Feature Extraction Model of Text Classification In National Defense Science And Technology Tracking

Kan Li

Teaching and research support center of Dalian naval vessel academy, Dalian, Liaoning, China
likandl@126.com

Abstract

National Defense Science and technology information is mostly distributed in the form of non classified information. It is an important reference unit for national defense science and technology development planning and technical route. It is an early information resource for national defense scientific research. National Defense Science and technology tracking is a process of measuring, analyzing and extracting national defense science and technology information in different states by using different data analysis technologies based on the information statistics related to national defense science and technology. At present, national defense science and technology tracking lacks effective tracking methods and quantitative analysis tools. This paper uses the method and idea of deep confidence network feature extraction to calculate the correlation between search engine and abstract, so as to monitor, identify, measure and preliminarily extract the texts in different states, such as scientific papers, conference materials, news trends, patent information, policies and regulations, economic and trade information and so on. The research results show that the use of the model can realize the accuracy of tracking, and provide a new idea to improve the tracking speed.

Keywords: *National defense technology, Information extraction, Feature classification, Text classification*

1.INTRODUCTION

Countries all over the world first apply the achievements of science and technology to the military field. National defense science and technology has the characteristics of technology intensive and rapid renewal. National defense science and technology has achieved unprecedented development with the rapid progress of science and technology in the whole world. On the one hand, the volume of national defense science and technology information has increased exponentially, and the disorder and excess of massive point information leads to intelligence disaster; On the other hand, due to reasons such as confidentiality, patent or industry competition, systematic information blockade will be formed between corresponding organizations and even between individuals within the same organization. The vast majority of Frontier achievements of national defense science and technology remove directional information such as models, codes and sensitive words. After being disassembled, implied and civilian, these messy information is contained in scientific and technological journals, scientific and technological conferences There are barriers to the transmission of information between scientific and technological organizations, as well as various information carriers in the network [1].

2.CURRENT SITUATION, TYPES AND CHARACTERISTICS OF ANTI SCIENTIFIC AND TECHNOLOGICAL INFORMATION TRACKING

In recent years, the military science and information research center has focused on the actual needs of national defense science and technology information, focused on the research and development of key technologies of national defense science and technology information big data, focused on the construction of national defense science and technology information auxiliary analysis system, based on the construction of big data resource system and intelligence object database, and supported by the construction of infrastructure such as big data server cluster, The ability to acquire, organize, mine and serve large-scale national defense science and technology information has been preliminarily formed. Effectively improve the accuracy of characteristic entity recognition in institutions, equipment, projects and other fields [2]; The research

A. El-Hashash et al. (Eds.): IEIT 2022, ACSR 100, pp. 355–360, 2023.
https://doi.org/10.2991/978-94-6463-058-9_59

and experiment of knowledge extraction technology have been carried out, and the extraction and integration of entity attributes, relationships and events such as institutions, equipment, projects [6], personnel and technology have been preliminarily realized.

The fundamental of tracking is to extract the text according to its characteristics on the basis of text classification. Based on the massive increase of detection data, it is necessary to select a more effective text detection model. The technical preparation of text detection is to locate the text area in the picture. As shown in Figure 1, the red box represents the "LAN" character ground truth (GT), and the green box represents detection box. When GT and detection box have the same IOU, the recognition results are very different. Therefore, detection has a great impact on subsequent recognition. At present, there are many text detection methods, including: EAST/CTPN/SegLink/PixelLink/TextBoxes/TextBoxes ++/TextSnake/MSR/...

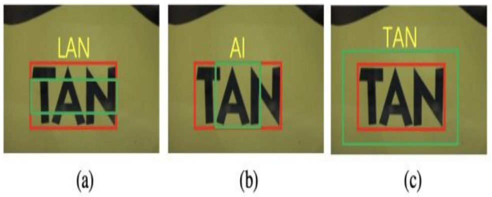

Figure 1: Area influence diagram of text detection

Firstly, semantic segmentation branches are added to the character recognition network to obtain the relative position of each character. Secondly, after obtaining the position of each character, the characters are classified to obtain character recognition information [4].

National Defense Science and technology information dissemination is an important part of science and technology dissemination. The exchange, dissemination and impartment of all scientific and technological knowledge belong to scientific and technological communication. In a broad sense, it is understood as all information flows in the process of scientific and technological information dissemination and exchange. It is a process of knowledge sharing in different personality spaces through the diffusion of scientific and technological knowledge information across time and space [3]. National Defense Science and technology information dissemination refers to the information exchange and sharing between different subjects through the diffusion of national defense science and technology information across time and space. It is the flow form of national defense science and technology information. It is a coherent and dynamic interaction process including communicators, communication contents, communication channels and receivers. The information flow is transmitted from the disseminator to the audience, and then the audience feeds back the information, forming an organic unity, in which all links are closely connected and act together on the final communication effect (see Figure 2).

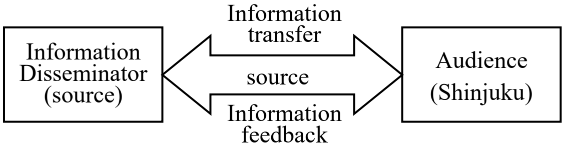

Figure 2: Flow chart of information dissemination

3. MAIN TECHNICAL PROBLEMS FACED BY NATIONAL DEFENSE SCIENCE AND TECHNOLOGY TRACKING

3.1. Insufficient information discovery and information collection technology

Papers, reports, patents and other scientific and technological literature resources have always been the main content of national defense scientific and technological information collection [5]. With the rapid development of the Internet, the importance of network information is highlighted, which contains more and more valuable intelligence. Internet information must be included in the information collection scope of national defense science and technology information institutions. Compared with scientific and technological literature, the types of network information are more diverse, the organization is more loose, the quantity is more huge, the value is more sparse and the update is faster. We must use big data technology to carry out systematic collection.

3.2. Insufficient maturity of effective information extraction technology

The content of Internet information is complex. It is necessary to clean the collected national defense science and technology information from different angles to improve the information quality and lay the foundation for further development and utilization. We need to develop technical means to automatically distinguish the types of web pages and improve the effect of subsequent utilization. In order to make more effective use of information, metadata such as title, publishing time and source should be extracted from web pages. There are a large number of websites, involving a variety of web page styles, which rely entirely on manual configuration rules, and the workload is huge. Technical means should be considered to automatically extract metadata or generate extraction rules to improve efficiency.

3.3. The technology of accurate classification of information content is insufficient

The processing degree of Internet information is generally not high. It is necessary to label the content of national defense science and technology information from the perspectives of language, classification, keyword, abstract, named entity and so on, so as to improve the degree of information disclosure. It should be guided by application, and the thickness should be moderate. Multiple classification tables can be set, such as revealing information from multiple dimensions such as industry and technical field. According to the characteristics and application requirements of national defense science and technology information, the named entities that need to be identified should be scientifically defined. Semantic calculation shall be carried out according to this context.

3.4. Lack of capability of multi-source heterogeneous information fusion and integration technology

Different types of national defense science and technology information have different information dissemination functions. To carry out the development and utilization of big data of national defense science and technology information, we must widely collect information, study the characteristics and laws of different types of information, formulate metadata specifications, develop technical means, fully integrate multi-source heterogeneous information, give full play to the overall benefits of information resources and provide maximum value for users.

4.FEATURE EXTRACTION OF DEEP CONFIDENCE NETWORK FOR TEXT CLASSIFICATION

The rapid development of information technology has brought about a blowout growth in the amount of data. In the face of different types of massive data, people hope to filter out useful information accurately and efficiently. Therefore, data mining technology has become a research hotspot in recent years. Text classification technology based on machine learning is a kind of information processing technology with strong innovation and high application value in this field. It usually includes text preprocessing, text representation and feature selection, training classifier and other modules. Text representation and feature selection are the key links, which directly affect the accuracy of classification.

4.1. Calculation of correlation between search engine and summary

The information on the Internet is vast, and the network resources are growing at a ten fold rate. It is difficult for a search engine to collect the network information of all topics. Even if the information topics are collected comprehensively, due to the wide range of topics, it is difficult to do all topics accurately and professionally, resulting in too much garbage in the search results. In this way, the vertical theme search engine occupies a position in all kinds of search engines with its high degree of targeting and specialization Vertical search engine is a new search engine service model proposed by the general search engine with large amount of information, inaccurate query and insufficient depth. It provides valuable information and related services for a specific field, a specific population or a specific demand. Its characteristic is "specialized, precise and deep", and has industry color. Compared with the disorder of massive information of general search engine, vertical search engine is more focused, specific and in-depth. Integrated search engine (all in one search page), also known as "multi engine synchronous retrieval system" (such as Baidu), is to link several independent search engines on a WWW page. When searching, you need to click or specify the search engine. One search input and multiple engines search at the same time, which is quite convenient to use. The integrated search engine has no self built database, does not need research and development support technology, and of course, it can not control and optimize the search results. However, the integrated search engine production and maintenance technology is simple, and the linked search engines can be added, deleted, adjusted and updated in time at any time, especially the integrated links of large-scale professional search engines (such as flash, MP3, etc.).

Web page summary and web page content relevance measurement index. How much does the abstract of an article represent the original content of the text, whether the abstract can ensure to provide information close to the actual content of the web page, and whether the keywords in the abstract can represent the text keywords. There is no certain basis for determining the quantitative relationship between the weight of web page title keywords and the weight of web page Abstract keywords, which are the basis of correlation analysis using web page abstract. Therefore, the research group uses statistical methods to investigate the correlation between the current web page title and abstract and the web page content, in order to determine the relationship between the web page title and abstract information and the actual content of the web page, and determine the reasonable weight setting, so as to improve the average accuracy of the search results of the meta search engine. In the process of information retrieval, the contribution of each index word in the document to the document content is different, so the importance of index words can be regarded as a clustering problem. Index clustering includes inter class similarity and intra class similarity. The intra class similarity is obtained by calculating the frequency of the word Ki in the document DJ. It is generally expressed by TF (term frequency). The larger TF, the stronger the ability of this word to express the content of the document. Inter class similarity is to calculate the inverse frequency of an index word in the whole document collection, which is generally expressed by IDF (inverse document request). IDF is used to determine whether the words representing the characteristics of the document obtained through TF really have the ability to distinguish this document from other documents and characterize the characteristics of this document, because a word appearing in other documents does not have the ability to distinguish this document from other irrelevant documents.

4.2. Feature extraction method of deep confidence network

Keywords are the selected words that best fit the expression of the text content, which can also be said to be the most representative characteristic words. Assuming that the number of keywords extracted from an article is n, the dimension of the corresponding word vector model is n×50, which greatly reduces the dimension of the input vector and ensures the high fitting of the text input. The improved classification based algorithm is used as the weight basis of keywords, and its weight expression is

$$W(t)=TM(t,C_i) \tag{1}$$

In model training, the energy model is introduced to describe the correlation between the variables of a deep network model, so the specific construction process of the model is the evaluation process of the optimal solution of the parameters in the energy model. Its energy function is defined as

$$E(v,\ h)=-\sum_{i=1}^{m}\sum_{j=1}^{n}w_{i,j}v_ih_j-\sum_{i=1}^{m}a_iv_i-\sum_{j=1}^{n}b_jh_j(\forall_{i,j},\ v_i\in\{0,1\},\ h_j\in\{0,1\}) \tag{2}$$

Equation (2) shows the dependence between the energy function and the nodes of each layer. Among them, m、n correspond to the number of nodes in the visible layer and hidden layer respectively. The parameters a_i、b_j represent the offset of the visible layer node i and the offset of the hidden layer node j respectively. $v_i$and h_j are the states of the visible layer node i and the hidden layer node j respectively, and $w_{i,j}$ is the connection weight between the visible layer node i and the hidden layer node j. By adding the energy of each layer through the energy function, the energy between the connecting structures can be finally obtained. It is required to solve the offset a_i、b_i and weight $w_{i,j}$ of each layer, and the energy function needs to be converted into a probability function.

In order to solve the problem of repeated calculation caused by undirected connection, the input unit is doubled and the weight between the visible layer and the hidden layer is constrained. The conditional probability distributions of the hidden layer node and the visible layer node are

$$p(h_j^1=1|v)=\sigma(\sum_i W_{ij}^1v_i+\sum_i W_{ij}^1v_i) \tag{3}$$

$$p(v_i=1|h^1)=\sigma(\sum_j W_{ij}^1h_j) \tag{4}$$

$$p(h_j^1=1|h^2)=\sigma(\sum_m W_{jm}^2h_m^2+\sum_m W_{jm}^2h_m^2) \tag{5}$$

$$p(h_m^2=1|h^1)=\sigma(\sum_j W_{jm}^2h_j^1) \tag{6}$$

Combine formula (5) and formula (6). For the input v, the conditional distribution of the hidden layer h^1 is

$$p(h_j^1=1|v,\ h^2)=\sigma(\sum_i W_{ij}^1v_i+\sum_m W_{jm}^2h_m^2) \tag{7}$$

The edge distribution $q(h_j^2=1|v)$ and sample data are used as the augmented input of the deep multilayer neural network, and then the whole model is fine tuned through standard backward propagation.

4.3. The experiment, results and analysis of a research group using the model

The data set compiled by network search is divided into 9 categories: scientific papers, conference materials, news trends, patent information, policies and regulations, economic and trade information, industrial information, emergency handling and dynamic standards of colleges and universities. In order to ensure that each information belongs to one category, the number of training samples is 5106 and the number of test samples is 3050. The detailed sample distribution is shown in Table 1.

Table 1: Distribution of training and test samples

category	Number of training samples	Number of test samples
Scientific papers	236	149
Meeting materials	784	446
news information	459	236
Patent information	398	241
Policies and regulations	605	311
Economic and trade information	712	436
Industrial information	537	320

Emergency treatment	663	392
University dynamics	721	519
Technical consultation	352	189
Standard release	277	149
Scientific research budget	558	326
Purchasing information	332	190

The training method and the number of output nodes will affect the classification results, and the accuracy of model text classification is 97.35 /%.

5.CONCLUSION

The purpose of tracking is to comprehensively and systematically collect and analyze the information data at home and abroad related to national defense science and technology, master the cutting-edge technology status and development trend in relevant fields, confirm new technology ideas, prevent low-level research, and provide strong decision-making information support for science and technology management. The key of national defense science and technology tracking is tracking technology and tracking information source, which is the process of intelligent analysis, identification and extraction of information source. It emphasizes understanding the individual and collection of information sources such as scientific and technological reports, scientific and technological papers, conference materials, patents, standards, achievements, reference books, yearbooks, regulations, planning plans, budgets, policies and regulations, technical consultation, news trends, blogs and microblogs from a dynamic, connected and systematic perspective, and labeling the content of national defense scientific and technological information from the perspectives of languages, classifications, keywords, abstracts and named entities, Improve the degree of information disclosure. After optimizing the feature extraction model of text classification, genetic simulated annealing algorithm can be adopted, which can effectively solve some new words and grammars in the search engine. That is, adding simulated annealing algorithm to genetic algorithm can solve the problem of falling into the optimal solution at the place of local convergence, that is, controlling jump out and using the strategy of multiple cycles to search, which improves the efficiency of search engine and the accuracy of search, Simulated annealing algorithm can make up for some defects of genetic algorithm and eliminate local convergence.

The demand for national defense and the development trend of science and technology determine the development direction of national defense science and technology. At present, the national defense science and technology tracking lacks stable data sources, effective tracking methods and quantitative analysis tools, resulting in the information demander at a loss. Aiming at the difficulty of obtaining national defense information and intelligence, this study selects a feature extraction model of text classification in national defense science and technology tracking, studies a new set of national defense science and technology tracking path, and popularizes the theory and application of the research results, so as to provide accurate tracking ideas and tracking tools for national defense science and technology practitioners, professional technology managers, various academic evaluation organizations and other technicians, In order to efficiently and accurately grasp the cutting-edge technology trends and development trends in relevant fields, confirm new technology ideas, avoid excess information, prevent low-level research, and improve the effectiveness of scientific and technological resources. Each period of social development will produce different national defense science and technology information and different ways of information dissemination, with an obvious sense of intergenerational. National Defense Science and technology information basically includes the most advanced and representative national defense science and technology of an era, and records the development history of national defense science and technology in this period. While meeting the information needs of recipients, it provides basis and support for their decision-making and deployment. The prerequisite for carrying out national defense pre research is to master domestic and international cutting-edge technologies. National defense science and technology tracking can provide pre reference for national defense pre research in a certain sense.

REFERENCES

[1] CAI, C. C. (2011).Web page classification technology based on incomplete data set. J. software guide.

[2] CAO, Y., JIA Y. C., WANG, Z. (2019). Research on semantic-based scientific and technological literature retrieval technology. Microcomputer Application, 35:16-18.

[3] LUO,W., TAN Y.S., LUO Z.C.(2018). Development and utilization of national defense science and technology information big data: problems, framework and practice. J. Iinformation theory and practice .

[4] TU, S. Z. & HUANG, M. L. (2016).Mining microblog user interests based on TextRank with TF-IDF factor. J. The Journal of China Universities of Posts and Telecommunications.

[5] WANG, C.(2018). Research on security classification method of documents. D. Master's

thesis of Beijing University of chemical technology .

[6] XU, T. H. & WU, M. L.(2020). Improvement of naive Bayesian algorithm based on TF-IDF J. Computer technology and development two thousand and twenty.
dience (Shinjuku).