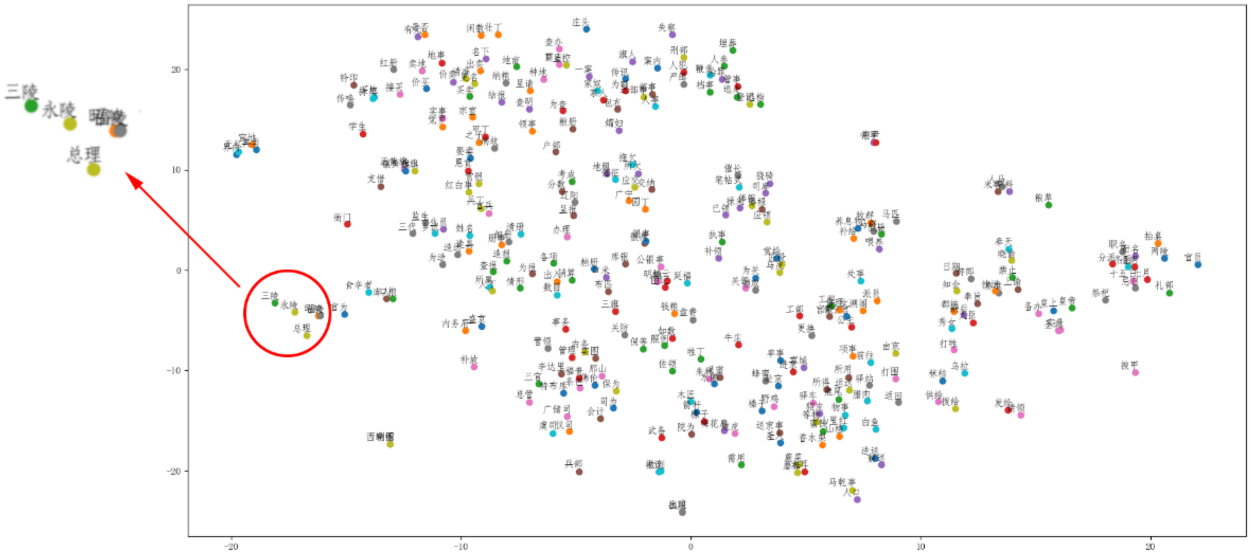


Figure 3. 2D tSNE visualization of word2vec vectors.

Analysis of the Relationship Between the Documents Received and Sent of the Institution

With the statistics of the text data obtained after word segmentation, we can find the quantitative relationship between the documents received and sent by the institution, using the Pearson correlation coefficient to judge whether there is a correlation between the number of documents received and the number of documents sent by the same institution.

$$\rho_{(r,s)} = \frac{cov(R,S)}{\sigma_r \sigma_s} \quad (3.1)$$

We suppose that the *Pearson correlation coefficient* between the number of documents received and the number of documents sent is *ρ(r,s)*, *R= {r1, r2, r3...r11}*. Here, *R* is the variable set of documents received from the institutional sample. Set *S= {s1, s2, s3…s11}* is the variable set of documents sent by the institutional sample. By dividing the covariance of *R* and *S* by the product of their respective standard deviations, we can obtain the value of the correlation coefficient of the documents sent and received by the same institution.

Mining the Relationship Between Institutions' Sending and Receiving Documents Based on Co-word Clustering

To mine the relationship between the institutions' sending and receiving documents, we adopt a co-word clustering algorithm to generate a visualized network map of institutional relationships. The global co-occurrence rate represents the probability of two words appearing together in all the data sets. In large-scale data sets, if two words often appear together in the text, these two words are considered to be strongly related to the semantics.[16] Clustering is a method that places objects into a group by similarity or dissimilarity. Thus, keywords with high correlation to each other tend to be placed in the same cluster. Social network analysis, which evaluates the unique structure of interrelationships among individuals, has been extensively used in social science, psychological science, management science, and scientometrics.[17] We can obtain a sociogram from the institutional function analysis. The main purpose of the sociogram is to provide information about the relationship between institutions' sending and receiving documents. In the sociogram, each member of a network is described by a "vertex" or "node." Vertices represent high-frequency words, and the sizes of the nodes indicate the occurrence frequency. The smaller the size of a node, the lower the occurrence frequency. Lines depict the relationships between two institutions. They exist between two keywords, indicating that they received or sent documents to each other. The thickness is proportional to the correlation between the keywords. The thicker the line between the two keywords, the stronger the connection. Using this rationale, the map visualization and network characteristics (centrality, density, core-periphery structure, strategic diagram, and network chart) were obtained by analyzing Pearson's correlation matrix or other similarity matrices.[18] In this study, we conducted network analysis on a binary matrix to display the relationships between the documents sent and received by the institutions in the Shengjing area during the Qing Dynasty recorded in the *Hetu Dangse*. Further, we extracted the receiving institution and issuing institution from each record of catalog data in the *Hetu Dangse*, and then we composed a new data set with the following data from the receiving institution: issuing institution and title content. We used Python to convert the new data set to EndNote format and import it into VOSviewer1.6.15 to calculate and draw a visual map of the new data set.

Van Eck and Waltman of the Netherlands' Leiden University developed VOSviewer, a metrological analysis software used for constructing and visualizing network graphs.[19] Although the software's development principle is based on documents' co-citation principles, it can be applied to the construction of data network knowledge graphs in various fields. Combined with the co-word clustering algorithm, we can create an entity connection network map for historical documents through VOSviewer software to reflect the recorded content.

Automatic Classification Method of Historical Archives Catalog Based on the SVM Model

We used the SVM model in machine learning for automatic classification. The SVM model has the advantages of strong generalization, low error rate, strong learning ability, and support for small sample data sets, making it suitable for historical archive catalog data samples with small sample characteristics. Therefore, we attempted to classify the catalog data set of *Hetu Dangse* using the SVM model. First, we divided the vectorized labeled data set into a training set and a testing set. The training set accounts for 70% of the data, and the testing set accounts for 30%. To ensure the accuracy of the model prediction, we adopted a random division method to avoid overfitting. Second, we used a linear kernel in the SVM model and grid search to find the best parameter. Various combinations of the penalty coefficient (C) and gamma parameter in the SVM model were tested based on their accuracy ranked from high to low. We then determined the best parameter combination. After the model was established, we validated the predictive performance of the model from multiple perspectives such as precision, recall, and F1 score to ensure the generalization ability and availability of the model.

We set the penalty coefficients to 10, 100, 200, and 300, while the gamma parameters are set to 0.1, 0.25, 0.5, and 0.75. We used the precision evaluation criteria to find the optimal parameter combination of the model and then imported them. The penalty coefficient is set to the X-axis, the gamma parameter set to the Y-axis, and the precision set to the Z-axis. We implemented the model to obtain the visualization that is shown in figure 4. Clearly, the optimal parameter combination is a penalty coefficient of 10 and a gamma parameter of 0.075.

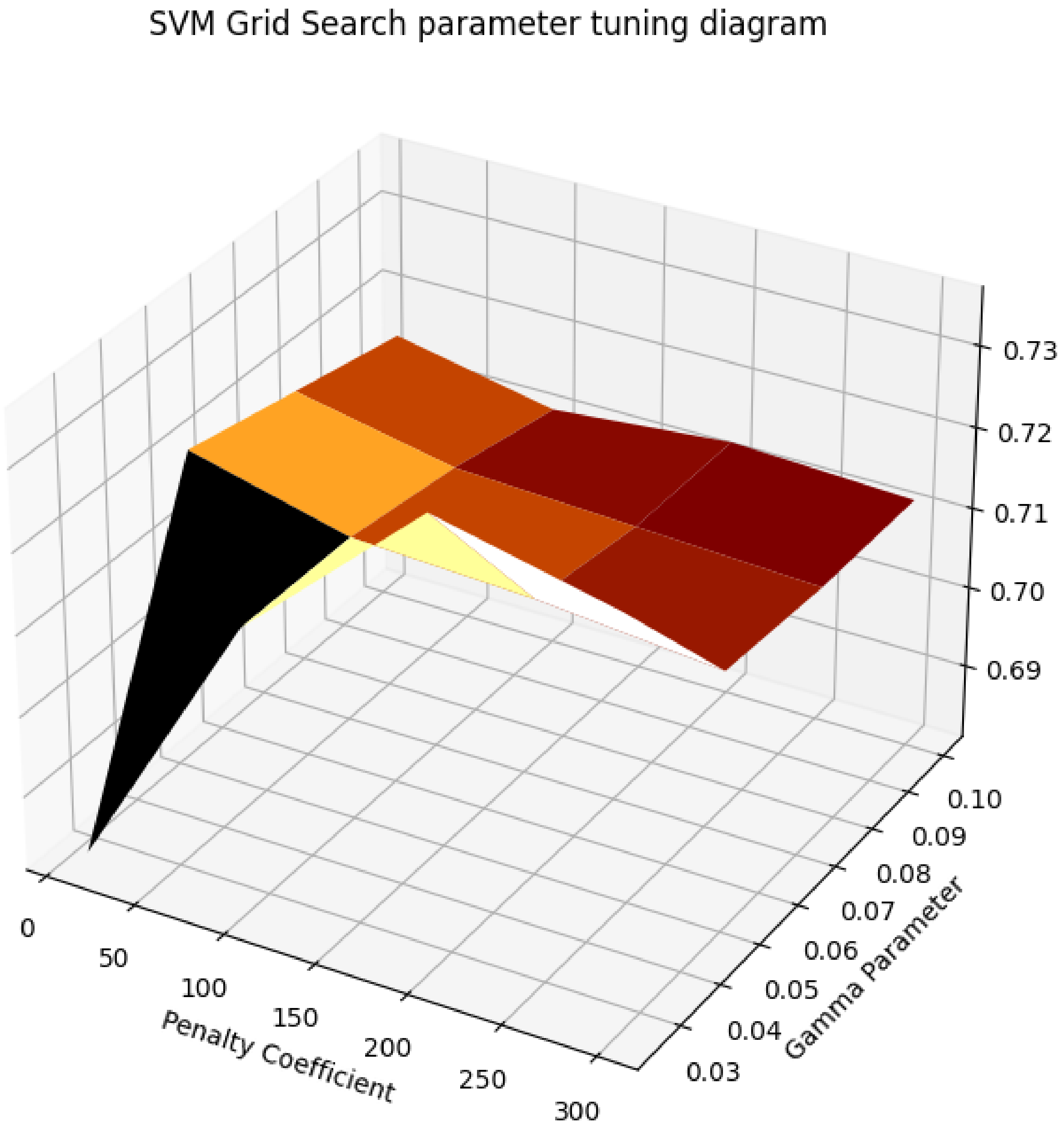


Figure 4. SVM grid search parameter tuning diagram.

DISCUSSION

The history of a nation is the foundation on which it is built. Historical documents are the witnesses and recorders of history. Through the study of historical documents, we can go back to the past, cherish the present, and look forward to the future. An increasing number of scholars have studied these documents in recent years due to their importance. The *Hetu Dangse* records the document communications between institutions in Shengjing (now Shenyang) and Beijing during the Qing Dynasty. It is an important historical document that cannot be ignored when studying the history of Northeast China during the Qing Dynasty. Here, we use the catalog data of the *Hetu Dangse* as the sample data to test the machine learning methods previously mentioned. We explore the results from the perspectives of institutional function, institutional relationship, and automatic classification to determine the feasibility of our methods.

Functions of Institutions

The number of institutions involved in the *Hetu Dangse* is over 150. These functional departments formed the governance system of the Shengjing area during the Qing Dynasty. To gain a deeper understanding of the Qing Dynasty's ruling system in the Shengjing area, the functions of these institutions should be examined. This study analyzes and studies the functions of the institutions in the Shengjing area through the number of documents and the frequency of content of the sending and receiving institutions.

Analysis of the Number of Documents Received and Sent by Institutions

By sorting and statistically analyzing the catalog data of *Hetu Dangse*, we obtained data on the number of documents received and sent by institutions in the Shengjing area recorded in the *Hetu Dangse*. We set the vertical axis as the total number of communicated documents, number of issued documents, and number of received documents. We set the horizontal axis as the names of the institutions and then drew a histogram. This study analyzes the number of institutional archives of the *Hetu Dangse* catalog from three perspectives: total number of sent and received documents, number of received documents, and number of issued documents to find the institutions with the highest research value in the Shengjing area.

In the histogram shown in figure 5(A), the top three institutions in total number of communicated documents are *Shengjing Internal Affairs Office*, *Shengjing Zuoling*, and *Shengjing Ministry of Revenue*. We can also observe that the top 10 institutions have different volumes of their respective documents received and sent by institutions. Therefore, the ranking of the total number of communicated documents is not directly related to the respective rankings of the number of documents received and the number of documents sent. In figure 5(B), we can observe that the top three institutions in number of documents received in the *Hetu Dangse* are *Shengjing Internal Affairs Office, Shengjing Ministry of Revenue,* and *Shengjing General Yamen*. Figure 5(C) shows the top three institutions in number of documents sent in the *Hetu Dangse* are *Shengjing Internal Affairs Office*, *Shengjing Zuoling*, and *Shengjing General Yamen*. The total number of communicated documents, number of documents sent, and number of documents received by the *Shengjing Internal Affairs Office* all rank first; this indicates that the *Shengjing Internal Affairs Office* is the most important department of the ruling system in the Qing Dynasty during the Shengjing area.

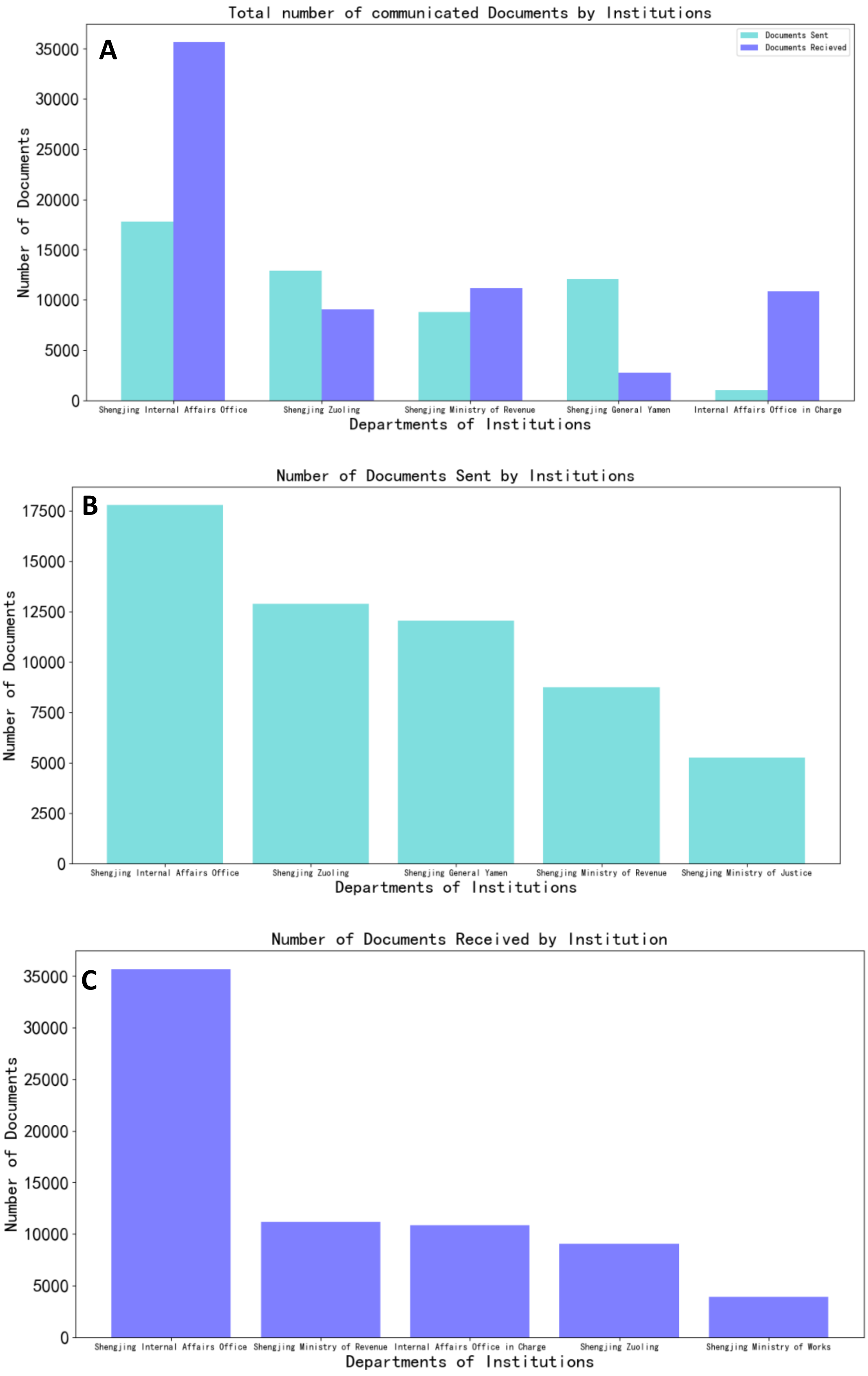


Figure 5. Number of documents received and sent by institutions.

By using the number of documents received and sent by the institutions, we calculated the Pearson correlation coefficient to determine if the number of documents received and sent by the same institution is relevant. As institutional samples, we selected the *Shengjing Internal Affairs Office*, *Shengjing Ministry of Revenue*, (*Beijing*) *Internal Affairs Office in Charge*, *Shengjing Zuoling*, *Shengjing Ministry of Works, Shengjing Ministry of Justice*, *Shengjing General Yamen*, *Shengjing Close Defense Zuoling, Shengjing Ministry of War, Fengtian General Yamen*, and *Shengjing Ministry of Rites*. Through calculation, the result of Pearson correlation coefficient is 0.69 (save two decimal places), so there is a correlation between the number of sent and received documents, as shown in figure 6.

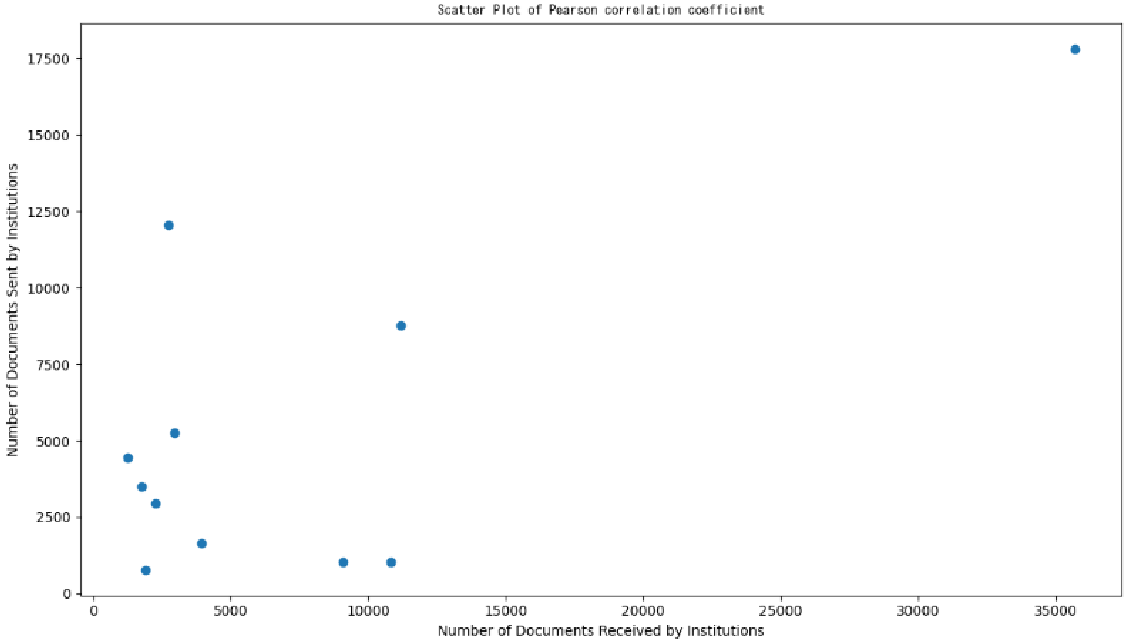


Figure 6. Scatter plot of Pearson correlation coefficient.

The *Hetu Dangse* is a copy of official documents dealing with the royal affairs of the *Shengjing Internal Affairs Office* during the Qing Dynasty. It contains the official documents between the *Shengjing Internal Affairs Office* and the *Beijing Internal Affairs Office in Charge*, the *Liubu*, etc. and the local *Shengjing General Yamen*, *Fengtian Office*, the *Wubu of Shengjing*, and other yamens.[16] Thus, there exist a large stock of documents with the *Shengjing Internal Affairs Office* as the sending and receiving agency. The *Wubu of Shengjing*, *Shengjing General Yamen*, *Shengjing Zuoling,* and other institutions are important hubs for the operation of institutions in Shengjing. They played an important role in maintaining and stabilizing the society of Shengjing. The number of documents is second in importance only to the *Shengjing Internal Affairs Office*.

Analysis of the Frequency of Documents Received and Sent by Institutions

To further explore the functions of institutions with research value, we extracted the contents of the catalogs from the top three institutions in total number of documents sent and received: *Shengjing Internal Affairs Office*, *Shengjing Ministry of Revenue*, and *Shengjing Zuoling*. We then classified the catalogs of the aforementioned institutions according to receipts and postings. Subsequently, we used word segmentation and word frequency statistics to process the two types of catalog information and draw comparison diagrams to explore their specific functions in the *Hetu Dangse.*

As shown in figure 7, we can roughly divide the obtained segmentation words into two categories. One is the name of the communicated official document institutions, such as *the Ministry of Revenue, the Ministry of Justice, and the Ministry of Rites* on the side of the word frequency (see fig. 7[A]). The other is the name of the official document content and the words *Zhuangtou (庄头)*, *Dimu (地亩)*, and *Zhuangding (壮丁)* on the side of the frequency of the words in the documents sent. Through a comparative analysis of the top 10 words received and sent by the same institution, we conclude that the institutions with a close relationship between receiving and sending documents are not the same. For example, the *Ministry of Revenue* of *Shengjing Internal Affairs Office* ranks first in the frequency of documents sent by institutions, while the *ShengJing Zuoling* ranks first for receiving institutions (see fig. 7[B]). The contents of documents sent and received by the same institution are different. Figure 7(C) shows how the affairs sent by *Shengjing Zuoling* to *Ula (乌拉)*, *Forage (粮草)*, and *License (执照)* differ from those represented by the *Zhuangtou (庄头)*, *Accounting (会计)*, and *Close Defense (关防)* in the frequency of documents sent and frequency of receipts, respectively.

Based on previous research on the functions of Shengjing's institutions, the *Shengjing Internal Affairs Office* was set up in the companion capital of Shengjing during the Qing Dynasty to be in charge of Shengjing cemetery, sacrifice, organization of staff transfer, and other matters.[20] This relates to the meaning of words such as *sacrifice (祭祀)* in figure 7(A). The functions of the *Shengjing Ministry of Revenue* were represented in Guangxu's *Great Qing huidian*. The cashiers in charge of taxation in Shengjing, number of annual losses in official villages, and *Banner Land* were carefully recorded. The expenditures were distinguished and the accounting obeyed the regulations according to the *Beijing Ministry of Revenue* at the end of the year.[21] This is related to the meaning of words, such as *Dimu (地亩)*, *land sale (卖地)*, and *money and grain (钱粮)* in figure 7(B). In Fu Yonggong and Guan Jialu's research of *Shengjing Zuoling's* functions, *Shengjing Zuoling* handled the transfer communicated documents; supervised and urged the various departments of *Guangchu, Duyu, Zhangyi, Accounting, Construction,* and *Qingfeng* to undertake matters; managed officials and various people; maintained the Shengjing palace and the warehouse; selected women to send to Beijing Inspect; heard all types of cases; undertook the emperor's general letter; managed the Ula people and tributes; and accepted the emperor or the *Internal Affairs Office in Charge*, among other tasks.[22] This is connected to the meaning of words such as *Ula (乌拉), Close Defense (关防)* and *License (执照)* in figure 7(C).

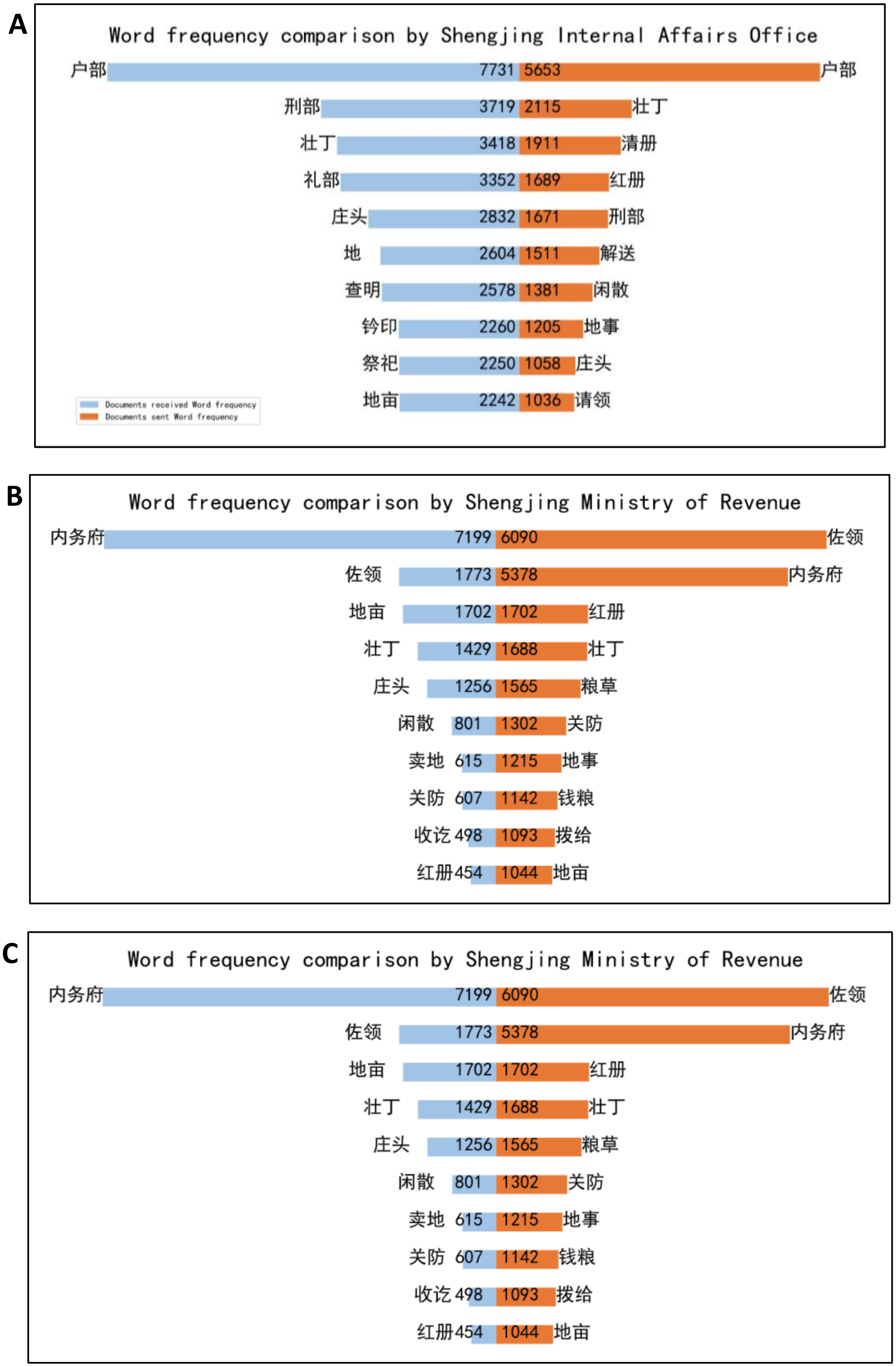


Figure 7. Word frequency comparison of documents received (in blue) and sent (in orange) by institutions.

Institutional Relationship Analysis

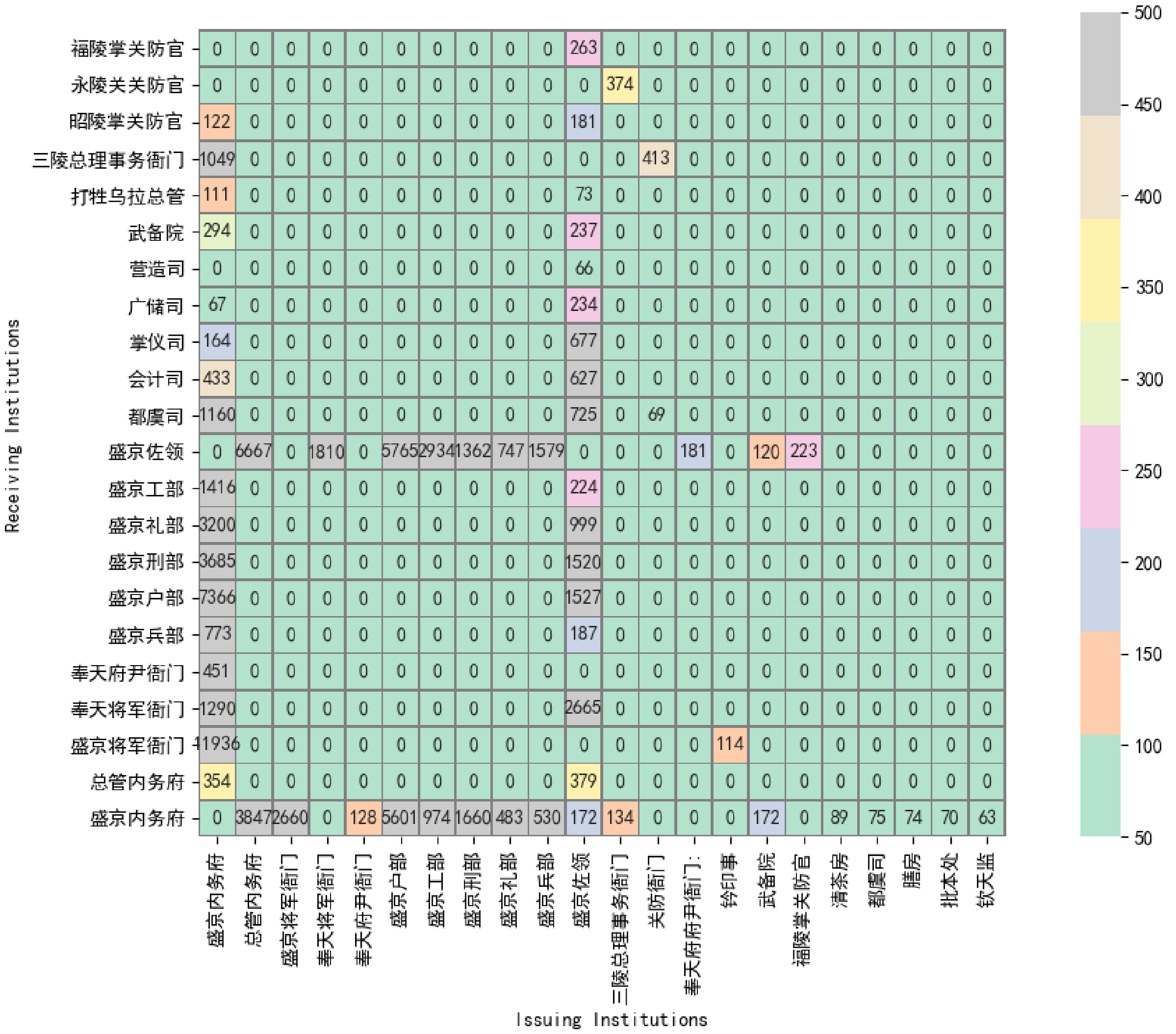
To further study the governance structure of the Shengjing area, we not only need to understand the functions of each institution but also explore the overlap between functions of institutions. The catalog data of the *Hetu Dangse* consist of three parts: receiving institutions, issuing institutions, and record title of the catalog. A document often includes two institutions, the receiving institution and the issuing institution, and it is certain that the content of a document relates closely to the functions between the two institutions. By observing the closeness between the number of institutions through visualizations, we conducted a quantitative analysis of consistent catalog data of the receiving and issuing institutions in the *Hetu Dangse* to provide reliable data for further research in the intersection of institutional functions in Shengjing area.

Results of Institutional Connection Analysis

Using the co-word clustering algorithm, we counted the number of archive catalog data consistent with the receiving and issuing institutions. We set the vertical axis as the issuing institution and the horizontal axis as the receiving institution to obtain figure 8. The numbers inside the boxes represent the quantity of catalog data that are consistent with the issuing institution. To facilitate measurements in the statistical process, records less than or equal to 50 communicated documents between the receiving institution and the issuing institution have been zeroed out.

As shown in figure 8, the institutions having close relations with the documents recorded in the *Hetu Dangse* are concentrated in the issuing institutions *Shengjing Zuoling* and *Shengjing Internal Affairs Office*, and the receiving institutions *Shengjing Internal Affairs Office* and *Shengjing Zuoling*. Among the receiving institutions, the number of documents received by the *Shengjing Internal Affairs Office* from *Shengjing General Yamen* reached as high as 11,936. The top three documents received by *ShengJing Zuoling* were *Fengtian General Yamen* (2,265 pieces), *Shengjing Ministry of Revenue* (1,527 pieces), and *Shengjing Ministry of Justice* (1,520 pieces). It is worth noting that there are less than 50 documents from *Shengjing Zuoling* in the *Shengjing Internal Affairs Office*.

The overlapping functions of the institutions in the Shengjing area enabled individual offices to play bureaucratic games, passing responsibility to other offices, leading to low efficiency in handling affairs. For example, the military and political power in the Shengjing area was jointly controlled by the *Shengjing General Office* and the *Shengjing Ministry of War*. The Shengjing area's tax power was controlled by the *Shengjing Ministry of Revenue* and *Fengtian Office* and their subordinate offices. This phenomenon ran through the entire Qing Dynasty. Research on the cross-functionality of institutions has always been a hot topic in Qing historiography. By analyzing the official documents between the institutional functions, we can further explore the overlap as well as the advantages and disadvantages of the Qing Dynasty Shengjing ruling system to study the history of Shengjing institutions in the Qing Dynasty more thoroughly providing a reference for the design of current institutions.



	盛京内务府	总管内务府	盛京将军衙门	奉天将军衙门	奉天府尹衙门	盛京户部	盛京工部	盛京刑部	盛京礼部	盛京兵部	盛京佐领	三陵总理事务衙门	关防衙门	奉天府府尹衙门:	钤印事	武备院	福陵掌关防官	清茶房	都虞司	膳房	批本处	钦天监
福陵掌关防官	0	0	0	0	0	0	0	0	0	0	263	0	0	0	0	0	0	0	0	0	0	0
永陵关关防官	0	0	0	0	0	0	0	0	0	0	0	374	0	0	0	0	0	0	0	0	0	0
昭陵掌关防官	122	0	0	0	0	0	0	0	0	0	181	0	0	0	0	0	0	0	0	0	0	0
三陵总理事务衙门	1049	0	0	0	0	0	0	0	0	0	0	0	413	0	0	0	0	0	0	0	0	0
打牲乌拉总管	111	0	0	0	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0	0	0	0
武备院	294	0	0	0	0	0	0	0	0	0	237	0	0	0	0	0	0	0	0	0	0	0
营造司	0	0	0	0	0	0	0	0	0	0	66	0	0	0	0	0	0	0	0	0	0	0
广储司	67	0	0	0	0	0	0	0	0	0	234	0	0	0	0	0	0	0	0	0	0	0
掌仪司	164	0	0	0	0	0	0	0	0	0	677	0	0	0	0	0	0	0	0	0	0	0
会计司	433	0	0	0	0	0	0	0	0	0	627	0	0	0	0	0	0	0	0	0	0	0
都虞司	1160	0	0	0	0	0	0	0	0	0	725	0	69	0	0	0	0	0	0	0	0	0
盛京佐领	0	6667	0	1810	0	5765	2934	1362	747	1579	0	0	0	181	0	120	223	0	0	0	0	0
盛京工部	1416	0	0	0	0	0	0	0	0	0	224	0	0	0	0	0	0	0	0	0	0	0
盛京礼部	3200	0	0	0	0	0	0	0	0	0	999	0	0	0	0	0	0	0	0	0	0	0
盛京刑部	3685	0	0	0	0	0	0	0	0	0	1520	0	0	0	0	0	0	0	0	0	0	0
盛京户部	7366	0	0	0	0	0	0	0	0	0	1527	0	0	0	0	0	0	0	0	0	0	0
盛京兵部	773	0	0	0	0	0	0	0	0	0	187	0	0	0	0	0	0	0	0	0	0	0
奉天府尹衙门	451	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
奉天将军衙门	1290	0	0	0	0	0	0	0	0	0	2665	0	0	0	0	0	0	0	0	0	0	0
盛京将军衙门	11936	0	0	0	0	0	0	0	0	0	0	0	0	0	114	0	0	0	0	0	0	0
总管内务府	354	0	0	0	0	0	0	0	0	0	379	0	0	0	0	0	0	0	0	0	0	0
盛京内务府	0	3847	2660	0	128	5601	974	1660	483	530	172	134	0	0	0	172	0	89	75	74	70	63

Figure 8. Relationship of communicated documents by the *Hetu Dangse* Institutions diagram.

Visualization of Institutional Network Map

We used the *Hetu Dangse* catalog as sample data and the co-word clustering algorithm to obtain the close relationship between institutions and the appearance frequency of institutions. We drew a visual network diagram by virtue of VOSviewer1.6.15 to obtain figure 9. In figure 9, institutions are represented by default as a circle with their names. The size of the label and the circle of an institution are determined by the weight of the item. The higher the weight of an item, the larger the label and the circle of the item. For some items, labels may not be displayed to avoid overlapping labels. The color of an institution is determined by the cluster the institutions belong to, and lines between items represent links.

As shown in figure 9, the relationships between the institutions and departments in the *Hetu Dangse* form three core groups: the *Shengjing Internal Affairs Office (in Charge)*, *Shengjing Zuoling*, and *Beijing Internal Affairs Office in Charge*. However, the relationships between the three groups are not similar; the distance between the group *(Beijing) Internal Affairs Office in Charg*e and the two other groups is relatively large. The group at the core of *Shengjing Internal Affairs Office* and the group at the core of *Shengjing Zuoling* are closely connected to each other through *the Wubu of*

Shengjing (*Shengjing Ministry of Revenue, Shengjing Ministry of Rites, Shengjing Ministry of War, Shengjing Ministry of Justice*, and *Shengjing Ministry of Works*). Further, there are two larger individuals: *Fengtian General Yamen* and *Shengjing General Yamen*. *Fengtian General Yamen* and *Shengjing Zuoling* are closely related to each other, and the relationship between *Shengjing General Yamen* and *Shengjing Internal Affairs Office* is relatively close.

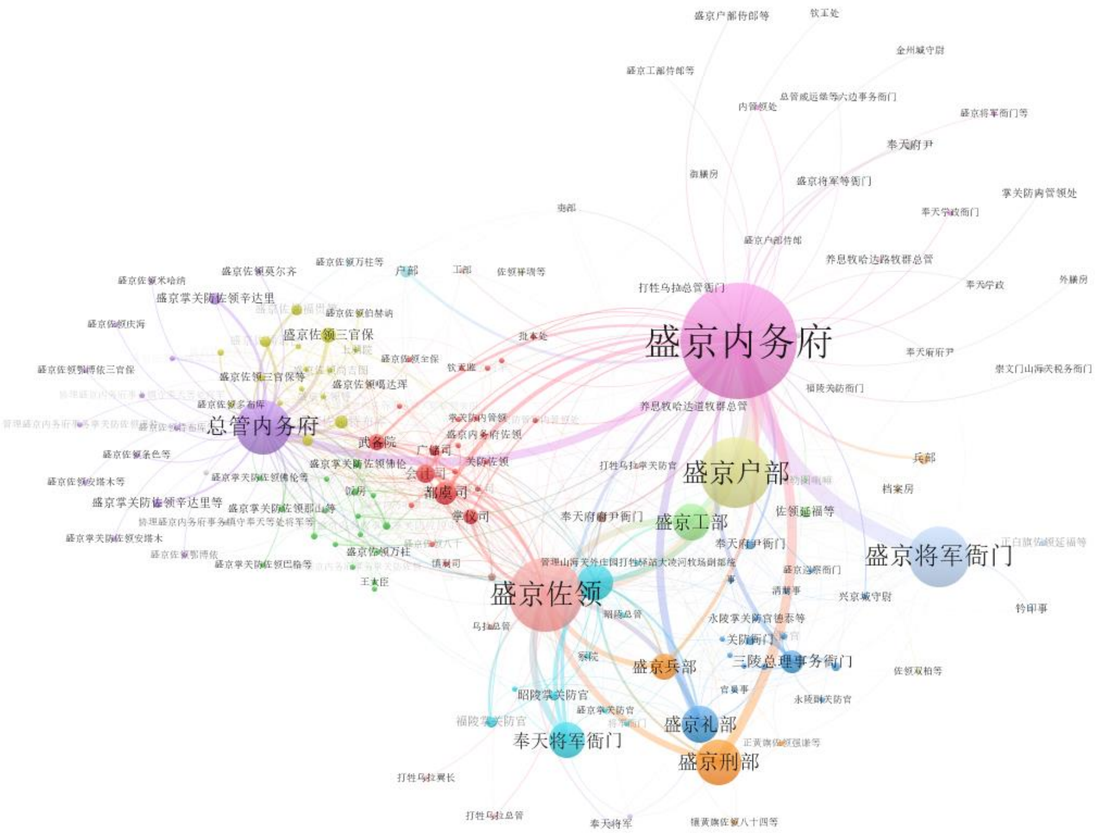


Figure 9. Co-occurrence of Institutions network map.

The city of Shengjing was the companion capital of the Qing Dynasty. The Qing government implemented special governance measures in these areas that differed greatly from those of direct inland provinces.[23] To ensure the stable rule of the Shengjing area, the Qing Dynasty performed the following tasks. First, the Qing Dynasty set up a general garrison as the highest military and political chief in the Shengjing area to be responsible for all military and political affairs within its jurisdiction. Second, they established the *Fengtian Office*, a capital of the same level as the *Shuntian Office*, to rule the *common people* of the Shengjing area. The states and counties, as well as the *Garrison Banner Officer*, which was under the rule of general garrison, were local administrative institutions under the *Fengtian Office*. These institutions implemented the dual management rule of the *Bannerman* and Common people. Third, as the companion capital, the Shengjing area followed the Ming Dynasty companion capital system to set up the *Wubu of Shengjing* to maintain power. In addition, the *Shengjing Internal Affairs Office*, which was in charge of palace affairs, communicated with the *Beijing Internal Affairs Office in Charge*.

Results of Automatic Classification Analysis

Catalogs are important information resources in the field of historical archives. The classification of archival catalogs can not only link relevant information in archives or archive fonds, improve researchers' utilization efficiency, and save time to search for required archives, but it can also be shown to readers in clusters. As the *Hetu Dangse* catalog is a series of historical documents stored for a long period of time, its original classification system does not suit well existing archival management methods. The *Hetu Dangse* has a total of 1,149 volumes and 127,000 pages. Each volume contains a different number of documents and the ink characters on Chinese art paper are in Manchu and Chinese. Reading and categorizing the full text of the *Hetu Dangse* not only requires a lot of manpower, material, and financial resources but also extremely high requirements for the classified staff. They need to possess a good knowledge of Manchu, archival science, document taxonomy, and other related disciplines. Therefore, sorting and organizing the content of the *Hetu Dangse* is an impractical task that relies on manual reading and comprehension. To address this problem, we used the SVM model of machine learning to automatically classify and explore the catalog data of the *Hetu Dangse*. This model further demonstrates the relevance of the knowledge between documents in the *Hetu Dangse* and facilitates an in-depth analysis.

We imported the vectorized labeled data set into the SVM model and selected the optimal parameter combination to run the model. To visualize the data results, the 50-dimensional word vector is reduced to a 2-dimensional word vector using the t-distributed random neighborhood embedding algorithm. We used the SVM model to establish a hyperplane visualized in 2-dimensional form. The legend only in figure 10 shows the data distribution of the six categories with the highest proportion owing to the large number of categorized data. To test the classification effect of the SVM model, we used precision and recall as metrics and calculated the F1 score to validate the model. The results are presented in table 3. Based on the created SVM model, 95,680 catalog data of the *Hetu Dangse* were predicted and classified. The results are shown in figure 11. Although there exist certain deficiencies in accuracy and other aspects, it a positive impact for the content research, management, utilization, and retrieval discovery of *Hetu Dangse*.

Table 3. SVM model validation parameters

	Result
Precision	0.736
Recall	0.717
F1 Score	0.716

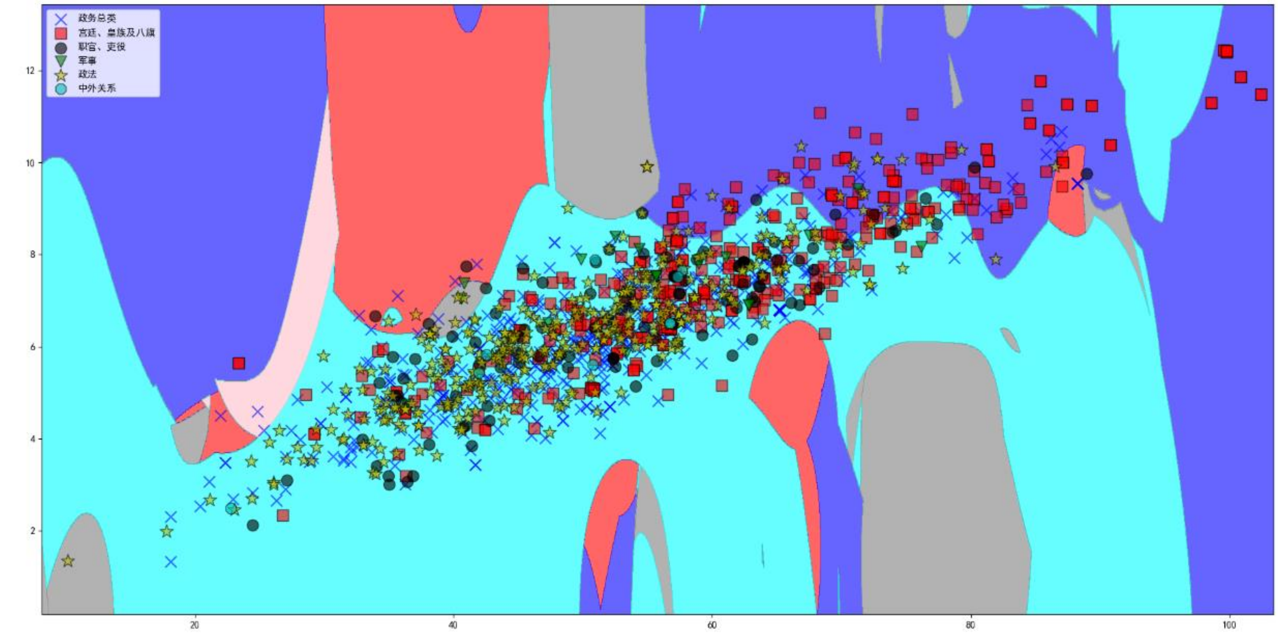


Figure 10. SVM decision region boundary.

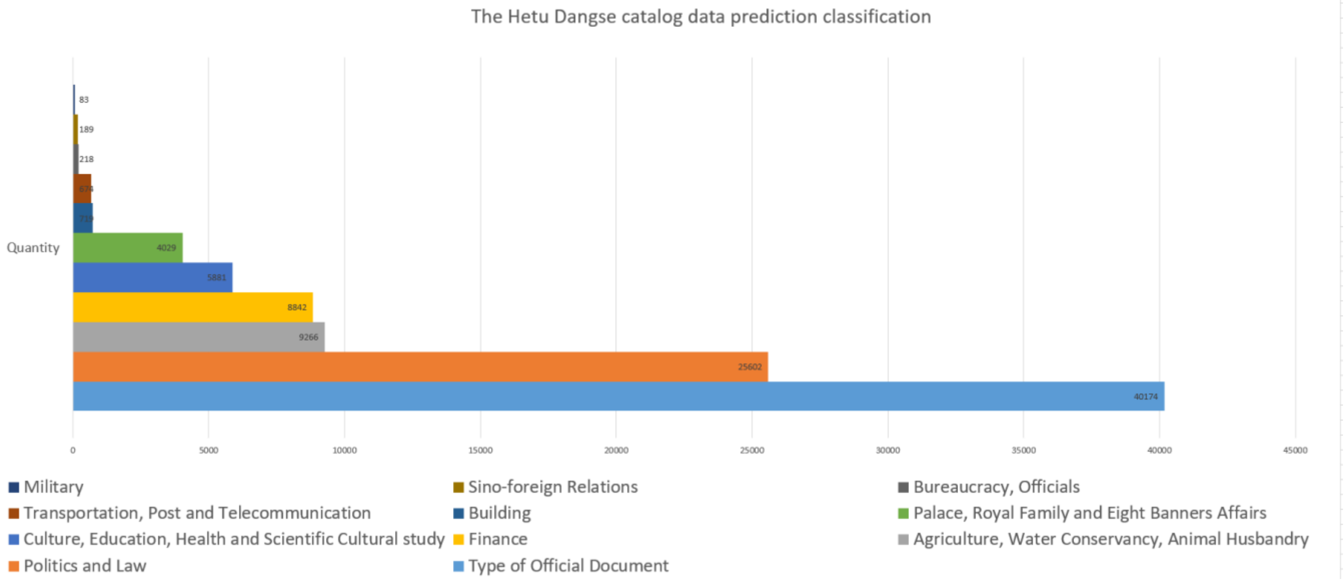


Figure 11. *Hetu Dangse* catalog data prediction classification.

CONCLUSION

In this study, we used machine learning to analyze and visualize the catalog data of the *Hetu Dangse*, revealing the functional relationship of the Qing Dynasty, Shengjing regional institutions recorded in this historical document, and showing the institutional communicated relationships. Using the SVM model, we achieved automatic classification of the *Hetu Dangse* catalog from the category perspective. Owing to the massive archives of historical materials in ancient China, the fonts of many historical materials cannot be recognized by computers or humans. The digitization of catalogs has become a digital bridge between researchers and historical documents. This not only achieves the concise summary and refinement of them but also greatly improves the utilization efficiency by researchers. The SVM model can "learn" through the labeled sample data and realize automatic classification of large amounts of unlabeled catalog data. By automatic classification of catalog data, historical data researchers and archive managers can use and manage a large number of historical documents and catalog data more effectively, greatly increasing their utilization. The co-occurrence algorithm can reveal the rules written by the catalog data itself, discover the distance between the catalog data, and form clusters providing a clearer direction for researchers to use historical documents. The algorithm also saves time for researchers to identify documents without purpose, making content presentation of historical documents to readers clearer. This paper improves archivists' awareness of archive data compilation and management. First, data is observed, topics are identified, and potential relationships between these are found and established to improve historical archives' compilation. Second, the visual presentation method and carrier is chosen, and via the web browser established relationships are visualized for the users to access and utilize. It can be said that scientometric research method can promote the transformation of historical research and archives management and compilation research from traditional explanatory scholarship to truth-seeking scholarship.

Currently, the application of machine learning technology has gradually extended from applied disciplines to traditional fields of literature, art, and sociology. However, there are still many opportunities in the field of historical research. This study used methods in the field of artificial intelligence to conduct text mining and visualize the presentation of historical archive document catalog data and proposes a new digital and intelligent solution for researching Chinese historical documents. With the development of science and technology, research methods for historical documents are undergoing constant changes from the traditional manual subjective analysis of historical data to relying on quantitative analysis represented by deep learning and data mining technology. It is an irreversible trend to research historical documents more comprehensively, accurately, and scientifically by means of artificial intelligence and other technologies on the scientific frontier.

For future work, we plan to conduct research on the Qing dynasty historical documents from a deeper semantic analysis level, construct a knowledge graph through the method of named entity recognition, and construct an ontological model transforming historical documents into a structured knowledge base to discover new knowledge from historical documents in an automated manner.

ACKNOWLEDGMENTS

Funding Statement

This work was supported by the General Program of the National Natural Science Foundation of China [grant number 72074060], the Research Foundation of the Ministry of Education of China [grant number 20JHQ012], and the National Social Science Fund of China [grant number 16BTQ089].

Data Accessibility

The data sets supporting this article have been uploaded as part of the Supplementary Material. https://drive.google.com/drive/folders/1bZs17otRUyvA_QKbShMF836yGDTi40y0?usp=sharing

Competing Interests
We have no competing interests.

ENDNOTES

[1] Wang Tao, "Data Mining of German Historical Documents in the 18th Century, Taking Topic Models as Examples," *Xuehai* 1, no. 20 (2017): 206–16, https://doi.org/10.16091/j.cnki.cn32-1308/c.2017.01.021.

[2] Kaixu Zhang and Yunqing Xia, "CRF-based Approach to Sentence Segmentation and Punctuation for Ancient Chinese Prose," *Journal of Tsinghua University (Science and Technology*) 10, no. 27 (2009): 39–49, https://doi.org/10.16511/j.cnki.qhdxxb.2009.10.027.

[3] Michael Stauffer, Andreas Fischer, and Kaspar Riesen, "Keyword Spotting in Historical Handwritten Documents Based on Graph Matching," *Pattern Recognition* 81 (2018): 240–53, https://doi.org/10.1016/j.patcog.2018.04.001; Wu Sihang et al., "Precise Detection of Chinese Characters in Historical Documents with Deep Reinforcement Learning," *Pattern Recognition* 107 (2020): 107503, https://doi.org/10.1016/j.patcog.2020.107503.

[4] Renata Solar and Dalibor Radovan, "Use of GIS for Presentation of the Map and Pictorial Collection of the National and University Library of Slovenia," *Information Technology and Libraries* 24, no. 4 (2005): 196–200, https://doi.org/10.6017/ital.v24i4.3385.

[5] Shaochun Dong et al., "Semantic Enhanced WebGIS Approach to Visualize Chinese Historical Natural Hazards," *Journal of Cultural Heritage* 14, no. 3 (2013): 181–89, https://doi.org/10.1016/j.culher.2012.06.009; Jakub Kuna and Łukasz Kowalski, "Exploring a Non-existent City via Historical GIS System by the Example of the Jewish District 'Podzamcze' in Lublin (Poland)," *Journal of Cultural Heritage* 46 (2020): 328–34, https://doi.org/10.1016/j.culher.2020.07.010.

[6] Aleksandrs Ivanovs and Aleksey Varfolomeyev, "Service-oriented Architecture of Intelligent Environment for Historical Records Studies," *Procedia Computer Science* 104 (2017): 57–64, http://doi.org/10.1016/j.procs.2017.01.062; Guus Schreiber et al., "Semantic Annotation and Search of Cultural-heritage Collections: The MultimediaN E-Culture Demonstrator," *Journal of Web Semantics* 6, no. 4 (2008): 243–49, https://doi.org/10.1016/j.websem.2008.08.001.

[7] M Kim et al., "Inference on Historical Factions Based on Multi-layered Network of Historical Figures," *Expert Systems with Applications* 161 (2020): 113703, http://doi.org/10.1016/j.eswa.2020.113703.

[8] Hobson Lane, Cole Howard, Hannes Hapke, *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python* (New York: Manning Publications, 2019), 165.

[9] Laurens Van der Maaten, Eric Postma, and Jaap van den Herik, "Dimensionality Reduction: A Comparative Review," Tilburg University Technical Report, TiCC-TR 2009-005 (2009), https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.

[10] Gavin Hackeling, *Mastering Machine Learning with Scikit-learn* (Birmingham: Packt Publishing, 2017).

[11] Richard Smiraglia, *Domain Analysis for Knowledge Organization: Tools for Ontology Extraction* (Oxford: Chandos Publishing, 2015).

[12] Kuo-Chung Chu, Hsin-Ke Lu, and Wen-I Liu, "Identifying Emerging Relationship in Healthcare Domain Journals via Citation Network Analysis," *Information Technology and Libraries* 37, no. 1 (2018): 39–51, https://doi.org/10.6017/ital.v37i1.9595.

[13] Archives of Liaoning Province in China, "The *Hetu Dangse* Series Archives Publication," *Qing History Research* 6, no. 2 (2009): 1.

[14] Amit Kumar Sharma, Sandeep Chaurasia, and Devesh Kumar Srivastava, "Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec," *Procedia Computer Science* 167 (2020): 1139–47, https://doi.org/10.1016/j.procs.2020.03.416.

[15] B Hongxi, "Research on the *Sanling* Management Institutions of the Qing Dynasty Outside the Pass," *Manchu Minority Research* 4, no. 12 (1997): 38–56.

[16] Guangli Zhu et al., "Building Multi-subtopic Bi-level Network for Micro-blog Hot Topic Based on Feature Co-occurrence and Semantic Community Division," *Journal of Network and Computer Applications* 170 (2020): 102815, https://doi.org/10.1016/j.jnca.2020.102815.

[17] S. Ravikumar, Ashutosh Agrahari, and S. N. Singh, "Mapping the Intellectual Structure of Scientometrics: A Co-word Analysis of the Journal *Scientometrics* (2005–2010)," *Scientometrics* 102 (2015): 929–55, https://doi.org/10.1007/s11192-014-1402-8.

[18] Jiming Hu and Yin Zhang, "Research Patterns and Trends of Recommendation System in China Using Co-word Analysis," *Information Processing and Management* 51, no. 4 (2015): 329–39, https://doi.org/10.1016/j.ipm.2015.02.002.

[19] Nees Jan Van Eck and Ludo Waltman, "Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping, *Scientometrics*, 84, no. 2 (2010): 523–38, https://doi.org/10.1007/s11192-009-0146-3.

[20] Z Yanchang and L Xinzhu, "The Study of the Function of Shengjing Office from the Use of the Official Communication — An Academic Investigation Based on *Hetu Dangse*," *Shanxi Archives* 8, no. 12 (2020): 179–88.

[21] ShengJing Ministry of Revenue, *Guangxu's Great Qing Huidian Volume 25* (Zhonghua Book Company, 1991), 211–12.

[22] F Yonggong and G Jialu, "Brief Introduction of Shengjing Upper Three Banners Baoyi Zuoling," *Historical Archives* 9, no. 30 (1992): 93–7.

[23] Wangyue, "Research on the Yamens and Their Affair Relationships in Shengjing Area," *Shenyang Palace Museum Journal* 1, no. 31 (2011): 67–77.