Application of Natural Language Processing Technology in Educational Resources Retrieval

Yibao Huang*

Guangdong Preschool Normal College in Maoming, Maoming, China

*Corresponding author e-mail: huangyibao@gdpnc.edu.cn

Abstract. According to the data structure characteristics of educational resources, natural language processing technology is used to process the original educational resources data from the aspects of word segmentation processing, named entity recognition, part of speech tagging, synonym analysis, word vector analysis, etc. For complex image data, OCR preprocessing and coding format conversion are used to reduce the complexity of the original data; according to the difficulty requirements of the user to retrieve data, key words or ontology are selected to further expand and enrich the semantics and improve the accuracy of the retrieval results.

Keywords: Natural Language Processing, Educational Resources Retrieval, OCR Pre-processing, Keywords, Ontology Retrieval

1. Introduction

With the rapid development of information technology and network technology, the total amount of information related to educational resources is gathering and expanding, and the database resources among schools are relatively independent [1, 2]. At present, there is no large range of school information resource sharing platform. The above factors jointly lead to the greater difficulties for users in the process of educational information retrieval [3, 4]. The non-disclosure and non-transparency of information, on the one hand, will lead to the waste of a large number of educational resources, which cannot give full play to the maximum effectiveness within its validity period; on the other hand, in the long run, it will also lead to the insufficient distribution of resources in the education system and unfair education [5]. The original education resource data is disordered and belongs to the typical multi-source heterogeneous data. However, the existing information retrieval methods, such as directory retrieval and full-text retrieval, have some problems, such as insufficient comprehensiveness, single retrieval method, lack of personalization, etc., and full-text retrieval will face the disadvantages of high cost. In recent years, with the promotion of information technology and big data cloud computing technology, AI technology and machine learning algorithm have made great progress. AI and machine learning have more advantages than traditional data acquisition and processing methods in dealing with massive, irregular and multi-source heterogeneous data sets. Therefore, on the basis of AI technology, this paper proposes an educational resource retrieval algorithm based on natural language processing. As one of the important branches of computer science and artificial intelligence, natural language processing (NPU) truly realizes the
interaction between human and computer, which is a milestone on the road of human exploration of AI. At present, natural language processing has become the core component of desktop computer and intelligent terminal chip. Natural language processing technology will play a very important role in processing massive educational resources information.

2. Classification of Natural Language Processing Technology

For the characteristics of Chinese characters and vocabulary, natural language processing technology is divided into word segmentation processing technology, named entity recognition, part of speech tagging, synonym analysis, word vector analysis, dependency grammar analysis, word position analysis, semantic normalization processing, text error correction technology, text similarity analysis technology, and document classification technology. Among them, word segmentation is the basis and one of the most important links of natural language processing, that is, to recombine words according to certain rules to help users better understand sentences or data. Named entity recognition is a basic task of natural language processing, that is to identify useful references from the text to be detected, which lays the foundation for subsequent task search and knowledge recognition. Taking the retrieval of educational resources as an example, using named entity recognition technology, we can retrieve the information needed by users from massive text resources by keywords. Part of speech tagging is also defined as part of speech tagging, which gives each object to be queried a specific label according to the part of speech of the word, and searches the task according to the label. Part of speech tagging methods mainly include fuzzy maximum entropy, statistical probability maximum method and so on.

There are many synonyms in Chinese vocabulary; too many synonyms will cause data redundancy in the process of retrieval [6]. Natural processing technology, local synonym analysis function, can judge the reference relationship between synonyms in different contexts. Word vector analysis is also one of the core technologies of natural language processing. Vector analysis transforms natural language and words into multi-dimensional vectors. For words with similar semantics and meaning, vector analysis can make more accurate output. Word vector generation methods include semantic based generation method and statistics based generation method. No matter which way to generate word meaning, we can use a variety of distributed representation of word vector.

The analysis of dependency grammar mainly analyzes the sentence structure from the perspective of the dependency relationship between words and sentence components, including subject predicate relationship, verb object relationship, and concurrent relationship and so on. Due to the special grammatical structure of Chinese, the position of words in sentences is very important, that is, word order will have an important impact on the expression of sentence meaning. For example, according to Chinese grammar, attributives, adverbials and other modifier words appear in front of the main components, so we can infer that the probability of keywords appearing in the middle or at the end of the sentence is significantly higher than that at the beginning of the sentence. In addition, the position of words in sentences is also helpful for resource retrieval and automatic extraction of article tags. The significance of semantic normalization is to eliminate the redundant words of repeated or similar expressions to the maximum extent. After normalization, the retrieval task is further digested, making the natural language expression more refined and indirect. Text error correction and tag extraction are the expansion functions of natural language processing [7]. Text tag extraction can quickly grasp the main content of retrieval. At the same time, text tags are also very helpful for the recognition of text lexical errors and semantic errors. In this paper, the tag contains specific and simplified language features. By observing the tag, we can identify whether most of the text contains error elements, and avoid too many text, language, and text editing and replacement errors. The similarity recognition of text is considered from the similarity of specific words and the similarity of text editing order. Because the grammar rules of modern text are not strict, the analysis of text similarity should be based on the normalization of text. In addition, the fuzzy information entropy algorithm can be used to calculate more complex text. When the subject model searches, it searches the required documents according to the keywords entered in advance. The subject retrieval has stronger purpose and relevance, and can strengthen the understanding of the subject of the retrieval content. Document classification is a
special natural language processing technology. When the content of the article to be retrieved is more complex, it is necessary to classify the complex content according to the text features obtained after the initial screening [8]. On the one hand, document classification can reduce the complexity of the overall retrieval content; on the other hand, it can improve the retrieval efficiency of key content.

3. The Application of Natural Language Processing Technology in the Retrieval of Educational Resources

Because the retrieval objects come from different types of documents, the natural language processing technology is used to preprocess the document with OCR and transform the coding format, and then normalize the document and transform other text formats.

3.1. OCR Processing

The educational resources in picture format are the most difficult to retrieve, query and use. When using natural language retrieval technology to query the corresponding picture materials, first do OCR recognition on the target document, and convert the picture format into text format. The implementation process of OCR is shown in Figure 1:

3.2. Semantic Analysis

Semantic analysis is the core of natural language processing technology [9]. The accuracy of analysis determines the final data retrieval accuracy. Processing text is to extract the detailed features of text data, and get the core information and content of the text according to the features. Firstly, feature extraction can transform unstructured original data into structured data output to meet the requirements of artificial intelligence and machine learning. Different feature extraction algorithms are suitable for different educational texts. In the task extraction process of project text, a sequence is
demarcated for each element, so that the input item is text data and the output item is sequence label. The final sequence label is used to decode to improve the efficiency and accuracy of educational resource retrieval. Deep learning algorithm in machine learning is a common algorithm in semantic analysis. Through the comparative learning of multi-layer neural network, more data can be processed in a shorter time interval. As an automatic way of feature extraction and learning, deep learning is divided into supervised learning mode and unsupervised learning mode. In intelligent text processing mode, supervised learning mode is mainly adopted, which can effectively assist document management through background parameter monitoring. The output result cannot be called directly. Finally, the format conversion and result filtering of the output text data are needed. The output result is more in line with the user's habits.

3.3. Retrieval Application of Natural Language Processing Technology
For information retrieval of educational resources with low complexity, information retrieval based on core keywords is usually adopted. The specific retrieval process is shown in Figure 2 below.

![Figure 2. Basic process of keyword based retrieval](image)

The choice of keywords can be a word or phrase, or multiple words or phrases. The more keywords you choose, the smaller the search target, but the more accurate the result is [10]. The less keywords you search, the more related terms you can retrieve. Keywords should not include words or phrases with the same or similar semantics. If retrieval users have a deviation in the semantic understanding of keywords, it will directly affect the final retrieval and query results.

When the educational information resources need to be searched are more complex or professional, it is often difficult to obtain better retrieval results by searching keywords. At this time, ontology retrieval should be used. The basic steps of ontology based retrieval are as follows:

Step1: according to the type of resource information source, build the related retrieval subject.

Step2: extract the basic features of the main information in the information source, and label the semantic sequence into the source database according to the user's retrieval target to form the ontology retrieval information database.

Step3: users summarize and abstract query statements, and use natural language processing technology and deep learning algorithm to filter target information. The statements of compound conditions in the source database are called into the ontology retrieval information database.

Step4: after searching, the result is returned to the user.

The process of ontology based semantic retrieval is shown in Figure 3
After semantic annotation and natural language processing, the document database constructed in ontology semantic retrieval has more feature matching coverage than keyword retrieval. In the process of feature extraction and concept processing of relational information, semantics is further expanded and enriched, and the retrieval results are more accurate.

4. Conclusion and Prospect
At present, natural language processing technology has been widely used in the field of information retrieval. Due to the small sample size of school education resource information and the complex data type, the effectiveness of natural language processing technology can be brought into full play [11]. Whether it is keyword based or ontology based retrieval, through text classification, OCR recognition and moral retrieval of natural language technology, we can get more accurate results. However, the scale of data in the field of education is still expanding. Natural language processing technology needs to further improve the semantic retrieval system, vector analysis performance, label automatic extraction efficiency and so on. In addition, we should constantly expand the overall scale of the basic resource database, and improve the analysis and evaluation function of natural language processing technology, so as to better serve the information retrieval work [12].

References
[1] Yao Hao, Ye Zhong, Shao Liping. An Empirical Study on the education quality gap of junior middle schools in China -- Based on the sampling analysis of 112 middle schools in CEPS (2013-2014). Education academic monthly, 2019 (08): 42-51
[2] Zhang Jinliang. Intellectualization: the inevitable trend of school education development and Reform in intelligent society. Education exploration, 2018, (01): 13-16
[3] He Xiangyang. Research on information resource allocation mode of education group based on "co creation and sharing". He Xiangyang. China audio visual education, 2018, 000 (001): 91-96105
[4] Wu Fang, Liu shanhuai. Space, obstacle and path of information technology to bridge the gap between urban and rural education development. China audio visual education, 2020, (02): 35-41
[5] Zhang Dong, Guo Nana. Problems and Countermeasures of sharing high quality education resources in Ideological and political courses of Application-oriented Universities. Education and career, 2020, (07): 98-104
[6] Tian Fangjie, Cheng Chengqi, Pu Guoliang. Massive photo organization retrieval method based on geosot spatiotemporal coding. Surveying and Mapping Science, 2018,43 (02): 77-82
[7] Zhang Quanming, Cao Kaijiang, Luo Yi, et al. Application of Lucene based full text retrieval technology in power project plan review. Science, technology and engineering, 2018, 18 (18): 193-196
[8] Zhang Xiaoyu. Application of fuzzy clustering in optimal decision making [D]. Southwest Jiaotong University, 2016

[9] Li Xiuge. Research on fuzzy clustering theory based on fuzzy equivalence matrix [D]. Liaoning University, 2015

[10] Xie Jing. Evaluation of quality engineering construction level based on fuzzy clustering maximum tree algorithm [J]. Journal of Hubei second normal university, 2014, 31 (08): 47-49

[11] Su Qingfang. Research on the construction of internal quality assurance system in application oriented universities [J]. Continuing education research, 2021 (02): 143-145

[12] Liu Qun. Research and application of intuitionistic fuzzy clustering analysis based on entropy