AGRICULTURAL CROSS LANGUAGES INFORMATION RETRIEVAL SCHEMA BASED ON MUTI-THESAURUS MAPPING

Chun Chang [1,*], Wenlin Lu [2]

[1] *Institute of Scientific and Technical Information of China, Beijing, P. R. China 100038*

[2] *Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing, P. R. China 100081*

[*] *Corresponding author, Address: Institute of Scientific and Technical Information of China, Beijing, 100038, P. R. China, Tel: +86-10-58882392, Fax: +86-10-58882312, Email: changchun@mail.istic.ac.cn*

Abstract: Based on the rapid development of Chinese agricultures, many English users are interesting on the Chinese agricultural information, and many Chinese users are interesting on English agricultural information too. This paper is a schema to design an agricultural cross languages engine, the core technology is the mapping between Chinese and English agricultural thesauri. The paper introduces the all rules of thesauri mapping, and give exact examples for these rules. With the mapping information, authors design a cross languages engine. English users can get Chinese agricultural information from web data by English descriptors; Chinese users can get English agricultural information from web data by Chinese descriptors

Keywords: cross languages search engine, mapping, thesaurus

1. INTRODUCTION

1.1 Background

Thesauri were born in 1950s, and, since then, they have been broadly used in different domains, both by humans and machines. In particular, they were

successfully used in information retrieval (Chang Chun, 2002). A thesaurus can be considered as a system for representing domain knowledge: it has some basic relationships between concepts/terms, and it is managed by information and domain specialists. From the '90s of 20 century, Internet has been used broadly: users can find enough information in Internet searching by keywords. Despite of that, sometimes it is possible to get so many results that it is difficult to select what is need. In order to avoid these problems, some information specilists research thesauri and ontology to resolve these problems (Chang Chun et al., 2004; Qin Jian, 2001).

The FAO developed a multilingual agricultural thesaurus, AGROVOC, it 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(Agricultural Information Management Standards, 2008). the Chinese Academy of Agricultural Sciences (CAAS) has developed the Chinese Agricultural Thesaurus (CAT) in Chinese with English translation (Information Institute of Ministry of Agriculture, 1994). As both of the thesauri 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. Based on this ideas, FAO and CAAS organized a mapping project from CAT to AGROVOC at the end of 2005(Chang Chun, 2006; Chang Chun 2007).

1.2 Work methods

The mapping schema was based on the SKOS rules (Miles, 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, 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.

2. USE PROTÉGÉ TO EXPRESS ALL RELATION RULES

OWL can express the mapping relations well, we use OWL to keep the mapping relation information.

2.1 The exactmatch relation

ExactMatch is one of the main relation of the mapping project. It is recorded as 'equivalentClass' in OWL. Such as '禾谷类作物' (English translation 'Cereal crop', Chinese termcode 17147) exact match with 'Cereal crops' (Chinese translation "禾谷类作物", English termcode 25512), we got the OWL document with Protégé, here is a part of the OWL document.

```
<rdf:Description rdf:about="http://www.caas.net.cn/2005/cat#c_17147_禾谷类作物_Cerealcrop">
  <owl:equivalentClass>
    <rdf:Description rdf:about="http://www.fao.org/aos/agrovoc/2005#c_25512_Cerealcrops_禾谷类作物">
      <owl:equivalentClass rdf:resource="http://www.caas.net.cn/2005/cat#c_17147_禾谷类作物_Cerealcrop"/>
    </rdf:Description>
  </owl:equivalentClass>
</rdf:Description>
```

2.2 The broadmatch relation

BroadMatch is another main relation of the mapping project. It is recorded as 'subClassOf' in OWL. Such as '普及教育' (Universal education, Chinese termcode 35234) broad match with 'Education' (Chinese translation "教育", English termcode 2488), the OWL document is as follow.

```
<rdf:Description rdf:about="http://www.caas.net.cn/2005/cat#c_35234_普及教育_Universaleducation">
  <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc/2005#c_2488_Education_教育"/>
</rdf:Description>
```

2.3 The narrowmatch relation

NarrowMatch seldom happen in our mapping project. It is also recorded as 'subClassOf' in OWL. Only need do it on the reversal way. Such as '岛屿' (Islands, Chinese termcode 8341) narrow match with 'Atolls' (Chinese translation "环礁", English termcode 695), the OWL document is as follow.

```
<rdf:Description rdf:about="http://www.fao.org/aos/agrovoc/2005#c_695_Atolls_环礁">
  <rdfs:subClassOf rdf:resource="http://www.caas.net.cn/2005/cat#c_8341_岛屿_Islands"/>
</rdf:Description>
```

2.4 the AND relation

There are some AND relations in the mapping project. It is recorded as 'intersectionOf' in OWL. Such as '自动标引' (English translation 'Automatic indexing', Chinese termcode 59683) exact match with the conception of 'Indexing of information' (Chinese translation "信息编目", English termcode 11729) AND 'Automation' (Chinese translation "自动化", English termcode 15855), the OWL document is as follow.

```
<rdf:Description rdf:about="http://www.caas.net.cn/2005/cat#c_59683_自动标引_Automaticindexing">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <rdf:Description  rdf:about="http://www.fao.org/aos/agrovoc/2005#c_11729_Indexingofinformation_信息编目
"/>
        <rdf:Description rdf:about="http://www.fao.org/aos/agrovoc/2005#c_15855_Automation_自动化"/>
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</rdf:Description>
```

2.5 the OR relation

There are some OR relations in the mapping project. It is recorded as 'unionOf' in OWL. Such as '大麦' (Barley, Chinese termcode 7536) exact match with the conception of 'Barley' (大麦, English termcode 823) OR 'Hordeum vulgare' (大麦植物, English termcode 3662), the OWL document is as follow.

```
<rdf:Description rdf:about="http://www.caas.net.cn/2005/cat#c_7536_大麦_Barley">
  <owl:equivalentClass>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
```

```
      <rdf:Description rdf:about="http://www.fao.org/aos/agrovoc/2005#c_823_Barley_大麦"/>
      <rdf:Description rdf:about="http://www.fao.org/aos/agrovoc/2005#c_3662_Hordeumvulgare_大麦植物"/>
     </owl:unionOf>
    </owl:Class>
   </owl:equivalentClass>
 </rdf:Description>
```

2.6 the NOT relation

There are some NOT relations in the mapping project. It is recorded as 'complementOf' in OWL. Such as '非传染性病害' (English translation 'Non-infectious diseases', Chinese termcode 12114) exact match with the conception of 'Plant diseases' (Chinese translation "植物病害", English termcode 5962) AND NOT 'Infectious diseases' (Chinese translation "侵染性病害", English termcode 34024), the OWL document is as follow.

```
<rdf:Description rdf:about="http://www.caas.net.cn/2005/cat#c_12114_非传染性病害_Non-infectiousdiseases">
   <owl:equivalentClass>
    <owl:Class>
    <owl:intersectionOf rdf:parseType="Collection">
     <rdf:Description rdf:about="http://www.fao.org/aos/agrovoc/2005#c_5962_Plantdiseases_植物病害"/>
     <owl:Class>
     <owl:complementOf rdf:resource="http://www.fao.org/aos/agrovoc/2005#c_34024_Infectiousdiseases_侵染性
病害"/>
     </owl:Class>
    </owl:intersectionOf>
   </owl:Class>
  </owl:equivalentClass>
</rdf:Description>
```

2.7 The comment of No Mapping relation

Some terms have no mapping relation with AGROVOC, we use the comment tool to express it. It is recorded as 'comment' in OWL. Such as '干扰' (English translation 'Interference', Chinese termcode 13867) has no mapping relation terms in AGROVOC, the OWL document is as follow.

```
<rdf:Description rdf:about="http://www.caas.net.cn/2005/cat#c_13867_干扰_Interference">
   <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
   >AGROVOC hasn't this concept</rdfs:comment>
</rdf:Description>
```

3. 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.

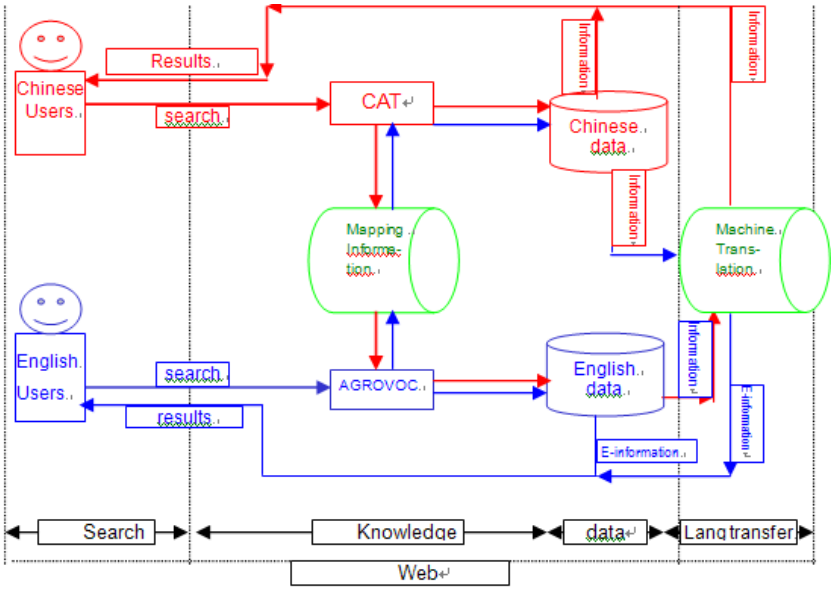

Fig. 1: Cross languages search engine architecture

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.

4. DISCUSSIONS

4.1 The development of data languages

The information of thesauri and mapping, can be keep with different data languages, human beings renew these languages continually. First we use RDF to keep these information, then we have OWL. Some years later, we will have other more new languages to repeat them. With the different languages, we have more chances to express new function. So there is a problem, it means we must develop tools to convert different languages documents, we need to keep original knowledge, or rebuilt these knowledge to new styles. The best method is to use conversion tool directly, as this research is quite new, normally we can't find standard tools.

4.2 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 these 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 main work is finished in Chinese Academy of Agricultural Sciences (CAAS), the project is supported by Food and Agriculture Organization (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.

REFERENCES

Agricultural Information Management Standards. http://www.fao.org/aims/ag_intro.htm

Chang Chun Lu Wenlin. The past and current situation and the future development of thesauri. Journal of Library and Information Science in Agriculture. 2002,(5):25-28 (in Chinese)

Chang Chun, Lu Wenlin. From Agricultural Thesauri to Ontologies. Fifth Agricultural Ontology Service (AOS) Workshop, 27-29 April 2004, Beijing. http://www.fao.org/aims/pub_aos5.jsp

Chang Chun. CAT-AGROVOC Mapping. http://www.fao.org/aims/pub_aos8.jsp

Chang Chun. Organizing and Implementing on the Thesauri Mapping Project. Seventh Agricultural Ontology Service (AOS) Workshop, 9-11 November 2006, India. http://www.afita2006.org/index_files/Page1119.htm

Information Institute of Ministry of Agriculture. Chinese Agricultural thesaurus. Chinese Agricultural Press, 1994(in Chinese)

Miles,A.,Brickley,D.SKOS Mapping Vocabulary Specification. http://www.w3. org/2004/02/skos/mapping/spec

Miles, A., Matthews, B. Inter-Thesaurus Mapping. http://www.w3c.rl.ac.uk/SWAD/deliverables/8.4.html

Qin Jian, Paling Stephen. Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, 2001,6(2) Available at: http://InformationR.net/ir/6-2/paper94 html