International Conference on Advances in Materials, Machinery, Electrical Engineering (AMMEE 2017)

The Research and Implementation of File Information Retrieval System Based on Big Data Semantic

Zebo Zhu[1] and Baochuan Lin[2]

[1]China Satellite Maritime Tracking and Controlling Department, Jiangyin 214431, China

[2]Zhejiang Agricultural Business College, Shaoxing 312000, China

Keywords: big data, semantic, information retrieval, distributed storage.

Abstract. There are many defects in traditional file service system, and this paper suggests a new system that based on the semantics of big data. It introduces some concepts, such as big data, semantic and information retrieval firstly. Secondly it introduces the key techniques. Then it analyzes and designs the new system from the aspects of system structure, database, and data flow. Finally it briefly introduces the implemented new system.

1. Introduction

In order to improve the success of spacecraft flight test and ensure the reliability of spacecraft flight test, a large number of flight test preparations have been carried out between various systems and subsystems. The number of documents, documents and documents is about 1000, Size up to 10GB or so. At present, these documents, the file is mainly stored in the file on the server, can not view the specific content of the document online, only through the download to learn the specific content, can not search the way to get the desired information, Discover the potential relationship between documents and information.

At the same time, with the advancement of space experiment task, the accumulated resources have become very large. This paper proposes a test document service system based on large data semantics, which aims to establish a system that can complete resource sharing and more intelligent, While separating the semantic information from the content of the document. This not only provides a new way of thinking for the improvement of test document management and information retrieval technology, but also provides a solid theoretical basis for it, so that the test documents of previous years can serve the existing tasks better.

2. A Survey of Large Data Semantic Information Retrieval

2.1 Big Data Overview

Large data refers to a collection of data that can not be effectively perceived, acquired, managed, processed, and serviced in a tolerable time with commonly used hardware and software tools.

Big data has four characteristics, namely four V features:

A) .Volume, which is its basic characteristics, mainly reflected in the Internet technology, sensor technology and other emerging technologies widely used, people get more and more convenient data, and the data closer to the original thing itself, resulting in the description of things Data dimension continues to increase, the amount of data is increasing.

B). Variety, which is an important feature, which is mainly reflected in the semi-structured, unstructured data in a large number of appear. Semi-structured, unstructured data There is no uniform data description method, it is difficult to use a fixed structure to represent. In the recording of semi-structured, unstructured data, but also need to record the corresponding data structure, which not only increases the amount of data storage, but also increase the difficulty of data processing.

C). Data processing is fast, which is the key difference between large data and Magnanimity Data. Large data in many cases must be processed in a short period of time, if not effectively processed, the subsequent data processing results will be outdated and invalid, the data will lose its due value, there is no meaning The

D). The data density is low because the large data contains a large amount of semi-structured, unstructured raw data, which not only retains the original appearance of its data, but does not handle any data. The advantage of this approach is that all the details of the data can be presented, and a more comprehensive analysis can be obtained. But there are corresponding shortcomings, namely the introduction of a large number of meaningless data, or even incorrect, wrong information, so the density of its data density is low.

2.2 An Overview of Semantics

Semantics, that is, Semantic, can be simply seen as a kind of information in the real world of things represented by the concept, it is actually the interpretation of things and logical representation.

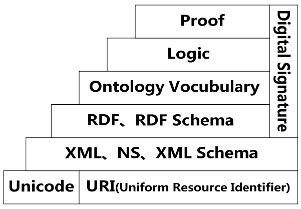

Figure 1. Web semantic architecture framework model

The above figure is the Web semantic architecture framework model presented at the 2000 XML conference. The model from the top to the lower level are Proof, Logic, Ontology Vocubulary, RDF, XML, Unicode and URI, Digital Signature runs through Proof, Logic, Ontology Vocubulary and RDF the whole process.

The main function of semantics is to make the computer more convenient and quick to deal with and find the data in different data sources, and to add information to the meaning of things in the data, modify the simple information into a complex, rich meaning, can be explained, Exchange and deal with the information, thus reasoning new facts.

2.3 Information Retrieval Overview

Information Retrieval, also known as Information Storage and Retrieval, refers to the process and technology of information being organized in a certain way and based on the needs of the information user.

With the emergence of semi-structured, unstructured data, traditional information retrieval techniques, such as PageRank, HITS, are no longer applicable. At the same time, the scope of information retrieval has changed from a single document information to a specific transaction, knowledge retrieval, and more and more attention to the semantics between things. Thus, the semantic retrieval method came into being. Before introducing the semantic retrieval method, first introduce RDF and SPARQL.

Simple Protocol and RDF Query Language is a standardized query language for RDF. It performs a few steps: to get a complete query; to convert a query into an abstract syntax; to tune the converted abstract syntax; to execute in a dataset After tuning the abstract syntax, get the results.

At present, the commonly used semantic retrieval methods are the following three kinds.

A). Concept semantic retrieval

As mentioned above, many concepts add relevant semantic information. Therefore, it is possible to use the semantic information contained in the information to search, thereby improving the accuracy of the search, which is the concept of semantic search.

B). Keyword semantic retrieval

In the keyword semantic retrieval, the RDF graphs and features are usually associated with the keyword search, and then the search results are sorted to obtain a sequence of information related to the degree of the keyword. The disadvantage of this method is that users must be familiar with the details of the data, while the query semantics are more familiar with the query to get the results you want.

C). Correlation analysis semantic retrieval

The association analysis is used to represent the sequence relationships between two interrelated entities. The essence is to use the correlation method to discover the correlation and complexity of the entities. Correlation analysis semantic retrieval, that is, correlation analysis for semantic retrieval, its advantage is the flexibility of the search, more scalable.

3. The Key Technology Research

3.1 Word Content Reading Method

A). Jocob

Jocob is actually a bridge, java and com is connected or com32 function of a middleware, it can not directly extract the contents of the word, the need for the corresponding dll file support. The advantage of this method is supported by the corresponding dll file, the need to write the processing function is relatively simple; drawback is not able to block extraction, more content, classification is not easy.

B). Apache POI

POI is a subproject of Apache that handles non-versioned office documents, including 2003 and 2007, Word, Excel, PowerPoint, and so on. Its advantages can support all kinds of office documents, and can be block read; the disadvantage is that the document can not read the picture, and the font support more trouble.

C). Docx4j

Docx4j project is mainly used to deal with office2007 related documents, its main functions include reading docx documents in the content, font support, compare document differences, word conversion to html or PDF and so on. It has the advantage of office2007 support is better and powerful; the disadvantage is relative to the Apache POI in office2003 document processing is relatively weak, fewer users, there may be unknown exceptions.

3.2 Research on Data Storage Method

A). Relational Data Storage

Relational data is a data that is described in a two-dimensional form in relation to a mathematical model. As the name suggests, relational data storage is to relational data for data storage units to store, usually in the form of two-dimensional table to store data, more common is the relational database, the current popular large-scale relational database DB2, Oracle, SQL Server, MySQL, SyBase, and so on. The advantage of this method is that the data storage content is intuitive, visualization is better, the disadvantage is not applicable to the storage of non-relational data, such as unstructured data, object-oriented data.

B). Object - oriented data storage

Object-oriented data storage method is a kind of object as the basic unit of a method of storage. Because the attributes of the object can be set according to the specific needs, and each object alone to maintain their own attributes, so the storage method is more flexible, while reducing the cost of data management, the disadvantage is not applicable to non-abstract data storage.

C). Distributed data storage

Distributed data storage is the main method to solve the large data storage, the main advantage is strong fault tolerance, scalability, consistency is better.

3.3 Research on Query Scheme Based on Semantic Data

When the client needs to query the name of the person in class X, the system will query the result in all the topologies and assign the name {x.aa, x.bb, x.cc} to class X as the result of the query and pass it to Map / Reduce after the calculation, the final results will be returned to the client.

This method is implemented by two queries and separates the semantic query from the distributed query. Through the semantic query to obtain the intrinsic link between the data, generate the corresponding data mapping file and RDF query results, and then query the results through Map / Reduce distributed computing, and ultimately get the query results.

4. System Design and Implementation

4.1 System Frame Design

The overall design framework of the test document information retrieval system based on large data semantics is divided into four layers, namely, presentation layer, control layer, business logic layer and data resource layer. The advantage of using the stratification is to make the software modules perform their duties, separate from each other; do not interfere with each other. If the user needs to change a layer, just in the corresponding layer to modify the code, without the need to modify in other layers. This approach is not only conducive to the realization of the system division of labor, but also conducive to future maintenance.

4.2 Database Design

According to the actual needs analysis, the paper uses the object-oriented method to analyze each module, and design the corresponding data table. The data table and its relationship are shown in Fig.

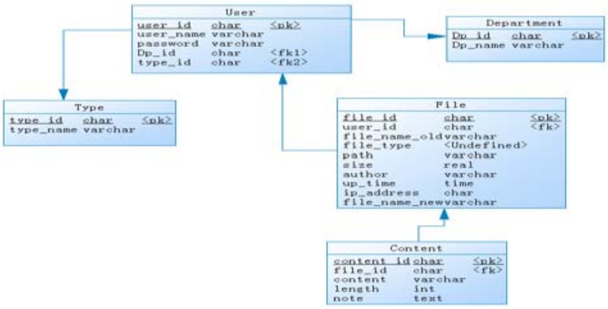

Figure 2. Data sheet and its relationship diagram

4.3 Data Flow Relationship

Figure 3 depicts the flow of data within the system. The user first initiates all kinds of requests through the browser. Spring DI calls the Struts interceptor to intercept the request and determines the model (Action processing class) to deal with the request, including the word document processing, storage file control, semantic query, Map / Reduce processing, call Hibernate and connection pool on the database for various operations, by the Spring DI decided by which view to display data, and through the Action Response feedback to the user.

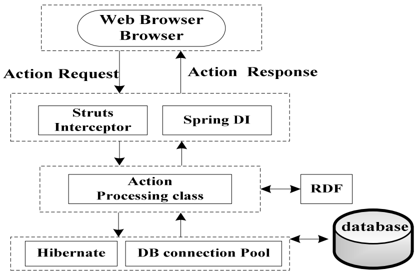

Figure 3. Data flow diagram

5. Summary

Because of the traditional test document service system there are many shortcomings, such as can not view the specific content of the document online, only through the download to learn the specific content, can not be online search way to get the information you want, but can not find the document Between the information and the potential relationship between the information, this paper presents a large data semantics based on the test document information retrieval system. In this paper, the concepts of large data, semantics and information retrieval are introduced, and the key technologies used in this paper are introduced in detail. Finally, the system architecture, database and data flow are analyzed and designed in detail. Test run to meet the design requirements.

References

[1]. Meng Xiaofeng, Ci Xiang. Large Data Management: Concepts, Technology and Challenges [J]. Computer Development and Research, 2013 (1): 146-169.

[2]. Yuan Jinping, Bao Aihua, Yao Li. Semantic Web technology and its logical basis [C]. Computer Engineering, 2008, 34 (24) 194-196.

[3]. Xu Jiabiang. Information retrieval Beijing: National Defense Industry Press, 2009.