Design and realization of Agricultural Information Intelligent Processing and Application Platform

Dan Wang

Institute of Agricultural Information，Chinese Academy of Agricultural Sciences，Beijing 100081，China
Key Laboratory of Digital Agricultural Early-warning Technology, Ministry of Agriculture, P. R. China
wd@mail.caas.net.cn

Abstract. The work of this paper aims to provide a practical software product, the intelligently gathering, processing and utilizing for the leaders of agricultural departments, research institutions, library and information systems and information consultancy departments, etc. The software product is an information processing platform integrated intelligent gathering, automatic indexing, and search and utilization of agriculture information. This software is tested to be accurate and can substitute people to process information to some extent.

Keywords: agricultural information, automatic collection, automatic indexing, information search, information management

1 Introduction

After 40 years in the development of the Internet and after 20 years in the development of the World Wide Web, we have entered into the era of digital information. There is a mass of real-time and valuable information and electronic documents scattered in various departments on the Internet. Currently the effective use of the digital information is affected by the three restricting factors. Firstly, how to crawl the desired information from the mass information on Internet in real time and collect electronic documents scattered in various departments to the local database. Secondly, how to classify and deal with the digital information from the local database, make it the standard information resources so that the information retrieval and utilization. Thirdly, how to retrieve the information they need from the local database easily, accurately and quickly，to achieve the purpose of efficient information exchanging, using and spreading.

To solve the above technical problems, we have developed the platform of the Agricultural Information Intelligence Collection, Processing and Utilization, also called information processing platform（IPP）. The IPP is a general platform which provides information intelligence collection, information ordering classification，

subject indexing and information retrieval for the leading departments, agricultural information and advisory organizations, the construction unit of agricultural information resources.

Agricultural information institute of the Chinese Academy of Agricultural Sciences (hereinafter referred to as AII）have made a lot of basic research work in the information organization and management of agricultural domain from 1990s, which have prepared "Agricultural Science Thesauri", "Chinese Library Classification - Agricultural Professional Classification" and "Multifunctional Agricultural Vocabulary", and developed "the Chinese information automatic indexing system". The AII has officially studied and developed IPP in order to enhance the original scientific and technological achievements with the development of the information age.

Fig. 1. Agricultural Information Intelligent Processing and Application Platform

2   Research Objectives and Content

R & D goal of the IPP is to develop agricultural information processing platform based on B/S structure for agricultural information intelligence collection, automatic classification, subject indexing and the use of agricultural information retrieval. The functions of the IPP can be classified the following parts: 1) the subsystem of integrated classification vocabulary of agricultural information management; 2) the subsystem of automatically grabbing information on the Internet and smart obtaining electronic document; 3) the subsystem of the automatic classification and subject indexing of agriculture information; 4) the subsystem of database definition and data storage; 5) the subsystem of retrieval information; 6) the subsystem of data import and export.

3 Architecture and Functional Design of the IPP

3.1 System Architecture

Information processing of the IPP is firstly a collection of agricultural information that is stored in the temporary database. Secondly, the information is automatic processed and stored in the formal database. Thirdly, information is retrieved and used. Based on the flow of information, the agricultural information Integrated classification vocabulary management subsystem, the database management subsystem, and system operating parameters participate in together to achieve the goal of the IPP.

3.2 System Functions

The main functions of the IPP can be concluded in the six following aspects:

3.2.1 Intelligent real-time acquisition of the agriculture network information

The IPP may intelligently collect personalized information real-time from the agricultural website. The IPP can collect requirement information on the Internet and store in the local database by way of user-set information acquisition parameters of generic and professional templates.

Fig. 2. The IPP may intelligently collect personalized information real-time

3.2.2 Information import function

Data import function of the IPP is to provide the collection function of other electronic documents and the collection function of network information. The main features are:

Fig. 3. A screen shot of information import function

3.2.2.1 XML format data

XML format data is read and analyzed by IPP after the data is imported into the related temporary database of the IPP.

3.2.2.2 TRS format data

TRS format data is the data format of import and export supported by full-text retrieval system which is developed by well-known domestic TRS company. Therefore, the IPP provides TRS format data import feature.

3.2.2.3 ISO2709 format data

ISO2709 format is the international standard of data format which is widely used in library and information sector for bibliographic information exchange. Firstly the IPP can read the data of the ISO format. Secondly the data can be converted to XML data. Thirdly the data is stored in the temporary database.

3.2.2.4 Electronic text format data

The IPP can invoke directly WORD processing software that read DOC, PDF and other document, and edit it. Then the data is imported into a temporary database.

3.2.2.5 Information based on Form
The IPP can automatically generate information form of the database, the user directly login and input the data which is stored in a temporary database.

The data collected from the website and import data from other electronic documents are stored in a temporary database. After information is automatic classified and subject indexed by the IPP, they are stored in a formal database of the IPP.

3.2.3 Information processing function

The IPP provides the function of information processing. The first step is that the data of the IPP is stored in the temporary database. Then, information is batch indexed on temporary database by the IPP selectively automatically, or which is entire database indexed or unattended indexed or artificial auxiliary indexed. At last, index result of the data is edited by the IPP. Specific indexing features:

①The IPP automatically produces classification codes (CLC, category code), the category names, geographical names, geographical codes and information keywords to pretreatment information.

②The IPP supports a variety of indexing methods, including the indexing of single articles, batch indexing, automatic indexing of the entire database as well as unattended indexing.

③The IPP supports the functions of modify and editing for automatic indexing results.

④Artificial auxiliary indexing functions are supported. When users index a message online, they may browse the information classified directory tree at any time and click required indexing category name. The category names and corresponding classification coding are automatic inputted the database by the IPP.

⑤Support the function of the word frequency statistics analysis and the function of updating vocabulary of data. When users may modify correction of the results of indexing keywords, if vocabulary words are not yet logged, the IPP will automatically added，and then the IPP will compute vocabulary frequency. When vocabulary frequency of the word exceeds the prescribed threshold, the word is updated automatically on the vocabulary of data.

⑥Flexible configuration in word field range for indexing, such as the title field, summary field, the body field.

⑦Flexible configuration for the position weight value of the keyword，such as keyword from the title field, from summary field, from the body field, from the first or the tail paragraph of the text, and from the first few words in the paragraph.

⑧The flexible configuration of indexing depth, that is, a literature can be marked with the number of keywords. According to the results of statistical analysis of artificial indexing, indexing depth is generally appropriate in 7-8 words.
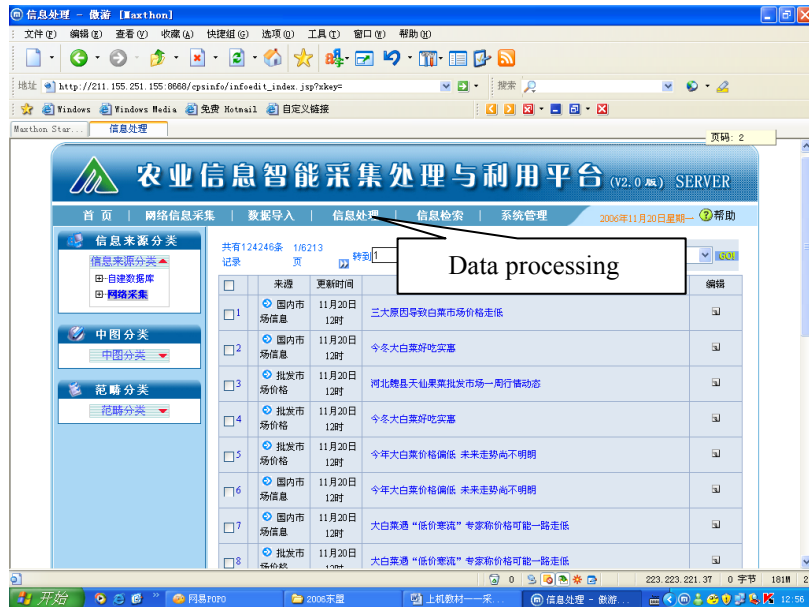
Fig. 4. A screen shot of information processing function

3.2.4 Information retrieval functions

The IPP can retrieval information which is stored in the formal database, the specific features can be summed up:

① The user can choose a database or multiple databases at the same time for retrieval. the IPP supports for multi-database searching.

② Support keyword search, the second search and text search in the limited field.

③ Support boolean logic expression to search, according to user's searching requests, composition of search expression with "and", "or", "non", as well as the search expression with operator priority retrieval.

④ Support the search function of the SDI. The IPP save the user's search expression, in order to retrieve the information repeatedly to form a thematic database.
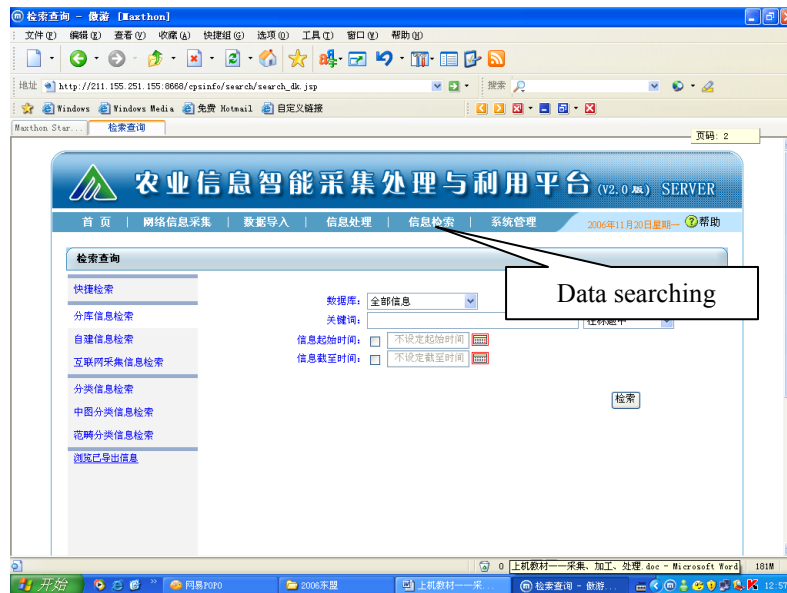
Fig. 5. A screen shot of information retrieval function

3.2.5 System management function

System management function contains the configuration and management operating parameters of Intelligence collection, processing and utilization from the agriculture network information, including management of users, databases of custom parameter configuration management, information collection indexing parameters of configuration management, information taxonomies, thesauri, regional tables and information areas of the table maintenance, and run log management.

①User management: It Supports the user's function of increasing, deleting and modifying; It sets up and manages the user name, password, roles and permissions.

②Custom database: Users can flexibly define databases structure in different type databases, describe the properties of the database to determine the taxonomic relationship between the different databases and create a user database.

③Vocabulary management: It contains maintenance and management of thesaurus of information, the scope of vocabulary, regional tables. Users can add, delete, change keywords according to the above vocabulary.

④The word frequency statistics and vocabulary Automatic Updates: The IPP computes word frequency for indexing word. If indexing words does not appear in any vocabulary database or word frequency in excess of the threshold, the system will automatically input into vocabulary database.

3.2.6 System management functions

The IPP provides the exchange with the outside world database interface that can output the information in XML format document, TRS format documents, ISO2709 format document output.
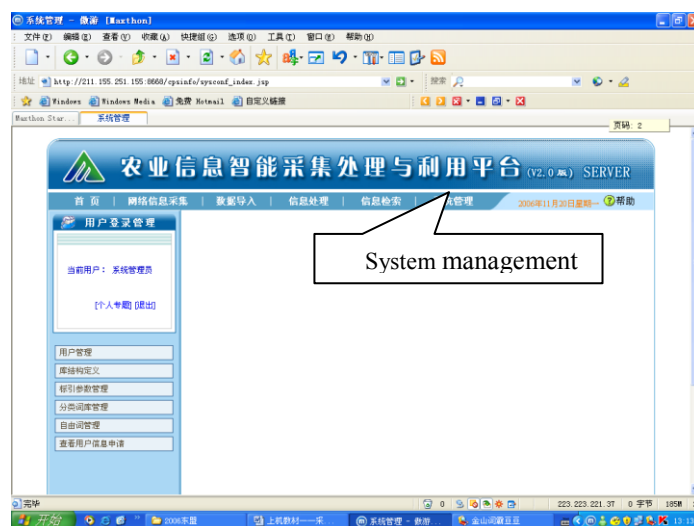

Fig. 6. A screen shot of system management functions

4 Operating environment

The platform is a software products of cross-platform operating system with JAVA language development based on B/S architecture, and operating environment is:

- hardware configuration: CPU Pentium 4 or above，CPU Clock Speed more than 1G
- Computer memory: more than 2G
- Hard disk: more than 20G
- Video card: resolution ratio support of 1028 * 768 pixels, True color refresh frequency 85HZ,   Video card memory more than 32M
- Indicator: Standard display, more than 15-inch
- Operating system: Windows2000, Windows XP, Redhat, or Linux9.0
- Database: ORACLE, MYSQL, or SQLSERVER
- Development software: JAVA 1.6
- Application : IE6.0 above

5 Conclusion

The user uses login name and password to enter the information processing platform, and click on the system menu to finish related operations. It automatically crawled and annotated 10000 pieces of documents from the websites and the documents provided by Chinese Academy of Agricultural Information Institute, China Agricultural University and other units, agricultural engineering, plant protection, soil nutrients, food economy, vegetable gardening, animal husbandry and veterinary literature information. We compare the annotated result with the results of manual annotation, which proved that the method is very effective.

References

1. FENG Xi-xi; ZHU Xue-fang: The Design and Implementation of Information Resources Extractor of Agricultural Websites Based on Spring. [J.] Informatization Research, vol.37, pp19-22, 2011(06)
2. Dong Shiwei, Yu Jun, Li Hong, Sun Danfeng: Development and Construction of Beijing Suburban Natural Resources and Socio-economic Information Management Platform. [J]. Chinese Agricultural Science Bulletin,pp 167-171, 2011(20)
3. ZHANG Wen-hui et al: Design of Agricultural Information Retrieval Platform Based on Chinese Word Segmentation. [J.] Journal of Anhui Agricultural Sciences, pp 12586-12587, 2011.39(20)
4. Zhang Youyi: On the Exploitation of Online Teaching Resources Abroad. [J.] New Century Library. pp78-79, 2012(05)
5. ZHENG Zhi-hong, XU Zhuo-jun, WEN Hong, ZHENG Zhi-gao: Improved Search Engine and Its Data Structure Design，[J.] Information Science, vol.30 No.2 pp200-205,2012(2)
6. KONG Wei-ting, YAN Hong-yin:The Design of Lucene-based Automatic Answering System [J.]Computer Development & Applications, vol.25 No.4 pp32-37, 2012(4)