


Deep Learning in Chinese Text Information Extraction Model for Coastal Biodiversity

Xiujuan Wang, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, China

Xuerong Li, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, China*

 <https://orcid.org/0000-0002-1754-3795>

ABSTRACT

In the coastal areas of China, scientists have collected nearly 500 species of coastal plants and seaweeds. The collected information includes species description, morphological characteristics, habitat distribution and resource value of plants in China. By effectively extracting Chinese text information, this article establishes a Chinese text information extraction model based on DL. This article is based on short-term and short-term memory artificial neural networks for short text classification. In addition, this article also integrates the L-MFCNN models of MFCNN for short text classification. Comparing the two methods with traditional text recognition algorithms, information extraction based on syntax analysis and deep learning, the results show that, compared with the comparison method, the recognition accuracy of Chinese text information of this neural network model can reach 96.69%. Through model training and parameter adjustment, Chinese text information of coastal biodiversity can be quickly extracted, and species categories or names can be identified.

KEYWORDS

Chinese Text, Coastal Biodiversity, Deep Learning, Information Extraction, Training Model

1. INTRODUCTION

Language is the most important tool for human communication. As one of the carriers of language, text, together with images and videos, constitutes the most important way of data storage. At present, climate change, natural disasters and other reasons lead to faster species extinction, and research on biodiversity conservation and sustainable use has increasingly become the focus of biodiversity research (Muluneh, 2021). Biological research is more important in the face of many biological and related global problems such as environmental degradation and endangered species. Therefore, the information extraction of the Chinese text of the biodiversity environment is more meaningful (Anne, 2012). Coastal biological species are one of the important contents in the field of biodiversity, and research on its diversity has attracted many researchers (Litjens, 2017). The extraction of textual information on coastal biodiversity is the starting point for coastal biological and ecological research. Due to the complexity of species, it is very difficult for researchers to quickly identify all these biological species. Search engines also cannot give accurate species information by species' descriptions (Wu Ying, 2019).

DOI: 10.4018/IJSWIS.331756

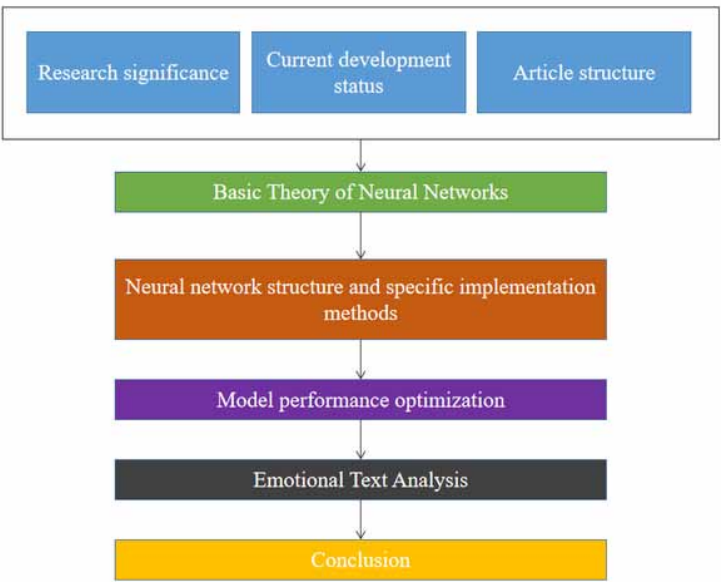
*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

At present, most of the research on biodiversity of Chinese texts focuses on dictionary modeling analysis and machine learning algorithm derivation of shallow learning (Chun, 2018). In 2019, the feature selection of text clustering and the improved krill swarm algorithm were proposed by scholars (Abuligah, L 2019). This paper presents the research results from the following aspects: the process of DL, the text characteristics and Chinese text information extraction for coastal biodiversity, the construction of Chinese text information extraction model, and the application of DL in species identification. Since 2015, the application of genetic algorithm in vector space model information retrieval has been put forward with relevant theories (Abuligah, 2015). Starting from 2017, the unsupervised text feature selection technology based on hybrid particle swarm genetic algorithm has been proposed by relevant scholars (Abuligah, 2017). In 2018, based on the hybrid clustering analysis of the improved krill-herd algorithm, relevant theories were studied by scholars (Abuligah, 2018). The traditional multi-classification text representation and classification method based on bag-of-words model features mainly extracts the low-level features of the text, which has the inherent disadvantages of high dimensionality and high sparseness of the text feature representation vectors. Therefore, the traditional multi-classification text representation and classification methods are difficult to achieve the expected performance. In view of this, it is of great value and practical significance to study the multi-classification text representation and classification methods that extract low-dimensional and dense high-level features of texts.

This paper first introduces the research background and significance of this paper, puts forward the research value of Chinese Sentiment analysis based on deep learning in the field of marine biological recognition, and analyzes the development and status quo of Sentiment analysis and deep learning research at home and abroad. In response to these issues, this article compares the three basic theories of CNN, LSTM, and RM, identifies their advantages and disadvantages, and chooses to use CNN theory to complete this paper experiment. Utilizing deep learning to automate the extraction of Chinese text features, utilizing the powerful classification function of vector machines, and using the proposed algorithm, a system for identifying and analyzing marine organisms is implemented. The article structure is shown in Figure 1.

Figure 1. Article structure



2. THE CONCEPT AND CHARACTERISTICS OF DL

2.1 The Concept of DL

The basis of DL is the distributed representation in the field of machine learning. The distributed representation sets the observations generated by the interaction of different factors. DL establishes a deep network structure to extract features by simulating the mechanism of human brain processing data, thereby gaining the ability to approximate human understanding and processing data. At present, mainstream models of DL include self-encoding, convolutional neural networks, deep confidence networks, and recurrent neural networks (Chen S 2019). There has been a lot of research in the field of natural language processing. The use of deep neural networks for language representation learning has gradually become a new research trend, and the method of DL has also been favored by more and more scholars. At present, the bottlenecks are broken because of the emergence of new training methods for deep neural networks, the great progress in machine learning algorithms, the explosive growth of training data sets, and the huge improvement of processing performance of computer chips, which promote the rise of DL. In the 1980s and 1990s, researchers attempted DL by means of random gradient descent and reverse propagation.

2.2 Characteristics of DL

DL model has a deep hierarchical structure, usually with five, six, or even 10 layers of hidden layer structure. Traditional machine learning models usually have only one layer of network structure, and the model trains single layer features without hierarchy. Among them, outstanding achievements have been made in computer vision, speech recognition, bioinformatics, natural language processing and other fields. The DL model transforms the sample features from the original space to the new feature space by layer-by-layer feature transformation, realizes autonomous feature learning, and then achieves more accurate classification or prediction. The theory of DL originates from the in-depth study of artificial neural networks. It is generally believed that the multi-layer neural network model with multiple hidden layers is a kind of DL model. The field of natural language processing covers a wide range of problems, covering different levels and different natures. This requires us to design corresponding DL models for different types of problems in order to better solve the problem.

3. CHINESE TEXT INFORMATION EXTRACTION MODEL OF COASTAL BIODIVERSITY BASED ON DL

3.1 Chinese Text Information Extraction Preprocessing Technology for Deep-Seated Coastal Biodiversity

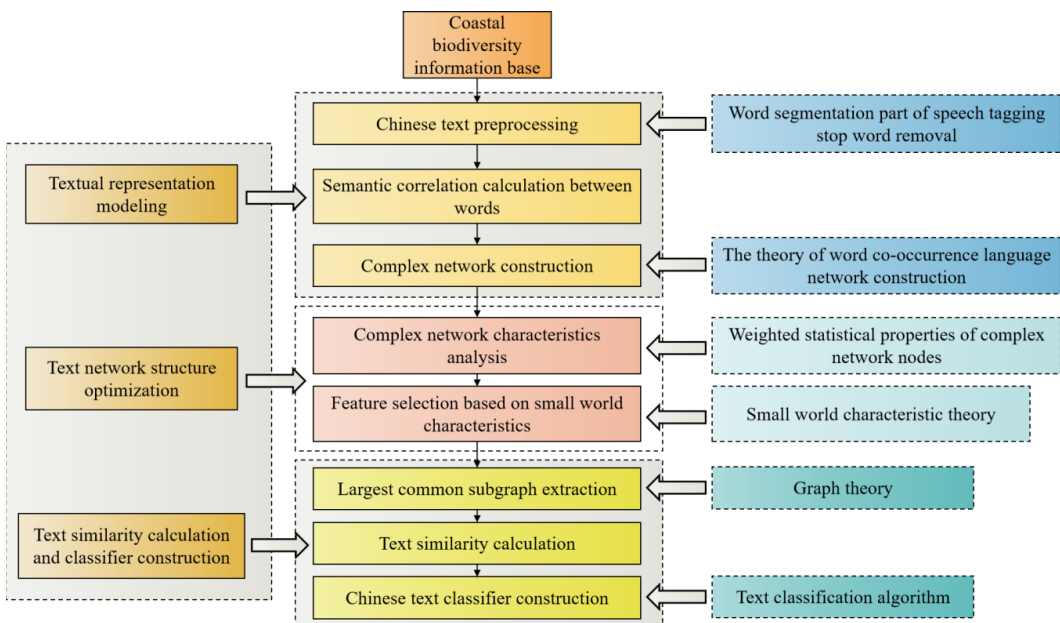
Text preprocessing is the initial step of Chinese text information extraction. The quality of preprocessing has a great influence on the result of information extraction. The depth features learned by the DL model can better represent the properties of the data. The Chinese text information of coastal biodiversity environment is usually unstructured or semi-structured information, which can not be recognized by classifiers. Therefore, Chinese texts in biodiverse environments need to be preprocessed to remove useless information and convert it into structured text. Considering that the model uses coarse-grained word segmentation results as the basic processing unit of the model, the system will have a coarse-grained word segmentation equipment for word segmentation processing in actual use. The words in coastal biodiversity environmental text information can be divided into two categories: functional words and content words. Function words mainly play a role in the rhetoric of the text content and have little to do with the theme of the text. Content words are the content words of text, which play an important role in reflecting the main body of text. Therefore, in order to segment the biodiversity text information, it is necessary to filter some functional words in the text. Preprocessing includes Chinese word segmentation, word segmentation filtering, vectorization and so on. After

word segmentation, Chinese texts usually contain a certain number of function words, such as “de, yes, in, le, ah”. These words have no real meaning, they are called stop words. Generally, in order to improve the efficiency of text classification, stop word filtering is used for text classification tasks.

At present, feature extraction using DL technology is widely used in the image field, and there are relatively few literature reports on text feature extraction, especially the research results of feature extraction for Chinese texts are scarce. At the same time, the feature dimensions extracted by traditional manual feature extraction methods are usually large, which makes the model training inefficient and consumes resources. Therefore, using DL method to extract features from Chinese long text data sets can reduce the difficulty of text feature extraction, improve the efficiency of model training, and at the same time, it can represent the semantic information of text more accurately. Parsing tree is an ordered tree with root nodes, which represents the syntactic structure of a string or a sentence. It was originally used in computational linguistics. Syntax tree reflects the grammar of the input language, and its construction is usually based on the component relation of component grammar or the dependency relation of dependency grammar, which can be divided into two specific syntax trees, namely component syntax tree and dependency syntax tree. The process of Chinese text classification based on complex network is shown in Figure 2.

In distributed database system, there are two levels of data sharing: one is local sharing, that is, users on the same site can share the data in the local database of this site, and have completed local application. The second is global sharing, that is, users on the distributed database system can share the data stored on each site of the distributed database system to complete the global application. With the increasing expansion of text tag set space, the output space of multi-tag text representation and classification tasks has increased exponentially, which finally leads to the difficulty of multi-tag text representation and classification model in predicting the tag set to which the text belongs. Assume that there are two nodes i and j in the network composed of N nodes. The distance between the two nodes is defined as the shortest path length from node i to node j , and this distance is defined as $l(i, j)$. For undirected graphs, the average path length can be defined as Equation 1:

Figure 2. Chinese text classification process based on complex network



$$i = \frac{1}{\frac{1}{2} N(N-1)} \sum_{i>j} l(i, j) \quad (1)$$

The clustering coefficient C_i of i node is defined as a connection relationship between nodes directly connected with node i in the network, which is the ratio of the number of edges between nodes i directly adjacent to nodes to the whole maximum possible number of edges. C_i is defined as Equation 2:

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \quad (2)$$

where k_i is the degree of node i , and e_i is the number of edges between nodes connected with node i . It can be seen that C_i describes local values. Therefore, the clustering coefficient C_i of a complex network can be expressed by the average of the clustering coefficients of all nodes(as show in Equation 3):

$$C = \sum_{i=1}^N C_i \quad (3)$$

k_i represents the number of nodes related to node i . Therefore, the average degree of the network can be calculated from the average of all node degrees. The average degree \bar{k} of the network is defined as Equation 4:

$$\bar{k} = \frac{1}{N} \sum_i k_i \quad (4)$$

The diffusion in nodes can be expressed by cumulative distribution function $p(k)$ (as show in Equation 5):

$$p(k) = \sum_{i=k}^{\infty} P(i) \quad (5)$$

The betweenness of i nodes is defined as the ratio of the shortest paths through i nodes to all the shortest paths in the network. Generally speaking, the greater the betweenness of a node, the greater the dependence of information on that node when it spreads in the network. Given an input sequence $x = \{x_1, x_2, x_3 \dots, x_t\}$, calculate the hidden state h_t of neural network feedback (as show in Equation 6):

$$h_t = f(h_{t-1}, x_t) \quad (6)$$

$f(x)$ is a nonlinear function, h_{t-1} is the hidden state of the previous moment, and its state information is determined by the hidden state of the previous moment and the current input. The sharing of biodiversity information involves different research and application units and organizational structures. The services shared by these application units and organizations may be developed based on different GIS platforms. In the process of sharing and using these services, the problem of heterogeneous platforms arises. Distributed database adopts the control mechanism combining centralization and autonomy. It has a hierarchical control structure based on global database administrators, but each local database administrator has a high degree of autonomy. Several nodes are connected together through a network. Each node is an independent spatial database system, and they all have their own databases, corresponding management systems and analysis tools.

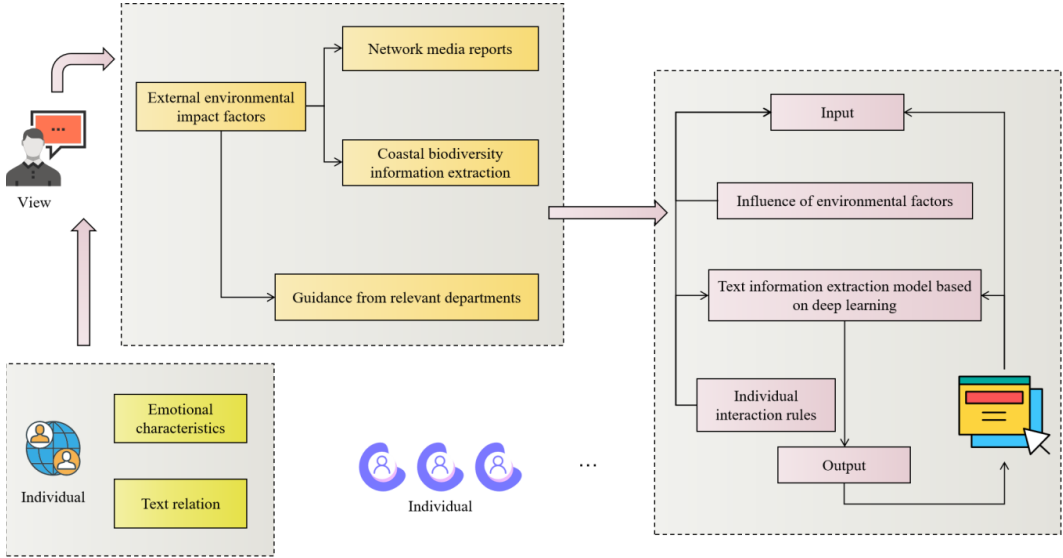
Although the amount of information in Chinese texts of coastal biodiversity is no longer the main contradiction restricting information extraction, it is both an opportunity and a challenge for information-driven DL. Early information extraction is based on the surface feature to select the correct answer. The surface feature is a simple and effective method of information extraction. It is absolutely impossible to collect textual information in coastal biodiversity environment manually in order to find the required textual information from the vast amount of information. We must rely on DL technology to collect the required text information. At present, the application of the DL method to extract the text information of the dark pocket biodiversity is mostly directed to a specific text information, and the scope of the model is mostly aimed at the aspects of interest of the researchers, that is, the establishment of a regional model. Even with the same training text information, different text information organization or input methods will produce different training results, which challenges the applicability of the model in other sea areas or more, that is, the versatility of the model needs to be upgraded. The multi-dimensional information classification and prediction module based on DL is the core module of the information analysis model. The main function is to classify the pre-processed Chinese text of the coastal biodiversity environment and predict its multi-dimensional information trend.

3.3 Application of DL in Text Information Extraction

The accuracy and spatiotemporal continuity of Chinese text Information extraction of coastal biodiversity is the key to Chinese biological data extraction in coastal waters. However, in practical applications, interpolation can cause significant information loss in data reconstruction, resulting in significant errors in the reconstruction results, which poses a huge challenge to existing reconstruction algorithms. On this basis, a DL based method for extracting spatial information of coastal biodiversity was proposed. In the construction process from low resolution to high resolution, unlike traditional extraction methods, current research mainly uses large samples to obtain the transformation rules from high to low resolution to construct models. At present, research on offshore biodiversity mainly focuses on a single piece of information. When text Information extraction is carried out for coastal biodiversity, first of all, it is necessary to identify the objects to be extracted for coastal biodiversity, that is, to establish a framework for semantic description of the knowledge required for coastal biodiversity, so as to carry out semantic description of the extracted knowledge, and effectively describe the extracted knowledge, so as to provide necessary support for the subsequent display of knowledge structure.

Before the text feature extraction system analyzes the data, it is necessary to clarify the content of public opinion information. With the rapid increase of network data, the classification performance of these methods will decline when dealing with large-scale data. The main reason is that massive text information and new words emerging from the continuous change and development of network vocabulary make it difficult for traditional machine learning methods to fully extract training text features. In order to overcome this shortcoming, an improved coastal biodiversity information identification scheme was put forward. The text feature extraction model based on DL technology is shown in Figure 3.

Figure 3. Text feature extraction model based on DL



When extracting word vector features, we can infer the meaning of a word through the context of the context, and further judge the emotional content of the word better. As for the word similarity calculation of statistical model, since the statistical model uses word vector to assign each word a multidimensional vector to represent its position in the lexicon, the word similarity of two words can be quantified by calculating the angle between two word vectors (as show in Equation 7):

$$sim_1(W_1, W_2) = \cos \theta = \frac{\vec{v}_1 * \vec{v}_2}{|\vec{v}_1| * |\vec{v}_2|} \quad (7)$$

where \vec{v}_1 and \vec{v}_2 are word vectors of W_1 and W_2 . When two words are synonyms or identical words, the angle between the two word vectors is 0, and the word similarity is 1. When the two words are completely different, the angle between the two words vectors is 90, and the similarity of words is close to 0. All sememe form a hierarchical system according to the hyponymic structure. In this system, the distance between words reflects the semantic similarity between words(as show in Equation 8):

$$\begin{cases} \lim_{d \rightarrow \infty} sim(W_1, W_2) = 0 \\ \lim_{d \rightarrow 0} sim(W_1, W_2) = 1 \end{cases} \quad (8)$$

The shorter the distance between words, the more similar they are, and the longer the distance between words, the less similar they are. Word similarity can be quantified by calculating semantic distance. Calculate the semantic distance between two sememe (as show in Equation 9):

$$sim_2(W_1, W_2) = \frac{\alpha}{d(W_1, W_2) + \alpha} \quad (9)$$

where W_1 and W_2 represent two sememes, and $d(W_1, W_2)$ represent the length of the shortest path of W_1 and W_2 in the sememe hierarchy. α indicates an adjustable parameter. With the process of simultaneous propagation of past and future information, we can effectively use the past characteristics, that is, the forward propagation state and the future characteristics, that is, the backward propagation state. Through the expanded forward and backward networks in an effective time, we can obtain all the required hidden state values.

When extracting Chinese textual information of coastal biodiversity based on DL, we first use this model to judge the information type of textual information, and then extract temporal and spatial attribute information of biodiversity texts based on the spatio-temporal attribute knowledge framework of the information type, using sentences as units. DL can learn complex models from a large number of sample data and control the efficiency of the models. It plays an important role in solving the problem of text information extraction from coastal biodiversity. Use the correct extraction results of the wrapper to learn the characteristic information of the text, including the pattern information of the biodiversity text, annotation information and possible hyperlinks. Applying DL to the extraction of biometric Chinese text information, the low resolution data can be resized by bicubic interpolation to make it the same as the number of pixels or grid of the target high resolution data, and as a network. Input. Therefore, the application of DL to the Chinese text information extraction of coastal biodiversity has very fast efficiency and good accuracy.

3.4 Text Classification Method Based on Semantic Web

The semantic web is machine understandable information, which is a data network or global database. The Internet Alliance defines the semantic web as the expression of data on the Internet, which is an extension of the current Internet, as information has clear and clear meanings, enabling better cooperation between humans and computers. The main component of a text classification system based on the semantic web is a set of words and terms that represent knowledge in a certain field. The compiler organizes these words and terms into hierarchical categories based on the structure of the knowledge field and provides more detailed definitions for some categories as needed.

Generally, support vector machine (VSM) methods are used. The average accuracy of the classification system based on traditional VSM is 71.3%, and the average matching rate is 78%; Even with the introduction of hierarchical weight coefficients, the improved VSM text classification system has an average accuracy of 85.6% and an average matching rate of 86% (as show in Table 1). Therefore, it can be seen that the semantic web is significantly higher than the Vector space model (VSM) in the accuracy of classification, and the matching rate is also better than the traditional VSM classification system in the field.

Table 1. Semantic Web Testing Experimental Results

| DDC class features | Semantic Web |
|--------------------|--------------|
| | Accuracy |
| DDC306 | 87.6% |
| DDC335 | 91.3% |
| DDC355 | 90.7% |
| DDC571 | 87.6% |

4. ANALYSIS AND DISCUSSION OF RESULTS

4.1 Experimental Dataset

The purpose of this experiment is to train three deep learning based models using sentiment labeled text datasets, so as to achieve good results in text sentiment classification. The word vector in this section is trained by word2vec on the basis of Chinese Wikipedia training corpus. The dimension of the word vector is 150 dimensions. The dataset consists of three parts: a dataset containing 3000 samples for entity class problems; The NLPCC 2015 Q&A evaluation task has a question set of 1000 samples and a question classification dataset of 9000 samples in total.

4.2 Experimental Results and Analysis

Due to the high dimensionality of features, the model training efficiency is low, resources are occupied heavily, and overfitting, fusion, and other issues are prone to occur, thereby affecting classification accuracy. On this basis, a Chinese text extraction model for coastal biodiversity was established using DL language. On this basis, further improve the extraction model to make it more suitable for extracting various types of text information. Systematic, complete, accurate, and timely biodiversity data and information are the data foundation for biodiversity conservation and sustainable utilization, as well as the key to scientific decision-making and management. They are also an important basis for the country to formulate biodiversity conservation strategies and plans, as well as relevant policies and regulations. By analyzing the data, a topic based topic mining method was proposed. Figure 4 shows the specific format of some data.

The component tree splits a sentence into sub-phrases. The dependency tree connects words according to their relationship. Each node in the tree represents a word, and the child nodes exist by the parent node. The connection marks the relationship between them. Traditional data mining algorithms have data loss, so it is necessary to conduct experimental analysis on the coastal biodiversity information model based on DL. The test results are shown in Figure 5.

It can be seen from Figure 4 that the proposed model basically converges when the training iteration reaches the 10th round, and the loss function value is only 0.26, so the training loss is small. This shows that the proposed model can achieve rapid convergence, and the training effect is ideal.

Figure 4. Time series diagram of coastal biodiversity topic

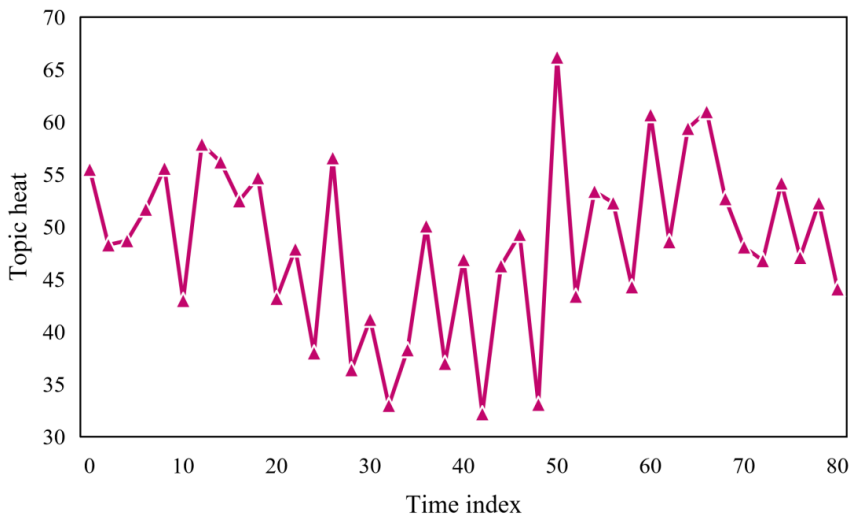
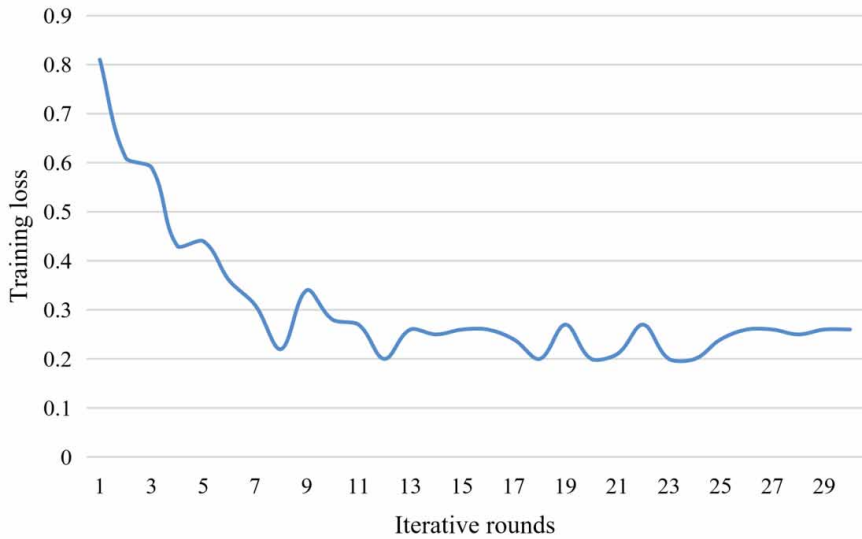


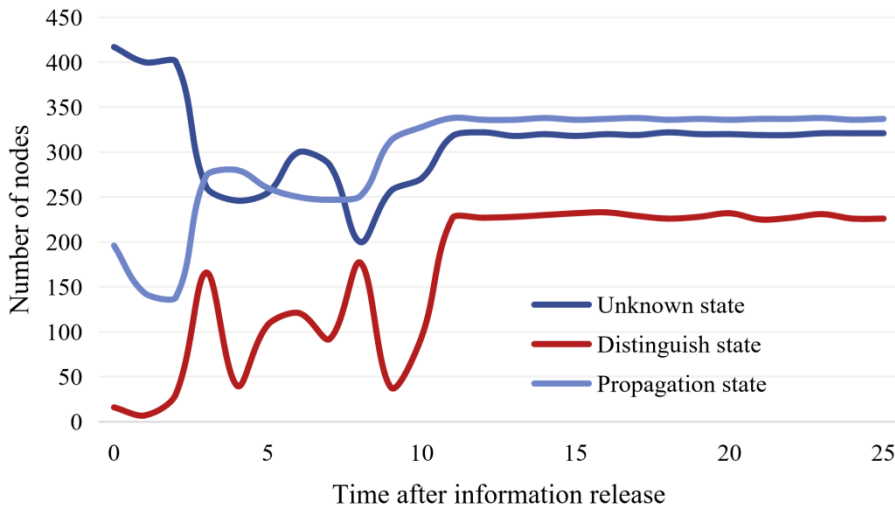
Figure 5. Training loss



In the experiment, the attenuation rate is fixed at 0.1, and the change of the number of nodes in each state during the propagation process is simulated, and the curve of the number of nodes in each state with time is obtained, as shown in Figure 6.

According to Figure 6, it can be seen that at the initial stage of coastal biodiversity information release, most user nodes are in an unknown state. At this time, the number of propagation nodes, discrimination nodes and immune nodes are all small. For the resolution of anaphora of relative pronouns, the two methods have the same experimental results, although this paper uses syntactic analysis rules and the fusion method uses a classifier model based on syntactic path features. This is because, compared with other types of anaphora, relative pronouns and their antecedents are usually

Figure 6. Changes of the number of nodes in each state with propagation time



located in the same sentence and close to each other, which makes it easy for both the rule method and the learning method to discover the potential relationship between them.

As can be seen from Figure 7, the running time of the proposed methods is lower than that of other comparison methods, and with the increase of data volume, the upper limit of analysis data volume appears in comparison methods. The coastal biodiversity information neural network model used in this paper can analyze and process the data of large flow, and has the shortest time and good comprehensive performance.

In order to solve the problem that the feature dimension is too large and the efficiency is low, we usually reduce the dimension of the features extracted by traditional methods, and select a subset of features that can best represent the text information and make the classification effect the best, but this will increase the workload of the classification task. Text usually contains words, phrases, sentences, paragraphs, spaces and other elements. In the text representation and classification algorithm, these elements or the combination of elements in the text are taken as the features of the text. However, not all elements can be extracted as effective features.

For the resolution of anaphora of relative pronouns, this paper uses the rules of syntactic analysis, while the fusion method uses the classifier model based on syntactic path features. This is because, compared with other types of anaphora, relative pronouns and their antecedents are usually located in the same sentence and close to each other, which makes it easy for both the rule method and the learning method to discover the potential relationship between them. In order to evaluate the analytical accuracy of the proposed model, it is compared with FCA and ID3, and the results are shown in Table 2.

As can be seen from Table 2, the analysis accuracy of the proposed method is higher than that of other comparison methods. The main reason is that the neural network algorithm can deeply excavate

Figure 7. Comparison of data size and time

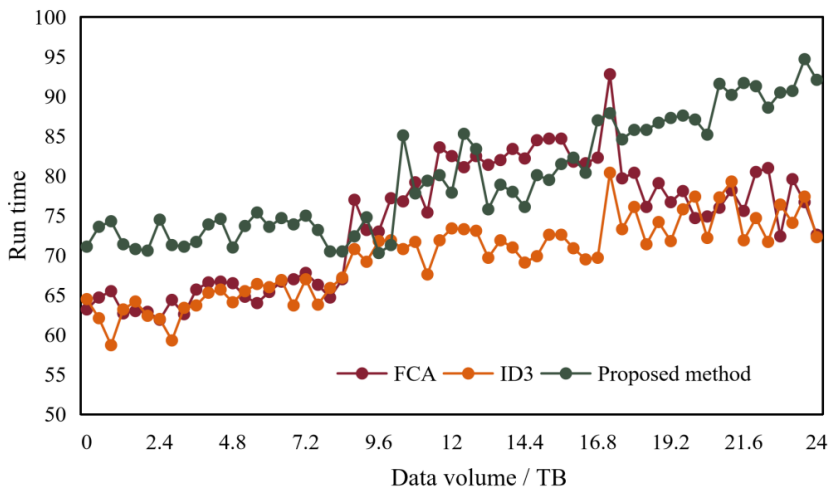


Table 2. Comparative analysis of accuracy of different methods

| Method | Data mining accuracy (%) | Accuracy of feature extraction of coastal biodiversity information (%) |
|-----------------|--------------------------|--|
| FCA | 86.55 | 86.86 |
| ID3 | 90.25 | 88.88 |
| Proposed method | 97.38 | 97.69 |

the internal relations existing in coastal biodiversity information, so as to achieve high accuracy analysis. FCA is only theoretically analyzed, lacking the test of big data, so the accuracy rate is not ideal. ID3 The outlier measurement method is used to mine data information, but the accuracy rate of data analysis with rapid growth is low. The comparison results of prediction accuracy are shown in Table 3 and Figure 8.

Through the experimental results in Table 3, it can be seen that in the analysis of coastal biodiversity information, the neural network model of coastal biodiversity information has achieved very good prediction results in accuracy, precision, recall rate and F1 value. Compared with the comparison method, the recognition accuracy of Chinese text information of this neural network model can reach 96.69%. Therefore, it can be explained that the construction of Chinese text information extraction model based on DL is reasonable, which is beneficial to the representation and classification of coastal biodiversity information. Through comparative analysis, the practicability of the model is verified. The text information extraction model based on DL has good text information extraction effect, improves the efficiency and accuracy of text information extraction and reduces the failure rate of text information extraction. Facts have proved that traditional text information extraction methods are often influenced by many factors, and can't achieve the expected results. The application of in-depth study brings new opportunities for the extraction of Chinese text information in coastal biodiversity environment.

5. CONCLUSION

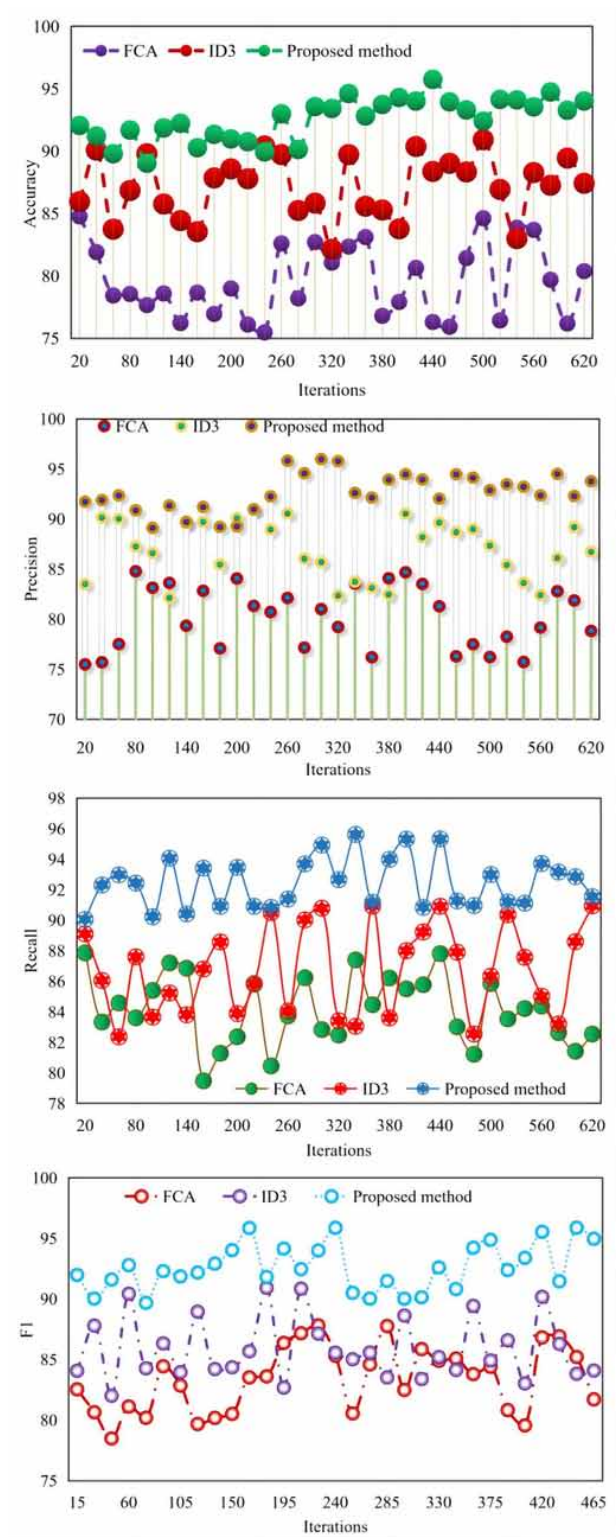
In this paper, DL is applied to Chinese text information extraction of coastal biodiversity environment, and a model of Chinese text information extraction of coastal biodiversity is proposed, and the adaptability of the extraction model to various types of text information extraction is tried to improve. The text information extraction model based on DL has good text information extraction effect, improves the efficiency and accuracy of text information extraction and reduces the failure rate of text information extraction. The amount of data has a great influence on the convergence of the DL model. By increasing the number of training samples, the reconstruction error of the model is reduced and the convergence speed of the model is improved. Facts have proved that traditional text information extraction methods are often influenced by many factors, and can't achieve the expected results. In the analysis of coastal biodiversity information, the neural network model of coastal biodiversity information has achieved good prediction results in accuracy, precision, recall rate and F1 value. Compared with the comparison method, the recognition accuracy of Chinese text information of this neural network model can reach 96.69%. Therefore, it can be explained that the construction of Chinese text information extraction model based on DL is reasonable, which is beneficial to the representation and classification of coastal biodiversity information. In some aspects, we can foresee the development trend of DL theory, that is, under the background of big data, the architecture of DL will quickly become larger and more complex, and these architectures will become a part of the future innovative architecture.

This paper proposes a deep learning method to optimize the shortcomings of current Information extraction methods. The method of extracting syntactic structure features using dependency

Table 3. Model comparison

| Method | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) |
|-----------------|--------------|---------------|------------|--------|
| FCA | 88.75 | 89.99 | 87.45 | 88.88 |
| ID3 | 93.75 | 90.77 | 87.22 | 88.77 |
| Proposed method | 97.79 | 95.79 | 92.77 | 91.69 |

Figure 8. Model comparison



syntax trees incorporates commonly used surface features, while using word vectors to represent understanding problem sentences and candidate information sentences.

Further research work in the future includes collecting more corpus and combining traditional feature extraction methods to reduce the impact of insufficient corpus and improve the accuracy of short text classification; Optimize the strategy of extracting candidate information sentences, and formulate corresponding Information extraction strategies for different problem types.

ACKNOWLEDGMENT

The work was supported by the Informatization Plan of Chinese Academy of Sciences, Grant No. CAS-WX2021SF-0408.

REFERENCES

- Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *The Journal of Supercomputing*, 73(11), 4773–4795. doi:10.1007/s11227-017-2046-2
- Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2017). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science*.
- Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). Hybrid clustering analysis using improved krill herd algorithm. *Applied Intelligence*, 48(11), 4047–4071. doi:10.1007/s10489-018-1190-6
- Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A Combination of Objective Functions and Hybrid Krill Herd Algorithm for Text Document Clustering Analysis. *Engineering Applications of Artificial Intelligence*, 73, 111–125. doi:10.1016/j.engappai.2018.05.003
- Abualigah, L. M. Q. (2019). Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering. *Studies in Computational Intelligence*.
- Abualigah, L. M. Q., & Hanandeh, E. S. (2015). Applying genetic algorithms to information retrieval using vector space model. *International Journal of Computer Science. Engineering and Applications*, 5(1), 19.
- Anne E. Thessen, Hong Cui, Dmitry Mozzherin.(2012). Applications of Natural Language Processing in Biodiversity Science[J]. *Advances in Bioinformatics*, vol. 2012, Article ID 391574, 17 pages. doi:10.1155/2012/391574
- Chen, S., & Demachi, K. (2019). Proposal of an insider sabotage detection method for nuclear security using deep learning [J]. *Journal of Nuclear Science and Technology*, 56(14), 1–9. doi:10.1080/00223131.2019.1611501
- Chun, M., & Xiao, H. (2018). Text information extraction algorithm of video images in multimedia environment [J]. *Journal of Discrete Mathematical Sciences and Cryptography*, 21(2), 305–310. doi:10.1080/09720529.2018.1449304
- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review [J]. *Neurocomputing*, 187(C), 27–48. doi:10.1016/j.neucom.2015.09.116
- Jung J, Sohn K (2017) Deep Learning Architecture to Forecast Destinations of Bus Passengers from Entry-only Smart-card Data[J]. *IET Intelligent Transport Systems* 11(6), 1.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis [J]. *Medical Image Analysis*, 42, 60–88. doi:10.1016/j.media.2017.07.005 PMID:28778026
- Muluneh, M. G. (2021). Impact of climate change on biodiversity and food security: A global perspective—a review article [J]. *Agriculture & Food Security*, 10, 36. doi:10.1186/s40066-021-00318-5
- Ying, W., Li, S., & Zhou, G. (2019). Classification of Film Review Speciality Based on LSTM and Multi-feature Combination [J]. *Computer Science*, (B06), 74–79.