Abstract—In order to build an efficient basic information system of science and technology experts based on Web mining, a novel system architecture for application is proposed in this paper. The proposed system architecture integrates spider module, local distributed storage and Mongo-DB. The basic experts information of science and technology appeared in the Websites are synthesized as two format and using two strategies to deal with it respectively. The normalized texts which extracted from Web page by URLs are suggested. The extracted results include the name, sex, birth, hometown and professional title of science and technology experts respectively. The data stream flow, the information management model for the users and science and technology experts, the target website URLs and URLs management model, and data processing module are introduced in detailed. The synchronization of multiple databases and replica sets architecture for sharing cluster architecture is proposed in application system. Experiments show that the application system obtains a very high efficiency. The results show as by proposed system architecture can satisfy the application requirements for the customer.

Keywords- system architecture, science and technology experts, distributed storage, sharing cluster architecture, Mongo-DB, spider module, Web mining

I. INTRODUCTION

In recent years, the rapid development of Internet brings geometric growth to web information, and the vast volume of information means it has the characteristic of polynary and redundant as well. Web pages cannot be directly made use of by traditional database systems for its semi-structured characteristic [1]. How to use the information better becomes the focus of attention. At present, most web pages are given to the semi-structured document in HTML form. Web documents can be represented as unstructured documents, semi-structured document and structured document. For the correct extraction of web information, a lot of work has been done at home and abroad. References [2] put forward a method that based on spider algorithm and DOM thinking; it is comprehensive and includes some advantages of traditional method. It will be more effective, convenient and accurate on the Web information extraction. References [3] proposed adopt hybrid conditional model for web information extraction. The model is a conditional probability model and each state has an exponential model as the conditional probability of the next state. The hybrid conditional model can use the characteristics of the contextual information based on different kinds of web pages, and that of overlapping and no independence to train parameters in different granularity. Reference [4] describes the design and implementation of a web information extraction system module which is based on R-G-B Algorithm.

For the system architecture, many researchers have proposed different ideal to deal with the lots of different information. Reference [5] proposed a architecture to fulfill the researchers’ requirements for the Arts and Humanities. Reference [6] proposed educational metropolitan area network structure. Reference [7] discussed how North Carolina State University virtual computing laboratory technology can be used to implement distributed reconfigurable data centers and information technology services in educational institutions. Reference [8] is about the architecture of scientific computing library portal for sequential algorithm which covers the four mentioned disciplines. Reference [9] shows an integration framework for information system based on web service. Reference [10] proposed a dynamic metadata query algorithm of High-speed Railway basic data based on heterogeneous integration, the pattern supports upwards integrated access interface, and also provides the description for semantic-level common metadata retrieval. Reference [11] proposed a security architecture design based on SOA and the core architecture design are given.

Based on our past work [12-15], we select technology expert information page from the well-known colleges and universities in China in order to build an efficient basic information system of science and technology experts based on Web mining and study the Multi-factor matching method for science and technology expert information extraction. The second part gives the system architecture for expert of science and technology expert information extraction. The third part we introduced the system management model. Target website URLs and URLs management model is proposed in the fourth part. The spider module is given in the fifth part and the data processing module is discussed in the sixth part. In the seventh part we introduced the system local distributed storage.

*Corresponding author. hyitzqy@126.com
II. SYSTEM ARCHITECTURE

The proposed system is divided into six modules of the user management module, the science and technology experts’ information management module, the spider module, the target website URLs management module, and the URLs management module, the data processing module. Among them, the user management module is divided into user registration, user login, and update user’s information. The science and technology experts’ information management module is aimed to manage the science and technology experts’ information including the addition, alter, delete of science and technology experts’ information. The spider module’s function is traversing each target website in order to find out the introduction pages of science and technology experts. The target website URLs management module is aimed to manage the URLs of target website. The URLs module is aimed to manage the URLs of the introduction pages of science and technology experts. At last but no least, the data processing module’s function is deal with the source web page according to the URLs of the introduction pages of science and technology experts. This module includes the identification of science and technology experts’ basic information like: name, sex, birthday, professional title.

The proposed system is a complex system. The proposed system contains two key parts. The first key part focuses on user authentication, the spider, data processing and so on. The second key part focuses on local distributed storage. Fig. 1 shows the full system architecture.

III. USERS AND SCIENCE AND TECHNOLOGY EXPERTS’ INFORMATION MANAGEMENT MODEL

System administrators can manage users, role assignment, and user review. But system administrators have no permission to operate the other functions in system. Normal users can register, login and manage personal information. In the proposed system also integrated the Web Service invoked to authenticate the user. Because of the Mongo-DB is document-based database, but we could define it as relational database model. Fig. 3 shows the system user model.

The fields of science and technology experts’ information management model mainly came by web mining. So there are lots of mistakes in them. The specific user assigned by administrators could manage the science and technology experts’ information. Because of the Mongo-DB is document-based database, but we could define it as relational database model. Fig. 4 shows the science and technology experts’ information management model.

IV. TARGET WEBSITE URLS AND URLS MANAGEMENT MODEL

The fields of target website URLs mainly came from web mining. It consists of the URLs of colleges and universities website, the URLs of hospital website, and the URLs of research institute website and so on. The specific user assigned by administrators could manage the target website URLs. Because of the Mongo-DB is document-based database, but we could define it as relational database model. Fig. 5 shows the target website URLs model.
The fields of URLs management model consist of the child URLs among the target website URLs. In the URLs lots of URLs have nothing to do with the Science and technology experts. Data processing module can identifications them automatically. The specific user assigned by administrators also could manage them artificially. Because of the MongoDB is document-based database, but we could define it as relational database model. Fig. 6 shows the URLs management model.

Figure 5. Target website URLs model  Figure 6. URLs management model

V. SPIDER MODULE

In order to discovery the science and technology experts' information completely, we must implement the spider to traverse the whole target website. We could get the most URLs of the target website. It is known to us all that the website could be regarded as directed graph. When we use Breadth-first algorithm, we could control the layer of traversal better. We also could get good effect on the fetch result at the same time. In the proposed system, we use Breadth-first algorithm. Fig. 7 gives a directed graph and the node traversal order in the Breadth-first algorithm shows in the Fig. 7.

Figure 7. The node traversal order in the Breadth-first algorithm

VI. DATA PROCESSING MODULE

Because of diversity of information on the website and the versatility of the data processing, we must add the data processing module in the proposed system in order to get the correct result what we expected. But the addition is not a simple data processing module. Through statistics on the most website of colleges and universities website, there are two kinds of style to introduce the science and technology experts:

- The first style is paragraph text showing. The translated web page shows in Fig. 8. (http://www.tsinghua.edu.cn/publish/th/6182/2011/2011012109573804469456/20110121095738044694565_.html).
- The second style is chart format showing. The translated web page shows in Fig. 9. (http://web.nuist.edu.cn/JSJ/szdw/sunxingming.asp).

Figure 8. The example of paragraph text showing

Xingxing Sun
Position: Professor
Affiliation: Beijing University of Information Engineering, Computer and Software Engineering

Figure 9. The example of chart showing

Through the statistics, we integrated Chinese word segmentation, Chinese semantic analysis, keyword matching, shortest distance matching and other methods in the data processing module in order to get the correct result what we expected.

VII. LOCAL DISTRIBUTED STORAGE

In order to solve the large mount of data, ensure the high availability of the proposed system, improve the proposed system efficiency and improve the proposed system security, we use the Mongo-DB to set up the local distributed storage.

Firstly, we use Replica Sets in Mongo-DB to ensure synchronization of multiple databases. In a physical machine, we deploy three databases. The first database is the main database, the second is the standby database and the third database is aimed to realize the separation between read and write. So the main database is only to write, the third database is only to read, the standby database is only to prevent sudden. In spite of that we assigned their respective mandates, but we could convenient to switch their role. The Replica Sets architecture shows in Fig.10.

Secondly, we use Sharing in Mongo-DB to set up a Sharing Cluster; every sharing in Sharing Cluster is Replica Sets architecture. It is known to us all that if there is only one database and the database is shut down without any sign, then it will impact on the service supported by the database hundred percent but if there are ten databases, it will only impact on the service supported by the database ten percent. In the proposed system, we use three physical machines to set
up a Shading Cluster. We put the data to the three databases respectively. When a client wants to request data from the three databases, The Mongo Routing process will take the request to the right database which stores the data. The Shading Cluster architecture shows in Fig.11.

![Figure 10. Replica Sets architecture](image1)

![Figure 11. Sharing Cluster architecture.](image2)

**VIII. CONCLUSION**

In this paper, the local distributed storage, sharing cluster architecture, Mongo-DB and spider module are used integrated for build an efficient basic information system of science and technology experts based on Web mining. Through using the local distributed storage, we find that the application system can be achieving very fast execution efficiency. The proposed system architecture can be obtained the application performance requirements.

**ACKNOWLEDGMENT**

This work is supported by the National Sparking Plan Project of China (2011GA690190), the fund of Huai'an Industry Science and Technology. China (HAG2011052, HAG2011045, HAG2010066).

**REFERENCES**


