Thesaurus and Ontology Technology for the Improvement of Agricultural Information Retrieval

Chun Chang, Guojian Xian and Guangda Li

Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing China 100081  changc@mail.caas.net.cn

Abstract

We have been in a web information stage, by new information management technologies, we can get better agricultural development. The paper introduces the research work on agricultural thesaurus and ontology; it could improve the agricultural information retrieval. Main work include to convert Chinese Agricultural Thesaurus (CAT) to the agricultural ontology, this can use traditional domain knowledge directly; research on the thesauri mapping between different languages, this can be help to develop cross languages search engine, and improve the ontologies mapping research; Based on all these works, the paper plans to develop cross languages search engine to index and search web information.

Keywords: Thesaurus, Ontology, Mapping

Based on the widely use of Internet, information management and technology are doing much powerful work for agriculture. It includes two parts of work, first to use the knowledge that human beings have gotten by new technology, or to invent new way to improve the relative work. For traditional organizations of agricultural information, such as agricultural libraries, they have done much work to help readers to get information and knowledge. At this new information stage, they can do more work to help users to get and use knowledge. As an national agricultural information organization, AII/CAAS (Agricultural Information Institute, Chinese Academy of Agricultural Sciences) has done much work to help agricultural readers, now we need do more work in this domain, in this paper, we introduce two parts of work, they are about what we have done on the ontology conversion and thesauri mapping.

1 the conversion work from thesaurus to ontology

As human beings are continually accumulating all kinds of knowledge, in agricultural information management domain, we have created agricultural thesaurus, it is helpful for us to index and retrieve agricultural information in computer data time. In Internet time, ontology will be the appropriate tools for index and retrieval. As agricultural thesaurus has included much domain knowledge, so we can convert this knowledge to ontology, it will be helpful for us to build ontology. Some researchers have done research work in this domain, such as Ms Qin et al, report to convert a controlled vocabulary into an ontology (2001); Information organization such as FAO (Food and Agriculture Organization), do some conversion work too. We converted Chinese Agricultural Thesaurus (CAT) into ontology (in RDF style) in 2004, and OWL style in 2006, now we are developing the intelligent search engine based on the ontology. Here is some research information about this work.
1.1 Introduction on CAT

In 1986, under the auspices of Chinese Academy of Agricultural Sciences (CAAS), nearly 100 domain specialists from more than 40 information institutes throughout China assembled together to collaborate on an CAT System. The CAT would be based on the international and national standards for the building of thesauri, and would make use of material and ideas from international agricultural thesauri such as AGROVOC of FAO, and the CAB thesaurus. After five years, the initial work on the project was complete. The first edition, which appeared in 1994, contained more than sixty thousand Chinese terms, including a corresponding English translation for every Chinese term, as well as Latin names for species terms. It includes biological taxonomy and geographical names; The CAT has been recommended as a national standard for the agricultural domain in China (Information Institute of Ministry of Agriculture, 1994).

1.2 the RDF document of ontology

In 2004, we convert the crop part of CAT into RDF style ontology. The data was transferred to an Access database; a JAVA program was used to generate a KAON file from the database. The file is based on RDF, and has the all relationships of the original thesaurus.

First, we selected the crop class, which includes 989 Chinese terms. After creating and running a system, we got an alphabetic list that corresponds to the taxonomy of the CAT.

Second, We wrote a FoxPro program to get 3 tables: the term number corresponding to the Chinese term name, the relationship file, the scope of terms. We changed the DBF file to an Access database. All relations of the terms have been transferred.

Third, Using a JAVA program, we converted the Access database to a KAON RDFS file. This file could be opened by the OI-Modeler, and showed all the relationships within the crop class.

1.3 the OWL document of ontology

In 2006, we convert CAT to OWL style ontology. We use a new program, it can convert Access document to OWL document, keep all information of thesaurus relationships, and ontology tools Protégé can read it. This time, we not only convert crop science part, but also convert all other parts of CAT.

OWL is a recommended language to present ontology by W3C, it have 3 styles, they are OWL Lite, OWL DL and OWL Full. As OWL DL not only can present hierarchical relations, but also can be full reasoned with properties. So we select OWL DL style as our conversion target. The main three relationships in thesaurus have corresponding relationships in OWL. USE/Use for relationship, normally convert to “<owl:equivalentClass>”; Hierarchical relationships are the relation of “rdfs:subClassOf” in OWL; Associative relation have different relationships in OWL, we just use a main relation to present this situation, the label is “<owl:someValuesFrom>”.

1.4 the application plan of agricultural ontology

Based on the OWL agricultural ontology document, we are developing intelligent search engine now. We have finished the system design, the system can add equivalent class automatically when we search with a concept, and system can also to expand the search
concept, or give more relative search concept based on the associative relations of the thesaurus.

2 the mapping work from CAT to AGROVOC

For CAT, we do another work on it; it is the mapping project from CAT to AGROVOC. We started this work on the end of 2005, and finished concepts mapping, we are developing cross search engine now.

2.1 Introduction on AGROVOC

AGROVOC is a multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment). AGROVOC is online, users can browse and down it on web. It has 17 languages now, include 5 FAO official languages, they are Arabic, Chinese, English, French, Spanish; each language has about 30 thousand terms.

2.2 the mapping work from CAT to AGROVOC

As both of the thesauri, AGROVOC and CAT, concerns the agricultural domain, and each of them has its own structure and particularities, the leading organizations decided to realize a mapping of the domain knowledge between these two thesauri; the resulting inter-thesaurus could then be used to develop applications from which both English and Chinese users can benefit. For example, we can develop a search engine to search databases containing different languages and the AOS initiative can be enriched by the experience of this mapping. Based on these ideas, FAO and CAAS organized a mapping project from CAT to AGROVOC at the end of 2005.

The mapping schema was based on the SKOS rules (Miles A, et al. 2004), which has been revised and adapted based on FAO and CAAS needs (e.g. the inexactMatch rule was used but not the majorMatch and minorMatch). All major rules and logical operators, such as exactMatch, broadMatch, narrowMatch, AND, OR, NOT, were included in the mapping mechanism. Based on our initial analysis, we also supposed that some CAT concepts/terms would not have a mapping in AGROVOC.

Part of the preparative work, was the conversion of the CAT Foxpro database to Microsoft Access, in order to have the same format as the AGROVOC thesaurus. Subsequently, both thesauri have been represented with the Ontology Web Language (OWL). In order to improve performances and allow distributed work, CAT concepts, have been split into categories. Each of the OWL files containing separate concepts grouped by category would have been given to a specific expert for performing the mapping. The tool used to realize the mapping is Protégé: all CAT OWL files would have been processed separately. In a second phase, all the mapping files would have been incorporate in a unique document. Now we have given all mapping information for CAT terms.

2.3 To develop the Chinese-English cross search engine based on mapping work

Based on the rapid development of Chinese agricultures, Chinese people have enough food for their life. How they resolve food problems for this 1.3 billion people, many English users are interesting on the Chinese agricultural information. As Chinese agricultural economy is
becoming one part of world economy, many Chinese users are interesting on English agricultural information too. Based on the mapping project, we have gotten a design schema, it is an agricultural cross languages engine, and the core technology is the mapping between Chinese and English agricultural thesauri. With the mapping information, English users can get Chinese agricultural information from web data by English descriptors, Chinese users can get English agricultural information form web data by Chinese descriptors. The system lines are in the Fig. 1.

We design to use machine translation system, as Chinese-English machine translation software can give enough translation information for users. Although there are some errors when users read the translation information, they can get more useful information. If users want to get native language information, the information organizations can give these kinds of service.

Fig. 1: Cross languages search engine architecture

3 Discussions on the conversion and mapping

3.1 To rebuild domain knowledge relationships

Thesaurus is useful to build ontology; we can convert main knowledge to ontology. But this conversion needs to be rebuilt by information and domain knowledge specialists. Thesaurus has its own national and international standard; it is impossible to save as ontology style
simply. Nearly all relationships and concepts need to be refined based on the operation of ontology. It includes basic modification and special modification. Basic modification, such as in ontology, ‘city’ is a class, ‘Beijing’ and ‘Shanghai’ are instances, but in CAT, ‘Beijing’ and ‘Shanghai’ are concepts under ‘China’, so when we do the conversion work, these relationships need to be refined artificially. In order to make the ontology more powerful, we need add more properties, more relationships, so the conversion work includes many other works, it has more work than to build a thesaurus, as this huge work, we can find more research work on conversion, it is difficult to find finished conversion ontology.

3.2 The main problem of cross languages search engine

Thesauri are design to used by human and computer, indexers choose some appropriate concepts in thesauri to index documents, readers can find these documents with concepts, some time this information is keep in computers. For our cross language search engine, it is based on Internet, users normally don’t know or no need to know thesauri, they more like to use natural languages, such as keywords, how these keywords can convert thesauri concepts? This is an important problem, thesauri can only to resolve some problems, we can give relationships artificially, but this is a huge work too, and seem no organizations can finish this work. For this problem, it is impossible to include all conversion information, we just try to make it useful to users, and this is the correct way to resolve the problem.

Acknowledgements

The mapping project was supported by FAO, this paper is the main work related with Chinese. FAO information officers, Dr Johannes Keizer, Ms Margherita Sini, did much work on the whole mapping project.

Reference

http://www.fao.org/aims/pub_aos5.jsp
http://www.afita2006.org/index_files/Page1119.htm
Miles, A., Matthews, B. Inter-Thesaurus Mapping.
http://www.w3c.rl.ac.uk/SWAD/deliverables/8.4.html
http://www.w3.org/2004/02/skos/mapping/spec