Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Statistical analysis of academic journal information based on CNKI Literature

Taizhi Lv [a], Jun Zhang [b]

School of Information Technology, Jiangsu Maritime Institute, Jiangsu Nanjing 211170, China

[a]lvtaizhi@163.com, [b]1052871890@qq.com

Abstract: The research and development of colleges and universities cannot be separated from scientific research. Mastering the current situation of scientific research in colleges and universities will help to strengthen the analysis and construction of scientific research ability of research academic institutions. Academic journals are not only an important index to measure the scientific research ability of colleges and universities, but also in order to objectively predict the development trend of the industry and count popular majors, a data statistical analysis based on the information of academic journals of China CNKI is designed and implemented. The system uses distributed crawler program to collect CNKI data, realizes data storage, cleaning and statistics through Hadoop platform, stores statistical results in Mongodb database, and realizes visual display of statistical results by using Flask web framework and Echarts technology. The statistical results include the publication trend of journal papers, the display map of journal hot spot index, the display map of funded academic research institutions, etc.

Keywords: Big data, China national knowledge infrastructure(CNKI), distributed crawler, Scrapy framework, Echarts library, flask framework, Spark Scala.

---

1. INTRODUCTION

The number of journal papers published can effectively represent the academic characteristics, the level of scientific research ability, and determine the scientific research direction and work development strategy. Most of the journal papers of ordinary scholars are collected by CNKI. They can use the CNKI journal data as the data source to make multi-dimensional analysis and understanding through the study of CNKI data, such as university teaching level, advantageous majors, academic research direction, social development, scientific research ability of academic research institutes and national strategy. In this paper, the data of CNKI journal papers are obtained through incremental distributed crawler, and the publication of journal papers is displayed through the front visualization technology processed by spark Scala platform, so as to provide data support for the analysis of domestic scientific research trend, domestic education trend, economic trend and scientific research level.

Academic journals are the main carrier of scientific research achievements. Analyzing the publishing situation and hot spots of domestic journals is conducive to quickly and accurately grasp the development direction of disciplines and disciplines and the academic development trends of scientific research institutions. Learning and analyzing the characteristics of excellent journals is very important to improve the quality and influence of journals [1].

The research on academic journals has always been concerned by all sectors of society. 238527 journals were searched on the CNKI according to the keyword big data. It can be seen that the research direction of big data mainly focuses on big data technology, data processing and prediction, the construction of data platform, and even needs to be associated with machine learning and artificial intelligence. On the basis of the above retrieval and the keyword "algorithm", 13919 journals were searched, and this manual retrieval is time-consuming and laborious. It is also necessary to provide another angle and method for academic research through the main information retrieval of the paper and the autonomy of funding institutions. The above methods are to analyze the scientific research capability of institutions, funds and other research academic institutions through investigation or manual data collection. The key words of published journal papers are generally 3 ~ 5 words, which cannot judge the technical proportion, or avoid being affected by subjective factors or untimely data, which cannot guarantee the objective reflection of the scientific research capability analysis of research institutions and collect data manually. Analyzing data is time-consuming and laborious. The required messages can be queried quickly through this system.

2. FUNCTIONAL REQUIREMENT

The journals included in China CNKI are important indicators that can reflect the scientific research ability of academic research institutions, social research direction and national development strategy. This paper studies and designs the data collection, data processing and visualization of CNKI journals. The functions of the system can be divided into the following three aspects.

First, it is very necessary to solve the problem of periodical data source and realize the efficient and comprehensive collection and storage of periodical data. The collected journal data mainly include: ID, detail page link, detail page name, publication time, times of citation, journal name, author, affiliated organization, keywords, financial support, DOI, album, topic and classification number, which provide effective support for the traceability of the following data analysis results. The main data source is CNKI.

Secondly, the cleaning and statistics of periodical data are realized, and the comparative statistics of publication time, cited times, financial support and affiliated institutions are designed to realize the multi angle mining of the number of periodicals. Including statistics by title, subject statistics and other functions. The system can meet the needs of users to analyze the scientific research ability of academic institutions

Third, the data information of domestic journals is complex, and the results of data analysis need to be clearly displayed. Therefore, it is necessary to realize the visual display of journal data in many aspects on the premise of data cleaning and statistics.

3. DATA ACQUISITION AND PROCESSING

According to the characteristics of CNKI website, a distributed crawler framework based on Python-scratch, MySQL, Redis is proposed to crawl and store the journals included in CNKI. The information related to journal papers: ID, detail page link, detail page name, publishing time, cited times, publication name, author, affiliated organization, keywords, financial support, DOI, album, topic, classification number, download quantity, etc. are the data sources of other functions of the system. In order to facilitate the subsequent functions to operate the data, the basic data processing shall be carried out before collecting the data information, so as to ensure the consistency of the data and store it in the Mongodb database. The acquisition process is as follows.

Step 1: data pre-analysis

From https://navi.cnki.net/knavi/Journal.html, the web page crawler crawls all journals of each discipline in the discipline navigation, and then saves them in segments from the starting time of CNKI journals to the current time. All journal names and year time periods are saved to the MySQL database. Using the double for loop, the journal names and year time are combined. The build request body is pushed to the Redis database.

Step 2: Data crawling

Start the Scrapy engine. The engine passes the required fixed URL to the scheduler through the scheduler, and then takes the proxy to the scheduler to request the post request body in Redis [2]. The scheduler carries the URL and its post request body and proxy to the downloader, and the downloader accesses the CNKI server through the proxy, and get the response body returned by CNKI server, pass the response body to spider, spider parses the response body, and the parsed data is returned to pipeline, which is saved to Mongodb database through Pymongo library [3].

The system rewrites make according to the domain name of CNKI academic journals_ request_ from_ Data method, use scratch Redis to read [url, form_data, meta] in Redis, and then send a post request. The query entry is https://kns.cnki.net/kns8? dbcode=CJFQ。As shown in Fig. 1, the developer tool opened by the browser determines the request parameters, converts them into variables according to the value representing time in the cookie value, splices the cookie value and sends it to the CNKI server for verification.

Fig.1 Request parameters diagram

Step 3: Web page parsing

Parsing CNKI web pages uses the XPath parsing library to find navigation elements and attributes in XML documents. In the journal list page, as shown in Fig. 2CNKI journal list, first find the class attribute of the specified label where the list is located, and then traverse all items a under the specified label. Through item a, analyze and obtain the relevant information such as classification number, topic, album, DOI, fund support, keyword, abstract, affiliated organization, author and journal name in the detailed page.

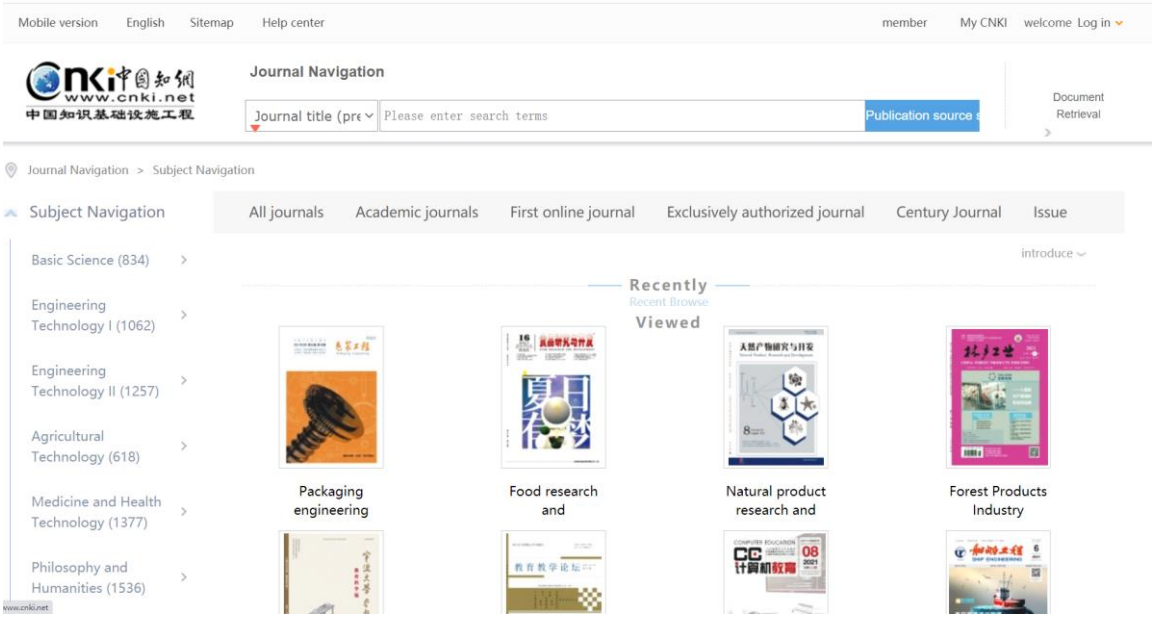

Fig. 2 List of CNKI journals

Step 4 data cleaning

In the data cleaning stage, the fields with null value, incomplete content and inconsistent content will be modified, corrected and eliminated. There may be abnormal data values in the original data, so it is necessary to eliminate format structures such as header through preprocessing data, so that it will not affect the results of data analysis [4].

The default value is cleared and filled according to the proportion of missing values. For example, if the journal citation is empty, fill in 0, for example, the collected summary text is too short to remove data, etc. The cleaning of format content includes the unification of date format, keyword and other formats. Unreasonable logic removes unreasonable values. For example, the number of pages is not a number, and the number of pages exceeds common sense.

Step 5 data statistics

Establish a spark connection on the spark platform to read the data source, and use the textfile method to read the file. In order to clean the data and avoid being affected, it is necessary to use the filter method to filter useless fields (such as empty in the organization list, illegal characters, etc.), and then use the map () method to flatten the mapping to extract the required field content (the content is converted into an array), The data is processed by the flatmap method for breaking up. In order to separate the data and facilitate subsequent grouping and aggregation processing, if the data is not broken up, the result data will be loaded into the same cell and subsequent operations cannot be carried out. By using the groupbykey() method to aggregate the same key values of CNKI data, the

method data is traversed, iterated and counted, mapvalues() method is used, and sortby function is used for descending arrangement to facilitate data visualization. Saveastextfile() method is used to save as a file, but there will be partition operation by default, so repartition() is required Method to re specify the number of partitions to uniformly save to the specified file [5].

Store the crawled original data and create the spark association table through Scala. By executing Scala statements on the association table and using Scala code, data slicing, statistics and classification are realized. SQL query extraction is commonly used in the data extraction stage. First query according to conditions from the data table, and then extract the required data by multi table Association, optimize SQL statements, and reduce resource waste by optimizing Association and other methods.

4. DATA VISUALIZATION

Flask framework has strong expansibility and can add custom components [6]. In addition to maintaining a simple style, it also has a rich plug-in library, which can realize personalized website customization and develop a stable, simple and safe website. Echarts is an open source visualization library based on JavaScript [7]. It is rich in charts from all walks of life, meets various needs, and is compatible with most of today's mainstream browsers. This paper realizes periodical data visualization based on Flask and Echarts.

Subject journal hotspot is a data analysis module based on CNKI journal data. The hotspot index can tell users: the proportion and scale of a keyword / hot word in the current industry, the trend of a period of time and the trend chart of related hotspots. Fig. 3 periodical hot spot index shows the proportion and weight of annual periodical papers of the discipline in 2020.

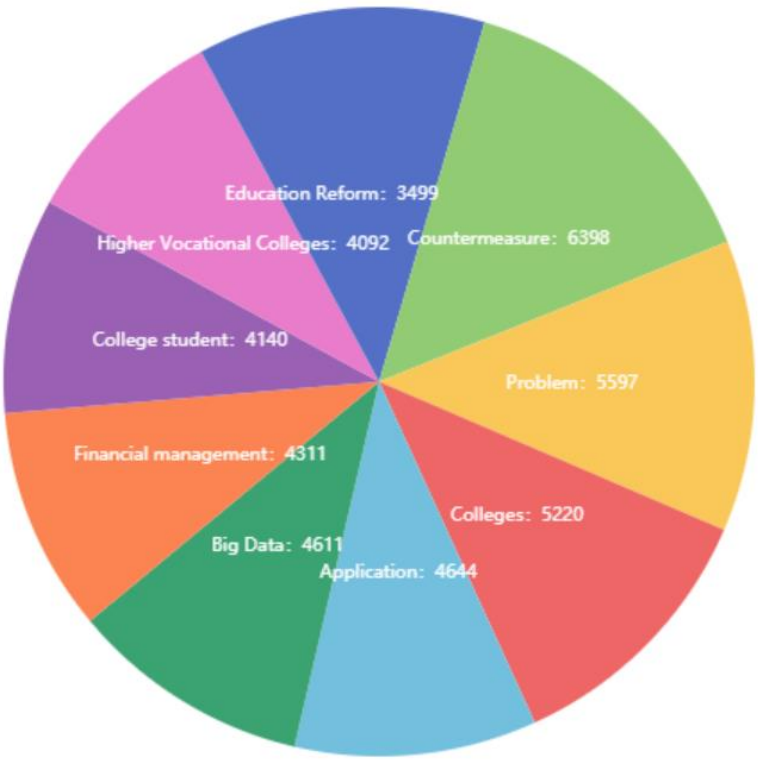

Fig. 3 Hotspot index of journals

The citation module of research academic institutions can analyze the influence of journals of each college by the number of citations of published papers. For some comprehensive or large research fields, the citation rate is relatively high due to the wide range of field research and many interdisciplinary disciplines. For example, financial, biological and Chemical Journals and major breakthroughs in science and technology will have an impact on the economy, so it is generally easy to have a greater impact. The cited number can represent its academic quality to a certain extent, but it is not linearly proportional to the academic quality. For example, finance is related to biology, chemistry and medicine, and financial journals cite papers on biology and medicine. Although it does not have the function of accurate quality evaluation of academic research, it can provide guidance from the macro perspective shown in the cited figure of papers of research academic institutions as shown in Fig. 4.

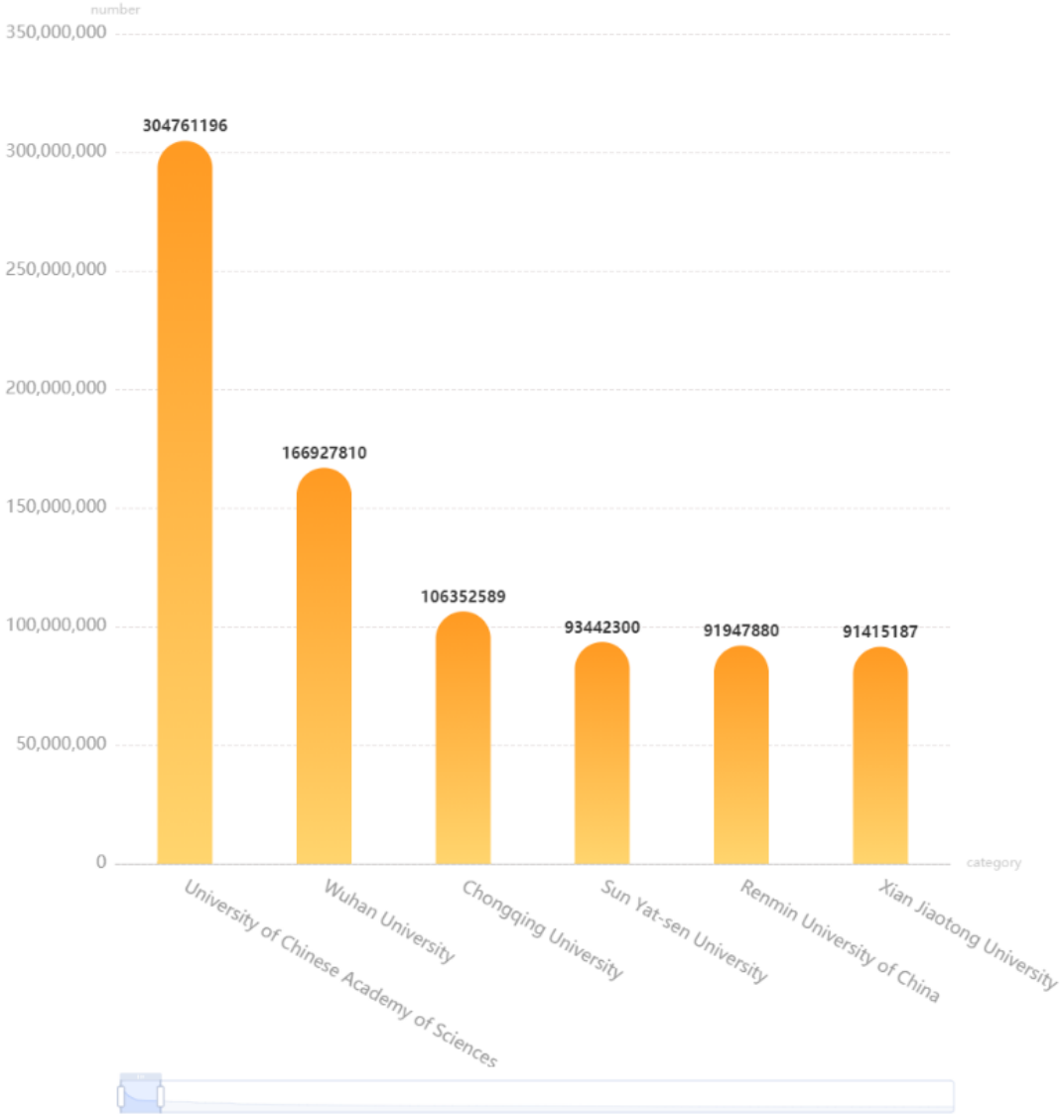

Fig. 4 Cited papers of research academic institutions

5. CONCLUSION

The data statistical analysis based on the academic journal information of China CNKI helps to mine and analyze the journal data by using crawler, data cleaning and data visualization, so as to speculate the future trend of science and technology, social trend and national development strategy. The

project has completed the basic data statistical analysis function. In the future, more idea models will be added to more intuitively understand the socio-economic trend, scientific and technological development trend, national development strategy and so on.

In the future, artificial intelligence (AI) can also be used to imitate human intelligence to perform tasks, update and improve the construction model and analysis model based on the collected information, understand the problems raised by the target object more quickly and efficiently, and provide more effective and persuasive analysis model and data prediction model.

ACKNOWLEDGEMENTS

This work was financially supported by the funding of Qianfan science and technology team of Jiangsu Maritime Institute(Big data analysis and application research team), the research project of computer basic education teaching from computer basic education institute of China colleges and Universities (Research, development and application of teacher scientific research ability analysis platform based on CNKI Literature, 2020-afcec-054), young academic leaders of Jiangsu colleges and universities QingLan project, excellent teaching team of Jiangsu colleges and universities QingLan project (Innovative teaching team of software technology specialty) , and the big data collaborative innovation center of Jiangsu Maritime Institute.

REFERENCES

[1] Tong, Zheng, et al. "Quality of randomized controlled trials of new generation antidepressants and antipsychotics identified in the China National Knowledge Infrastructure (CNKI): a literature and telephone interview study." BMC medical research methodology 18.1 (2018): 1-11.

[2] Kaiying, Deng, Chen Senpeng, and Deng Jingwei. "On optimisation of web crawler system on Scrapy framework." International Journal of Wireless and Mobile Computing 18.4 (2020): 332-338.

[3] Gupta, Eeshan, et al. "Attribute-Based Access Control for NoSQL Databases." Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy. 2021: 317-319.

[4] Gudivada, Venkat, Amy Apon, and Junhua Ding. "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations." International Journal on Advances in Software 10.1 (2017): 1-20.

[5] Omar, Hoger Khayrolla, and Alaa Khalil Jumaa. "Big data analysis using apache spark mllib and hadoop HDFS with scala and java." Kurdistan Journal of Applied Research 4.1 (2019): 7-14.

[6] Vogel, Patrick, et al. "A low-effort analytics platform for visualizing evolving Flask-based Python web services." 2017 IEEE Working Conference on Software Visualization (VISSOFT). IEEE, 2017: 109-113.

[7] Li, Deqing, et al. "ECharts: a declarative framework for rapid construction of web-based visualization." Visual Informatics 2.2 (2018): 136-146.