The Design and Development of Comprehensive Analysis System of Investigation and Interrogation Text Based on Data Mining

Hong Zhou(✉) and Yipeng Yang

Sichuan Judicial and Police Officers Professional College, Deyang 618000, Sichuan, China
549735999@qq.com

Abstract. Faced with the massive electronic information of police investigation and interrogation texts, police officers lack scientific and rational means of deep-level information mining and analysis, resulting in low work efficiency and insufficient information utilization. In this paper, the investigation and interrogation text will be taken as the research object, the data mining technology as the core, and the comprehensive analysis system of investigation and interrogation text will be built in Python language development environment, combined with Django framework. The system will use class libraries such as Jieba and NLTK under Python system to realize operations such as word segmentation, feature word extraction, classification and cluster mining of investigation and interrogation texts. Finally, the visual chart results of NBM algorithm and K-means algorithm will enhance the judgment ability of investigation and interrogation texts, promote the efficient and three-dimensional development of investigation work, and make a beneficial attempt for the construction of intelligent policing in the new era.

Keywords: Data Mining (DM) · Python · Investigative Interrogation Text · Digitized Information

1 Introduction

As the informatization and networking of police work is carried out, special electronic transcripts for investigation and interrogation are gradually popularized. Although the "paperless" case-handling mode can improve work efficiency, a large amount of electronic text information is piled up, while the conventional processing mode can only realize single inquiry or centralized storage, ignoring the significance of rediscovering the investigation and interrogation text data related to gang crimes, serial and parallel cases, crime prediction, etc. [1].

Therefore, in the current digital information age, the electronic investigation and interrogation text can be transformed into quantifiable and structured data information that can be recognized, understood and processed by computers. With the help of the application advantages of data mining technology in electronic text processing, this paper adopts Python language and Django framework to complete the construction of comprehensive analysis system for investigation and interrogation text. The application

V. Balakrishnan et al. (Eds.): ICMEIM 2022, AHSSEH 6, pp. 450–454, 2022.
https://doi.org/10.2991/978-94-6463-044-2_57

process of the system will include electronic text data extraction, text preprocessing, text key feature word selection, text analysis and mining. Police officers can realize the functions of case element extraction, case classification and cluster analysis and other investigation and interrogation text information judgment by simple operation on the Web application according to the actual case requirements.

2 Overview of Key Technologies

2.1 Data Mining Technology

As a kind of computer science and technology, data mining aims at the complex process of finding valuable patterns and extracting useful data from a large number of data sets. The general data mining objects include structured relational databases, which also support semi-structured data and even heterogeneous data of text, multimedia data, spatial temporal data, Web dynamic data [2]. Among them, the data mining of text content belongs to the field of artificial intelligence in essence, and there is a difference between the overall process and general data mining. It is necessary to use linguistic principles to complete feature selection and mining engineering design.

2.2 Python

The Python is a high-level scripting language that combines interpretive, compiler, interactive and object-oriented [3].

In the process of Web application development, Python language will combine Django development framework to complete the design and development of system server according to MVC pattern. But Django framework is different from the standard MVC pattern. Under Django framework, more attention is paid to Model, Template and Views, also known as MTV pattern.

2.3 Development Process

According to the system development requirements and the use requirements of the above key technologies, complete the configuration and deployment of the development environment.

The electronic text information data mining process mainly includes the steps of text data extraction, text preprocessing, text key feature word selection, text analysis and mining, etc. Firstly, in the process of text data extraction, the system can directly call the database content under the special electronic record management system for investigation and interrogation through a special data service interface. In the text preprocessing stage, Chinese word segmentation and cleaning will be performed on the original text data. In this research, the system will choose Python Jieba class library and police-specific terminology thesaurus to complete the Chinese word segmentation task. In the cleaning process, the system will automatically remove stop words, remove single words, and remove some onomatopoeia, prepositions, conjunctions, etc. according to part-of-speech tagging [4].

```
def tfidf_top(trade_list, doc_list, max_df, topn):
    vectorizer = TfidfVectorizer(max_df=max_df)
    matrix = vectorizer.fit_transform(doc_list)
    feature_dict = {v: k for k, v in vectorizer.vocabulary_.items()}    # index ->
feature_name
    top_n_matrix = np.argsort(-matrix.todense())[:, :topn]    # top tf-idf words for each row
    df = pd.DataFrame(np.vectorize(feature_dict.get)(top_n_matrix), index=trade_list)    #
convert matrix to df
    return df
```

Fig. 1. Calculating the key word TF-IDF value code (original)

In the stage of selecting key words of text, the system will use TF-IDF algorithm to calculate the preprocessed text. The system uses the API interface of TF-IDF under scikit-learn package to realize the statistics of TF-IDF value of each word, and the key implementation code is shown in Fig. 1.

In the stage of text analysis and mining, the system mainly supports text named entity recognition and extraction, text classification and mining, and text clustering extraction. The system mainly relies on the named entity identification interface under NTLK class library to realize two functions of entity boundary identification and entity category confirmation, that is, the special nouns contained in the investigation and interrogation text are automatically extracted. For text classification mining, the system will select naive Bayes classifier according to the TF-IDF value of text content to complete the classification of cases. Besides, for text clustering mining, the system will use K-means algorithm to complete the clustering of cases, and extract the topics of this category to facilitate the similarity analysis of cases and the processing of serial and parallel cases.

For the development of the whole server side of the system, the operating system is Windows10.0. The Web server is Nginx server, the version is Nginx/Windows -1.12.2, the project development language is Python 3.6.6, the development tool is PyCharm 2018.3.1 x64, and the database is MySQL5.7 to complete the construction and support of the system database system. The whole server is implemented by Django2.0.1 framework, while the visualization of the final results of text analysis and mining depends on Echarts4.2.1. Through the introduction of the above key technical theories, the overall environment of the system development, the configuration of related software and tools are determined, and the technical feasibility of the overall project of the investigation and interrogation text comprehensive analysis system is also clarified.

3 Function Realization

3.1 Data Retrieval

In the data retrieval function module, the system supports users to search and call the contents of investigation and interrogation texts through different case elements. According to the system text content naming entity recognition result, the text content is tagged, and the key code is shown in Fig. 2. Based on this, the corresponding thesaurus can be formed, and the retrieval efficiency of text content can be accelerated.

```
import nltk
newfile = open('news.txt')
text = newfile.read()
tokens = nltk.word_tokenize(text)
print("tokens",tokens)
tagged = nltk.pos_tag(tokens)
print("tagged",tagged)
entities = nltk.chunk.ne_chunk(tagged)
a1=str(entities)
file_object = open('out.txt', 'w')
file_object.write(a1)
file_object.close( )
print (entities )
```

Fig. 2. Python NTLK class library realizes the key code of named entity identification (Original)

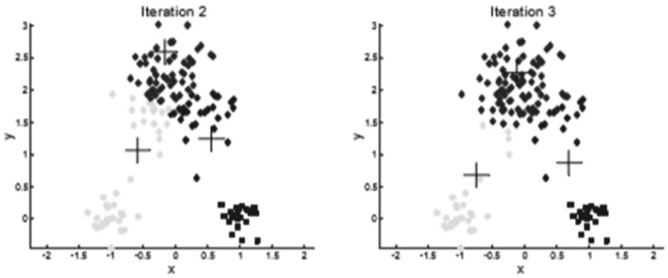

Fig. 3. Case Cluster Analysis Results (Network)

3.2 Comprehensive Analysis and Judgment of Cases

In this function module, the system will support case classification and case clustering. Under case classification, users are supported to complete case classification through the vector expression input into the classifier model. After the classification, users will intuitively obtain structured data display, and support the integration and analysis of investigation and interrogation texts. For the case clustering analysis, the system will also input the vector expression of the key words of the text into the clustering algorithm model to complete the case clustering. According to the clustering results, users can quickly obtain the similarity analysis of different cases, and show the related cases in the form of atlas, which can be used to intuitively discover the implied relationship between cases and realize the serial and parallel processing of cases [5].

3.3 Visual Display

The visual drawing tool of Echarts is used to complete the rendering of electronic text classification results. There are many kinds of charts of Echarts, including line charts, pie charts, radar charts, etc., as shown in Fig. 3, the case clustering analysis effect chart.

4 Conclusion

In this paper, the investigation and interrogation text is taken as the research object, aiming at the lack of scientific and rational deep-level information mining and analysis means for the current massive electronic police investigation and interrogation text information, with the help of digital mining technology, Python language combined with Django framework is adopted to complete the construction of comprehensive analysis system for investigation and interrogation text. The system will enhance the judgment ability of investigation and interrogation texts by visualizing the chart results, promote the efficient and three-dimensional development of investigation work, and promote the development of intelligent policing construction in the new era.

References

1. Wei Wenyan, Lv Xin and so on. Application of Text Mining Technology in Case Analysis of Public Security Field [J]. Journal of Hunan Public Academy.2017.06:98–104
2. Hu Jiming, Tian Peilin. Topic Mining and Evolution Analysis of Text Intelligent Computing Research [J]. Journal of Information.2021.04:140–146
3. Fang Ji, Xie Huimin. The Application of Python in Big Data Mining and Analysis [J]. Digital Technology and Application.2020.09:75–76
4. Zhu Yongzhi, Jing Jing and so on. Research on Chinese Word Separation Technology Based on Python Language [J]. Communications Technology.2019.07:1612–1618
5. Gao Yating. The Study and Implementation of Text Mining System Based on the Elements of Criminal Cases [D]. Chang'an University .2019.06