A Semantic Search Engine Based on SKOS Model Ontology in Agriculture

Yong Yang，Jinhui Xiong，Shuyan Wang

School of Information and Electrical Engineering, Shengyang Agricultural University, 110866 Shenyang, China
E-mail: yangsyau@163.com

Abstract. A simple agriculture ontology system was constructed under extended SKOS model in this paper. A theme relevance algorithm based on terms' distances in ontology system was tested and applied in improving the Pagerank evaluating. And also an online agricultural semantic search engine named as Sonong was implemented and deployed for service on internet. This online engine provides semantic hierarchy inference with the ontology system and a satisfying ranking list of retrieved information.

1. Introduction

The application of IT in agriculture in china is still in its infancy, the number of agricultural websites and rural community users has steeply increased in recent years [1]. Liu etc. introduced a theme filter in general search engine, and improved accuracy and completeness in agricultural information retrieval by adopting algorithm of keyword oriented vector space model [2]. Xian etc. employed agricultural ontology in index system to capture semantic relations between terms and implemented a prototype search system [3]. Zhou etc. reviewed agricultural semantic search on system structure, functions, and key algorithms, and practiced a structural indexing on Chinese literal web pages by introducing semantic relationships under the SDD algorithm [4]. Zhou etc. constructed an agricultural search engine from Nutch architecture, and improved the accuracy by using agricultural lexicon, theme filtering and ranking techniques [5]. These achievements have furthered the research on agricultural domain search engine. This paper presents the methods of conducting a simple agriculture ontology system under the extended SKOS model and a theme relevance algorithm based on terms' distances in ontology system. An online agricultural semantic search engine named as Sonong is developed and deployed for service at www.sonong.com.

2. Agriculture ontology

Ontology is, from a philosophic viewpoint, the study of existence, of all kinds of entities - abstract and concrete - that make up of the world [6]. A formalized ontology is defined as an explicit specification of shared concepts and theories [7][8]. This formal representation enables computer operations as well as aiding human comprehension [9]. A linguistic ontology contains a list of terms in a glossary for a specific domain and relationships between terms. A mixed ontology is made up of a concept hierarchy called TBOX in knowledge base, consists of terms with generalization or specification relationships [10].

We mapped the Chinese Agricultural Thesaurus into a light ontology system under an expanded SKOS model. SKOS is an area of work developing specifications and standards in supporting the use of Knowledge Organization Systems (KOS). It provides a standard way to represent knowledge organization systems by using the Resource Description Framework (RDF). Agricultural ontology consists of a concept hierarchy. Each concept has RDF attributes such as preferred terms, non-preferred terms of synonymic terms, hierarchical relationships, and associative relationships [11][12]. The SKOS model was expanded to formalize term relations such as prefer-of, nonprefer-of, subclass-of, superclass-of and related-of in the Chinese Agricultural Thesaurus (Fig.1). The expanded model defined four classes such as Subject, ConceptScheme, Concept and TopConcept, and six attributes such as skos:inScheme, skos: prefLabel, skos:altLabel, skos: broaderTransitive, skos: narrowerTransitive, skos: related, skos: memberof (Fig.2). The class Subject and attribute skos: memberof are the expanded model elements.

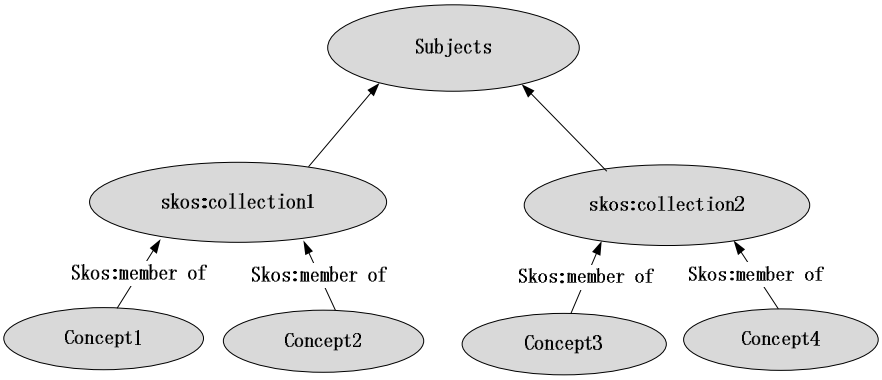

Fig. 1. Expanded SKOS Model

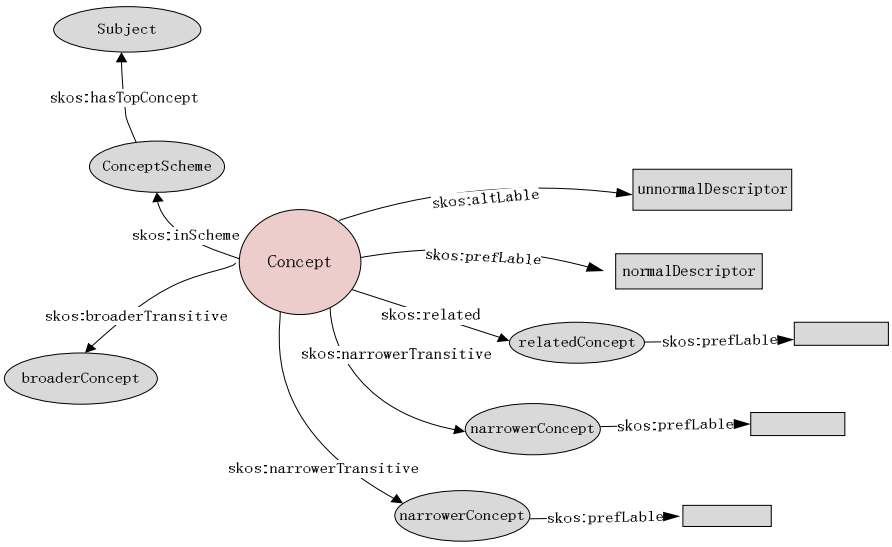

Fig. 2. SKOS Concept Model

Jena inference engine was integrated in Sonong system to identify implied relationships under concept hierarchy [13]. Also the consistency of the hierarchy was checked, for example, to verify cases of shared subclasses between different concepts.

3. System architecture

Fig.3 shows a three-component system architecture including information retrieval, reprocessing and indexing. It is the Sonong system implementing model.

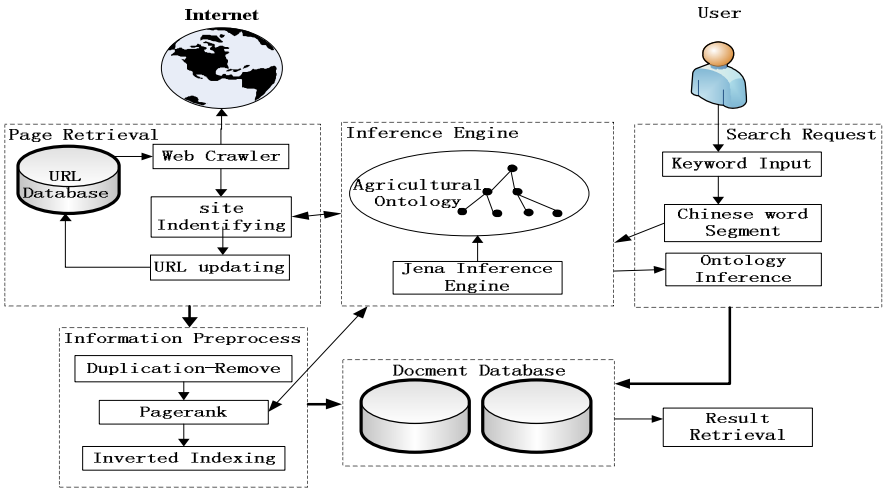

Fig. 3. Architecture of Sonong system

3.1. Webpage retrieval

Webpage retrieval module consists of crawlers, site theme identifier and URL database updating procedures with responsibilities for retrieving web pages from internet, identifying agricultural related pages and sites, filtering less related ones, and updating the URL database. When a page is retrieved, the identifier checks its theme, digs out URLs embedded, picks out less related themes and pages, saves related ones in document database, and updates the URL database with URLs dug out. This process keeps running at given intervals.

3.1.1. Distributed web crawlers. Pages on internet are captured by hyperlinks referring to each other. Following the link relations the crawlers collect pages from sites and databases automatically, and dig out fresh links from the collected pages. A distributed crawler set was deployed for page retrieval in Sonong system. Each has a URL queue. The process starts with an initial URL set. These URLs are allotted into each crawler's URL queue by the controller. When a page was downloaded, the crawler would dig out links in the page for URL database, and its queue would be filled with new URLs from database.

3.1.2 Site identifier. Site identifier serves as a theme filter with combination of agricultural ontology system. Ontology is applied to improve theme relevance algorithm by checking distances between terms. Site identifier parses web pages, analyzes URLs

and structure of the pages, evaluates the importance of the URLs, and filters pages by thresholds. It also conducts crawlers to their next destinations. Site identifier recognizes an agricultural site through its ratio of URLs remained.

3.2 Information preprocess

Information preprocess includes duplicates removing, PageRank calculating and inverted indexing.

3.2.1 Duplicates removing. The development of internet has resulted in the flooding of numerous copies of web documents. By eliminating the duplicate URL, clustering procedures and signature analyzing, the duplicate documents may be ignored. Sonong system adopted a method based on feature codes. Page text is identified by primary code and auxiliary code. The primary code is derived from paragraph structure while the auxiliary code from the content. The primary codes are clustered first, the auxiliary codes are matched. The algorithm has been proved efficient.

3.2.2 Page rank calculating. The most popular method to evaluate the importance of a page is PageRank algorithm which considers it a vote if a page is referred to by a link. Votes do mean importance. Meanwhile, the importance of the page which refers to is also taken into account. High PageRank of a page donates the voted page more importance. The Sonong system furthered this idea. The theme relevance factor is considered in PageRank calculating to result in more accurate ranking list.

3.2.3 Inverted indexing. Inverted indexing is well known efficient for full text indexing. The Sonong index system is founded on the apache Lucene program which is considered a perfect architecture for full text indexing. But Chinese is not support in Lucene. So the language parser and full text index tool kit of Lucene must be improved to support Chinese indexing functions.

3.3 Query interface

Query interface of Sonong system is a kind of semantic parser for request strings posted by users. It analyzes the posted strings with a Chinese word segment agent and generates a keyword set. Keywords generated are inferred with Jenna inference machinery in agriculture ontology represented in OWL. A hierarchy of keywords will be retrieved with terms related to keywords. This hierarchy helps users refine their requests effectively.

4. Key algorithms

4.1 Ontology oriented theme relevance algorithm

The site identifier evaluates the theme relevance of pages with ontology inference, and filters the pages according to the evaluation. Vector Space Model (VSM) is a well known model for theme relevance algorithms. To improve the accuracy of theme filter, the weight of related terms inferred from agriculture ontology was taken in account in calculation of eigenvector of page text.

The inferred terms are divided into two sets $keyword$ and $\overline{keyword}$. A term and its synonymies are elements of $keyword$ set, general terms, narrower terms, and related terms are contained in $\overline{keyword}$ set. The distance of a term from itself and its synonymies is 0, therefore the relevance is 1. Closer the distance is between two terms stronger relevance they are. Formula 1 shows this idea.

$$W = e^{-l(Dis(keyword,\overline{keyword}))} \quad (1)$$

Where W is the weight of terms, it is an exponential function, l is a predefined constant， $Dis(keyword,\overline{keyword})$ is the distance of terms。

So for arbitrary two terms from $\overline{keyword}$ and $keyword$, suppose $\forall x$ ， $x \in \overline{keyword}$ ， $y \in keyword$, the weight of x is:

$$W' = e^{-l(Dis(y,x))} \quad (2)$$

In general, terms in title of a document are more expressive than those in content. So the document is usually divided into two parts, and assigns title with higher weight account. Thereby for a given theme D, the relevance of a page with D can be calculated with formula 3.

$$sim(D,p) = a_T sim(D,T) + a_c sim(D,C) \quad (3)$$
$$= a_T \sum_{t \in T} e^{-l(Dis(k,t))} + a_C \sum_{c \in C} e^{-l(Dis(k,c))}$$

Where, D defines a theme, and $k \in D$. P is a page, and T, C is its title and content respectively, and terms $t \in T$, $c \in C$. a_T is weight of title, a_c is weight of content, and $a_T + a_c = 1$.

Considering the cardinality of *T* and *C*, the relevance should be calculated by formula 4, which is an application of reference 24.

$$sim(D,p)' = \frac{a_T \sum_{t \in T} e^{-l(Dis(k,t))} + a_C \sum_{c \in C} e^{-l(Dis(k,c))}}{a_T N_T + a_c N_c} \quad (4)$$

Where N_T is the cardinality of set *T*, N_C cardinality of set *C*.

4.2 Improvement of PageRak

The in-degree of a web page is an important indicator in evaluating its PageRank. General PageRank algorithms are basically static beyond any query requests. So PageRank of pages could be calculated in silent. This ensures efficiency of search and response. However, there are numerous improvements about general PageRank algorithm. The rank list of pages retrieved for users in Sonong system is based such an improvement that enables the relevance of linked pages being calculated in a PageRank evaluation. Suppose v is a page that links to page u, and B(u) is a collection of v. F(v) is a set contains all pages linked to v. w is an element of F(v). PR(v) is the PageRank of page v calculated by general PageRank algorithm. S(v) is the relevance between v and u, and can be calculated by formula 4.

In addition to the contribution of PageRank that v brings to u, the relevance of v and u would be taken into account as $PR(v) \times S(v)$. However, page v has a collection of pages viz. F(v) linked to, they also share contribution from v. Taken relevance between v and elements in F(v) as S(w), the actual contribution that v brings to u can be calculated by formula 5.

$$PR(u) = \frac{PR(v) \times S(v)}{\sum_{w \in F(v)} S(w)} \quad (5)$$

Considering all pages linked to u, PageRank of page u with respect to the relevance between pages can be calculated by formula 6.

$$PR(u)' = \sum_{v \in B(u)} \frac{PR(v) \times S(v)}{\sum_{w \in F(v)} S(w)} \quad (6)$$

5. Tests on algorithms and using scenarios of system

5.1 Tests on algorithms

Table 1 presents the test result of theme relevance algorithm. Under the experiences of pretest with sample data, a_T, a_C and l is set to 0.6, 0.4 and 0.7 respectively. Among 200 tested pages, half are selected agricultural document. The rest are retrieved randomly from internet by crawlers.

Table 1. Result of theme relevance test

Relevance	Pages	Relevance	Pages
0.9-1	13	0.4-0.5	17
0.8-0.9	49	0.3-0.4	20
0.7-0.8	35	0.2-0.3	39
0.6-0.7	12	0.1-0.2	6
0.5-0.6	7	0-0.1	2

It shows that there are 116 pages of which the relevance is more than 0.5. Personal check shows that there are 96 percent of the selected 100 agricultural pages of which the relevance is more than 0.5, 62 percent of which the relevance is more than 0.8.

The improved PageRank algorithm is tested with 17000 pages collected by crawlers. Taking theme “agriculture” as the query request, 631 pages retrieved under general PageRank algorithm, and 96 pages with improved algorithm. Rank list of retrieved pages shows pages with high PageRank have also high theme relevance.

5.2 Scenarios of system

It listed out all related terms inferred from ontology system and presented their relationship in a tree structure. Users can refine their query requests by navigating the hierarchy, and then post their refined request to Sonong sever, final search result with page rank list is presented in Fig.4 and Fig.5.

Fig. 4. Hierarchy retrieved

Fig. 5. Pages retrieved

6. Conclusion

Based on ontology and search engine development techniques, a theme relevance algorithm was proposed and applied in improvement of PageRank algorithm. A semantic search engine for agriculture was implemented and deployed for online services. The search engine provides semantic hierarchy inference with the ontology system and a satisfying ranking lists of retrieved information. Further research of agriculture semantic search engine will mainly concern the evolvement and refinement of agriculture ontology, and improvement of key algorithms.

References

1. J. H. Xiong, L. Xiao, et al, "The Situation and Evaluation of China Agriculture Information Website Development", Agriculture Network Information, 2006(2), pp. 4-7.
2. H. L. Liu, L.F. Guo, et al, "Design and Implementation of Chinese Focused Search Engine for Agriculture", Journal of Zhengzhou University (Natural Science Edition), 2007.39(2), pp. 74-77.
3. G.J. Xian, X.X. Meng and C. Chang, "The Design and Realization of Intelligent Retrieval Prototype System Based on Agricultural Ontology', Chinese Agricultural Science Bulletin, 2008. 24(6), pp. 470-474.
4. G. M. Zhou, J. C. Fan and Y. T. Zhou, "Design and Implementation of Chinese Agricultural Search Engine Based on SDD", Journal of Library and Information Sciences in Agriculture, 2008. 20(11), pp. 48-50.
5. P. Zhou, H. R. Wu, et al, "Research and design of agficulture search engine based on Nutch", Computer Engineering and Design, 2009.30(3), pp. 610-612.
6. S. John.F, Knowledge Representation: Logical, Philosophical, and Computational Foundations, 2000, New York: Brooks/Cole, 2000.
7. T.R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 1993, pp. 199-220.
8. N. Guarino, Formal Ontology and Information Systems, IOS Press: Amsterdam, 1998, pp. 3-15.
9. B. Smith, Basic Concepts of Formal Ontology in Formal Ontology in Information Systems, IOS Press: Amsterdam, 1998, pp. 19-28.
10. D. Nardi, and R.J. Brachman, An Introduction to Description Logic in the Description Logic Handbook, Cambridge University Press, 2002, pp. 17-19.
11. A. Miles, "SKOS: requirements for standardization", Proceedings of the 2006 international conference on Dublin Core and Metadata Applications, 2006, pp. 55-64.
12. V.A. Mark, M. Veronique, et al, "A method to convert thesauri to skos", The 3rd European Semantic Web Conference, 2006, pp. 95-109.
13. C. Jeremy J., D. Lan, et al, Jena: implementing the semantic web recommendations, in www'2004:Proceedings of the 13th International World Wide Web Conference, 2004: New York, pp. 74-83