Research on Intelligent Retrieval System for agricultural information resources based on ontology

Xiaojing Qin, Hui Zhang and Huaiguo Zheng*
Institute of Agricultural Information and Economics, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
Corresponding author’s e-mail: xiaojing_qin@163.com

Abstract. With the rapid development of computer technology and Internet, agricultural information resources are exploding. The traditional information retrieval based on keyword matching lacks semantic analysis ability and semantic extension ability, resulting in a large number of missed and false detection of retrieval resources and low retrieval efficiency. Based on the framework of Scientific and Technological Knowledge Organization (STKOS), this paper constructs the ontology of agricultural domain and applies domain ontology in semantic annotation, query extension and result sorting. On this basis, the paper constructs an intelligent retrieval system of agricultural information resources based on domain ontology, in order to improve the retrieval efficiency of agricultural information resources and improve user satisfaction.

1. Introduction

With the rapid development of computer technology and Internet, the network information resources are exploding. Information retrieval has become the main approach to gain knowledge out of the massive information resources and solve the problems. The essence of traditional keyword matching-based information retrieval is to use advanced algorithms to retrieve information from the Internet, establish the mapping relationship between keywords and information, and then sort the retrieval results by using algorithms. Users search and extract the knowledge used to solve practical problems from the retrieved results[1]. The traditional retrieval method is based on simple keyword literal matching, not the meaning or concept that keywords express. It severely separates the semantic association between words and words. In addition, Polysemy and synonym are widespread in natural language. The retrieval results simply return the web page containing the keywords. As a result, a large number of missed and misdiagnosed resources are retrieved. It is difficult for users to find resources closely related to their needs from the retrieved results, which cannot meet the needs of accurate retrieval and semantic retrieval in the new generation of semantic web environment[2-3]. The key to solve these problems is how to carry out semantic analysis and semantic extension of user's retrieval expressions, and how to improve the traditional keyword-weighted information retrieval methods to the knowledge-based and semantic-based level to achieve semantic retrieval[3].

Ontology, as a knowledge modeling tool, has attracted the attention of many researchers at home and abroad since it was proposed. Because ontology can well describe concepts and the relationship between concepts and concepts, and has a good concept hierarchy and support for logical reasoning, introducing ontology into information retrieval technology can realize the role of concept definition, query model and reasoning foundation, and improve recall and precision[4]. Therefore, the research of ontology based intelligent retrieval is of great significance. Because of the strong semantic
representation ability of ontology, scholars have carried out in-depth research on Ontology construction, ontology-based user query processing methods, semantic annotation and index, ontology-based information retrieval methods and models, and ontology-based intelligent retrieval system applications, and have achieved a lot of research results[5-6]. In the field of agriculture, Guo Jian Xian and Xian Jian Meng designed and implemented an intelligent retrieval prototype system based on the agricultural ontology transformed from the thesaurus of agricultural science, and achieved good retrieval results[7].

On the basis of previous studies, this paper constructs agricultural domain ontology based on the framework of Scientific and Technological Knowledge Organization (STKOS), and applies domain ontology in semantic annotation, query extension and result sorting. On this basis, the paper constructs an intelligent retrieval system of agricultural information resources based on domain ontology, in order to improve the retrieval efficiency of agricultural information resources and improve user satisfaction.

2. Ontology based intelligent retrieval system model for agricultural information resources

The agricultural information resource retrieval system, which is based on ontology, utilize the organizational advantages of agricultural information resources in the agricultural domain ontology as well as natural language processing technology and similarity algorithm to finally construct the intelligent information retrieval, improve the efficiency of agricultural information resource retrieval and improve user satisfaction. The retrieval system model is as follows:

An intelligent retrieval model of agricultural information resources based on domain ontology:

Step 1: Based on domain ontology, the system carries out semantic annotation of domain information resources and builds index library.

Step 2: The user input query content, the system based on the Ontology vocabulary and stop vocabulary word processing word.

Step 3: The system carries on the ontology concept matching to the segmentation result, and divides the matching result into the ontology concept set and the non-ontology concept set. Through the keywords extracted from the literature and the synonymous relation words in the domain ontology, the two sets are expanded and the query set is generated. In the process of user input, the retrieval
suggestion module of Solr is used to match the retrieval words with the index library in real time, which provides faster retrieval recommendation for users.

Step 4: The system submits the query set to the index library, selects Top-N as the query result, and presents the result to the user via intelligent sorting.

3. Key technologies of intelligent retrieval system.

3.1 Building the Ontology

Based on the framework of Scientific and Technological Knowledge Organization (STKOS), agricultural domain ontology knowledge base combs, selects and classifies the core concepts of STKOS on agricultural domain vocabulary. At the same time, according to the mapping rules of Thesaurus transformation, this paper refers to the AGROVOC, forestry domain ontology, crop cultivation domain ontology, fishery domain ontology, agricultural scientific thesaurus, agricultural thesaurus and other fields ontology or glossary, and unifies the relevant field researcher's guidance. Finally, the system realizes the organization of the core concept of agriculture and the description of key information sources.

The knowledge in the agricultural domain ontology is organized around the category. The category is a tree structure. There are 1,854 categories, including 5 top categories: agricultural basic science / Agronomy / Forestry Science / Aquatic products, fisheries, hunting / Animal husbandry science. Concept is the basic unit of describing knowledge. The same concept can belong to many categories, and the concepts have subordinate and subordinate relations. There are 70,232 concepts, 86,421 relationships between concepts and categories, and 53,403 epistemic relations among concepts. The same concept can also have multiple synonyms and multiple concept definitions. At present, there are 190,408 synonyms and 13,365 definitions in the system.

Figure 2. Agricultural domain ontology

3.2 Semantic indexing

Semantic annotation and indexing are carried out by using domain ontology concept. It includes the pretreatment of documents, the extraction of semantic vectors, and the establishment of indexes.

Document pre-processing: Based on the existing agricultural vocabulary, a keyword extraction model has been built by means of semi-supervised learning keyword extraction. Using the trained model to extract keywords from the document and through manual audit, complete the extraction model and form a high-quality core word set.

Semantic vector extraction: The cosine similarity of vectors is used to represent the semantic similarity between words. The pre-processed core vocabulary sets are matched with the concepts in domain ontology to form a mapping relationship between documents and domain ontology. When the matching is unsuccessful, fuzzy matching is performed.

Finally, the relationship between data and data is submitted to Solr, and the document is indexed and stored in a relational database for retrieval.
3.3 Semantic retrieval
Based on Solr full-text index server, this paper establishes an intelligent retrieval system for agricultural information resources by configuring domain lexicon and ontology library. Solr is a top-level open source project under Apache, developed in Java, and is a full-text search server based on lucene. Solr provides a rich query language and configurable, scalable and other features to achieve a simple and rapid retrieval of resources.

3.3.1 Search terms segmentation
This retrieval system uses IK Analyzer word segmentation as the basis to process the user input keywords. By adding the stop word library and matching domain ontology library, the corresponding ontology concepts of the retrieval words are found, and the domain ontology concepts set are formed. The unsuccessful words are not matched to form the non-ontology concepts set.

3.3.2 Extension of search terms
By collecting the key words in the domain literature, we can further improve, supplement and update the domain vocabulary and form the domain extended lexicon. At the same time, we can expand the ontology concept set and non-ontology concept set after the segmentation by combining the synonym concept in the domain ontology. Based on the two extended sets, the system matches the index information stored in the index library to obtain a list of query results related to the search terms.

3.3.3 Retrieval suggestion
Through the real-time acquisition of user input keywords, the system based on Solr's Suggest module and semantic index library, will be closely related to the search terms of the resource title directly displayed to the user, so that users can quickly find the relevant information content they need.

3.3.4 Sorting of retrieval results
According to the results of the click, the importance of resources, the extension of the inspection resources and the location of matching search terms, different weights are given to the retrieval results, and the final retrieval results are sorted according to the weights.

Acknowledgement
The funding of this research is funded by the Youth Science Foundation of Beijing Academy of Agriculture and Forestry Sciences QNJJ201617.
References

[1] Tan Xiaobo, Zhai Xiapu. (2018) Research on knowledge retrieval based on ontology knowledge set Research on Library Science, 1: 60-66.

[2] Meng Hongwei, Zhang Zhiping, Zhang Xiaodan. (2013) Research on Intelligent Information Retrieval Model Based on Domain Ontology. Journal Of Intelligence, 9: 180-184.

[3] Zheng weiqing. (2013) Automatic Semantic Retrieval and Visualization Model Based on Ontology Integration. Information Science, 05(31): 77-83.

[4] Yang YH, Du JP, Ping Y. (2015) Ontology-Based intelligent information retrieval system. Ruan Jian Xue Bao/Journal of Software, 26(7): 1675–1687.

[5] Morente Molinera J A, Pérez I J, et al. (2016) Creating Knowledge Databases for Storing and Sharing People Knowledge Automatically Using Group Decision Making and Fuzzy Ontologies Information Sciences, C: 418−434.

[6] Castells P, Fernandez M, Vallet D. (2007) An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval .Knowledge and Data Engineering, 19( 2) : 261−272.

[7] Xian Guojian, Meng Xianxue, Chang Chun. (2008) The Design and Realization of Intelligent Retrieval Prototype System Based on Agricultural Ontology. Chinese Agricultural Science, 6(24): 470-474.