Journal of Artificial Intelligence Practice (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2023.060202
ISSN 2371-8412 Vol. 6 Num. 2

A Study and Implementation of an Optimized University Library Book Recommendation System Based on Artificial Intelligence and Python Crawler Scraping Technology

Ke Luo

Institute of Library, Shaoyang University, Shaoyang, Hunan, China

Keywords: Book recommendation, python crawler technology, artificial intelligence, accurate purchasing model

Abstract: Given the limitations of space and funding in libraries, achieving maximum efficiency has become a major challenge in the library world. With the continuous development of artificial intelligence technology, the degree of automation in book recommendation systems has also increased, requiring traditional procurement methods to be continuously optimized. By combining Python web crawling technology and precision procurement models, a book recommendation mechanism was constructed to transform book recommendations from traditional manual collection to automated assistance and prediction through the precise procurement model based on the recommendation list, thereby achieving the automation of book procurement. This mechanism improves the efficiency and accuracy of book procurement, provides a solution for the limited resources and funding of libraries, and also provides better services for readers.

1. Introduction

At present, most libraries still adhere to the traditional method of book acquisition, relying solely on the recommendations of faculty, staff, and students and the availability of suppliers as a reference for book procurement. However, this overreliance on manual recommendations has resulted in limited coverage of diverse areas, making it challenging to provide a more comprehensive range of books for patrons to access [1]. The major challenge with traditional book purchasing methods is that they are laborious and time-consuming, and borrowing rates may not always match expectations, leading to the problem of wasted space in the collection. As library resources shift from storage space to learning space, libraries need to optimize their collection space by selecting books that best suit the readers' preferences from the perspective of the most appropriate readers. It is essential to explore how to provide more books efficiently to reduce book review time and accelerate the pace of new books coming into the library. This strategy ensures that the collection resources can keep up with the publishers, resulting in lower labor costs and higher productivity. These issues are critical for libraries to consider and explore, as they strive to enhance the patron's experience and provide access to an extensive range of resources.

In order to cope with the problems of traditional book procurement methods, libraries must adopt

a more specialized and scientific book procurement model to meet the diverse needs of readers. In this paper, by referring to the book purchasing models of domestic universities[2-4], we propose an automated book recommendation mechanism by combining crawler technology to automatically mine the latest book sources and using an integrated learning model for accurate purchasing prediction, and replace the traditional manual recommendation method with automatic system recommendation. This mechanism can improve book acquisition efficiency, reduce labor costs, and upgrade the library service model to better meet the needs of readers.

2. Related Works

2.1. Python Crawler Technology

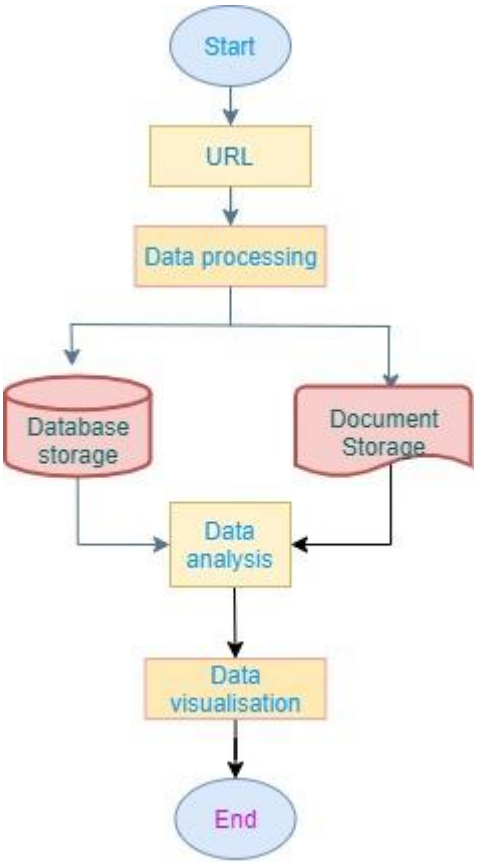

Figure 1: Flow chart of web crawler data processing

A web crawler (also known as a spider or robot) is a program that automatically navigates web pages and extracts data [5]. It is commonly used for data mining, information processing, and indexing in search engines. Web crawlers start with a list of URLs to visit, follow the links on each page to find new pages, and repeat the process until a certain criterion is met.

Crawler technology is suitable for rapidly acquiring important information from a large amount of data, bringing the acquired information together, and analyzing it to improve efficiency. The process of crawler data processing is shown in Figure 1, where the target web page is first accessed through a URL and the web data is crawled, then the collected data is processed, such as removing missing values, and then the data storage method is selected, either in the form of a database or a file, for subsequent data analysis and visualization.

2.2. Stacking Integration Learning

Stacking is an integrated learning technique that combines several different learners through a meta-model [6], each of which outputs multiple predictions after training. The predictions from each learner are used as input to the meta-model and trained to obtain an output, as shown in Figure 2. The advantage of stacking is that it can use multiple different models to capture different features in the data and combine them into more powerful models, more powerful models to combine them; the disadvantage is that it requires more computational resources and time to train and test the models. In 2019, Duan Jidong and Liu Shuangrong took a dataset of microblog forum website comments for text sentiment analysis and used the stacking integrated learning technique to fuse and train multiple models to get up to 93% accuracy, which significantly improves the prediction compared to a single learner [7].

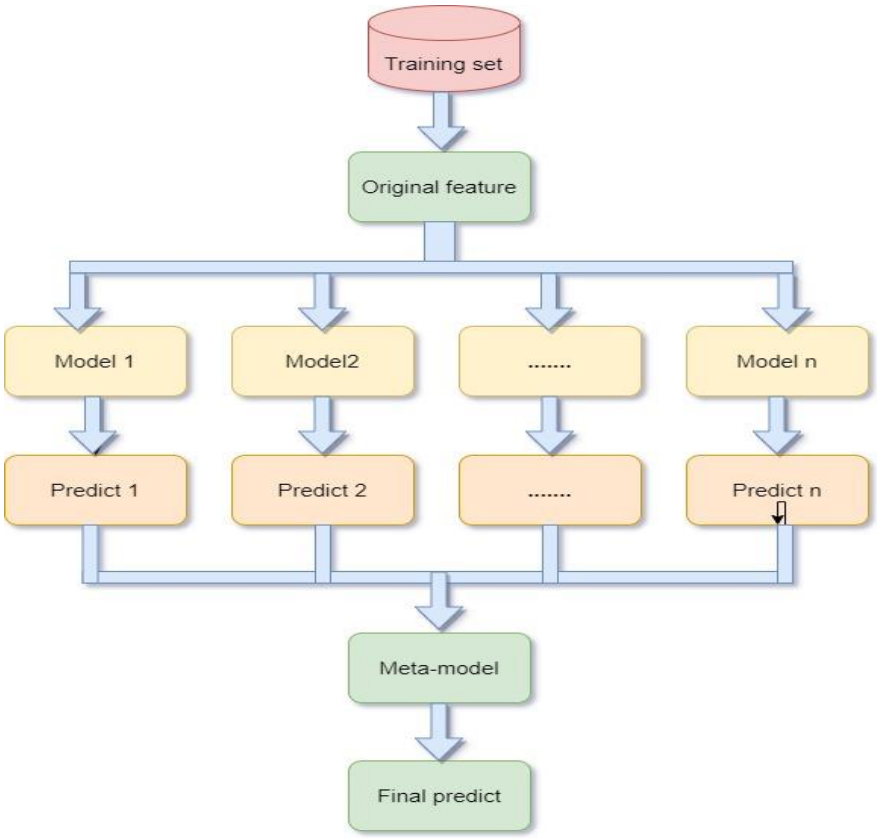

Figure 2: Integrated learning stacking architecture diagram

3. Book Recommendation System Based on Artificial Intelligence and Python Crawler Technology

3.1. General Process Framework of Book Recommendation System

Python provides a large number of class libraries to implement crawling technology. Based on the principle of web crawling, this paper uses Python crawling technology to collect a large amount of book information data on Dangdang.com as a dataset. The overall process is as follows: first, crawl the book information data from Dangdang.com according to the input book types and recommended volumes; second, pre-process the book dataset, using Jieba in the Python library for Chinese word separation, and using text frequency and inverse document frequency (TF-IDF) for text processing; again, put the extracted book text features into the accurate purchase prediction model; then, book

recommendations are made according to the prediction results, and the classification results predicted by the model are divided into two types: recommended and not recommended; finally, a list of book purchase recommendations is provided according to the classification results. The specific situation is shown in Figure 3.

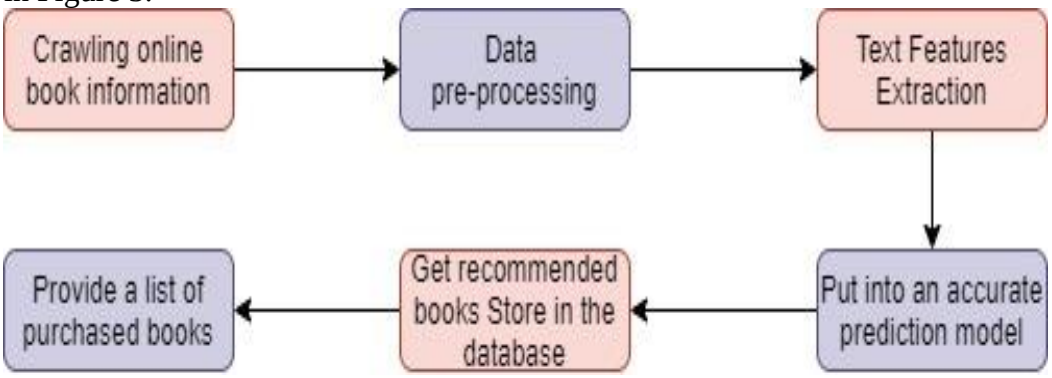

Figure 3: General flow of the book recommendation system

3.2. Automated Online Collection of Book Recommendations

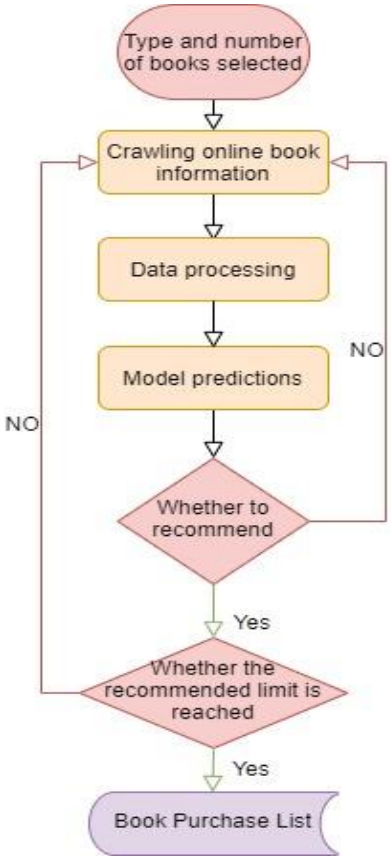

Figure 4: Automated book recommendation process

In order to realize automated book purchase recommendations, this paper designs an automatic book data collection process, as shown in Figure 4. The specific steps of this process are: the librarian enters the type of set titles and the number of recommended books, and through web crawling technology, obtains the latest book data from online bookstores, and analyzes and predicts them. This information is then stored in a database for subsequent purchasing decisions. The implementation of

this process involves a variety of technical tools, including data mining, web crawling, machine learning, and so on. Web crawlers are used to obtain data sources, data mining techniques are used to process large amounts of book data, and machine learning is used to analyze and predict the data, ultimately realizing automated book purchase recommendations.

3.3. Constructing a Book Recommendation Prediction Model

The book prediction model used is a classification model constructed by applying machine learning and deep learning techniques to text analysis, and combining integrated learning methods to improve accuracy. The purpose of the model is to identify the characteristics of books that are likely to be borrowed and to build a model that matches the preferences of readers' borrowing behavior to predict the likelihood of the book being borrowed in the future. The research sample for model training is the library borrowing records, and the text content will be first processed with text features using BM25[8] to calculate the important keyword words, and then LightGBM [9], Logistic Regression, and Nave Bayes algorithms in machine learning will be used as the first layer of the integrated learning method architecture, while the second layer is the convolutional neural network algorithm in deep learning, and the model flowchart is shown in Figure 5.

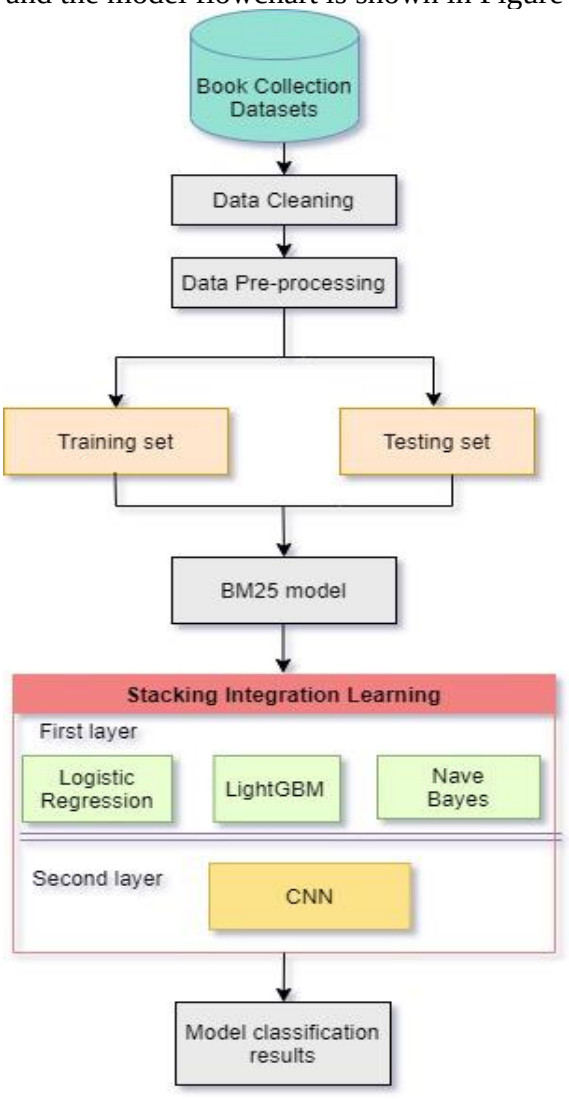

Figure 5: Book recommendation prediction flow diagram

4. Experiments and Analysis

4.1. Experimental Environment

The aim of this experiment is to build an automated book-purchasing recommendation platform to test the accuracy of online book data in prediction models. The platform was developed in a Windows Server 2012 R2 server environment, based on Python 3.9, and using the Flask framework and an Oracle database. The collection book information dataset was pre-processed and formatted to suit the needs of the recommendation algorithm, implemented using machine learning algorithms from the Scikit-learn library. In addition, the performance of the model was evaluated using a Windows 10 Professional OS computer with an Intel Core i5-10210U CPU and NVIDIA GeForce RTX 3080 VGA card to verify its feasibility and effectiveness for book purchase recommendations.

4.2. Book Data Sources and Text Pre-processing

The book data covered in this paper comes from DangDang Books. DangDang Books not only provides abundant book resources, but also provides bestseller lists based on readers' feedback, so that readers can keep abreast of the most popular current reading reviews and book selection guidelines. Through web crawler technology, key information such as book title, author, publisher, year of publication, and comments can be collected from the Dangdang Books website, and a dataset containing a large amount of book data is constructed. We put these datasets into the accurate purchase prediction model and came up with the prediction results. The accurate acquisition prediction model is a classification model, and the required sample is the collection lending records of university libraries from 2017 to 2021. After data collation, a dataset containing nearly 200,000 borrowing records was obtained, which contains fields such as book number, book title, author, year of publication, publisher, and a total number of borrowings. This method can not only provide an accurate prediction model for university library, but also help librarians to better understand readers' needs and preferences for book purchasing and management. At the same time, due to the authority and authenticity of the data source, the model prediction results we obtained are more accurate and reliable, which helps to improve the efficiency of the library and the satisfaction of readers.

Data pre-processing is a very important step in the machine learning process, which can effectively improve the prediction accuracy and robustness of the models. In the pre-processing process, different methods and tools are needed to clean and organize the data according to the specific situation, so that it can meet the requirements of the machine learning algorithms and thus better carry out the subsequent analysis and modeling. In this study, data pre-processing for book data is performed using the following methods:

1) Deactivation word processing: We used a list of common deactivation words and processed the text data to remove common words that occur frequently but have no practical meaning, such as: had, of, is, while, and, etc. This can effectively reduce the data noise and improve the prediction accuracy of the model.

2) Non-text data processing: The original data contains some non-text data, such as: blank areas, punctuation marks, special symbols, etc. We used a text processing tool to remove this non-text data. This can ensure that the data we analyze are somewhat standardized and comparable, and it can also avoid the influence of these non-text data on the model prediction results.

4.3. Experimental Methods and Results

In this paper, we take the DangDang online bookstore as the target web page for crawling work, crawl the book title, author, publisher, publication year, reviews, and other information in the book

through the book category classification defined by DangDang, put these book data into the trained model for prediction, and come up with the result of 0 or 1. The book with a prediction result of 1 is considered recommendable, while the book with a prediction result of 0 is considered not recommendable. In the prediction process, the information of books with a prediction result of 1 is stored in the database, and this information includes important items such as book category, title, author, price, year of publication, and publisher. This information not only helps college libraries purchase books according to readers' needs, but also can be used as the basis of the book recommendation system to help readers find books they are interested in quickly in a large number of books. It should be emphasized that this paper adopts efficient and automated crawler technology to build the dataset and uses the trained model to make predictions, which in turn enable book recommendations and purchases. This approach can not only greatly reduce the workload of librarians but also improve the efficiency of libraries and meet the needs of patrons.

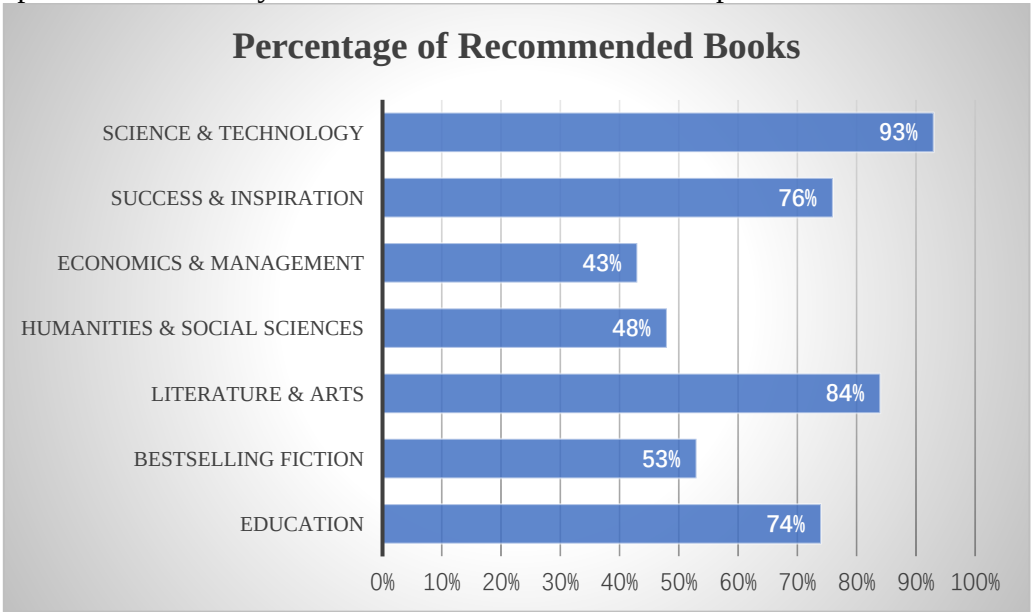

Figure 6: Percentage of recommended books in the predictions

In this paper, the prediction results of the model are further analyzed. In order to get a deeper understanding of readers' interest preferences, the book categories defined by Dangdang.com were used, seven different categories of books were selected for crawling, and 100 different books were put into the model for prediction. We counted the proportion of books predicted to be recommended in each category to all books, and this proportion can reflect the readers' acceptance of books in that category. The statistical results (Figure 6) show that readers are most interested in books in the categories of science and technology and literature and arts, with 93% and 84%, respectively. This is followed by education, success, and motivation, as well as best-selling fictions, with 74%, 76%, and 53%, respectively. The acceptance level for economic management and humanities and social sciences is relatively low, at 43% and 48%, respectively. These findings suggest that the research and interest areas of university faculty and students are biased toward engineering. Librarians can also use the results of these analyses to map out the profile of patrons' interests as a basis for assigning the proportion of book categories to be purchased. For university libraries, science technology, and literature should be the key acquisition targets, while increasing the number of educational books, successful and inspirational books, and best-selling novels can be considered. The purchase quantity of economic management and humanities and social sciences can be reduced appropriately to better meet the needs of readers.

4.4. Evaluation of Experimental Results

To evaluate the accuracy of the automated book purchase recommendation mechanism, the top 10 best-selling books on Amazon from 2017 to 2022 were selected as the objects of evaluation. These books have high market sales and popularity, which can better reflect the applicability of the automated book purchase recommendation mechanism. The testing set data of the model was compared with books in university libraries and 34 of the same books were found. These books were then put into the model for prediction, and the prediction results were compared with the real book borrowing status. The experimental results showed that the number of books predicted correctly was 25, with an accuracy rate of 73.5%, where the correct book prediction means that the book was recommended and actually checked out. The detailed results of the comparison are shown in Figure 7.

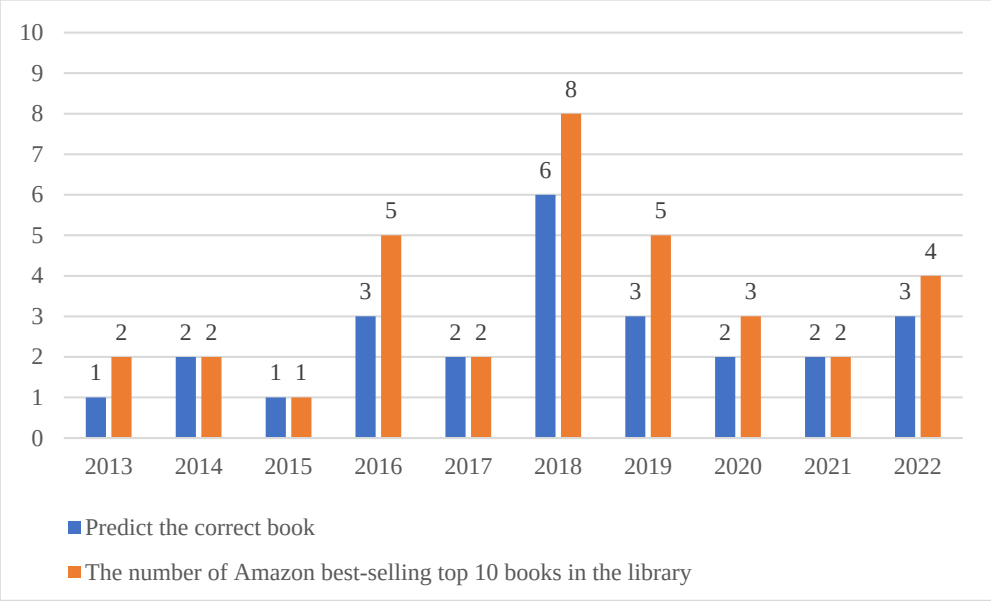

Figure 7: Comparison between model predictions and real book collections

5. Conclusion

In this paper, a new automated book recommendation mechanism is proposed by combining crawler technology with accurate book-purchasing to provide a more convenient and efficient way for libraries to make purchases. The mechanism collects books from online bookstores through crawler technology, adds the latest and hottest sources of book selection, and provides patrons with appropriate book purchasing criteria based on model predictions. In addition, it helps librarians reduce the workload and time required in selection and review work, thus shortening the speed of new books in the library, increasing the book borrowing rate, and achieving the purpose of automated book purchase recommendation and precision purchasing. In conclusion, the precise book recommendation mechanism proposed in this paper provides a new book purchasing strategy for libraries to provide better and more diversified services to readers. In the future, the mechanism can be further improved, such as by adding more data sources, improving the accuracy of model predictions, and performing more detailed data processing, to achieve better results.

Acknowledgements

This paper is funded by Hunan Provincial Philosophy and Social Science fund project "Research

on the Application of Artificial Intelligence Technology in Accurate Book Procurement in Universities under Smart Library "(21YBA179).

References

[1] Ameen K, Haider J S. Book selection strategies in university libraries of pakistan: an analysis. Library Collections, Acquisitions and Technical Services, 2007, 31(3):208-219.
[2] Xie Ling. An Empirical Research on Recommendation System of Wuhan University Library. Research on Library Science, 2016 (8):74-78.
[3] Xu Xinqiao, Liu Hua, Zhang Xinyun. An Empirical Research of Recommendation System of Shanghai University Library. Research on Library Science, 2014 (24):5-9.
[4] Xie Ling. Development Status and Improvement Measures of Literature Recommendation System in "985 Project" Universities Libraries. Library Work in Colleges and Universities, 2015, 35(6):41-44.
[5] Chen Cong, Zhou Lizhen. Tracing and Filtering of Fake Data Based On Python Crawler Technology. Computer Simulation, 2021, 38(3):346-350.
[6] Zhang Liu, Chen Yifei, Yuan Jiawei, Pei Ziquan, Mei Pengjiang. Application of Stacking Ensemble Learning Model in Blended Performance Classification and Prediction. Computer Systems and Applications, 2022, 31(7):325−332.
[7] Duan Jidong, Liu Shuangrong, Ma Kun, Sun Runyuan. Text Sentiment Classification Method Based on Ensemble Learning. Journal of University of Jinan (Science and Technology, 2019, 33(6):483-488.
[8] Stephen E.Robertson, Hugo Zaragoza. The probabilistic relevance framework: bm25 and Beyond. Foundations and Trends in Information Retrieval, 2009, 3(4):333-389.
[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in neural information processing systems, 2017: 3146-3154.