Design and Implementation of Search Engine Based on JAVA Technology

Yanyan Yang

Nanyang Institute of Technology, Nanyang, China, 473004

Keywords: Search Engine; JAVA; System Design

Abstract: Search engine that the Internet information retrieval system, the use of search engines can search in the network, crawling a lot of information, and its intelligent extraction, quality analysis, indexing, loading index database, and then according to the user's query request in a certain algorithm support from the index data to find information, and finally return to include all the matching keywords of the page. Search engines make a variety of special algorithms involved in the process to correlate the degree of information to the client in the order from high to low. JAVA technology innovation for the development of search engines has brought new impetus to promote a higher level of development. This paper puts forward the design and implementation of search engine based on JAVA technology.

The Composition of the Search Engine

Search engine is essentially a class of database, its work mode, including automatic information collection and regular search, such as Google engine will be in a certain period of time using spiders to actively search and found that the new site will extract the relevant information stored in the database, by This shows that the continuous update of the search engine database can continue to expand its scope of application to improve the user's convenience.

Specifically, the search engine consists of parsers, indexes and searches, and Web servers. The main function of the analytic program is to parse html, pdf, word, excel and other documents, the document preprocessing process is not only simply read the characters from the file, but also according to its special format to extract the relevant content, the application of the corresponding The open source parsing module gets the text information. In the search engine using Jdbc way to document the title, author, keyword and other attributes written to the database, write before get Nextld method to obtain the ID number to be inserted, and then return to the user with the system structure method, the user will the ID number is transmitted to the Lucene index, which corresponds to the database record. In the access to the page after the temporary storage in the temporary database, then need to establish the index in accordance with the inverted file format to store, in order to improve the efficiency of query information. The user enters the search condition in the search program, which retrieves through the index database, and then classifies the search results according to certain criteria and returns them to the user. Users through the browser query information, Web server connection index database and the user input query conditions, Web server to receive the user's query conditions in the index database query, sort, and then return to the user to complete the search. Search engine workflow includes four links, first in the network to crawl the web page, the establishment of index database, in the index database to retrieve information and finally the search results are processed and sorted, and feedback to the client.

The Advantages of JAVA Technology

Compared with other assembly language, JAVA advantage is mainly reflected in the following aspects: First, security. In the network environment, the security of JAVA technology is of great significance, and its security mechanism can effectively attack the malicious code, and ensure the security of the information to the greatest extent. Second, it is the mandatory. JAVA technology object-oriented process generally only supports one-way inheritance between classes, so to carry out multiple must have multiple interfaces, so JAVA has a mandatory feature. Again, it is the

dynamic. JAVA language can match the dynamic environment changes, so it is for a variety of systems, software has a good compatibility, especially it is easy to upgrade applications and JAVA dynamic advantage is more prominent. Finally, it is the multithreading features. JAVA multi-threaded features make the relevant applications, performance is better guaranteed, greatly improving the quality of the developer and the user's service.

The Design and Implementation of Search Engine Based on JAVA Technology

JAVA-based search engine design mainly includes network spiders, indexers and searchers in three parts to better improve the basic functions of search engines, so based on JAVA technology search engine design and implementation include the following aspects:

Software Development Environment. Internet connection based on JAVA technology. In the search engine design process application network spider is mainly to communicate with the Web server, better crawl web information, and download information. JAVA technology can provide a variety of Internet connection class, commonly used include socket class and URL class two. Socket class is in all connected to the network computer has a socket to promote the computer program to take effect, under normal circumstances these sockets are called the port, and set the number; any computer must specify a port number to connect Server, where the client must also specify a port number to complete the understanding of the request, even if many clients can connect to the same server port, but in fact a server program can only be used to listen to the same port. HTTP default port 80, the port is very important. JAVA technology is mainly defined for the Socket and Server Socket, are part of the socket program, which Socket class applied to the client, mainly for the client socket late; Server Socket for server-side socket statement.

In the formal establishment of Socket before the establishment of point-to-point to be a party to listen to the other side of the request, the formal establishment of the connection through the client and the server through the completion of communication, after the success of the client and the server side there is no difference, can promote the two-way data transmission, thus using the socket to read and write data. Application URL class can not only resolve the URL, you can also create the object after the completion of the host name and path to solve the board, and the URL class can even open an address from the URL to obtain the ability to retrieve information.

JAVA technology Chinese processing. In the crawling of the page, the search results often appear in Chinese characters distortion, the main reason for this problem is the character encoding, application JAVA technology as long as the pre-set the correct code, you can do the support of Chinese characters. Chinese characters are double-byte, China's mandatory provisions of the Han code for the GB2312, so the basic treatment of all Chinese applications are supported GB2312, which not only a secondary Chinese characters but also 9 symbols. In addition there is GBK code, but it is not mandatory for the specification, GBK, including GB2312.

JAVA multi-threaded mechanism. Application of multi-threaded mechanism to crawl web pages, indexing and search work can greatly improve the efficiency of the work of the background is similar to the background, the establishment of JAVA technology to clear the background code execution code, the code included in the JAVA thread run. Thread operations can be achieved through two methods, one is to determine the inherited thread object after the completion of the packaging thread work, through the thread class to develop a specific thread code, because JAVA does not support multiple inheritance, so the practicality of the method is not strong; Another is determined according to JAVA Runnbale interface to determine the run method, because JAVA support multi-interface operation, so fewer restrictions.

JDBC applications. Only through a reasonable way to complete the drive site queue in order to achieve a large number of network spiders access to the site, so you need to maintain the list of pages through the DBMS, under the action of JDBC to complete the SQL submission to control the form of the database. The use of JDBC operation is through the importajva.sql. And then determine the statement object, which belongs to the connection object, is relatively independent of the SQL statement, in the network spider program only use prepared remarks class can be better to improve efficiency, with prepared statement SQL statement is repeated to write the most important SQL

statement of the part, to the maximum extent possible to eliminate the need to write a number of issues. Finally, we should pay attention to Results, its main role is to save and return SQL command results.

The Realization of Network Spiders. The preparation of network spiders requires the construction of interfaces and classes. The web spiders include three classes and two interfaces. There are three types of spiders that are part of the spider interface to provide the spider method, which involves managing the pool of threads. The spider object reports to the discovery page and identifies the part of the spider job completion time that belongs to it. In addition, the spider operation promotes the use of two additional classes of Ispider Reportable and Iworkloadstorable. In the Ispider Reportable interface, the spider sends a partial event to the controller, defines it, processes it, retrieves it in the page, completes the page processing; the main function of Iworkloadstorable is to constrain the behavior of the spider. Spider's main job is to organize the access point list service, as long as the object can be identified in the list to complete the page storage and retrieval, when the spider program download is complete, spider worker the main page link back to the job, start spider program You can create spider worker class, which is very similar to the thread pool, you can ensure that the various threads at the same time operation, but also in a good spider and find the page.

The Realization of Lucene. Lucene is a high-performance and easy to extend the JAVA class library, which through the JAVA class can be completed in the program needs a variety of indexes and search, this time will be applied to Lucene JAVA technology can be better cross-platform. Lucene can be used in the application to add indexing and search functions, which can search any text format data can be converted, just the data source, language, etc. can be converted to text. In the remote Web server pages stored in the local file system documents, whether it is simple text or a specific format of the text, you can retrieve the text information.

Conclusion

In summary, in the massive network information, the application of search engine has greatly improved the efficiency and convenience of user information retrieval. This study analyzes the composition of search engine and the advantages of JAVA technology, and puts forward the design and implementation of search engine based on JAVA technology method. JAVA for the search engine provides an important technical support, especially in the search engine intelligent development environment and JAVA technology applications greatly improve the search engine functionality.

Acknowledgements

Fund Project: Project of Henan Province Science and Technology Project (Project Name: Research on Data Fusion Technology of Multi - source Heterogeneity, Project No: 162102210358)

References

[1] CHU Li-li.Application of Java-based search engine technology in Web information mining [J]. Journal of Liaoning Technical University (Natural Science Edition), 2016 (5): 1006-1008

[2] Research and Implementation of Web Vertical Search Engine Technology Based on JAVA + LUCENE + HERITRIX [D]. Hebei University of Technology, 2014

[3] Zhu Maosheng, Wang Bin, Cheng Xueqi. Meta Search Engine and Its Realization [J]. Computer Engineering, 2016, 28 (11): 11-12.

[4] Wang Shi, Gao Wen. Clustering method in data mining [J] .Computer Science, 2015,27 (04): 42-45.

[5] Chen Ning, Zhou Longxiang. Data mining in the Internet application [J] .Computer Science, 2016, 26 (07): 44-49.

[6] Xu Qian. Network information retrieval intelligent trend[J]. Library Theory and Practice, 2016 (02): 63-65.

[7] Wang Shi. Web-based access to information mining recommended method research [D]. Beijing: Graduate School of Chinese Academy of Sciences, 2011.

[8] Xie Bin. Personalized meta-search engine research and implementation[D]. Inner Mongolia University of Science and Technology Graduate School, 2014