ISSN: 2688-9323

Application of Text Classification in Information Retrieval

Xinyu Zhang, Yanzhihui Cao

Anhui University of Finance and Economics, Bengbu 233000, China

Abstract

In recent years, with the continuous development and wide application of new media technology, search engine has become one of the main ways to obtain information. However, with the explosive growth of information, how to quickly and accurately obtain the required information has become a major problem. In order to solve this problem, text classification technology came into being. The content of this paper is the text classification in the field of information retrieval. Text categorization is the process of automatically categorizing a text into a predefined text category. It can solve the problem of messy information to a great extent, and help us find the needed information more quickly. At the same time, it also provides more efficient search strategy and more effective search results for information retrieval, improving the user's search experience. In the research of text classification technology, how to select appropriate features has a great impact on the effect of text classification. At present, the common feature selection methods include chi-square test, mutual information and information gain. In addition, machine learning algorithms are also widely used in text classification, such as Naive Bayes, Support Vector Machines and so on. To sum up, text categorization is of great value in information retrieval. With the continuous development of the Internet, text classification technology is constantly developing and improving, providing us with more efficient and accurate information retrieval services.

Keywords

Information Retrieval; Text Classification; Application Necessity.

1. Introduction

With the rapid development of computer technology, the popularity of the Internet and the wide use of the network, people can access more digital information, but it takes more time to process these information. Therefore, people began to use computers to achieve automatic text classification. Character classification refers to the automatic recognition of characters related to characters through the analysis of characters in a certain classification system. Text classification is a subject that combines artificial intelligence and information search technology. In recent years, a lot of research work has been done on text classification and its related information retrieval and extraction. This paper makes an in-depth study of text classification. According to the textual features, the frequency, meaning and order of words can be classified. At the same time, this method can also be used in other fields, such as emotional analysis, spam filtering and so on. At the beginning of the research, automatic text categorization is based on the knowledge engineering method and the experience of the experts in the field. With the continuous development of machine learning, scholars have begun to study various classification algorithms based on machine learning, including supervised learning and unsupervised learning.

ISSN: 2688-9323

2. Definition of Text Classification

Text classification is the process of automatically dividing natural text files into one or more predefined categories according to their content. This is a kind of instructional learning, through the marked training document set, find the relationship model between the characteristics of the document and the document category. Text classification is the technical basis of information filtering, information retrieval, text database, digital library and mail classification. It has a broad application prospect. Text classification technology has been widely used in spam filtering, emotion analysis, public opinion monitoring, news classification and other fields. With the development of deep learning technology, the application effect of deep learning in text categorization is getting better and better. In addition, with the development of big data and artificial intelligence, the application of text categorization will be more extensive. Therefore, it can be foreseen that the development of text classification technology will bring more opportunities and challenges in various fields in the future.

3. The Concept of Information Retrieval

In a broad sense, information retrieval refers to the process and technology of organizing and storing information in a certain way and finding out relevant information according to users' needs. Narrow sense of information retrieval is only refers to from a certain information set to find out the required information process, equivalent to what people usually call information query. Information retrieval involves database technology, library and information science, artificial intelligence, natural language processing, machine learning and so on. The main aim of information retrieval is to present, store and organize information, and make it easier for users to get the information they need or are interested in. Information retrieval system generally consists of the following parts: data collection, information preprocessing, index construction, query processing, result sorting and display. Among them, index construction is one of the core technologies of information retrieval system. The common index structures are inverted index and vector space model. In addition, information retrieval also faces some challenges, such as language ambiguity, query expansion, result evaluation and so on, which need to be explored and studied continuously.

The demand of information retrieval for text and information organization is mainly shown in the following aspects.

Text is mined through topic structure. There are two ways to classify document sets in text information organization: taxonomy and topic taxonomy. Taxonomy is a method of sequencing information according to the characteristics of the subject system of information, which is based on the system and category relationship of the Chinese Library Taxonomy and other national standards, and is applicable to the static or changing information organization and management, such as the classification of library books, research reports and papers. Automatic text classification. Text classification using computer technology is an important task in information retrieval. It can automatically classify the text into different categories, easy for users to retrieve and browse. This method is suitable for dynamic information management, such as news, blog, social media, etc.

Subject rule is a method of organizing and arranging information according to its subject features. With the development of the Internet, the structure of the data set may change at any time. Static classification can not cover all documents, so dynamic subject classification must be used as the classification standard. However, such a huge set of text data has gone far beyond the ability of human understanding, and it is impossible to accurately obtain the classification structure system by purely artificial means.

ISSN: 2688-9323

Automatic text categorization. After determining the criteria for categorization of a text set, it is an important task in text information organization to quickly categorize documents into categories. Information retrieval technology is widely used, including but not limited to document retrieval, web search, information classification and filtering, intelligent Q&A, social media analysis. In recent years, with the development of big data and artificial intelligence technology, information retrieval technology is constantly innovating and improving, which brings great convenience and benefits. More importantly, with the rapid growth of text information and the changing structure of text sets, text categorization can not meet the needs of efficient information retrieval by mere manual work.

4. Promotion of Text Classification for Information Retrieval

Text categorization belongs to supervised machine learning. Generally speaking, the process of text categorization is as follows. Training text set: The training text set consists of a set of preprocessed text feature vectors, each of which has a category label. Text categorization is an effective information retrieval method. It uses the training text set to train the initial classification model and get the classification discrimination model. These models can be used to classify other texts. The main task of text classification is to map other unlabeled text objects in the text set to the preset category by learning the labeled text set under the preset category system. This technology can meet the second requirement of information retrieval for text information organization. In addition, text categorization has many other applications. In the aspect of spam filtering, this technology can effectively identify and filter out spam, and provide better protection for people's email experience. In emotional analysis, text categorization technology can help people better understand and analyze user emotions on social media, thus better understand user needs and market trends. In addition, text categorization technology can also be applied to the field of intelligent customer service, through the analysis and understanding of users' problems and feedback. The application prospect of text classification technology is very broad. At present, we have seen the application of text categorization technology in many fields, such as smart customer service, spam filtering, emotional analysis and so on. In the future, we can expect more new scenarios, such as automatic summarization, automatic translation, etc. These application scenarios will further promote the development and innovation of text categorization technology.

Automatic text categorization and information retrieval are complementary to each other. Automatic text classification technology improves the precision and speed of information retrieval by classifying text into different categories. But the information retrieval has put forward the higher request for the automatic text classification technology. For example, information retrieval requires an understanding of the semantics of the text to classify it more accurately. Therefore, the combination of the two will bring more extensive application prospects. Such as the increase of the amount of information on the Internet, the increase of content for the automatic text classification of new research topics. The aim of automatic text categorization is to organize text sets in an orderly manner, and to group related or similar texts together. As a knowledge organization tool, it provides more efficient search strategy and more effective search results for information retrieval. Efficiency and effectiveness are two important aspects of automatic text categorization. Efficiency means that users can determine the possible categories of queries in advance, thereby reducing the amount of text to be queried further. Validity refers to similar text may be related to the same query, thus improving the recall and precision of the query. With the increasing of network information, the demand for efficient and fast information processing is becoming higher and higher. This provides a broader stage for the research of automatic text classification. Therefore, researchers are constantly proposing new algorithms and techniques to improve the performance of automatic text

categorization. Some of these new techniques include deep learning and natural language processing, which can identify semantic and contextual information more accurately and improve the accuracy and efficiency of classification. Text information organization is an important link to solve the problem of efficient management and utilization of massive data under the current rapid growth of text information. Text classification is one of the key techniques in text information organization, which plays a vital role in text structure mining, document automatic classification and efficient Chinese text indexing.

5. Conclusion

To sum up, with the accelerated development of the information age, emerging technologies continue to emerge, more and more technologies are widely used in all walks of life. In the aspect of information retrieval, under the impetus of modern technology, information retrieval becomes more perfect and more convenient. As an important function, information retrieval aims at simplifying the user's information retrieval process and improving the user's experience. Therefore, it needs to rely on innovative ideas and new technologies to support. Especially in the current era of information explosion, the information retrieval program should strengthen its own technological innovation and provide users with diversified and personalized information retrieval services in combination with the latest technological advantages and innovative service concepts, so as to realize the overall support for social development. Therefore, we should actively explore new technologies and apply them to information retrieval to better meet the needs of users. At the same time, we should also strengthen information retrieval services, to provide users with more choices, so that users feel better service. In a word, information retrieval is a constantly developing field, we should constantly explore and innovate to better meet our needs.

References

[1] Feng Chenghao, Xie Zhenping, Ding Bowen. Methods for minimizing test cases of Chinese text error correction software [J/OL]. Small-scale microcomputer systems: 1-12 [2023-08-24].

[2] Summer Wuji, Huang Heming, more wording, etc. Summary of extractive texts based on unsupervised and supervised learning [J/OL]. Computer Application: 1-17 [2023-08-24].

[3] Zizhijie, Dorje. Selection of Primitives for Tibetan Text Classification [J]. Journal of Chinese Informatics, 2023,37 (01): 64-70.

[4] Dai Jiayang, Zhou Dong. A Study of Cross-Lingual Information Retrieval Based on Multitasking [J]. Journal of Guangxi Normal University (natural science edition), 2022, 40 (06): 69-81.

[5] Yi Xiaoyu, Yi Mianzhu. Unstructured Data Text Classification Algorithm Based on SWOT Analysis [J]. Technological Innovation and Application, 2022, 12 (29): 25-28 + 33.

[6] Gao Chenchen, Liu Qiyan, Wang Juncheng et al. Construction and Empirical Study of Interdisciplinary Patent Retrieval Strategies [J]. Information Engineering, 2020, 6 (05): 90-99.

[7] Wang Ding. Analysis and Research of Text Classification Techniques Based on Machine Learning [J]. Science and Technology Innovation Guide, 2020, 17 (08): 90 + 92.

[8] Xie Zichao. Research and Implementation of Automatic Classification and Retrieval Platform for Unstructured Texts [J]. Software, 2015, 36 (11): 112-114 + 119.