A SCHEME FOR SHARE AND EXPLOITATION OF NETWORK AGRICULTURAL INFORMATION BASED ON B/S STRUCTURE

Huitao Liu [1,2], Limei Tan [1,2], Yuan Yao [1,2], Qing Wang [1,2], Hongsheng Zhang [1,2], Guanglu Zhang [1], Jintong Liu [1,*]

[1] *Center for Agricultural Resources Research, Institute of Genetic and Developmental Biology, CAS, Shijiazhuang, China;* [2]*Graduate School, CAS, Beijing, China*

[*] *Corresponding Author: Center for Agricultural Resouces Research, CAS, 286 Huaizhong Road, Shijiazhuang, 050021, China. Tel: 86-0311-85871749, Email: jtliu@sjziam.ac.cn*

Abstract: The features of network agricultural information (NAI) are summarized by analyzing Chinese typical, famous agricultural information websites. The features include the storage-scattered of agricultural informations, high frequency of updating, non-uniform data format and a lots of raw information existing. It is pointed out that the features have disadvantages for sharing and exploiting the NAI efficiently. Then a scheme is proposed to try to overcome the disadvantages of the features, which is mainly based on brower/server structure and information extraction technology from webpages. Also the key technologies to implement the scheme are described in the paper. In the end, an example application of the scheme is carried out to demonstrate the concrete steps to develop practical applications based on the scheme.

Key words: network agricultural information, information extraction technology, brower/server structure

1. INTRODUCTION

Practices and studies have proved that agricultural information share and development can be an effective way to reduce the pressure of natural resources and promote rational allocation of natural resources (Chen Pinde, 1993). In the face of the massive agricultural information (Gudivada VN,

1997), how to rational and efficient use and sharing of them are becoming increasingly important. Bayesian methods and boosting algorithm has been proposed in respect of data mining (Tang Chunsheng, 2003). And artificial intelligence, mathematical statistics and the sample-specific training methods (O. Etzioni, 1995; Y.Yang, 1999; Pui Y Lee, 2002) are used in information retrieval. Wai Lam put forward an automatic text categorization according to the information content and it is used in retrieval of text (Wai Lam, 1999). Zheng LiuYue (2003) proposed the W3C DOM (Document object model), metadata and XML-based network information extraction model (Liu Zhengyi, 2003). However, the implementation of those techniques is a complex project and not suitable for general users (T.Radecki, 1997) and these techniques and methods are obviously poor or limit to TXT files or XML. The common defect of the techniuqes and methods is that they cannot accurate positioning information, have no function to exploit the information retrieved to find out new knowledges.

In this paper, an endeavor be done to try to find a scheme, whick not only can efficiently retrieve and share network agricutrual information, but can deeply exploit the retrieved information.

2. FEATURES ANALYSIS OF NAI

2.1 Features of network agricultural information

To gain the features of NAI, this paper select 50 representative, well-knowed agricultural information websites as study sample. By analyzing its producing, storing and dissemination, we get the features as following:

Agricultural Information has a scattered storage. The same type of agricultural information is often distributed in a number of websites.

Agricultural information data have a high frequency of update and the amount of data increases rapidly.

The data formats of agricultural information are not uniform, which vary from website to webite. For prices, a website may use "U.S. dollars / ton", while another may use "Yuan / kg".

Most of NAI is raw information, which is simply stacked into a website, and through which we can’t find the internal rules. The deep development is urgent to be implemented. For example, by browsing the items of supply and demand information on the website, it is difficult to get the situation of supply and demand, or to forecast the prices of agricultural products or the seasonal fluctuation of price.

2.2 Disadvantages of the features of NAI

Since the agricultural information storage is dispersed, for a complete grasp of certain type of agricultural information, we have to visit as much websites as possible, so the workload is large and the efficiency is low. Meanwhile, it is a mechanical, monotonous work, for the agricultural information update frequently; and in order to grasp the latest information, we have to do the same job every day (or frequently): visiting web site as much as possible to obtain the latest information. However, it's prone to errors when comparing the non-uniform format data, for example, to convert the "dollar / ton" and "Yuan/kg" to same unit. As NAI is mostly raw information, the use of information is at low-level and cannot meet the high-level demand. For instance, policy maker want to know the price changes with seasons.

In summary, the features of NAI have severely restricted the high efficient share and deep exploitation of NAI.

3. A SCHEME FOR HIGH EFFICIENT SHARE AND DEEP EXPLOITATION OF NAI

To overcome the disadvantages of the featurs, a scheme is proposed here for realization of high efficient share and deep exploitation (HESDE) of NAI, which is mainly based on browser/server structure. The scheme is described as *Figure 1*.

The contents in dotted-line box are the main body of the scheme. According to its functions, the main body can be divided into three main modules: information extraction module, deep exploitation module and management-query interface module. In fact, it can also be understood as three functional module procedures running on a server.

3.1 Information extraction module

This module is mainly to solve the former three disadvantages of the features of NAI. Further, the model can be divided into three propcedures: data extraction procedure, format standardizing procedure and data store procedure, which respectively completes the automatic information extraction and collection, data format standardization and centralized data repository.

In the light of user's requirement, data extraction procedure extracts specific category information from website1, website2, and so on. At the same time, format standardizing procedure uniform the data formats to a

standard one. Then data store procedure saves the uniform datum to database and completes centralized data repository. This module provides basic datum for the next phase of deep exploitation module.

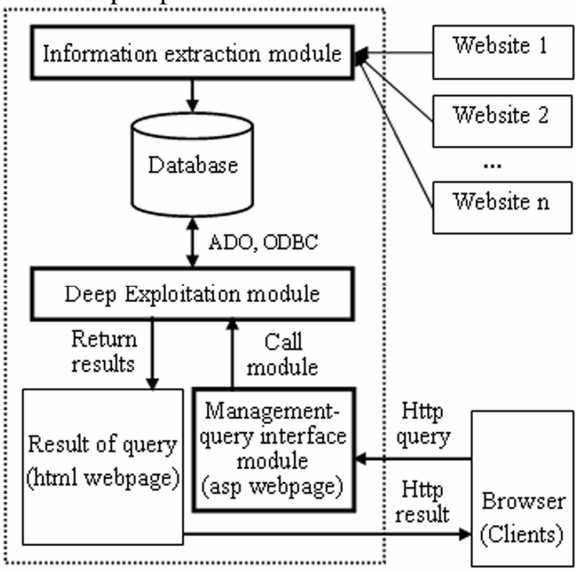

Figure 1. Sketch map for the sheme of HESDE of NAI

As for data extraction procedure, we can use Webbrowser control provided in VB6 (and above) for the browsing the websites, use its Navigate2 method for information location. Then, according to the specific properties of data set, such as the index (related with DOM) of the table and the DOM (Document Object Model) model, we can locate information resources in webpages and get access to the interest data by acquiring object attributes such as InnerText. After uniforming the datum extracted to same format (for example, all to Yuan/kg), data store procedure save them to database by the use of ADO/ODBC database manipulation technology.

3.2 Deep exploitation module

The module mainly implements the deep exploitation and development of raw information of NAI to find out new knowledge, for example, inherent laws or implicit rules. The deep exploitatin module is an open subsystem to the scheme. In other words, users can add, modify and delete deep exploitation programs in the system to meet the needs; also, the deep exploitation programs can be added or deleted remotely. Once a deep exploitation program is uploaded by one user, it can be used by others in the users group. This realizes the sharing of the deep exploitation method and avoids repeat development of procedures with the same function. This

module can be viewed as a package of user transaction handling process and can be managed and called from management-query interface module and return results in html webpages.

In fact, deep exploitation program can be developed by Java, PHP, Perl, Javascript, and VC and so on. Also it can be released in many forms, such as scripting language program, .COM, .EXE and .DLL form.

3.3 Management-query interface module

This module mainly provides the interface of background management and foreground call. Through this module users can manage information source websites, set the freqency of data extraction procedure, upload and manage the deep exploitation program, and call certain uploaded deep exploitation program to execute data analysis. Still through this module, the analyzing results will be returned in text, graphics and photographs forms.

Actually, the above three modules are on the server end of brower/server structure. There are many operation systems and information techniques can be choosed to build the interface of the module, the B/S system and backgroud database. For example, we can choose Unix or Windows as the operating system platform, choose Apache or IIS to build the B/S structure, use Oracle or SQL to the develop a database and use different programming languages (VC, Java, VBscript, Jscript, etc.) to develop and produce the deep development programs, system interface and background management procedures.

According the analysis of functions of the above three main modules, we can make a conclusion that the scheme can overcome the disadvantages and is a solution of high efficient share and deep exploitation of NAI. Up to this point, we know the scheme is a framework of utility of NAI, which has the feasibility of implementation by using existing technologies.

4. AN EXAMPLE APPLICATION OF THE SCHEME

In order to further explain the scheme and demonstrate how to build a specific application, an example application of the scheme is given here.

The application is built mainly by Windows 2000 Server, IIS5.0, MS SQL and VB. And it is designed to find out the fluctuations of the price of vegetables and the relationship between supply and demand by analysis of the wholesale prices of vegetables in six markets in Beijing.

In the case, three websites are selected as information source website (see *Figure 2*), all of which have vegetable price information, but in different format. A simple database with the fields of "data source", "vegetable

varieties", "market", "price" and "Date" is build by MS SQL. Data extraction procedure, format standardizing procedure and data store procedure are develped and released to DLL files respectively by using VB and its built-in WebBrowser control, by using the DOM to collect information data and by using ADO/ ODBC to visit database. Format standardizing procedure uses "yuan/kg" as uniform unit; "fanqie" and "tomato" are uniformed as "tomato"; publishing date are uniformed as "yyyy-mm-dd". Data store procedure save the uniform format datum into the simple database.

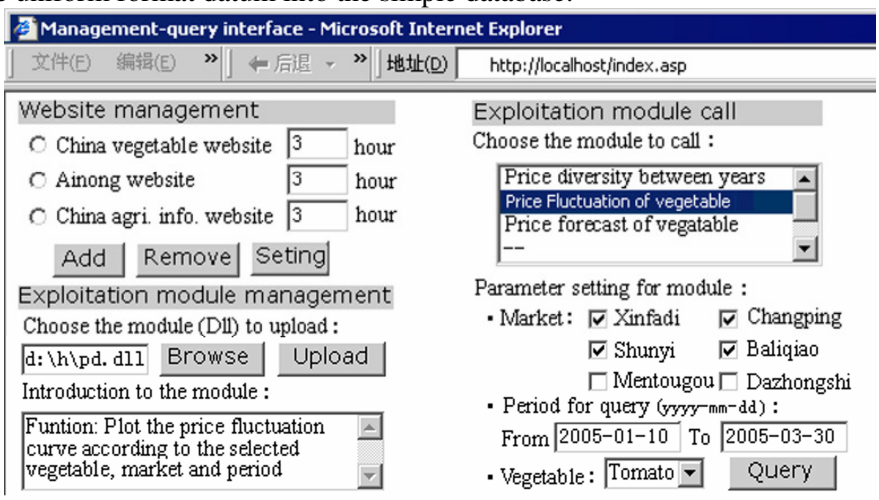

Figure 2. Interface ofr user management and aquery of example application

In this example, we developed three programs (see *Figure 2*). The program named "Price flactuation of vegatable" can output the given type of vegetable price volatility curve and give the mean price of different market in given period. The program takes a ten-day as a unit to calculate the mean price, and uses DbtoChart components to generate data chart, which be returned to the user in the form of web pages using ASP and VBscript script.

Management-query interface module use Html, VBscript and ActiveX to develop. It is necessary to explain that each deep exploitation program is added to deep exploitation module, there will be a reaction that its name will be listed in textbox under “Exploitation module call”(see *Figure 2*); Once the user to choose a program in the module list, the bottom of the list will show corresponding parameter options. In this example, this background management procedure is also produced by using VB 6.0, published by DLL files, which are convenient for it’s the uploading, remote registration and deletion. The function of uploading file is realized partly by using SA-FileUP components of Artisans. A SA-FileUP component is an ActiveX DLL server component and easily integrated into the ASP website; In ASP pages, we use VBscript to call CreateObject to generate WScript.Shell, and

then call regsvr32.exe by WScript.Shell to finish the procedure of uploading and registration module.

Through the interface, the interval time of extracting data from three website is set to 3 hours. Users can upload deep exploitation program in the section of Exploitation module management. Here, three programs have been uploaded, which are listed in the section of Exploitation module call. In this paper, "Price flactuation of vegetable" is called (corresponding parameters had been set and was shown in *Figure 2*), and the result return to the uers in the form of webpages (see *Figure 3*).

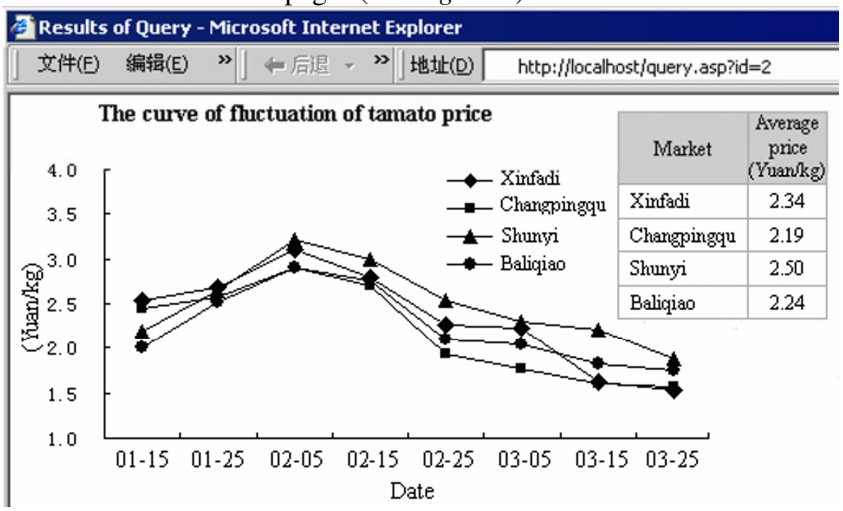

Figure 3. The result of the program of "Price flactuation of vegatable"

Figure 3 indicates that Beijing's vegetable prices in late January and in early February (that is during the Spring Festival) rose rapidly and reached its peak at Feb. 5th, then began to go down steadily.We must notice the information conveyed from *Figure 3* cannot be easily to get by simply browsing original imformation in the three websites.

5. CONCLUSION & DISCUSSION

The features of NAI are analyzed and summed up as following: NAI has a scattered storage, have a high frequency of update and the amount of data increases rapidly; the data formats of NAI are not uniform, and most of NAI is raw information.

The disadvantages of the features of NAI are pointed out; and they hinder the realization of high efficient share and deep exploitation of NAI.

In order to overcome the disadvantages of the features of NAI, a scheme is put forward. By analyzing the functions of the scheme, it was theoretically

proved that the scheme can greatly promote the high-efficient share and deep exploitation of NAI. In addition, this program is applicable to XML format and a large number of web HTM format.

An example application of the scheme is built up to demonstrate the concrete processes of a practical application of the scheme. Users can build more complex and practical applications by using it as a template.

However, for the limited space, we do not discuss the system building and technical details. In addition, although the scheme is proposed on the basis of NAI, it is equally applicable to other types of network information, which indicates the good expandability and practicality of the scheme.

REFERENCES

Chen Pinde. Development of web-based information system. Computer Engineering, 1993, 24(3): 7~11

Gudivada VN. Information retrieval on the World Wide Web. IEEE Internet Computing, 1997, 12(5):58~68

Liu Zhengyi. DOM-based and metadata-based information extraction for web sources. Computer and Modernization, 2003, (10):81~83.

Niu Zhenguo, Fu Haifang, Cui Weihong. Multilevel-users-oriented agricultural information classification. Resources Science, 2003, 25(2):20~25

O. Etzioni, D.S.Weld. Intelligent agents on the Internet: Fact, Fiction, and Forecast. IEEE Expert, 1995, 10(4):44~49

Pui Y Lee, Siu C. Hui, Alvis Cheuk M. Fong. Neural networks for web content filtering, IEEE Expert, 2002, 17(5):48~57

Tang Chunsheng, Jin Yihui. A Multiple Classifiers Integration Method Based on Full Information Matrix. Journal of Software, 2003, 14(6):1103~1109

Wai Lam, Miguel Ruiz, Padmini Srinivasan. Automatic text categorization and its application to text retrieval. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(6):865~879

Zhao Chunjiang, Wu Huarui, Yang Baozhu. Development platform for agricultural intelligent system based on techno-componentware model. Transactions of the Chinese Society of Agricultural Engineering, 2004, 20(2):140~143