Available online at www.sciencedirect.com
ScienceDirect
Journal of Integrative Agriculture
2012, 11(5): 823-830
May 2012
RESEARCH ARTICLE

Design and Implementation of the ZH/EN Bilingual Retrieval System Based on the CAT/AGROVOC Mapping

SUN Wei[1], ZHANG Xue-fu[1], Ahsan Morshed[2], Gudrun Johannsen[2], Johannes Keizer[2], Yves Jaques[2], Stefano Anibald[2], LI Nan[1] and LIU Jia-yi[1]

[1] *Institute of Agriculture Information, Chinese Academy of Agricultural Sciences, Beijing 100081, P.R.China*
[2] *Office of Knowledge Exchange, Research and Extension, Food and Agriculture Organization of the United Nations, Rome 00153, Italy*

Abstract

For the users' convenience of accessing the AGRIS resources quickly and using them fully, the paper decomposes the structure of AGRIS Search net, analyzes the users' requirement met for conducting a bilingual (ZH/EN) retrieval, the system function extensions based on AGRIS English retrieval system and the key issues which the core function module should resolve. Derived by the application requirement, the paper also puts forward to a bilingual retrieval model on the basis of CAT/AGROVOC mapping, designs and realizes the ZH/EN bilingual retrieval prototype system.

Key words: bilingual retrieval, AGRIS retrieval, AGROVOC, CAT, mapping

INTRODUCTION

The AGRIS (International Information System for Agricultural Science and Technology) (FAO 2010b; FAO, OEKC 2010c) database of FAO is one of the three biggest agricultural databases in the area of agriculture literature in the world, covering research information of many areas including forestry, domestic management, hydrophilic science and fishery and so on. There are more than 100 thousands new records added to the database every year, which are expressed in many languages including English, French and Spanish. The database provides services for the users all over the world, but the AGRIS Search System of FAO only provides English retrieval. For the reason, it's difficult for Chinese users to use AGRIS (FAO 2010c). Therefore it's necessary to build a bridge between the Chinese users and AGRIS database through modern information technologies to help Chinese users to conquer the language barrier.

AGRIS SEARCH WEB PAGE STRUCTURE ANALYSIS

To enable the AGRIS search to conduct an ZH/EN bilingual retrieval, the first task is to analyze the function structure of AGRIS search and the information processing flow according to the analysis of its web page structure, then to focus on the function which should be extended based on the original system, and to find out the key technologies which should be further resolved, and to realize the ZH/EN bilingual retrieval

Received 28 June, 2010 Accepted 1 September, 2010
Correspondence SUN Wei, E-mail: sunwei@caas.net.cn; ZHANG Xue-fu, E-mail: zhangxf@caas.net.cn

with higher efficiency at the condition that the existing system is stable.

Fig. 1 shows the web structure of AGRIS search. The whole web service including AGRIS retrieval, AGROVOC retrieval and results display, is based on the AGRIS database and AGROVOC database. Thereinto, AGRIS search is the core function of the website; AGROVOC search is mainly used to assist users to conduct AGRIS retrieval by providing users with the preferred terms which are candidates for building a retrieval expression. These two parts both can interact with users. Each function's workflow is depicted in detail as flows:

AGRIS retrieval

AGRIS retrieval is mainly based on the existing AGRIS English database and meets users' different level demands by different ways. The retrieval ways mainly includes the ordinary retrieval and the advanced retrieval. Ordinary retrieval is used for KEYWORD search. By the ordinary retrieval, users can conduct a retrieval by inputting keywords directly; and also can use "and", "or" and "not" to connect the keywords to enlarge the subject retrieval range, and advanced retrieval can be performed based on search engine "Baidu" and "Google". Meanwhile, users can adjust the retrieval expression according to the keywords provided by AGROVOC in this page, then conduct a more accurate extended subject retrieval. In advanced retrieval, by choosing the range of literatures' country, time and other properties in the search assistant, users can conduct a multi-field advanced grouped retrieval, certainly can do a more accurate extended subject retrieval with the reminding of AGROVOC's search results. Although AGRIS search system has a relatively strong retrieval function, it's confined to English literature retrievals. There is no Chinese literature retrieval service in the system.

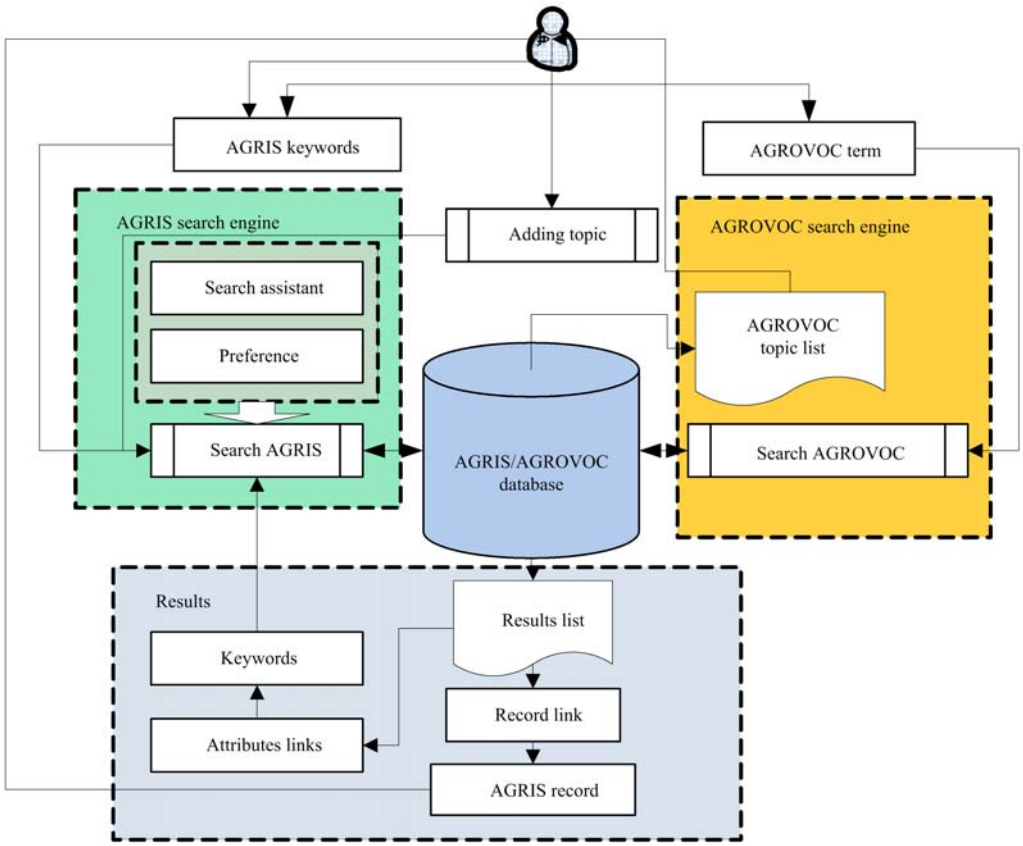

Fig. 1 AGRIS search web structure.

AGROVOC retrieval

AGROVOC is an international multilingual thesaurus belonging to FAO (FAO 2010a, 2011). It covers agriculture, forestry, fishery, food and other related areas, conforms to the standard criterion of multilingual thesaurus and is used to index and retrieve the related literature information. FAO maintains it and updates it with a general cycle of 3 months. AGROVOC is used all over the world and expressed in 5 languages: English, French, Spanish, Chinese, and Arabic, at the same time, Czech, Portuguese, and Thai, other languages such as German, Italian, Korean, Japanese, Hungarian, Slovak, Polish, Hindi, are also being translated. English is the AGROVOC's original language. AGROVOC Search has a powerful thesaurus retrieval function and thesaurus relationship extension mechanism, supporting thesaurus retrieval with means of BT (broader term), NT (narrower term), RT (related term), UF (non-descriptor). AGROVOC also provides an alphabetic browsing. By browsing the thesaurus showed in the AGROVOC view box, users can construct a more exact retrieval expression assisting to extended retrieval. As AGROVOC doesn't support Chinese thesaurus retrieval, it can't provide the retrieval assistant services for AGRIS Chinese retrieval.

The retrieval result display

The result display function is mainly used for displaying users' retrieval results normatively. The content to be displayed includes literature's title, keywords, abstract, etc., which users can download and save. But the displayed results only contain English results.

According to these introduction and analysis of the AGRIS retrieval web structure, it is known that if one wants to extend the English AGRIS retrieval system to bilingual one, it needs to add the function that provides Chinese thesaurus candidates for retrieving, and build a Chinese assistant communication mechanism between AGRIS search and retrieval of candidate thesaurus to assist Chinese extended retrieval in AGRIS, and add the function that analyses the users' input terms to distinguish Chinese term and English term, and build a ZH-EN mapping, so as to realize AGRIS ZH/EN bilingual retrieval.

THE ZH/EN BILINGUAL RETRIEVAL MODEL BASED ON CAT/AGROVOC MAPPING

Through the analysis mentioned above, it is known that the key to realize AGRIS ZH/EN bilingual retrieval is to resolve the mapping problem between Chinese thesauri and English ones.

CAT thesaurus

"Chinese Agricultural Thesaurus" (CAT) (CAAS 2007a) is a CAAS dean fund project organized by the Ministry of Agriculture of China, which involved in nearly 40 departments and more than 100 professional. CAT conforms to the international and national standards and completed by adopting the facet classification method and the computer editing table and other technologies. CAT covers more than 40 subjects such as agriculture, forestry, biology, and contains 63 thousands ZH/EN thesauruses. It's the largest agriculture thesaurus. Obviously, retrieving the CAT, the results can provide candidate thesauri to assist users to rebuild retrieval expressions.

CAT/AGROVOC mapping

CAT/AGROVOC (Liang *et al*. 2005; CAAS 2007b) is the production of the mapping between CAT and AGROVOC, and conforms to the SKOS (W3C 2009) mapping standard which was published by W3C in 2009 as a recommendation. The details are given as follows:

(1) The three main kinds of match relationship: exactMatch, means that the two concepts are exactly same; broadMatch or narrowMatch, means that the source concept is more abstract or specific than the target concept; majorMatch or minorMatch, they both mean two concepts are near-synonyms, while the former means that the two concepts are mostly same, the latter means that the two concepts are partly same.

(2) Three main kinds group methods of match are: ① AND, means all the concepts presenting in the AND expression will work in this retrieval; ② OR, means at least one of the concepts presenting in the OR expression will work in this retrieval; ③ NOT, means the concepts presenting in the NOT expression will be elimi-

nated in this retrieval.

(3) The term which has no related thesaurus will be specially labeled.

So the CAT/AGROVOC mapping results produced by these relative complete mapping rules above can assist AGRIS Search in analyzing Chinese retrieval terms, which is so called ZH/EN mapping.

Retrieval model

In view of the analysis of the AGRIS search web structure, based on CAT, AGROVOC, and CAT/AGROVOC mapping, we construct a ZH/EN bilingual retrieval model (see Fig. 2) based on CAT/AGROVOC. The retrieval model mainly includes four core modules: term analysis module, AGRIS search module, CAT Search module, and AGROVOC Search module. It can distinguish Chinese and English terms and the unidentified Chinese terms, and realize the auto-rebuilding of query expressions to retrieve data respectively from Chinese or English index library, providing users with the ZH/EN bilingual retrieval service. The details are given below:

Term analysis module This term analysis module mainly resolves the problems of users' input terms such as identifying Chinese and English terms, processing the unpreferred terms, and auto-updating the preferred terms, etc. It's depicted as Fig. 3. The details are given below:

(1) On the basis of present analysis technologies of mixed English and Chinese (Yakushiji *et al.* 2003; Multilingual Statistical Parsing Engine 2009; Apache Lucene – Query Parser Syntax 2010; ICTCLAS 2010), take ICTCLAS as the basis parser, and add Chinese and English recognition algorithm to it. Taking the CAT and AGROVOC Thesaurus as one of user dictionaries, analyze user input terms, and identify the input terms following the ZH/EN identification rules in the recognition algorithm, if they are English, produce English preferred term, go to step (4); if not, go to step (2).

(2) Identify Chinese preferred terms according to CAT/AGROVOC. If there are no corresponding Chi-

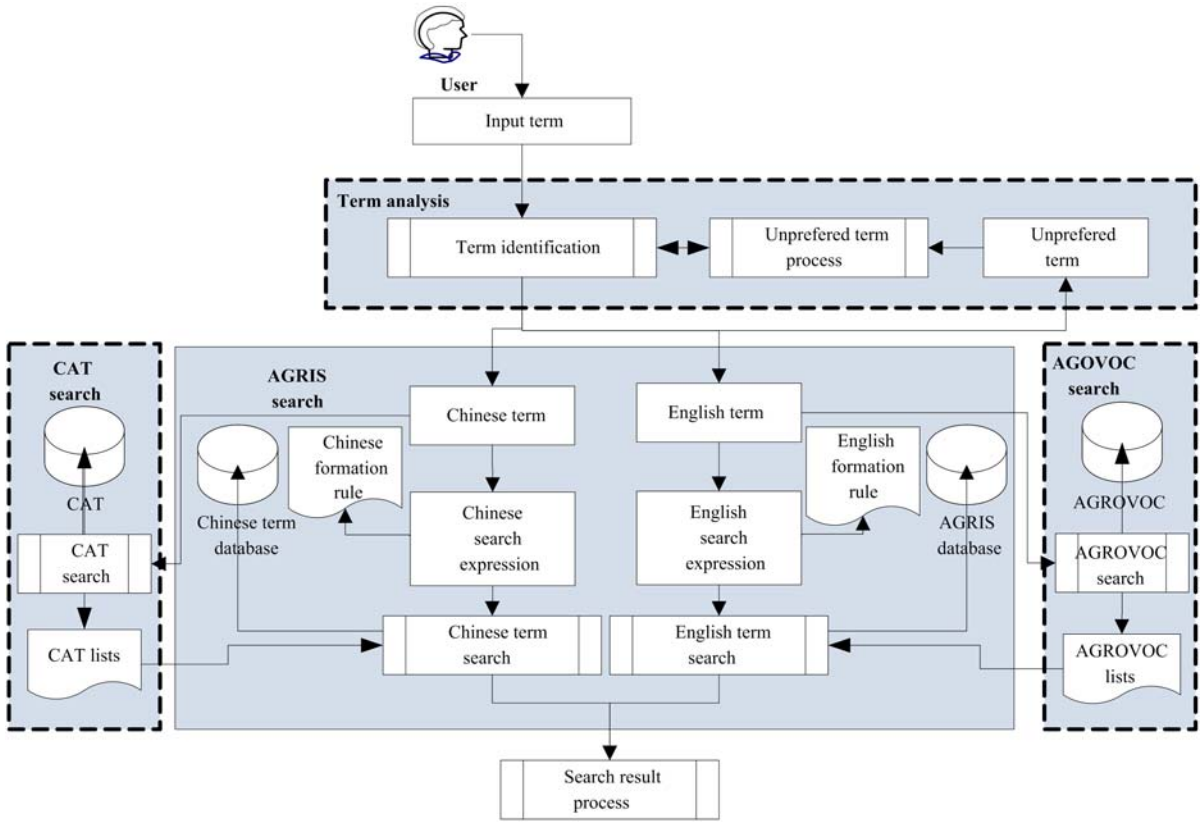

Fig. 2 Search model based on CAT/AGROVOC mapping.

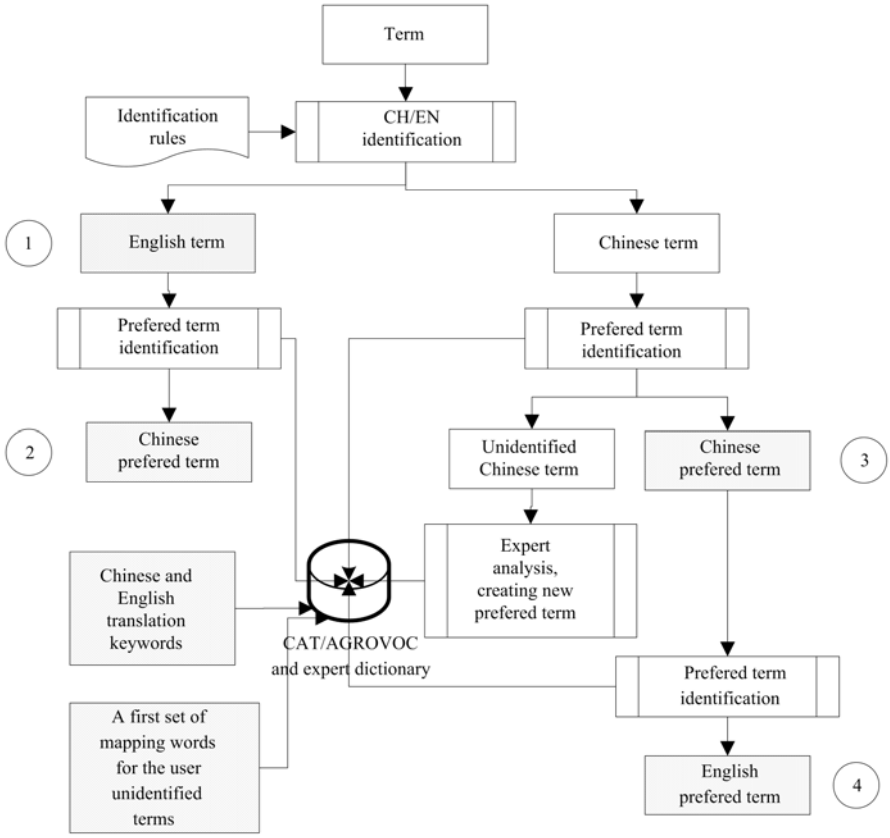

Fig. 3 Term analysis flowchart.

nese preferred terms in CAT, they are unidentified Chinese terms, then, send them to expert analysis and learning processing module. In the module, on the one hand, we have introduced some literature databases in advance, such as web of science, and have extracted English and Chinese translation keywords from agricultural literature, and then have saved them into expert dictionary. On the other hand, by selecting a certain amount of representative agricultural literature as the textual corpus, we have produced a first set of unidentified terms with the words filtering based on CAT/AGROVOC mapping, natural language processing and other technical operations. And through experts mapping analysis, the first set of Chinese and English mapping words for unidentified terms are formed and are also input into expert dictionary. In addition, expert will make Chinese and English mapping analysis for the user unidentified terms regularly and continuously in a short-term, and add Chinese and English mapping words into expert dictionary (Sheng *et al*. 2001; Guo *et al*. 2002); if there are corresponding preferred term(s) in CAT, go to step (3).

(3) Get the corresponding Chinese preferred term(s) from CAT, map them to AGROVOC according to CAT/AGROVOC mapping table to get English preferred term.

(4) According to CAT/AGROVOC, map English preferred term to CAT to get Chinese preferred term.

Terms (1), (2), (3), and (4) can be used directly as processed terms.

CAT search module This module chiefly performs the CAT database retrieval function, and displays the alphabetic results in order to select more exact Chinese preferred terms to assist rebuilding a Chinese retrieval expression.

AGROVOC search module This module chiefly performs the AGROVOC database retrieval function, and displays the alphabetic results in order to select more exact English preferred terms to assist rebuilding an English retrieval expression

AGRIS search module Take preferred terms (2), (3) produced by term analysis module as Chinese preferred terms, construct Chinese query expressions according to query expression constructing rules, retrieve the

AGRIS Chinese database, and form the Chinese retrieval result; take preferred terms (1), (4) produced by term analysis module as English preferred terms, construct English query expressions according to query expression constructing rules, retrieve the AGRIS English database, and form the English retrieval result; users can adjust the query expressions with the assistant of the corresponding ZH/EN thesaurus retrieval results list every time they do a retrieval, no matter Chinese retrieval or English one.

Finally, through retrieval result processing, the alphabetic results are displayed, the ZH/EN bilingual semantic retrieval is completed.

THE REALIZATION OF THE CN/EN BILINGUAL INFORMATION RETRIEVAL SYSTEM BASED ON THE CAT/AGROVOC MAPPING

According to the proposed ZH/EN bilingual information retrieval model based on the CAT/AGROVOC mapping, and on the basis of the present developing technologies of mixed English and Chinese research system (Zhang *et al*. 2005; Xiao *et al*. 2009; Drupal 2010; The Apache Software Foundation 2010; Yuan and Zhong 2010), we constructed ZH/EN bilingual information retrieval prototype system (Chang and Lu 2010; FAO, OEKC 2010a, b).

System structure

The system structure is generally divided into four layers: data layer, function layer, application layer, and user & feedback layer (see Fig. 4). The data layer is the basic data resources of AGRIS retrieval system; the function layer is the functions that a cross-language retrieval system should hold; the application layer faces user application; the user & feedback layer is an interactive layer for users and the system. Details are given below:

Data layer As the data basis of the whole AGRIS bilingual retrieval prototype system, data layer serves as a data supporter providing data support for function layer and application layer. The data source mainly includes: AGRIS Chinese library, AGRIS Chinese library, CAT database, AGROVOC database, and CAT/AGROVOC mapping library. Thereinto, AGRIS Chinese library and AGRIS English library are respectively used to support AGRIS ZH/EN data retrieval, CAT database and AGROVOC database are respectively used

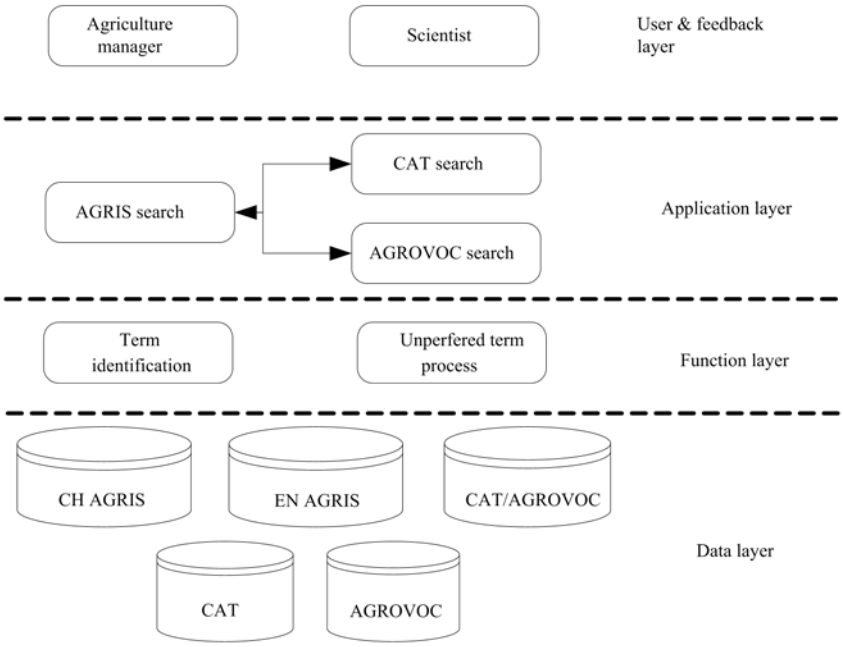

Fig. 4 System structure of the multilingual search system based on the CAT/AGROVOC mapping

to support the preferred terms candidate for AGRIS ZH/EN retrieval; CAT/AGROVOC mapping library is used to support the users' input term analysis.

Function layer Function layer is the special function layer of AGRIS bilingual retrieval system, providing application layer with special function support according to the data provided by the data layer. It is responsible for identifying the ZH/EN preferred terms and unpreferred terms, and processing unpreferred terms. For Chinese users and English ones, it will produce corresponding EN/ZH bilingual preferred terms, and then provides application layer with the produced preferred terms as retrieval expression constructors.

Application layer Application layer takes the preferred term provided by data layer as data source basis, and takes the special function provided by function layer as the function basis, and provides users in user layer with services including AGRIS ZH/EN bilingual retrieval, CAT retrieval and AGROVOC retrieval. CAT retrieval and AGROVOC retrieval have a mechanism for preferred terms communication with AGRIS ZH/EN bilingual retrieval, so as to assist AGRIS ZH/EN bilingual retrieval and improve the retrieval efficiency.

User & feedback layer The service object of AGRIS ZH/EN bilingual retrieval prototype system includes managers and researchers in agriculture area. In user layer, users submit the retrieval information, then modify the retrieval expression with the assistant of the CAT retrieval or AGROVOC retrieval result, repeat the same procedure until get a satisfied results. The whole process forms a feedback mechanism.

AGRIS ZH/EN bilingual retrieval system test

Fig. 5 shows an interface fragment of ZH/EN bilingual information retrieval system based on the CAT/AGROVOC mapping. The center input box is the core part of AGRIS, where users can input terms to conduct a bilingual retrieval. The high right corner input box is the thesaurus retrieval box, where users can find the related preferred terms candidate for modifying query expression. The low right corner windows below the thesaurus retrieval part displays the list of news which is related to the users' input terms. In Fig. 5, user inputs the Chinese term "水位" ('water level' in English) into the center input box, then the related ZH/EN records are selected out, in the thesaurus retrieval box, the narrow term of "水位" ("water level" in English), "水位调节" is displayed for user to conduct a more accurate retrieval.

In the AGROVOC and CAT mapping database, Chinese "水土保持" is an exactMatch term for English "soil conservation". Here we take a simple example for query expression analysis, taking "水土保持" and "soil conservation" as query expressions to illustrate the query expression analysis result of the prototype system. Ex-

Fig. 5 The instance for the interface of CAT/AGROVOC mapping based ZH/EN bilingual information retrieval system.

cept for AGRIS query expression analysis rules, in the system, the retrieval result should also be in accordance with the analyzed expression in Table.

Table Query expression analysis in AGRIS ZH/EN bilingual retrieval system[1)]

No.	Query expression	Query expression results
1	水土保持	"水土" or "保持" or "soil conservation"
2	Soil conservation	"Soil" or "conservation" or "水土保持"
3	"水土保持"	"水土保持" or "soil conservation"
4	"Soil conservation"	"Soil conservation" or "水土保持"

[1)] "水土保持" is "soil conservation" in English. "水土" is "soil" in English. "保持" is "conservation" in English.

CONCLUSION AND OUTLOOK

After a simple test on the system retrieval function, we conclude that the retrieval results of the prototype retrieval system are accordance with the analyzed expression above, and have generally realized ZH/EN bilingual retrieval function of AGRIS, and can meet users' ZH/EN bilingual retrieval requirement to some extent. But presently, the unpreferred terms user input in rebuilding a retrieval expression hasn't been considered; the preferred term mapped by an expert is lagged behind the rebuilding of the retrieval expression; also the system's retrieval efficiency still needs to be improved; all these are the further work that should be resolved in next step.

Acknowledgements

The paper subsidized by the 2010 Central Public-Interest Scientific Institution Basal Research Foundation, China (10211).

References

Apache Lucene – Query Parser Syntax. 2010. [2010-10-20]. http://lucene.apache.org/java/2_4_0/queryparsersynatax.html

CAAS. 2007a. *Chinese Agricultural Thesaurus*. CAAS, Beijing. (in Chinese)

CAAS. 2007b. Mapping database between CAT and AGROVOC.

Chang C, Lu W. 2010. Design about agriculture multi-language search system based on thesaurus. [2010-09-30]. http://www.paper.edu.cn/index.php/default/releasepaper/content/200803-469

Drupal. 2010. A content management system (CMS). [2010-08-20]. http://drupal.org/

FAO. 2010a. AGROVOC term info. [2010-09-10]. http://aims.fao.org/agrovoc-term-info?mytermcode=24929

FAO. 2010b. The last three years of AGRIS XML data (including those from CAAS). [2010-09-30]. ftp://ext-ftp.fao.org/GI/Reserved/Agris/AgrisData/AGRIS_XML/

FAO. 2010c. AGRIS. [2010-10-1]. http://agris.fao.org/

FAO. 2011. AGROVOC thesaurus. [2011-8-20]. http://aims.fao.org/zh-hans/standards/agrovoc

FAO, OEKC. 2010a. Search help of AGRIS. [2010-10-2]. http://agris.fao.org/agris-search/search/search_help.html

FAO, OEKC. 2010b. The AGRIS search engine source code and structure of the AGRIS repository. [2010-09-30]. ftp://ext-ftp.fao.org/GI/Reserved/Agris/Software/AGRIS_Search/

FAO, OEKC. 2010c. The structure of the AGRIS database. [2010-09-30]. http://purl.org/agmes/agrisap/dtd

Guo M, Sun H, Huang T. 2002. Research of knowledge-base organization and knowledge-base maintenance techniques in expert system. *High Technology Letters*, **12**, 1-4, 9.

ICTCLAS. 2010. ICTCLAS Chinese segmentation system. [2010-08-20]. http://ictclas.org/

Liang A C, Sini M, Chang C, Li S, Lu W, He C, Keizer J. 2005. The mapping schema from Chinese agricultural thesaurus to AGROVOC. In: *Proceedings of the 6th Agricultural Ontology Service(AOS) Workshop on Ontologies: The More Practical Issues and Experiences*. Vila Real, Portugal. [2009-12-10]. ftp://ftp.fao.org/docrep/fao/008/af241e/af241e00.pdf

Multilingual Statistical Parsing Engine. 2009. Software. [2009-10-17]. http://www.cis.upenn.edu/~dbikel/software.html

Sheng Z, Zhao W, Chen G. 2001. Expert-oriented optimization of knowledge base. *Journal of Management Sciences in China*, **4**, 40-45.

The Apache Software Foundation. 2010. Lucene. [2010-09-20]. http://lucene.apache.org/

W3C. 2009. SKOS simple knowledge organization system reference. [2010-10-2]. http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Xiao J, Chen X, He H, Cui S. 2009. Design and implementation of web application based on AJAX and struts. *Computer Engineering and Design*, **30**, 1934-1937.

Yakushiji A, Tateisi Y, Miyao Y, Yoshinaga N, Tsujii J. 2003. A Debug tool for practical grammar development. In: *Proceedings of the 41st Annal Meeting on Association for Computerational Linguistics*. Sapporo, Japan. pp. 173-176.

Yuan D, Zhong N, 2010. Initiative retrieval of web information. *CAAI Transactions on Intelligent Systems*, **5**, 112-116.

Zhang Y, Fu H, Wang X. 2005. Research and design of multilanguage webSite based on dynamic technology. *Modern Computer*, **31**, 80-83.

(Managing editor ZHANG Juan)