2020 5th International Conference on Education and Social Development (ICESD 2020)
ISBN: 978-1-60595-676-3

Research and Realization on Big Data of Science and Technology Resources Search Engine Based on SOLR

Lin MU[a], Ye-Yang ZHANG[*] and Yang HAN[b]

Institute of Scientific and Technical Information of China, Beijing 100038, China

[a]mulin@istic.ac.cn, [*]zhangyy@istic.ac.cn, [b]hany@istic.ac.cn

Keywords: SOLR; Science and technology resource; Big data; Search engine.

Abstract. Since the reform and opening to the world, the development of science and technology has made great progress. A large amount of scientific and technological resources has been accumulated in China, the storage forms of which can be divided into two categories: structured data and unstructured data. SOLR playing an important role in this field is a means for us to obtain information and data fast and precisely. In this paper, we import the stored data to SOLR, and establish the index of relevant content. Then the retrieval results data could be presented quickly on the basis of index.

Introduction

Since the reform and opening to the world, the development of science and technology has made great progress. In 2017, the main quantitative indicators of scientific and technological innovation have leaped to the top 3 in the world. China has become an influential country of innovation. The international ranking of 10 indicators of scientific and technological innovation, such as personnel, funds, papers and invention patents, has all entered the top three in the world. A large amount of scientific and technological resources has been accumulated in China, the storage forms of which are also rich and diverse. Generally speaking, it can be divided into two categories: structured data (i.e. line data, stored in ordinary databases, which can be logically expressed in a two-dimensional table structure) and unstructured data (referring to all formats of XML, office documents, HTML, text, pictures, various reports, video, audio information, etc., including completely unstructured data and semi-structured data Chemical data).

How to find the useful information from the massive science scientific and technological resources efficiently becomes a prose because traditional ways are expensive and difficult to extend. The search engine which plays an important role in this field is a means for us to obtain information and data convenient and fast. An excellent search engine framework requires features such as efficient indexing, timely search responses, and reliable system services. In order to solve these problems, this paper proposes a large-scale full-text visualization analysis system based on SOLR [1, 2, 3, 4]. It integrates the science and technology resources in different resource databases and realizes quick search on this basis. The main reason why SOLR is adopted in this paper is that it is a relatively mature solution in the field of information retrieval, and it also has high efficiency and speed for large-scale data.

SOLR is a high-performance full-text retrieval tool based on Java, which is widely used in Enterprises. SOLR is built based on Lucene [5, 6], which provides a complete search function. The SOLR server accepts the HTTP data request, and then passes the return data results through XML, JSON and other formats. SOLR can data from different types of databases and different formats are uniformly indexed, and carry on the full text retrieval.

Technical Framework

SOLR search engine can effectively solve the problem of full-text search and improve the search (query) performance of the system. Its technical structure is as follows:

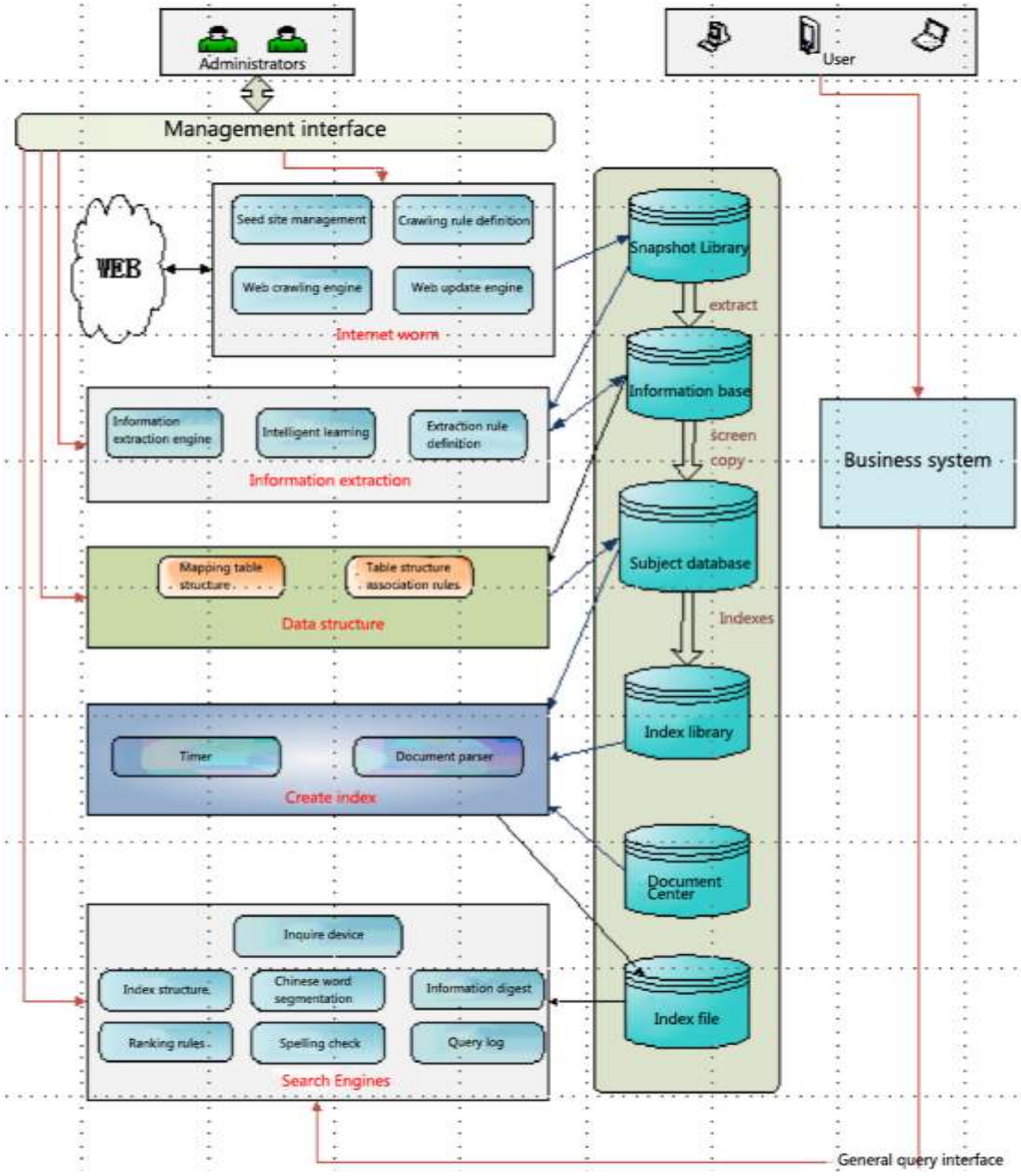

Figure 1. Technical Structure Based on SOLR.

In the portal of golden education project or other projects, the main search includes two categories: one is all kinds of information on the web page crawled from the internal network; the other is business data based on the specification and from the business system, such as retrieving the information of a certain student, teacher and school.

Experiment

Simulation environment:

CPU: Intel (R) Xeon (R) CPU e5-2660 V2 @ 2.20GHz 32 core; Memory: 128G; Hard disk: 10T;

Concurrent number: 500; Increase mode of concurrent users: initialize 5 concurrent users per second.

Data volume: about 0.5billion, 1.3TB (including project, personnel, organization, paper, patent, audio, video and other data).

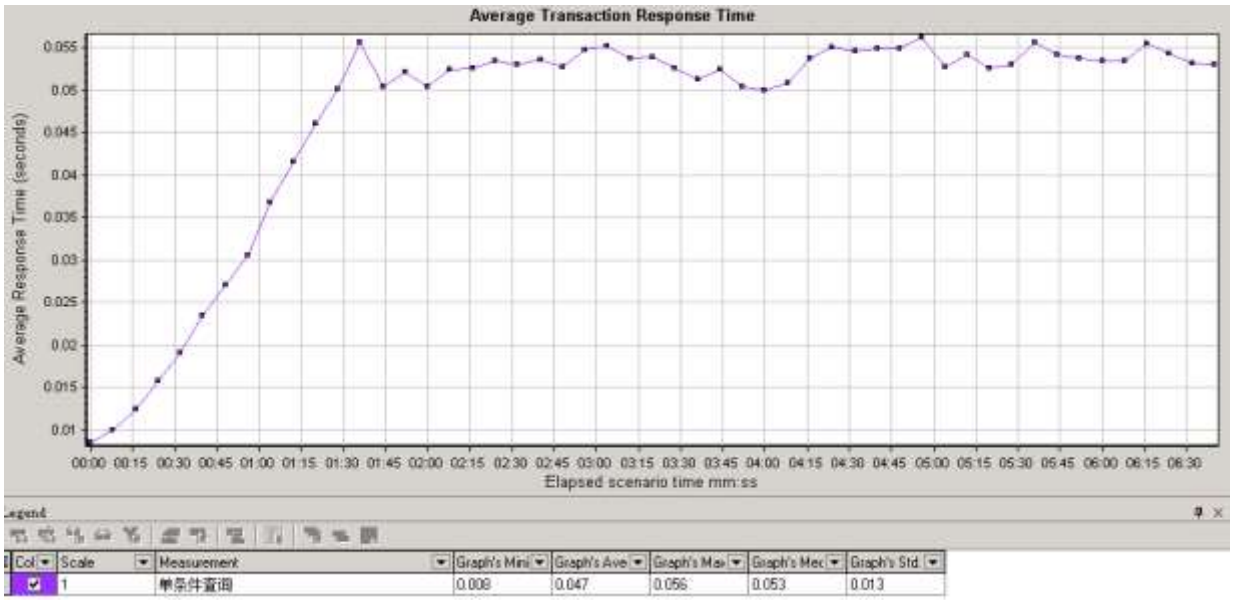

Figure 2. Average Transaction Response Time.

Figure 3. Transactions per Second.

The results show that the average processing power of personnel association query is 8946.626 transactions/second and the average response time is 0.047 seconds.

Conclusion and Future Work

The enterprise-level open source retrieval platform SOLR has powerful functions, low acquisition cost, high secondary development efficiency, and strong scalability. In the face of massive data, SOLR performs well, and its concurrent performance is good. At the same time, cache resources can reach a high utilization rate. Applying SOLR's functional characteristics to the construction of technology resource platforms has obvious advantages and broad application prospects.

With the development of artificial intelligence, more and more intelligent algorithms and tools will emerge. In the next step, we will pay more attention to some other tools, such as Hermes and Zebra [6, 7, 8].

Acknowledgement

This research was financially supported by the Innovation Research Fund of Institute of Scientific and Technical Information of China in 2019 (Project No. ZD2019-14 and ZC2019-07).

References

[1] Shahi D. Apache SOLR [M]. Apress, 2016.

[2] http://en.wikipedia.org/wiki/solr.

[3] Yadav, D., Sanchez-Cuadrado, S., Morato, J., et al. An Approach for Spatial Search Using SOLR [C]. Confluence 2013: The Next Generation Information Technology Summit (4th International Conference), 202-208, 2013.

[4] Mei, J.Z. Research on Parallel Indexing and Cache of Searching with Massive Data based on Solr [D]. Central China Normal University, 2016.

[5] http://lucene.apache.org/.

[6] Stephen, E. LUCENE/SOLR Now Ready for the Big Leagues [J]. Online: 2012 36 (6): 40-43.