THE ACQUISITION OF CLASS DEFINITIONS IN THE COMMODITY ONTOLOGY OF AGRICULTURAL MEANS OF PRODUCTION

Lu Zhang [1] , Li Kang [1,*] , Xinrong Cheng [1] , Guowu Jiang [1] , Zhijia Niu [2]

[1] *College of Information and Electrical Engineering, China Agricultural University, Beijing, P. R.. China 100083*

[2] *Agricultural bank of China, Beijing, P. R..China 100073*

[*] *Corresponding author, Address: College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, P. R.. China, Tel: +86-13691086472, Email: kangli.cau@gmail.com*

Abstract: The agricultural means of production is also called the means of agricultural production. The previous work focused on constructing the means of agricultural production commodities ontology taxonomy. After finishing constructing the ontology taxonomy, adding the detail information to the ontology are the following work. The detail information includes classes' definitions, properties, relations, instances and axioms. The classes' definitions are the concrete domain knowledge manifestation. Therefore they can be used to learn class's properties and relations. This paper focuses on obtaining these definitions. To this end, the most important work is to determine where to obtain class's definitions. This paper discusses the selective process. According to authority, completeness, accuracy, practicability and computerization, compares three kinds of knowledge sources, and then selects online "Encyclopedia of China" as the knowledge source. Analyzes the encyclopedia website and its entries, and then proposes an automatic method to get class definitions. The experiment shows that nearly 70% of classes can get their definitions. Using Jena API to add the definitions to the ontology model represented in OWL format.

Keywords: means of agricultural production commodities, ontology, class definitions

1. THE METHODOLOGY FOR BUILDING THE COMMODITY ONTOLOGY OF AGRICULTURAL MEANS OF PRODUCTION

Choose the seven-step method (Noy et at., 2001) to construct the ontology. The previous work has completed the ontology taxonomy. Under the seven- step method, the next step is for the ontology to add attributes, relationships, instances and axioms. The implementation of this step must have knowledge sources, which contain the knowledge of the means of agricultural production commodities. In fact, knowledge source is equivalent to class definitions in the ontology. Therefore getting class definitions is the prerequisite for the future work. In order to obtain definitions, where to find the definitions is the first issue to be considered. Because the total number of classes in the ontology was over 700, it is not suitable for manual input definitions. So the next step is to identify suitable knowledge sources from which we can get definitions automatically. Section II discusses the knowledge source used by the process of getting definitions and its selective process.

2. THE CHOICE OF KNOWLEDGE SOURCES

2.1 The characteristics of knowledge sources in this paper

The commodity ontology of agricultural means of production will be applied for intelligent agricultural information systems used by peasants. So the knowledge described by the ontology should be recoganized generally to be authoritative and practicable within the subject and circulation domain of the means of agricultural production. Besides, the completeness and accuracy of the knowledge contained in the ontology are other factors that decide qualities of services provided by the agricultural information systems. So related knowledge sources from which we get the class definitions should have these characteristics above. For the purpose of getting definitions automatically, it is necessary to consider whether the knowledge source could be processed by computers conveniently. In a word, the characteristics of the knowledge source in this paper are authoritative, complete, practicable, accurate and digitized.

2.2 The selective process of the knowledge sources

If we see sci-tech documents on this domain (the means of agricultural iaotangshan Station. production commodities) as knowledge sources (Wang Qian, 2004), the domain experts must find out suitable ones among lots of documents. It is time-consuming. Besides, although a large number of sci-tech documents stored in database are easy to access, domain knowledge is widely distributed and thus all of them requires amount of time to be find out.

Professional books include textbooks and monographs. Their content not only could reflect author's academic level, but also is a more systematical explanation to the domain knowledge than sci-tech documents'. However, there are many different levels books in a field, so ontology engineers have to spend a lot of time on finding out the books with authoritativeness and accuracy. Besides, electronic versions of some books are hard to access, and thus ontology engineers will have to enter a large amount of text when they find class definitions in these books.

Encyclopedias and dictionaries both cover and contain everything. They are related to all human subject and knowledge (Jin Changzheng, 2007). Therefore encyclopedias could give a comprehensive explanation to all domain knowledge. Furthermore persons usually say that encyclopedias are "a university without walls" (Jin Changzheng, 2007), so encyclopedia is also an educational book for persons and the knowledge explained by it could be authoritative, accurate and practicable. At present some online encyclopedias could be looked up easily and it is probable that a program based on the structure of online encyclopedias can get every entry's explanation automatically.

"Encyclopedia of China" is the first large and comprehensive encyclopedia (Encyclopedia of China Publishing House, 1980), and it is also a large-scale encyclopedia in the world. From 1978 to 1993, the chief editor committee for "Encyclopedia of China" and the Encyclopedia of China Publishing House organized more 20,000 experts and scholars to complete this book. "Encyclopedia of China" is composed of different volumes explaining different subject knowledge. This paper chose the online "Encyclopedia of China" as the knowledge source to get class definitions.

By these consideration above, we have laid solid foundation for the future work of acquiring class definitions.

3. THE ACQUISITION OF CLASS DEFINITIONS

In some volumes of online "Encyclopedia of China", the names of some entries are the same as the classes of ontology and thus these entries' explanation can be seen as the definitions of corresponding classes. The aim of this paper is to get these entries' explanations and add them to rdfs:comment element of corresponding classes in the ontology model represented in OWL format.

The development platform used is Eclipse, and the program language is Java.

Table 1. The characteristics of three knowledge sources

The characteristics of knowledge sources	Knowledge sources		
	Encyclopedias	Sci-tech documents	Professional books (textbooks and monographs)
Authoritativeness	Educational books and written by domain experts, knowledge contained is authoritative	Some are written by domain experts, high authoritativeness; it will take ontology engineers some time to find them	Some are written by domain experts, high authoritativeness; it will take ontology engineers some time to find them
Completeness	Contain all domain knowledge about any subject	A paper usually focuses on a small part of a domain, and it will take ontology engineers a lot of time to find all	Better than sci-tech documents, not as good as encyclopedias, and it will take ontology engineers some time to find all
Accuracy	Educational books and written by domain experts, knowledge contained is accurate	Some are written by domain experts, high accuracy; it will take ontology engineers some time to find them	Some are written by domain experts, high accuracy; it will take ontology engineers some time to find them
Practicability	Educational books for persons, knowledge contained is practicable	It will take ontology engineers some time to find the documents which are related to practical application	It will take ontology engineers some time to find the books which are related to practical application
Computerization	Online versions could be processed by computers	Electronic versions are easy to processed by computers	Difficulty to be processed in the absence of electronic versions

Process of getting definitions is divided into three phases based on the type of work that is being done. In the first phase, write a program allowing obtaining explanations of entries in the form of HTML from online "Encyclopedia of China". In the second phase, HTML tags can be removed from HTML files with the benefit of HTML parser (Oswald et at., 2006) to get pure text content of explanations. In the third phase, with the help of Jena API (Battle et at., 2001), the explanations will be processed and added to rdfs:comment element of corresponding classes which don't have class definitions in ontology.

Get explanations of entries in the form of HTML

Online "Encyclopedia of China" is organized by different volumes. Type http://202.112.118.40:918/search?ChannelID=2 in browser and the browser will show a list of entries in the form of hyperlink. Every hyperlink pointed to a file whose content is an explanation of current entry. There are 78203 entries and 3911 pages on online "Encyclopedia of China". These entries are showed according to the sequence of volumes. It is shown in Fig.1.

Click 'view source' in browser to view the HTML source code of the web page from http://202.112.118.40:918/search?ChannelID=2. They are shown in Fig.2-4.

In Fig.3, the FORM element whose name is "OutlineForm" represents the choice of pages from page 1 to page 3911. In this form, the INPUT element whose name is "pagetext" is an input box where write page number (In Fig.1, it is 1). A JavaScript function "Outline_onsubmit( )" in Fig.3 save the number in "pagetext" to the INPUT called "page" and then submit this form. When submitting this form, according to the action attribute of the "OutlineForm" form, the browser opens another URL representing another page which contains a few other entries. The action attribute in Fig.3 is "/outline?ChannelID=2&randno=3960" and it is also called "page base URL" because every new page URL is composed of the "page base URL" and a page number. The new URL's format is "host + page base URL + &page=pagenumber". For example if when I wanted to open the URL of page 12, the wanted URL is

http://202.112.118.40:918/outline?ChannelID=2&randno=3960&page=12

In Fig.4, "DetailForm" form shows every entry's explanation. Every entry is represented in the form of an "<a>" element. A JavaScript function "javascript:gotorec('1','32037')" in Fig.2 submit the "DetailForm" form with the first parameter whose value is 1 in Fig.4. In Fig.2, the function of "gotorec" saves the number of clicked entry to an INPUT tag called "record". The same as opening another page URL above, the URL for opening an entry's explanation is http://202.112.118.40:918/detail?ChannelID=2&randno=19265&&record=1

The string of "/detail?ChannelID=2&randno=19265" is used to indicate the constant part of URL for opening an entry explanation, which is also

called “detailInfoFormBaseURI” in program written for the first phase.
The first phase’s flowchart is shown in Fig.5.

Fig.1: Display a list of entries in a browser

```
function gotorec(currec,rand)
{
        document.DetailForm.record.value=currec;
        document.DetailForm.submit();
}

function gotopage(type)
{
        var curpage=1;
        var pagenum=3911;

        if(type=="head"){
                curpage=1;
        }
        if(type=="tail"){
                curpage=pagenum;
        }
        if(type=="prev"){
                curpage--;
                if(curpage<1)return;
        }
        if(type=="next"){
                curpage++;
                if(curpage>pagenum)return;
        }
        document.OutlineForm.page.value=curpage;

        document.OutlineForm.submit();
}
```

Fig.2: Part of source code from http://202.112.118.40:918/search?ChannelID=2

```
<form method="post" name="OutlineForm" action="/outline?ChannelID=2&randno=3960"
target=_self>
<input type="hidden" name="presearchword" value="">
<input type="hidden" name="presortfield" value="">
<input type="hidden" name="preextension" value="">
<input type="hidden" name="page" >
       <td nowrap><a href="javascript:gotopage('head')"><img src="\web\gotohead.bmp"
border=0 alt="首页"></a></td>
       <td nowrap><a href="javascript:gotopage('prev')"><img src="\web\gotoup.bmp"
border=0 alt="上一页"></a></td>
       <td nowrap><input type="text" name="pagetext" size="4" value="1"
onkeydown="Outline_onsubmit()"></td>
       <td nowrap><a href="javascript:gotopage('next')"><img src="\web\gotodown.bmp"
border=0 alt="下一页"></a></td>
       <td nowrap><a href="javascript:gotopage('tail')"><img src="\web\gototail.bmp"
border=0 alt="尾页"></a></td>
</form>
```

Fig.3: Part of source code from http://202.112.118.40:918/search?ChannelID=2

```
<table cellpadding="1" cellspacing="0" >
<form method="post" name="DetailForm" action="/detail?ChannelID=2&randno=19265"
target="main">
<input type="hidden" name="presearchword" value="">
<input type="hidden" name="presortfield" value="">
<input type="hidden" name="preextension" value="">
<input type="hidden" name="record" >

  <tr align=left>
    <td valign=top width=1 align=left>
       <a href="javascript:gotorec('1','27980')" style="text-decoration: none; font-size:
9.0pt;"><img src="\web\page.gif" border=0></a>
    </td>
    <td valign=top align=left>
       <a href="javascript:gotorec('1','32037')" style="text-decoration: none; font-size:
9.0pt;"><font size=3 style="font-family: ZYSongDbk">财政</font></a>
    </td>
  </tr>
```

Fig.4: Part of source code from http://202.112.118.40:918/search?ChannelID=2

We try to get entries' explanation from five volumes including agriculture, modern medicine, chemical engineering, mechanical engineering and Chinese traditional medicine. In online "Encyclopedia of China", the start page number and total page number of volume on Agriculture are 1563 and 119, on Modern Medicine are 2947 and 88, on Chemical Engineering are 565 and 67.

Finally, we get files whose file name is entries' name and the file content is corresponding entry's explanation. The volumes and their total number of files are shown in Table 2.

2 Extract pure text content from HTML-format text files

The second phase removes some extra information from files generated by the first phase, such as the icons of print, and then extracts pure text content from HTML-format text files with the help of HTML Parser

API(Oswald et at., 2006). The obtained pure text content replaces the HTML-format content generated by the first phase.

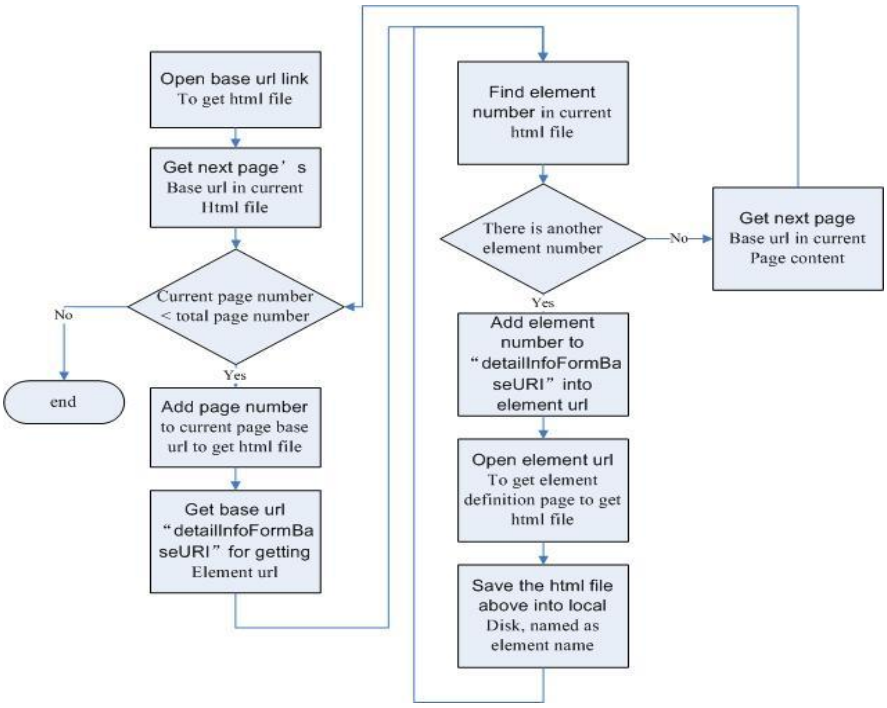

Fig.5: The flowchart of the first phase

Table 2. The volumes and their total number of files

Volume name	The total number of files
Agriculture	2391
Modern medicine	1759
Chemical engineering	287
Mechanical engineering	1438
Chinese traditional medicine	306

3 Process the content of files and add them to ontology

The third phase starts to extract class definitions from files generated by the second phase. Three methods are used to finish this task. If a class's name of ontology is equal to a file's name then the file's content must be the class's definition, which is the first method. If a class of ontology doesn't find any definition with the first method, the second method will be applied to check whether a class name and a file name have intersection. For example, if a class's name is ABC and a file's name is BC, then the file's content may be the class's definition. The third method will work under the situation when the second method doesn't find any definition. This method is designed to regard a paragraph of the file as the class's definition when the class name appears in that paragraph.

Besides, definitions acquisition programs could get subclass definitions in the class definition with the help of regular expressions based on language rules.

水溶性磷肥(water soluble phosphorous fertilizer)即能溶解于水的磷肥。主要品种有：

①过磷酸钙(ordinary super phosphate)，由磷矿粉经硫酸处理而成，为磷酸一钙、石膏及少量游离酸的混合物，有效磷(P2O5)含量 12～20%，易吸湿、结块。② ...

The content above is a definition of water soluble phosphorous fertilizer. Ordinary super phosphate is a subclass of water soluble phosphorous fertilizer. Some other class definitions also have this pattern. Corresponding regular expression is "[\u2460-\u2473]([\u4e00-\u9fa5]+)，([[[\u4e00-\u9fa5\u002d-\u0039\uff08-\uff09]+ 、]* ，]*[\u4e00-\u9fa5\u002d-\u0039\uff08-\uff09]+)". Before using this regular expression to extract subclass definitions, definitions require a preprocessing to replace some punctuation and then only "，" and "、" can appear in subclass definition.

4. RESULTS AND DISCUSSIONS

Table 3. The results of experiments

First level class	All	Has def	First level class	All	Has def
veterinary drug	109	63	fertilizer and manure	19	16
pesticide	46	28	small and middle farm implements	78	72
feedstuff	16	7	seed	13	11
land	13	10	agricultural machinery	414	287
Breeder & breeder bird	20	18	agricultural film	27	4

In Table 3, first level classes include 10 classes under the root of the ontology. For a first level class, the total number of its subclasses is recorded under the column of "all" and the total number of its subclasses which have class definition is recorded under the column of "has def". Both statistics count the first level classes themselves. The total number under the column "all" is 755, under the column "has def" is 516. Nearly 70 percent classes of ontology have their definitions.

5. CONCLUSIONS

The work of this paper focuses on the acquisition of class definitions for the constructing of the commodity ontology of agricultural means of

production. First of all, various knowledge sources are comprehensively discussed according to the characteristics of authoritativeness, accuracy, completeness, practicability and computerization. Finally online "Encyclopedia of China" is chose as the knowledge source for getting class definitions automatically. Experiment shows that the definitions of most classes (nearly 70%) in the ontology can be obtained automatically. It will save ontology engineers a lot of time and labor and provide basic data to support following extracting work.

The future work is to add class definitions for classes whose definitions don't exist in online "Encyclopedia of China". Then different from traditional methods that rely on domain experts to manually build the ontology, a highly efficient and intelligent method will be used to mine properties, relations, instances and axioms from class definitions.

ACKNOWLEDGEMENTS

Funding for this research was provided by the sub-topics of the National Science and Technology Support Plan of China (Grant name: The Study on Constructing Technology of the Commodity Ontology of Agricultural Means of Production. Grant Number: 2006BAD10A050103.)

REFERENCES

Derrick oswald, somik raha, ian macfarlane, david walters.

Encyclopedia of china publishing house. Encyclopedia of china,

Html parser,http://htmlparser.sourceforge.net/ ,2006

Http://protege.stanford.edu/publications/ontology_development/ontology101.pdf, 2001

Introduction: http://202.112.118.40:918/web/bzxx.htm#bkqs, 1980(in chinese)

Jin changzheng. Encyclopedics and a science of compiling encyclopedia. Editors monthly,2001,(5):24-25(in chinese)

Natalya f. Noy and deborah l. Mcguinness. Ontology development 101: a guide to creating your first ontology. 2001.

Online version: http://202.204.214.134:918/web/index.htm, 1980(in chinese)

Steve battle et at. .jena, http://jena.sourceforge.net, 2001

Wang qian. Research on approach to construct dynamic ontology based on text mining[doctor's degree paper,jun,2007]. Beijing: china agricultural university, 2007(in chinese)