Research on Judicial Big Data Text Mining and Sentencing Prediction Model

Juan Xu
Sichuan Judicial and Police Officers Professional College, DeYang, Sichuan, China
618000
2014326215@dlvtc.edu.cn

Abstract—In the era of big data, computer data has grown tremendously, especially in unstructured text data, and its growth rate is amazing. For this reason, in the application of judicial big data, people need to apply automated text mining technology to process unstructured data and predict the specific value of information. This article summarizes the research design and technical background content based on previous work experience. The author discusses the analysis of specific research results from three aspects: pre-processing of research data, establishment of sentencing prediction models, model testing and evaluation.

1. INTRODUCTION
Up to now, big data has become a basic strategic resource for the development of many countries and societies, and it can drive the entire world towards the data age. How to apply and discover massive amounts of data and turn it into valuable information for human society is the current focus of various industries. In actual development, data mining and text mining are the most common technologies for processing big data. Among them, data mining deals with structured data, while text mining deals with unstructured data. Nowadays, domestic and foreign big data technology is widely used in crime investigation, judicial sentencing, etc. and has achieved good application effects.

2. CHARACTERISTICS OF JUDICIAL BIG DATA

2.1. Mass and Openness
With the implementation of judicial disclosure, judgment documents and information on trial procedures have been disclosed successively. This has led to a sharp increase in the amount of judicial big data, and the agglomeration effect is obvious. Relevant data shows that as of May 2017, the people's courts big data management and service platform has gathered 108 million case information from courts across the country, making it the world's largest trial information resource library. Besides, judicial big data is not simply a superposition of various data, but is presented in different forms on the basis of the original classification. The content involved includes court organizational structure, work dynamics, and litigation guidelines, etc., as well as live court case data and execution case data.

2.2. Authenticity and Openness
The emergence of judicial big data is mainly the process of recording judicial activities. The objectivity of judicial activities itself directly affects the authenticity of subsequent judicial big data. The massive case information data mainly displays the records of disputes, litigation and judicial judgment activities. Among them, the authenticity of judicial big data is the basic prerequisite for its function. All judicial
organs need to collect and process data in a comprehensive manner to prevent the authenticity of the
data from being affected, let alone subjective assumptions or false constructs. From the perspective of
big data development, it mainly involves two aspects. Firstly, the openness of judicial big data itself. In
general, judicial data is not a closed system, and digital resources are gradually increasing. Secondly,
judicial big data itself is public data and needs to be open to the public and share information. In
addition, judicial data is a "mirror" for testing judicial work, and understanding the specific operation of
its system and working channels can provide various courts and judges with an information platform for
learning and reference [1].

3. THE FUNCTION OF JUDICIAL BIG DATA

3.1. Improve the Quality and Effectiveness of Trials
At this stage, the main service objects of judicial big data developed by the court system are trial
execution and trial management. With the aid of judicial big data, real-time analysis of the
characteristics of the parties to the case, the number of cases, and the trial situation can be carried out,
and thematic reports of different case types can be established. This can push relevant laws and
regulations, reference cases and other documents to judges handling cases, reduce the work pressure of
judges, and provide them with comprehensive intellectual support. More importantly, judicial big data
can not only analyze whether similar cases can be judged at the same time, but also help courts at all
levels formulate uniform judgment rules and standards for similar cases, so as to avoid problems such
as arbitrary judgments and misjudgments. Otherwise, people can also use the judicial big data
information sharing platform in the Internet era to completely change the traditional execution mode
and method, and lock the specific property information and activities of the person being executed.
People can use real-time monitoring, online inquiry, etc., to find relevant clues faster to provide strong
support for the final settlement. Staff can also use network process reengineering to ensure automatic
eye early warning and intelligent supervision of case handling, and to measure the workload of each judicial
officer and the specific difficulty of handling cases. This can make subsequent resource allocation and
performance appraisal work more objective.

3.2. Promote Judicial Openness
Judicial openness is not only the basic requirement of the people for judicial work, but also a channel to
demonstrate judicial justice. At this stage, the total number of effective judgment documents published
by the China Judgment Documents Network at all levels of courts across the country has reached 42.6
million, with a total of 13.2 billion visits. These visits came from more than 210 countries and regions
around the world, making the website the world's largest website for judgment documents. For this
reason, relevant departments need to use technological and informatization methods to strengthen
cooperation with information technology companies, and use this to strengthen data analysis and
processing capabilities. This is also an inevitable choice for innovative judicial public formulation. As
the pace of people's lives accelerates, the time cost of public attention and supervision of judicial affairs
has increased significantly. For this reason, relevant departments need to break free from the limitations
of traditional thinking and use modern information technology to build new smart courts. This can
ensure that citizens can use multiple information channels to achieve judicial supervision, protect the
people's right to know and participate, and improve judicial disclosure procedures [2].

4. TECHNICAL BACKGROUND AND RESEARCH METHODS

4.1. Technical Background

4.1.1. Text Mining
In big data processing, data mining and text mining are key technical contents. Generally speaking, the
basis of data mining is mainly numerical morphological data, which is suitable for application in
text-type narrative data. In contrast, data mining itself has obvious structural advantages, and it can build regular tables or databases. For one thing, text mining does not have any specifications, and the length of the content is also inconsistent. It is text data with no defined nature. It can be a specific language in people's daily life, or it can be an online text. It can also be seen from this that text mining is a dynamic analysis of other unstructured data. Generally speaking, the key to text mining is to cut the hidden information between texts into independent information units to achieve comprehensive processing of various information. This also involves a lot of synonym confirmation. For another, the existing text mining software applications involve synonymous character libraries, as well as synonymous but different words confirmation, and finally all converted into processable information units. Before actual text mining, it is necessary to apply specific programs to remove texts that do not have analytical value in the file, and perform subsequent analysis operations with the help of mining tools.

It can also be seen from the application of big data technology that unstructured text mining is more difficult than structured data mining, and workers need to dig out data related to human behavior patterns from text. At this stage, text mining has been widely used in automatic classification, automatic classification, etc. People can use text mining technology to achieve effective classification of documents, or to predict sentencing. If text mining technology can be combined with judicial documents, various electronic document data can be used to reasonably predict future crime trends, and then provide assistance for judicial criminal investigations [3].

4.1.2. Automated File Classification
In general, computers do not have the ability of humans to read files. Therefore, people need to convert the document into a format that can be analyzed by the system, and divide the unstructured sentence into multiple words. Then, we can use these words to do in-depth analysis to ensure the effective selection of feature values. Secondly, we can implement supervised learning operations in machine learning with the help of known file training data. This can ensure that the classification and counting can automatically learn the knowledge of file classification, highlight the file classification rules, and achieve automatic file classification. In general, automatic file classification mainly involves two aspects. First, the feature value of the target file is selected. Second, we can realize automatic file classification through machine learning. Traditional documents are often manual-based in the classification process, which is subjective and has limited work efficiency. In the design of automatic document classification, the classification rules are summarized from the entire document data set by means of the supervised learning method in machine learning. Afterwards, we can refer to the classification model to develop new file category labels.

File classification often uses a large number of feature words to create a new "feature vector pattern" expression file. In this process, genetic algorithms have become more common in recent years. They often apply the "font method" in the representation of actual document attributes, but this method will increase the feature dimension of the document. When the classification algorithm processes this kind of data, there will be a lot of resistance, and people call this kind of scene a "dimensional disaster". When the actual file dimensions are improved, the first thing the researchers need to do is to determine the difficulty of possible attribute combinations and the time required to avoid classification errors. Secondly, most of the features on high-dimensional file data will not show any help for classification, and may even have an impact on classification work. Hence, people need to select a set of keywords from the original document and reduce the feature dimension to improve the performance of the classification algorithm.

In the previous research process, the staff mainly used feature selection methods to perform file classification operations. The staff will also use statistical algorithms as a basis to understand the specific frequency of words, including document frequency, information profitability, and so on. In the research of this paper, we mainly analyze the key words in the judgment with the help of the law of relevance, and finally can realize the classification of the characteristic values of the judgments of different drug trafficking crimes and formulate stable test conditions.
4.2 Research Method

4.2.1 Word Segmentation Technology in Text Mining

In general, words are the smallest language unit. In the actual operation of the language processing system, first of all, we need to analyze the words in the text. Only in this way can the subsequent processing operations continue, which also makes the automatic Chinese word segmentation work an important technical content in language processing. The actual word segmentation technology can help the staff distinguish between the thesaurus method and the non-thesaurus method. Among them, the non-thesaurus method is mainly used to segment a large amount of corpus content. If the number of occurrences of related words is higher than the threshold value, people can use it as an independent vocabulary. In the recognition of new words, it will not be restricted by the traditional thesaurus method. In contrast to the application of the lexicon method, it is mainly through the previously constructed lexicon to compare the document word by word, and then understand the hidden vocabulary information [4].

Using statistical methods to carry out word segmentation operations requires a large number of files and corpus training operations. If it is a non-thesaurus method, it is necessary to obtain enough word frequency words to keep fit with the actual data processing. In this way, there is no need for experts to define the vocabulary in advance, reducing the chance of related problems. The non-thesaurus method is mainly to understand the specific frequency of words. It can mechanically cut a string of N characters as a unit, and find new words from it. When the non-thesaurus method is combined with natural language analysis, each word has a relationship with the previous word. People call this model the Markov model of N-1. For example, the term "fixed-term imprisonment" can be divided into index word forms such as "period-term", "term imprisonment" and "imprisonment". With the help of this kind of algorithm, people can judge the content of specific legal clauses and words in the judgment, and make it clear that there is no mitigation of sentence. This has positive significance for the development of follow-up statistical analysis.

4.2.2 Keyword and Weight Technology in Text Mining

Due to the different importance of different words in the entire document, relevant staff have put forward concepts such as word frequency and reverse document frequency. For example, "tablet computers" appear more frequently in information and technology documents, but not many times in some social documents, and their importance is not consistent. To this end, researchers need to use TF and IDF to weigh specific feature values for words. In the process of information retrieval and text mining, TF-IDF is a common weighting technique. It belongs to the category of statistical methods, which can measure the importance of a single word in a text. The word frequency mainly represents the specific number of occurrences of a word. The more occurrences of a word, the higher its importance. The numerator represents the number of occurrences of the word in the document, and the denominator is the total number of occurrences of all words in the document. However, since the word frequency can only indicate the number of times the text appears, some more frequent words will be scattered in the file. Moreover, it can treat representative words as unimportant words, but we cannot apply this indicator alone to make any judgments.

4.2.3 Related Technical Analysis

The purpose of actual correlation analysis work is to clarify the correlations between projects and events, and present the causal relationship in the law. It can express the relationship between data through clear management rules. For example, this technology needs to be applied when analyzing the same consumer or shopping basket in the same stage. When data mining operations are carried out through the association law, the simultaneous relationship and frequency analysis in the transaction can be realized to obtain the association status of the product. Furthermore, the main function of association analysis technology is to show the connections between data. It can be applied in many fields, including judicial data research. When using data sets, people can establish new association analysis rules, and
then conduct comprehensive knowledge exploration operations. Among the actual association rules, each analyst needs to filter out useful rules. This can ensure the in-depth study of this type of rules, and clarify whether the causal relationship in the rules really exists [5].

5. RESULT ANALYSIS

5.1. Pre-processing of Research Data
In the actual relevance analysis and keyword selection, its main purpose is to determine the legal causality and set up effective association rules. In this way, the relationship between the data can be expressed, and the specific conditions are shown in Table 1. For example, in Article 67 of the Criminal Law, the law contains the word surrender. In the actual text mining of the judgment, the text of the law itself and "surrender" can become keywords that affect the sentence.

| Word       | Related Laws                  | Increase or Reduce Sentence | Variable Code |
|------------|--------------------------------|----------------------------|---------------|
| Article 67 | Article 67 of the Criminal Law | Commutation                | L67           |
| Surrender  | Article 67 of the Criminal Law | Commutation                | W67           |
| Article 68 | Article 68 of the Criminal Law | Mitigation or Exemption     | L68           |
| Expose     | Article 68 of the Criminal Law | Mitigation or Exemption     | W68           |
| Article 27 | Article 27 of the Criminal Law | Discretionary Commutation   | L27           |
| Accessory  | Article 27 of the Criminal Law | Discretionary Commutation   | W27           |

5.2. Establish Sentencing Prediction Model
In order to ensure the completeness of the variables in the evaluation of the sentencing prediction model, the staff needs to collect the specific results of the court's decision. This includes the first trial and the second trial to see if there are specific differences. First of all, in order to clearly determine whether there is a difference between the litigation process and the sentencing period during the research, people can implement independent sample testing. For example, in the processing of training data, the basic training data is mainly based on drug judgments. People can analyze with the help of statistical regression. Comprehensive data mining tools can ensure that the training data is toward linear regression and neural network data modeling, and then provide a basis for subsequent training results. In this process, the staff can choose linear regression research to observe the basic situation of the model establishment [6].

5.3. Model Testing and Evaluation
In order to facilitate research, the staff randomly selected 78 completed drug trafficking judgments from the Judgment Written Network in 2018, and classified them according to specific research requirements. After that, people have to convert classification judgments into basic data. People can calculate it according to the sentencing formula to get the model's predicted average sentence and comparison of the actual sentence of the case. This can clarify the predictive ability of the actual sentencing prediction model, and provide favorable conditions for the entire judicial big data text mining work.

6. CONCLUSION
In summary, actual court judgment text mining and automated model construction can provide effective references for subsequent judicial big data applications. Due to the different types of cases, relevant staff should overcome the high-dimensional characteristics and tight coupling characteristics of data in the judicial field to achieve automatic classification of various types of data. Relevant staff should use
statistical linear regression to establish various types of case sentencing databases to strengthen the real-time characteristics of the main service work.

REFERENCES

[1] Shu Honghui. Research on Judicial Big Data Text Mining and Sentencing Prediction Model [J]. Law Science, 2020(07): 113-129.

[2] Wang Jiayun. Judicial Big Data and the Realization of Judicial Justice[J]. Journal of Jishou University (Social Science Edition), 2020, 41(02):137-143.

[3] Wang Lusheng. Jurisprudence conflict and value balance in the application of judicial big data——From the French judicial big data ban[J]. Comparative Law Research, 2020(02):133-146.

[4] Tu Fan. The past, reality and controversy of the application of judicial data in the United States-and the enlightenment to the application of judicial big data in my country [J]. Science and Technology and Law, 2020(01): 57-67.

[5] Zheng Ge. Coordination and Integration of Judicial Science and Technology[J]. Application of Law, 2020(01): 3-11.

[6] Hu Xiangyang, Zhang Wei. Data mining and analysis of investigation and interrogation text based on big data[J]. Journal of Chinese People’s Public Security University (Social Science Edition), 2019, 35(06): 35-43.