Agricultural Library Information Retrieval Based on Improved Semantic Algorithm

Xie Meiling

Library of Agriculture University of Hebei, Baoding, China
hebauxie@hotmail.com

Abstract. To support users to quickly access information they need from the agricultural library's vast information and to improve the low intelligence query service, a model for intelligent library information retrieval was constructed. The semantic web mode was introduced and the information retrieval framework was designed. The model structure consisted of three parts: Information data integration, user interface and information retrieval match. The key method supporting retrieval was designed. The traditional semantic similarity algorithm was improved according to its shortages. An algorithm based on semantic distance was designed and tested. The results can improve the recall ratio and precision of information retrieval, improving information retrieval performance.

Keywords: Improved semantic algorithm, Agricultural library, Information retrieval.

1 Introduction

Agricultural digital library was the carrier of agricultural information resources. The current digital library retrieval query service there are some disadvantages, including low query intelligence, independent results and low degree of shared information data. The traditional pattern of information query was based on the keyword query. This kind of query can't reflect the deep meaning of user query demand[1]. Semantic web query mode was a hot research topic in recent years. In semantic query mode, user's semantics of natural language can be understood by the machine to a certain extent and the semantic level of retrieval was implemented finally. How to build the information query model based on semantic technology was a subject to be solved in which the most key problem was the semantic similarity computation. Traditional semantic similarity algorithms include algorithm based on semantic distance, algorithm based on the concept features and algorithm based on the amount of information. The disadvantage of the above algorithms was easy to lead to errors[2]. Some other researchers studied the information retrieval based on semantic retrieval[3-6]. In this paper, aiming at the shortcomings of the similarity calculation in traditional semantic algorithm, the traditional semantic algorithm was improved. Using the similarity algorithm based on semantic distance, the model of information query based on semantic technology was built and information retrieval precision was increased.

D. Li and Y. Chen (Eds.): CCTA 2014, IFIP AICT 452, pp. 589–594, 2015.
DOI: 10.1007/978-3-319-19620-6_65

2 Information Query Architecture and Improved Similarity Algorithm

2.1 Information Query Architecture of Digital Library

In the human-computer interaction environment, the information resources in the digital library can be combined with the user semantic body through the semantic web technology. The specific meaning of the language user used in specific environment can be accurately expressed. Using standardized semantics, understanding between the user and the system was implemented and user information demand was accurately obtained from the perspective of semantics. And finally the information the user needed was obtained through information retrieval. Digital library retrieval model based on the semantic technology was constructed, as shown in the figure below:

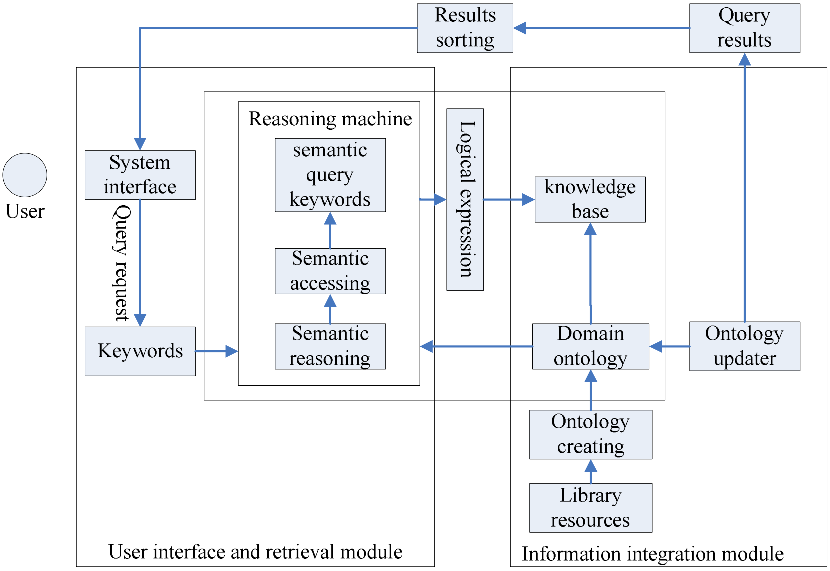

Fig. 1. The model of information retrieval based on semantic web technology

The figure 1 showed that in the semantic web technology, the retrieval model was divided into three modules.

(1) User Interface and Retrieval Processing Module
Using semantic web technology, the module processed query conditions submitted by the user in the interface based on human natural language. Query keywords was pretreated and converted to ontology query. According to the domain knowledge ontology, natural language query submitted by the user was resolved by the reasoning machine. Semantic reasoning based on semantic similarity calculation was executed.

(2) Information Integration Module
According to the resources in the semantic web annotation, this module created the ontology model for the information of the digital library. And domain ontology was formed According to the evolution of the information, ontology updater was used to expand and refresh the domain ontology periodically. Using XML technology, the information be standardized processed. Metadata information with high relevance to user requirements was acquired and stored in knowledge base.

(3) Match and Output Module
According to the search keywords pretreated, this module searched results in the domain ontology knowledge base. Semantic reasoning was conducted based on ontology by reasoning machine unit according to the search keywords. Logical expression of user query and retrieval was constructed and submitted. The result set complied with the expression was searched out by the system. The result was sorted using semantic similarity algorithm. Then the system returns the user interface.

With the support of these modules, the user retrieval process can be described as:

(1) key information in the sources of was acquired. Using the semantic web technology such as XML (Extensible Markup Language), RDF (Resource Description Framework), these key information was standardized processed according the metadata such as MARC(Machine Readable Catalog) ,DC (Dublin Core Element Set). The key information was converted into metadata and was stored in metadata database.
(2) According to the data in the metadata database, the relationship between the concepts were derived using logical reasoning. Domain ontology was constructed and stored in the knowledge base.
(3) Search conditions entered by the user were read. Based on the domain ontology constructed in (2), user's search conditions were converted into standard formats. Semantic extraction operations were conducted according to semantic similarity. Preliminary treatment of user query needs was finished.
(4) Through user query needs to build query expressions, system knowledge base was traversed. Matching information collection was searched out. All elements in the set were screened using domain Ontology-based semantic similarity. Query results are sorted and displayed according to semantic similarity.

3 Improved Semantic Similarity Algorithm

Key of library information retrieval was to determine the semantic similarity between concepts. The higher the degree of similarity, indicating the query results returned to the user and the user query needs more match. Semantic similarity algorithm based on distance described the connections between the concepts using hierarchical network. The geometric distance of the concept in the hierarchical network was as semantic distance. The traditional method to measure the similarity between the concepts was:

$$similar(v_1, v_2) = \frac{2(len-1) - dis(v_1, v_2)}{2(len-1)} \qquad (1)$$

Where len was the depth of the hierarchical network possessed. $dis(v_1, v_2)$ was the directed edge number contained in the minimum value of paths between the two concepts. The traditional similarity algorithm based on semantic distance had some disadvantages. Only the lengths of the paths between the nodes were paid attention. Semantic distance effects on similarity were ignored. Node hierarchy of semantic effect was not considered. These led to the difference between the calculated results and the actual situation[7]. Ontology hierarchy in the following as an example:

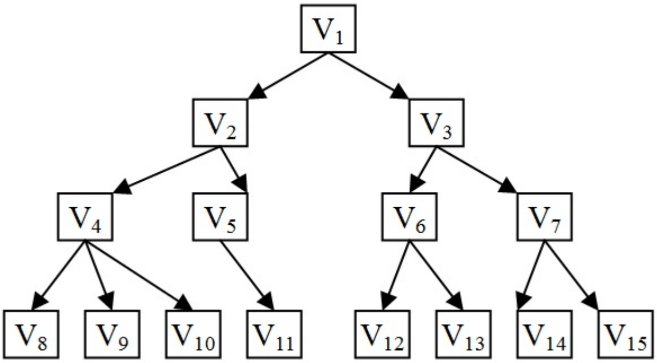

Fig. 2. Ontology hierarchy chart example

Assuming the instance belonged to the complete structure of domain ontology, the semantic similarities between nodes v6 and v12, v6 and v3 were computed respectively using traditional semantic similarity algorithm. The results were both 0.83. In the complete structure of domain ontology, the closer it got to the ontology layer, the lower the semantic distance between the elements values should be. So the truth should be the semantic similarity between the node v6 and node v12 was greater than the semantic similarity between the node v6 and node v3. The traditional method did not reflect the actual situation and needed to be improved.

In order to overcome the shortage of the traditional methods, the improved semantic distance similarity algorithm was designed in this paper. Semantic depth as an important factor was introduced to the semantic distance similarity calculation. If the number of connected path between two nodes were equal, the semantic similarity between nodes and ontology hierarchy were related. The higher the sum of two conceptual levels, the greater the similarity was. The greater the difference between the hierarchies of the two concepts, the smaller the similarity was. This can be represented by the following expression:

$$\frac{|\,depth(v_1) - depth(v_2)\,|}{depth(v_1) + depth(v_2)} \tag{2}$$

This expression was a reduction function. Therefore, the depth of the semantic ontology nodes was expressed by α:

$$J \tag{3}$$

Where $depth(v_1)$ and $depth(v_2)$ were the numbers of layers (depth) of node v1 and node v2 respectively. The depth of the root node of the ontology hierarchy network can be regarded as the value 1. Along with the increase of the layer, the depth was also increasing. According to (3), the semantic similarity between nodes v6 and v12 was 0.86 and the semantic similarity between nodes v6 and v3 was 0.80.

The position of the node in the structure should be considered in semantic similarity calculation based on the semantic distance. Generally, if the two nodes belong to parent-child relationship, the parent node contained all the properties of the child nodes and the child nodes contained some properties of the parent node. Similarity between high-level and low-level nodes should be lower than the similarity between low-level and high-level nodes. J was defined as a status factors of the node in the network hierarchy, expressed as:

$$j = \frac{1}{2}(1 + \frac{depth(v_1) - depth(v_2)}{len_{\max}}) \tag{4}$$

Where $len_{\max}$ was the max depth of the hierarchical network. The semantic similarities between nodes v6 and v3, v3 and v6 were computed according to (4) and the results were 0.63 and 0.38 respectively. Based on the above analysis shows that the semantic similarity between the nodes affected by the 3 kinds of properties: path length, semantic depth and node hierarchy. Namely:

$$similar(v_1, v_2) = (\frac{2(len-1) - dis(v_1, v_2)}{2(len-1)} * i * j)^{\frac{1}{2}} \tag{5}$$

4 Results and Discussion

Using the improved algorithm and traditional algorithm to calculate the similarity between different nodes respectively, the results were shown in table 1.

Table 1. Composition of the experimental samples

	Traditional algorithm	i	j	Improved algorithm
semantic similarity between v12 and v6	0.83	0.86	0.63	0.67
semantic similarity between v6 and v3	0.83	0.79	0.63	0.65
semantic similarity between v3 and v6	0.83	0.79	0.38	0.49

The chart showed that using the traditional semantic similarity calculation method to calculate the similarity between different nodes, the results were the same and did not reflect the actual situation. Table 1 showed that using the algorithm presented in this paper, the semantic similarity(0.67) between node v12 and v6 was higher than the semantic similarity(0.65) between the node v6 and node v3. The semantic similarity(0.65) between node v6 and v3 was higher than the semantic similarity(0.49) between the node v3 and node v6.

The semantic similarity between node v12 and v6 was the largest, followed by the semantic similarity between node v6 and v3. The semantic similarity between node v3 and v6 was the smallest. This was consistent with the actual situation. This showed that the optimization algorithm designed in this paper can more accurately reflect the actual situation.

5 Conclusions

With the improvement of the level of agricultural information, digital libraries should also improve their level of query to provide users with information services in this trend. .A model for intelligent library information retrieval was constructed. The semantic web mode was introduced and the information retrieval framework was designed. The model structure consisted of three parts: Information data integration, user interface and information retrieval match. The key method supporting retrieval was designed. The traditional semantic similarity algorithm was improved according to its shortages. An algorithm based on semantic distance was designed and tested. This showed that the optimization algorithm designed in this paper can more accurately reflect the actual situation. There were no algorithm can fully realize the semantic similarity search. Therefore, this study also needs further improvement.

References

1. Sheng, X.: Knowledge organization of digital library. Library and Information Service (3), 26–29 (2011) (in Chinese)
2. Qin, J.: Semantic Web and Ontologies. In: 2nd Joint Adcanced Workshop in Digital Libraries, Beijing (May 2012)
3. Shi, X., Niu, Z., Song, H., et al.: Intelligent agent-based system for digital library information retrieval. Journal of Beijing Institute of Technology 12(4), 450–454 (2003)
4. Wang, J., Yang, X.: A distributed cooperative approach to Web information retrieval using metadata and Z38.50. Journal of Software (4), 620–627 (2001)
5. Li, H., Wu, W., Zhang, C.: Implementation research on information retrieval based on semantic. Journal of Gansu Lianhe University (Natural Sciences) 22(2), 86–89 (2008)
6. Klyuev, V., Oleshchuk, V.: Semantic Retrieval of Text Documents. In: Proceedings of the 7th IEEE International Conference on Computer and Information Technology, pp. 189–193 (2007)
7. Li, S.: Research of Relevancy between Sentences Based on Semantic Computation. Computer Engineering and Applications 38(7), 75–76 (2012)