An Intelligent Retrieval Platform for Distributional Agriculture Science and Technology Data

Xiaorong Yang[1,*], Wensheng Wang[1,*], Qingtian Zeng[2], Nengfu Xie[1,*]

1 Agriculture Information Institute, Chinese Academy of Agriculture sciences, Beijing, P. R. China

*Key Laboratory of Digital Agricultural Early-warning Technology（2006-2010）, Ministry of Agriculture, The People's Republic of China

2 Shandong Science and Technology University, Shandong Province, P. R. China

Abstract. In the agricultural domain, the variety of data used by organizations is increasing rapidly. Also, there is an increasing demand for accessing these data. Now, the problem of the digital divide causes serious problems in manipulating the distributed information. Based on this condition, this paper presents the intelligent retrieval architecture of distributional agriculture science and technology data which focuses on research of the integration support technology, the concept extending retrial technology based on agricultural ontology and the personalization retrieval technique based on the user model. In the experiment, the intelligent data application platform provided by the paper proves that the architecture is effective.

Keywords: Agriculture science and technology data, Data integration, Agriculture ontology, Intelligent retrieval

1. Introduction

With the development of computer and network technology, the amount of data which are collected, saved, processed and transmitted has grown rapidly. Many sharing and serving platforms of agriculture science and technology information are constructed by different departments throughout the country. But these platforms lack a unified plan and management in the important implementation techniques and storage

technology. The heterogeneity and dynamic distribution become basic features of these systems at present. Particularly the heterogeneity in semantics results in data sharing difficulty. An intelligent data application platform should be constructed to make full use of different distributed heterogeneous data resources. The platform can provide a public and unified data access interface of different distributed data sources for users. Users needn't consider the problem of data extracting and data combining. So the unified and high-efficiency access of data can be achieved.

2. The Architecture of the Intelligent Retrieval Platform of Distributional Agriculture Science and Technology Data

2.1 The Logical Architecture of the Intelligent Retrieval Platform of Distributional Agriculture Science and Technology Data

A traditional retrieval system of distributional agriculture science and technology data includes distributional data integration module, data category module and data retrieval module. To improve retrieval intelligence and satisfy users' individualized need, the intelligent retrieval module and personalization service module are designed based on the traditional retrieval architecture of distributional agriculture science and technology data. The intelligent retrieval module uses domain ontology to support the information retrieval based on different languages, synonyms and related information resources. The personalization service module can record and mine users' historical data to discovery users' interest. It can recommend information according to users' interest. Fig 1 shows the logical architecture of the intelligent retrieval platform of distributional agriculture science and technology data.

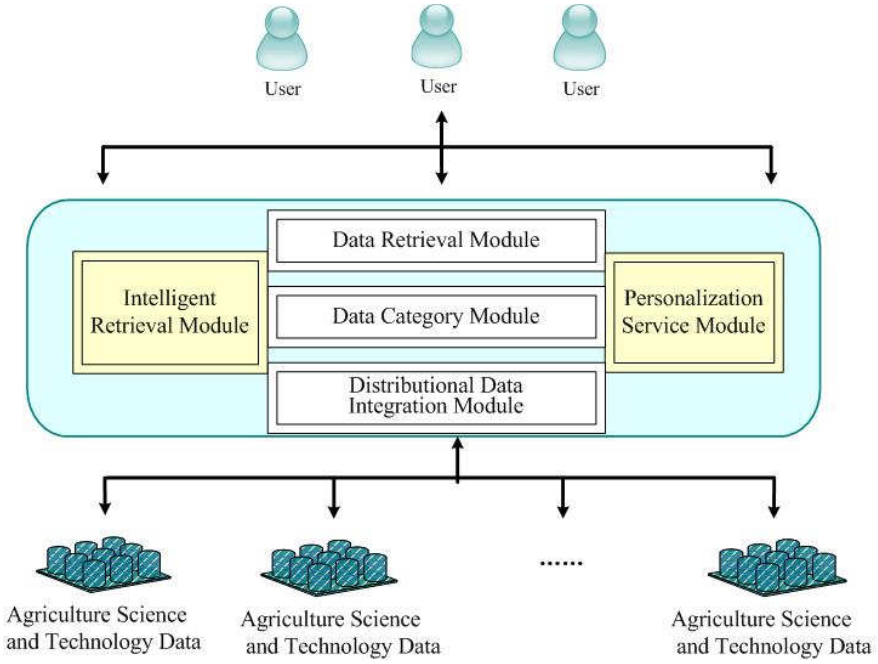

Fig 1: Intelligent Retrieval Logical Architecture of Distributional Agriculture Science and Technology Data

2.2 The Function Architecture of the Intelligent Retrieval Platform of Distributional Agriculture Science and Technology Data

According to the logical architecture the detailed function architecture of the intelligent retrieval platform of distributional agriculture science and technology data (Fig 2) is designed. The function architecture includes the management and retrieval of data sources layer, the central metadata mapping and intelligent retrieval layer and the system interface layer.

The management and retrieval of data sources layer consists of the node metadata management module and the web retrieval module. The node metadata management module manages bottom database sources and the web retrieval module can accept query parameters from upper layer, access database and return retrieval results.

The central metadata mapping and intelligent retrieval layer consists of the intelligent retrieval module, the central metadata mapping database and the central metadata manager. The intelligent retrieval module can accept query parameters from system interface. According to the central metadata mapping table it can find corresponding

data source and submit this query parameters to corresponding web retrieval module. The central metadata manager can manage metadata and mapping relation between metadata and data sources.

The system interface layer provides classification retrieval and keywords retrieval for users. The platform completes semantic extension for query condition which a user inputs and submits them to the bottom web retrieval module.

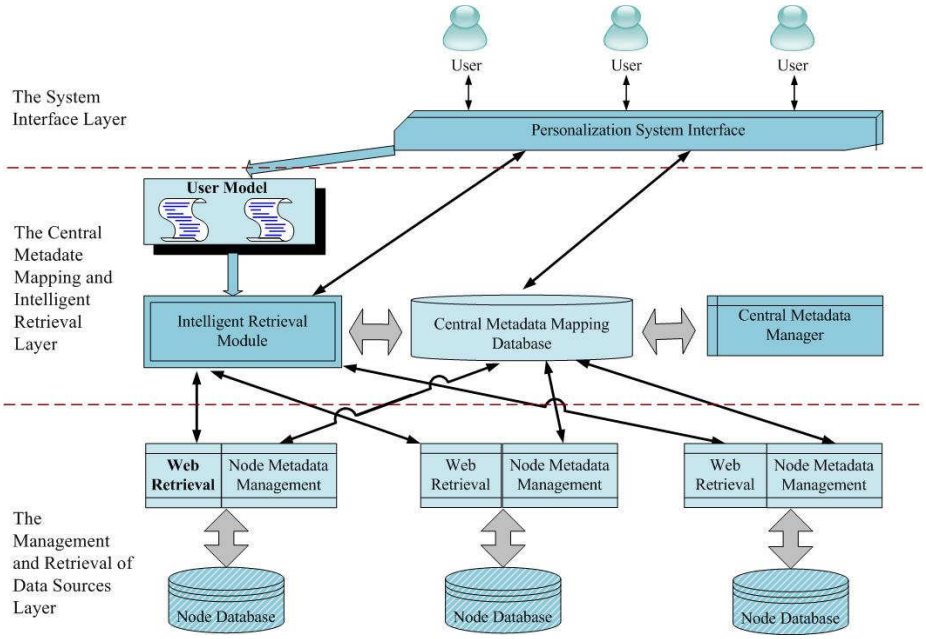

Fig 2: Detailed Function Architecture of the Intelligent Retrieval Platform of Distributional Agriculture Science and Technology Data

3. The Key Technology about Intelligent Retrieval of Distributional Agriculture Science and Technology Data

3.1 The Integration Supporting Technology of Distributional Agriculture Science and Technology Data

Integration and category are main function of the integration of distributional agriculture science and technology data. This study adopts middleware technology to solve the integration of distributed heterogeneous data. A middle layer is developed

between users and distributional agriculture science and technology data sources. It can provide a unified data access interface for distributed heterogeneous data sources. It also defines classification standards for data resources. Then the information is classified and displayed to users (Song Lan et al. 2010). Fig 3 shows the logical architecture of Integration and category of distributional agriculture science and technology data. Because node administrators know more about node database, this study adopts metadata technique to descript resources. Metadata of bottom resources are described by node administrators. The middle layer uses the metadata to manage different node data sources. It administers collectively the metadata of different node database and sets up a unified metadata mapping table. So all heterogeneous database can be operated as a simple database. The unified metadata mapping table can organize and access heterogeneous network information resources (Li Jianhui. 2007) (Song Xiaoyu et al. 2008). User layer establishes query performance according to the information classification to submit it to the data integration layer. And the data integration layer searches the classification mapping table and metadata mapping table and locate the corresponding data source. This study presents own metadata standard according to database structure based on Dublin metadata standard.

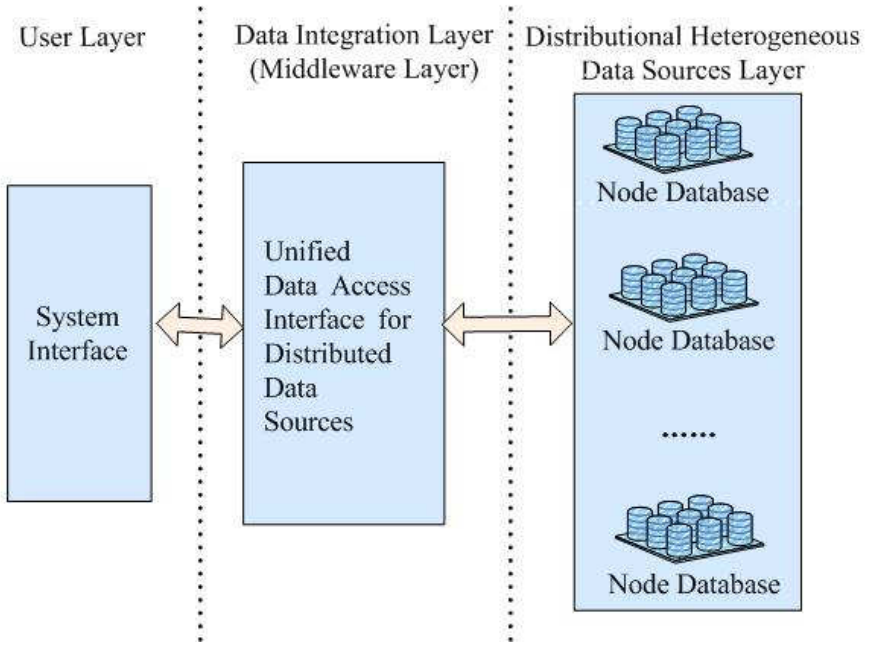

Fig 3: Logical Architecture of Integration and Category of

3.2 The Concept Expansion Retrieval Technology Based on Domain Ontology

Domain ontology technology is adopted to standardize retrieval keywords in order to reach united comprehension to information between human and human or Machine. The concept-based retrieval technology can improve retrieval efficiency and speed. This study adopts description logic to establish domain ontology model and analyses the reasons of the heterogeneity among information systems from the angle of ontology. The semantic integration framework of the heterogeneous systems is established based on ontology mapping. Domain ontology can be set up in two methods (Cao Yukun et al. 2010). Semantic extension keywords set and some metadata set of node data sources inputted by node administrators are main domain ontology keywords. And the domain keywords extracted from the specialized websites are ancillary sources.

This study constructs ontology by using the graphical interfaces of the ontology edit tool Protégé. According to ontology rules, node administrators extend the keywords in the semantic dictionary from four properties of synonym, abbreviation, English language and Chinese pinyin. Thus a unified semantic dictionary is set up in the center database and becomes more and more abundant (Song Lan et al. 2009). It can improve the recall ratio of information retrieval. Then the center administrators delete repetitive and ambiguous words. The precision ratio of the platform can be implemented (Chen Lihua. 2010).

By semantic analysis of ontology concept and retrieval keywords, retrieval association and expansion are completed step by step. This technology supports synonym retrieval, information retrieval of different languages and recommendation of related information resources. For example, if a user inputs the keyword crop, Chinese and English information about crop can be searched and information about rice and wheat can be searched.

3.3 The Personalized Recommending Technology Based on User Model

User interest model is set up according to user's explicit demand and implicit demand. It maintains uses' history behavior information and personal information. It provides different comprehension of same keywords from different users in depth and scope. User feedback process based on users' opinions makes retrieval service more accurate and friendly. User interest model can analyse a user's behavior and record and mine users' hidden interest. Because the user's interest changes, user interest model self-studies continuously to improve itself (Fei Hongxiao et al. 2009). The platform sets higher priority to the information which are often accessed by the user. user interest model can forecast a user's interest and demand to implement personal information retrieval and recommending. Firstly, this study implements the dynamic sort of information resources according to a user's interest. When a user accesses some information resources, the system records his behavior and analyses his interest in classified information resources. When the user retrieves information again, the data resources which are often accessed by him will be displayed ahead. Secondly, the system can customize personal fields of database. The fields can be defined as the language and words needed by a user in order to satisfy his usage pattern. Finally, the system can record information accessed by a user. The user can operate the accessed records and define if they are useful to him. By calculating the probability of the accessed information, the user's interest in information resources of some sort can be gotten.

4. Application Case

To evaluate the intelligent retrieval platform for distributional agriculture science and technology data, the platform is applied in the management of Tibet science and technology information resource. In Tibet, all kinds of information resources are saved in different database and websites. These systems don't communicate each other because of the independence in the design and deployment. By applying the intelligent retrieval technology, the platform integrates, maintains and shares the distributional agriculture science and technology data in Tibet. The platform provides a unified data access interface for distributed heterogeneous data sources. Through the interface users access the needed information conveniently and needn't consider the

problem of data extracting and data combining. Not only the information which meets the inputted keywords can be searched, but also the information about the synonym, English language and related information of the inputted keywords can be found. And the information are displayed according to users' interest priority. The retrieval intelligence and individualized service of the platform satisfy users' demand.

5. Conclusion

To integrate and share distributional heterogeneous agriculture science and technology data, this study designs and implements the intelligent retrieval platform for distributional agriculture science and technology data. This paper introduces the logic and function architecture of the platform and the integration supporting technology of distributional agriculture science and technology data, the concept expansion retrieval technology based on domain ontology and the personalized recommending technology based on user model. Finally, as an application case, the platform has been applied to manage Tibet science and technology information resource to verify the performance of the management platform.

Acknowledgements

The work is supported by the Academy of Science and Technology for Development fund project "intelligent search-based Tibet science & technology information resource sharing technology", the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2009ZX03001-019-01), and the special fund project for Basic Science Research Business Fee, AII (No. 2010-J-07).

References

1. Cao Yukun, Ding Mingwei., "Discovering Model of Semantic Web Service Based on Ontology", Computer Systems & Applications, 2010(19) 04, pp.98~102

2. Chen Lihua, “Comment on Latent Semantic Analysis of Retrieval Precision Rate Factors Based on the Impact of Natural Language”, Journal of Modern Information, 2010(30)03, pp.26~31
3. Song Lan, Lei Lixia, Wang Hong, “A Study of Intelligent Semantic Information Processing System Based on Ontology”., Journal of East China Jiaotong University, 2009(26)05, pp. 31~34
4. Wang Xiaogang, “Semantic-based Query in Heterogeneous Information Integration Environment”, Doctor Degree Dissertation of Huazhong University of Science & Technology, 2006
5. Li Jianhui, “Key Problems Research on Metadata Oriented to Scientific Data Sharing”, Doctor Degree Dissertation of the Chinese Academy of Science, 2007
6. Song Xiaoyu, Wang Yonghui, Data Integration and Integration Application., The Chinese Publishing Press of Water Conservancy and Hydroelectric Power, 2008
7. Fei Hongxiao, Tan Siming, Li Wenxing, Li Qinxiu, Dong Xin, “Web User Clustering Based on Interest”, Computer Systems & Applications, 2010(19)04, pp. 62~65