Digital Archive Management Based on Lucene Full-text Search Engine

Daqing Wang1,*

1Liaocheng Guangming Eye Hospital, Liaocheng, Shandong, China, 252000

*Corresponding author e-mail: wangdaqing@eye0635.com

Abstract. With the rapid development of society, people have an increasing higher demand for information. As one of the vital information resources, archives are facing the significant impact of the global information tide. The original manual management method for archives can no longer meet the needs of people for their efficient management and utilization. How to transform archive management from manual to digital mode based on advanced information technology has become a hot issue of concern and research. Based on the digital application of archive management, supported by the advanced computer, communication, network, database, and multimedia technologies, we designed and developed a general digital system of archive management. The topological structure of the client/browser/server is used in the system, with the cross-platform characteristics.

Keyword: Full-Text Retrieval, Engine, Digital Archive Management, Network Technology

1. Introduction
The traditional search engine can be divided into three information retrieval methods according to various ways of work: full-text search engine catalog search engine, and metasearch engine [1-2]. Although this type of search engine has brought great convenience to users in terms of retrieval, it still cannot understand people's query request well, distinguish the synonyms of keywords associate related synonyms, or consider the special language environment of keywords [3-6]. It has been only a few years since Lucene full-text search engine was developed, but it leads to the new trend of network information retrieval with its magnificent natural language understanding and knowledge processing ability and shows the incomparable advantages of traditional search engines. Generally speaking, it has the characteristics of intelligent natural language technology, personalized information service, extensive file supportability, powerful indicator function, support for multilingual search, etc.

2. Design of lucene full-text search engine system based on the knowledge base
The intelligent searcher responds to the user's retrieval request, queries in the inverted indicator database, sorts the relevance, and feeds back the sorting results to the user. The synonymy is taken as an example. When users input keywords, they first annotate them according to the general dictionary,
then analyze their various related relationships according to the subject dictionary, finally, retrieve the corresponding results according to the indicator library, and submit them to users. These synonyms have the same identifier, which is the classification identifier obtained by the searcher. Next, the searcher searches according to this identifier in the inverted indicator library, and then the URL corresponding to all synonyms can be retrieved and fed back to the user.

The intelligent indexer allows us to understand the information searched by the intelligent searcher, extract indicator items from it, establish indicator tables to represent documents, and generate document libraries. In addition to the objective indicator items such as file number, file unit, industry, full file number, etc., the intelligent indexer can also indicate the content indicator items such as keywords, their weights, phrases, etc., which are used to reflect the content of documents, to ensure the accuracy and relevance of the description of search results. Web spiders get a URL first, and determine whether the URL has been processed according to the records of the database. If not, download the corresponding web page, analyze the web page, get other links referenced, and join the queue. Then, according to a specific traversal algorithm, submit the next URL to the main program, and so on, until no URL is available. The function of the user interface is to input user query conditions, display search results, and provide relevance feedback. When designing and implementing the user interface of the Lucene full-text search engine, we should fully consider the user's thinking habits and behavior.

As an idea applied in the field of artificial intelligence, Ontology has its characteristics and advantages and is suitable for knowledge representation. Therefore, ontology technology is used to represent knowledge and build an advanced manufacturing technology knowledge base. Owl (Ontology Web Language) is selected as an ontology description language. A kind of OOA method is used to build the prototype system of domain ontology. For the collection of domain knowledge, with the help of domain experts, and with reference to the authoritative works in this field, the knowledge of advanced manufacturing technology is obtained. The knowledge base in this system consists of two parts: one is the class base, which is the ontology description of scientific and technological achievements. The second is the instance library, which is the concrete instance collection of the class in the class library. It is the primary carrier of information. Through the algorithm of artificial intelligence, the net spider can have the function of learning, and take the most effective search strategy, choose the best time to get the information collected and organized automatically from the Internet, and automatically complete the indicator of online information. The search engine can mine and get information as much as possible in any place and at any time of the Internet or intranet. To improve the search speed, the search engine uses a focused algorithm to capture relevant web resources. The idea is to build a mathematical model, calculate the downloaded Web pages, get an evaluation weight, rank the downloaded Web pages according to the weight, and then download them in order. A new algorithm is based on link content evaluation. Content-based evaluation mainly evaluates the value of links based on the similarity between topics (such as keywords) and linked texts.

Assuming that the maximum depth of the large data set domain is N, and the search engine corresponding to the spatial grid into which \( N \leq n \) is finally divided to \( q = q_1q_2\ldots q_N \), \( q_1, q_2, \ldots, q_N \in \{0,1,L,7\} \). Hence, \( q_1q_2\ldots q_N \) can be obtained. This form of search engine uniquely represents any one of the smallest space Grid unit. Moreover, this spatial search engine has a very good property. For any point in the space with known coordinates, it can directly find the search engine of its grid. At the same time, if any spatial network is known. The grid's search engine can also directly identify the coordinates of the front left corner. For example, for a point \( p(x,y,z) \) in space, an integer is first taken for \( x, y, z \) to obtain the corresponding binary representation.
\[ x = i_i L \ i_n, \]
\[ y = j_j j_j L \ j_n, \quad i, j, k \in \{0, 1\}, \ i = 1, 2, \ldots, n \]
\[ z = k_k k_k L k_n, \]  

Subsequently, the cube grid search engine where \( p \) is located is \( q_i q_j q_k \), where
\[ q_i = i_i + 2j_i + 4k_i, \quad i, j, k = 1, 2, \ldots, N \]

If a spatial grid retrieval engine is known as \( q_i q_j q_k \), the coordinates of the lower left corner of the spatial grid are
\[ x' = i_i L \ i_n 0 \ L 0, \] the number of 0s followed by \( n - N \)
\[ y' = j_j j_j L \ j_n 0 \ L 0, \] the number of 0s followed by \( n - N \)
\[ z' = k_k k_k L k_n 0 \ L 0, \] the number of 0s followed by \( n - N \)

For example, compared with direct partition, the direct partition method needs six times to judge whether a point is located in a specific area, while the digital file information algorithm only needs one time to judge, which takes about \( 1/6 \) of the direct partition method. It dramatically accelerates the speed of the fast adaptive algorithm, which is the source of the fast and efficient digital file information algorithm.

**3. Application of lucene full-text search engine technology in the management of digital archives**

The so-called digital archive management refers to the integrated service that digital archives use computer networks, artificial intelligence, and many other information technologies to obtain and analyze the background, habits, preferences, and requirements of each user, to provide different users with full satisfaction of their individual information needs. That is to say, the personalization of service time and space provides services to users at the time and place they want; the personalization of service mode can develop services according to users' personal interests and characteristics; the personalization of service content provides services that are no longer the same, but take what they need and get what they want. The following describes the application of Lucene full-text search engine technology in the management of digital archives.

**Table 1. Evaluation indicator system of archives protection level of digital archive management**

| Target | First-level indicator | Second-level indicator |
|--------|-----------------------|------------------------|
| Archives protection level of Digital Archive management | Digital archive management | Filing of Digital Archive management projects |
| | | Government supervision and Implementation |
| | | Capital investment |
| | | Storage level of Archives |
| | | Evaluation and identification of inheritors |
| | | Material rewards of inheritors |
| | | Number of inheritors |
| | | Inheritor's knowledge and culture level |
| | Social environment | Awareness of mass protection |
| | | The level and strength of folk propaganda |
| | Science and technology | Public participation |
| | | Scientific and technological means of Archival Protection |
| | | Construction of information platform |
| | | Academic support and cooperation in Colleges and Universities |
A total of 20 questionnaires were distributed, and 15 valid questionnaires were recovered. Each level of indicators has a judgment matrix evaluated by 15 experts. The following lists the matrix formed by expert B on the indicator level of "Digital Archive management" in the expert consultation table, as shown in Table 2 below.

**Table 2. "Digital archive management" indicator hierarchy matrix of experts**

| Indicator | Filing of non-genetic projects | Government regulation | Capital investment | Archives keeping level |
|-----------|-------------------------------|-----------------------|-------------------|-----------------------|
| Filing of non-genetic projects | 1/3 | 3 | 1/4 | 1/3 |
| Government regulation | 1/3 | 1 | 1/6 | 1/5 |
| Capital investment | 4 | 6 | 1 | 4 |
| Archives keeping level | 3 | 5 | 1/4 | 1 |

The intelligent fault-tolerant technology of the file information intelligent retrieval service: due to the lack of unified standards for Chinese information and the diversity of file data storage; also due to the user's habits, the user's understanding of the data is not deep enough, which greatly affects the accuracy and recall rate of the file retrieval. For example: name of person Li Peng, and Li Peng, Li Peng; number 112 and one two, 112, full angle and half-angle in brackets, and various formats in date are considered to be different data in the computer, but users may all think that they are the same data. Hence, intelligent fault-tolerant technology in file retrieval can increase the flexibility and affinity of personal retrieval.

When the file user enters a search term, the system will automatically compare the search term with the relevant word in the internal vocabulary, and display the phrase logically related to the word on the page, and ask the file user whether to search the keyword, which can help the file user selects the keyword closer to their own search goal to complete the search and improve the information query rate. The management of digital archives is a key part of the whole digital archives system. It takes users as the center and satisfies users' personalized value pursuit as the goal. In the management system, the knowledge base is the foundation and core of the Lucene full-text search engine. It is refined and expanded based on the information resource base of digital archives. It is the judgment, extraction, analysis, and generalization of the information resource base of digital archives. It is also the processing of images to better meet the needs of users.

4. Conclusion

Lucene full-text search engine is a personalized intelligent search engine, a subdivision, and extension of the traditional search engines. It has integrated the latest artificial intelligence, computer network, database, data mining, machine learning, and other technologies. Due to its advantages, Lucene full-text search engine, as a crucial part of the whole digital archives system, plays an irreplaceable role in the construction of digital archives. The establishment of a digital archive management system based on Lucene full-text search engine can not only reflect the advantages of many functions integrated in the system but also significantly improve the working efficiency of the system, accelerate the operation of business processes, and provide great convenience for users to search the Internet information based on its intelligent, personalized functions.

**References**

[1] Cheng Ziyu, & An Haining. (2012). The building of digital archives personalized service website based on web 2.0. Physics Procedia, 25, 2096-2102.

[2] Isto Huvila. (2016). Affective capitalism of knowing and the society of search engine. Aslib Journal of Information Management, 68(5), 566-588.

[3] Ronilda Lacson, Sonali Desai, Adam Landman, Randall Proctor, & Ramin Khorasani. (2017). Impact of a health information technology intervention on the follow-up management of
pulmonary nodules. Journal of Digital Imaging, 31(6), 1-7.

[4] Xuexiang Li, Fulin Bian, & Yongge Shi. (2012). System integration of digital real estate management based on service. Physics Procedia, 24, 1012 – 1017.

[5] Davies R. J. (2016). Digital computing techniques in the manufacture and operation of engine management systems. Aeronautical Journal, 79(776), 349-353.

[6] M. Kardoš. (2013). Methods of digital photogrammetry in forest management in slovakia. Journal of Forest Science, 59(2), 54-63.