Design and Implementation of Media Manuscript Retrieval System New

Peng Chen
Shandong Institute of Commerce and Technology, Jinan, China

Keywords: New Media, Index, Full-Text Search.

Abstract: With the rise of the Internet and mobile Internet in recent years, new media has also achieved vigorous development, and new media articles and manuscripts have also shown an explosive growth trend. In the face of massive and multi-format new media data information, how to quickly and accurately find the required manuscript information in such large-scale data information has become a problem faced by users of we-media. According to the above problems and requirements, this paper designs and develops the architecture of Spring+SpringMVC+Hibemate, combines Solr search engine service and Baidu speech recognition tool, and proposes a new media manuscript retrieval system with B/S architecture. The system uses Java as the development language to implement. This paper focuses on the analysis of the key technologies and strategies used in the system architecture design, and develops and designs a new media manuscript retrieval system based on Solr, which mainly includes pre-processing, building solr system, user query and database. Based on the open source search engine Solr as the core of the system, this paper studies the implementation principle of the core technology index of search engine. In order to ensure the efficiency and quality of word segmentation, the algorithm of word segmentation and the performance comparison of various Chinese word segmentation are studied. In order to facilitate Solr to use text to build index, the text conversion method of non-text files is studied.

1 INTRODUCTION

In recent years, with the development of the Internet and mobile Internet, people's production and life have undergone great changes. People use Weibo and we chat to send messages and share their lives through short videos. More and more people use the Internet to study and work (Li ZY, 2021). We are ushering in an era of information explosion, which generates a large amount of information and data every day. The concept of "new media" was first put forward by the American P. Goldmark in 1967(Chung, S. W., 2020). Immediately afterwards, the new media appeared many times in the report that Chairman Rostow submitted to U.S. President Richard Nixon. Since then, the term "new media" has become popular in American society and gradually spread to the world (Longepe, N., 2022).

Although the term "new media" can be seen everywhere nowadays, and the study of new media has become one of the hot topics studied by scholars at home and abroad, so far, the academic circle believes that new media is a relative concept, and there is no unified definition (Yao, X. X., 2021). It is undeniable that new media is a kind of media communication technology based on the current rapid development of computer technology and Internet technology. This technology breaks through the limitation of time and space and allows a large amount of information to spread rapidly on the Internet, narrowing the distance between people, and becoming a major milestone in the history of human media development. Jiang Hong and others believe that new media is developed on the basis of traditional media, but it is fundamentally different from traditional media (Huang, B., 2023). Especially with the advent of artificial intelligence and big data era, various new technologies have reshaped the media industry and accelerated the transformation and reconstruction of new media ecology.

How to quickly obtain the information they need in the massive information wave of the Internet has become a problem in front of Internet users, in this case, people increasingly need to obtain information through search engines. With the rapid development of new media, all kinds of articles and consulting news fill people's lives every day, and there is a rapid growth trend. For Internet users, how to query

Chen, P.
Design and Implementation of New Media Manuscript Retrieval System.
DOI: 10.5220/0012282900003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 333-337
ISBN: 978-989-758-677-4

manuscripts accurately and quickly in the massive manuscript library becomes particularly important (Zhang, K. X., 2017). When facing the search request, the general processing method is to query the database by keywords. With the increasing amount of data in the database, the query speed through keywords will become slow, and it is more and more difficult to meet the demand for quick query. Therefore, it is urgent to design a separate search system. Therefore, this paper designs Solr retrieval system for enterprise manuscript retrieval, which can provide users with quick response query function. In this paper, an enterprise-oriented new media manuscript retrieval system is designed, which uses Solr search engine framework to provide fast retrieval service (Fairbairn, D., 2019).

From the perspective of the development history of search engines, it can be roughly divided into three stages of development. The first stage is the first-generation search engine represented by Yahoo and InfoSeek, which is based on the World Wide Web and supports natural language search and advanced grammar search for the first time, requiring manual directory sorting. At this stage, the input information resources are limited, the number of indexes is not large, and the index query speed is slow; The second stage is mainly the second generation of search engine technology represented by Google Browser, which is based on data(Pereira-Sanchez, V., 2022). Mining technology and website rating technology and machine retrieval through keywords, using distributed service technology, so the retrieval speed and accuracy have been greatly improved; The third stage is the third generation of search engine technology represented by the "technology-driven" search engine concept proposed by Microsoft Corporation (Nogueira, M. S., 2021). The second generation of search engine technology has been greatly upgraded and improved, which will provide users with more quality search services and search experience. At the moment.

The first-class search engine companies in the world include Google, Microsoft, Baidu and Yahoo, etc. The mainstream search companies in China include Baidu, 360 and Sogou. These companies are leading the trend to provide Internet users with high-quality search services. For most small and medium-sized enterprises in China, it is necessary to carry out information management, quickly and accurately retrieve enterprise product information and improve work efficiency. Enterprises can customize the enterprise search engine through the services provided by mature search engine companies, but the enterprise personalized search is limited and the function scalability is not strong, which cannot conduct in-depth analysis according to the fields of different enterprises, resulting in low retrieval accuracy and slow query speed. Therefore, it is necessary to build a set of enterprise personalized search engine (Zhao, X., 2020).

2 METHODS

2.1 Search Engine

Search engine is a technology that searches out the information and data with high matching degree from the huge information data of the Internet by adopting certain computer algorithms and technical means. Search engines need a lot of computer technology as technical support, in order to meet the Internet users fast search, high matching search needs and user experience. At present, the computer technology related to search engines includes big data, web crawler, index sorting, natural language processing and other technologies. With the advent of the 5G era, search engines will combine advanced technologies such as big data, artificial intelligence, and pattern recognition to provide Internet users with more high-quality and humanized services (Ufer, N., 2021).

The basic working principle of search engines is to use web crawlers to continuously obtain a large number of web resources from various websites on the Internet, collect these web resources into local databases, and then process them through web technology to remove useless interfering information and further extract key information from useful web information to build indexes. After the index is successfully constructed, it is stored in the index database. When the user uses the search engine of the browser to query information, it will quickly search through the index database of the search engine to find the index and web page information with high similarity and matching degree with the keywords entered by the user, and sort the search results by relevant sorting algorithms. The value is returned to the user in the order of matching degree from high to low (Qin, P., 2017).

Search engine contains a variety of types of search methods, search methods according to the different characteristics of collecting and querying information can be divided into the following four ways, including full-text search, meta search, vertical search and directory search. Among them, full-text search engine is a search method that obtains a large number of web page resources on the Internet through a crawler

program, extracts web page information and sets up an index library for retrieval.

2.2 Search Engine Framework

After determining the full-text search mode as the search mode of this project, the search engine framework used by this project is selected. The common search engine services can be roughly divided into three categories, Lucene, Solr and ElasticSearch (Qin, P., 2017).

Through the introduction and comparison of the above search engines, Lucene is a search engine architecture, which cannot directly run and query information. ElasticSearch is mainly characterized by strong usability and scalability, which is mainly used for distributed query. This system uses Solr as an open source search engine, which has powerful function, simple configuration, low development cost, and can meet the needs of enterprise search.Figure 1shows the system structure.

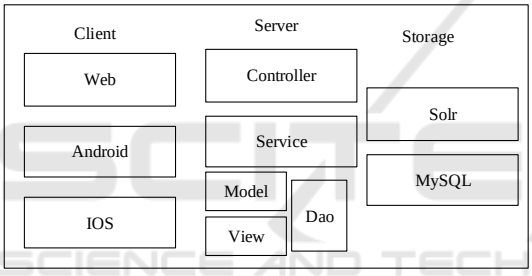

Figure 1. The system structure.

The system adopts the Spring+SpringMVC+Hibernate system framework to build the system framework, and divides the whole system into three layers. View is responsible for presenting the view, and the input box in the front end accepts the user's query request, including a variety of input forms, such as audio. Images and other query parameters are passed to the Controller layer. Cortoller invokes Service for logical business processing. It mainly writes query statements according to query parameters and invokes the Dao layer encapsulated by Hibermate framework for data persistence. Finally, it returns to the View layer and displays the manuscript information to the user in the form of a web page list.

3 SYSTEM IMPLEMENTATION

3.1 Solr Retrieval Module

In order to index article information and facilitate quick retrieval by Solr, it is necessary to use Chinese word segmentation. This system uses ANSJ word divider, and its configuration is as follows: Download ANSJ word divider jar package ansj_seg-5.1.6.jar, put it in the lib directory of webapp, and modify the managed-shchema file to add the configuration at the end of the file.

After starting the Solr service, you can view information about the ANSJ classifier on the web page. Enter a piece of text information, click the word segmentation button, the Solr system will use the integrated word segmentation, and the result of the segmentation will be displayed.

Solr management page provides rich management methods, including Chinese word segmentation, adding index, querying index, modifying index and deleting index. Indexes can be added in full or incremental mode. You can choose the two methods based on the actual situation. You can also use the two methods together. Query index is the core function of Solr system, which supports very rich query functions, including filtering, setting gravity, highlighting, etc., which can be set by parameters. You can delete an index by id or query condition. You can delete an index by using Documents on the management page. The function of adding, deleting, modifying and checking the Solr system can be managed through the management page, or it can be operated through the code. Figure 2 shows the management page.

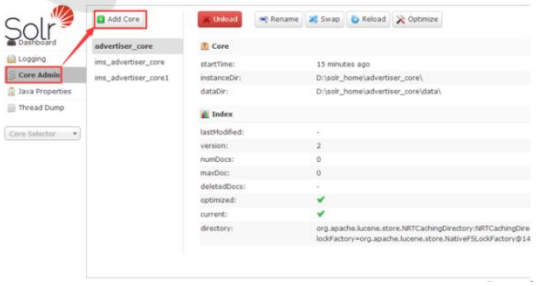

Figure 2. The management page.

Sla provides a management page. You can use the SLA Web management page for direct query. Figure 3 shows the query screenshot:

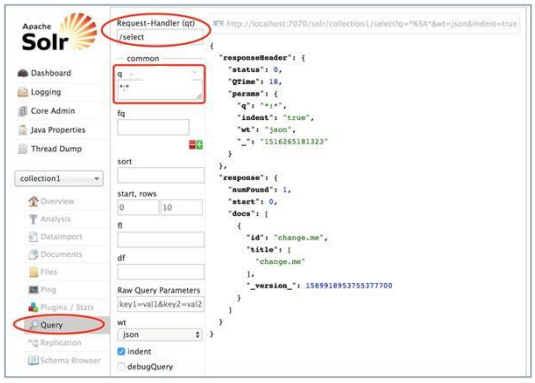

Figure 3. The query screenshot

Index import includes full index and incremental index, where full import is to import all the cable bows at one time, and incremental index means that only the new index is imported each time. Generally, you can import full data for the first time and incremental data for subsequent imports. You can also use full import to overwrite the original index. In general, you can use the Solr management page to import database data and create indexes. The process of adding indexes is shown below:

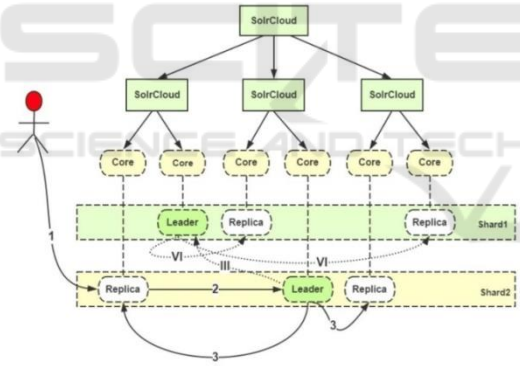

Figure 4. Add an index.

(1) The user submits the document to be added to any Replica (either a master copy or a slave copy);

(2) If the Replica receiving the request is not a Leader, it will forward the request to the Leader in the same Shard;

(3) The Leader routes the document to all other replicas of the Shard;

(III) If according to the routing rules, the current document does not belong to the current Shard, the Leader will forward it to the Leader of the corresponding Shard;

(VI) The corresponding Leader routes the file to each Replica in the Shard to complete the add operation.

The main process of the query module is to query information, and then return the query result to the user. This module contains the following parts: query classification, request query, query result sorting and so on. The query categories include text query, picture query, audio query and video query. The request query is responsible for querying the SRR service, and setting the proportion of the query conditions. The query results are sorted comprehensively according to the proportion of the query conditions, and finally returned to the user.

4 CONCLUSION

After the page test, the overall function test and the compatibility test of the manuscript retrieval system respectively, the display of each component of the page is normal, the button click function is normal, the paging function is normal, and the interface can return the response data normally, including the prompt data of the response success and the response failure. In the compatibility test, the windows and mac computers are tested under Internet Explorer, Firefox, and Google Chrome. The page is compatible with each browser and responds normally under different browsers. Through the system function test, the functions of all parts of the system can run normally, and the browser page can respond normally, which basically meets the needs of enterprises for searching.

This paper introduces Solr search engine in detail. By analyzing and comparing the advantages and disadvantages of Solr, Luence and ES search engines, Solr is selected as search engine to build search system. Moreover, the realization principle of Solr is studied in depth, and the realization of its core principle index is analyzed in depth. Through the in-depth understanding of Solr engine, the enterprise retrieval system is better designed and constructed.

The speech recognition function of Baidu is to convert the speech into text, which is convenient for subsequent word segmentation, index construction and query. However, due to the unclear pronunciation of the speech or the interference of environmental sounds, the translated text is wrong, which will affect the accuracy of the conversion and the matching degree of the query. The accuracy of the speech recognition function still needs to be further improved.

REFERENCES

Li ZY, Xu XM, Zhang D, Zhang P. Cross-Modal Hashing Retrieval Based on Deep Residual Network(J). *Computer Systems Science and Engineering*. 2021; 36(2):383-405. https://doi.org/ 10.32604/csse.2021.014563

Chung, S. W., Chung, J. S., & Kang, H. G. Perfect Match: Self-Supervised Embeddings for Cross-Modal Retrieval (J). *IEEE Journal of Selected Topics in Signal Processing*, 2020, 14(3), 568-576. http://doi.org/10.1109/jstsp.2020.2987720

Longepe, N., Mouche, A. A., Ferro-Famil, L., & Husson, R. Co-Cross-Polarization Coherence Over the Sea Surface From Sentinel-1 SAR Data: Perspectives for Mission Calibration and Wind Field Retrieval (J). *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60-66. http://doi.org/10.1109/tgrs.2021.3055979

Yao, X. X., Zhao, S. C., Lai, Y. K., She, D. Y., Liang, J., & Yang, J. F. APSE: Attention-Aware Polarity-Sensitive Embedding for Emotion-Based Image Retrieval (J). *IEEE Transactions on Multimedia*, 2021, 23, 4469-4482. http://doi.org/10.1109/tmm.2020.3042664

Huang, B., Pagowski, M., Trahan, S., Martin, C. R., Tangborn, A., Kondragunta, S., & Kleist, D. T. JEDI-Based Three-Dimensional Ensemble-Variational Data Assimilation System for Global Aerosol Forecasting at NCEP (J). *Journal of Advances in Modeling Earth Systems*, 2023, 15(4), 102-109. http://doi.org/10.1029/2022ms003232

Zhang, K. X., Zhou, L. H., Goldberg, M., Liu, X. P., Wolf, W., Tan, C. Y., & Liu, Q. H. A Methodology to Adjust ATMS Observations for Limb Effect and Its Applications (J). *Journal of Geophysical Research-Atmospheres*, 2017, 122(21), 11347-11356. http://doi.org/10.1002/2017jd026820

Fairbairn, D., de Rosnay, P., & Browne, P. A. The New Stand-Alone Surface Analysis at ECMWF: Implications for Land-Atmosphere DA Coupling (J). *Journal of Hydrometeorology*, 2019, 20(10), 2023-2042. http://doi.org/10.1175/jhm-d-19-0074.1

Pereira-Sanchez, V., Alvarez-Mon, M. A., Horinouchi, T., Kawagishi, R., Tan, M. P. J., Hooker, E. R., Teo, A. R. Examining Tweet Content and Engagement of Users With Tweets About Hikikomori in Japanese: Mixed Methods Study of Social Withdrawal(J). *Journal of Medical Internet Research*, 2022, 24(1), 331-338. http://doi.org/10.2196/31175

Nogueira, M. S., Raju, M., Gunther, J., Maryam, S., Amissah, M., Lu, H. H., Andersson-Engels, S. Tissue biomolecular and microstructure profiles in optical colorectal cancer delineation (J). *Journal of Physics D-Applied Physics*, 2021, 54(45), 132-139. http://doi.org/10.1088/1361-6463/ac1137

Zhao, X., Wang, S. J., Yu, W. W., Wei, H., Wei, C. L., Wang, B. C. Metrology of Time-Domain Soft X-Ray Attosecond Pulses and Reevaluation of Pulse Durations of Three Recent Experiments(J). *Physical Review Applied*, 2020, 13(3), 157-163. http://doi.org/10.1103/PhysRevApplied.13.034 043

Ufer, N., Simon, M., Lang, S., & Ommer, B. Large-scale interactive retrieval in art collections using multi-style feature aggregation (J). *Plos One*, 2021, 16(11), 69-72. http://doi.org/10.1371/journal.pone.0259718

Qin, P., Simis, S. G. H., & Tilstone, G. H. Radiometric validation of atmospheric correction for MERIS in the Baltic Sea based on continuous observations from ships and AERONET-OC (J). *Remote Sensing of Environment*, 2017, 200, 263-280. http://doi.org/10.1016/j.rse.2017.08.024