Research on Semantic Information Retrieval Model of Bamboo & Rattan Domain Based on Query Extension

Lin Peng, Ming-ming Lai, Xin Zhang
College of Big Data, Yunnan Agricultural University, Kunming 650201
TContact: Lin Peng (penglin2286351@163.com)

Abstract: The existing domain semantic information retrieval models generally do not consider the wide range of domain knowledge involved, and lack the ability to query multi-level domain knowledge. Based on the semantic expression of the knowledge about bamboo & rattan via ontology, this paper mainly studies the measurement of semantic relevancy in the field of bamboo and rattan, and finally proposes a Semantic Information Retrieval Model in this field which is based on relevancy. Experimental results show that the semantic query extension based on the concept of similarity proposed in this paper can improve the recall rate and F value of retrieval model significantly.

1. Introduction
At present, a large number of data related to bamboo & rattan have been digitized, laying a foundation for rapid identification and semantic retrieval of bamboo & rattan [1-4]. However, the existing semantic information retrieval models don’t take into account the wide range of domain knowledge involved in bamboo & rattan domain knowledge and lack the ability to query multi-level domain knowledge. Based on the construction of the database of bamboo and rattan ontology, this paper introduces the query extension technology, combines the semantic characteristics of the knowledge of bamboo and rattan domain, and proposes a query extension based semantic information retrieval model of bamboo and rattan domain.

2. Query expansions
In order to improve retrieval Recall, the method of re-query by recombining original query conditions and adding new keywords and it is called query extension. This method can assist users to generate new query keywords, effectively improve the recall rate of retrieval, and better understand the actual retrieval intention of users. The core of this method is the automatic generation algorithm of new query words. At present, query extension is generally realized through relevant feedback technology, semantic word derivation technology, keyword co-occurrence and ontology technology. Based on ontology hierarchy, this paper depends on similarity of ontology concepts to realize query extension in semantic retrieval of bamboo & rattan domain.

2.1. Semantic query extension
In the query extension method of this paper, the query extension based on concept similarity in ontology is adopted. The greatest difference between the method and the similar word extension method in the early days is that the generation of extended vocabulary is no longer dependent on the query document collection itself, but based on the domain knowledge ontology, and the extension content is independent
of the query content. At the same time, because of the relationship between concepts in the same domain is relatively stable, extended vocabulary can be generated more accurately. Meanwhile, hierarchical conceptual structure in ontology can effectively avoid word meaning confusion.

When semantic correlation is extended by using hierarchical relations such as the composition and inclusion of ontology concepts, the extended pattern directly affects the recall rate and precision rate about the retrieval results. When the extension word is the combination of the hyponym, the homonym and the partite of the query word, it can improve the recall rate and has little influence on the precision rate. The same effect can be achieved through the extension of the upper word, while other combinations of extended relationships, including sibling relationships, have a more damaging effect on accuracy than on recall. Therefore, semantic query extension based on concepts in ontology is carried out according to bamboo-rattan domain ontology, mainly including sub-words and overall word extension of query words according to concepts. The concrete implementation is shown in table 1.

| NO. | Query word | Conceptual semantic query expanded                                      |
|-----|------------|------------------------------------------------------------------------|
| 1   | leaf       | Leaf sheath, ear, tongue, leaf blade, petiole                          |
| 2   | glumelle   | Stamen, pistil, lemma, lemma, scale coat, ovary, stigma, anther        |
| 3   | pole       | Stem, stem base, internodes, handle Tuo, section, section, section, internodes tip |
| 4   | tongue     | Ligule, Tuo tongue                                                     |
| 5   | sheath     | Leaf sheath, Tuo sheath                                                |
| 6   | flower     | Stamen, pistil, lemma, lemma, scale coat, ovary, stigma, anther        |
| 7   | Podozamites| Bamboo, stalk, sheath, ear, leaf, tongue, stem and hair                |

2.2. Conceptual similarity calculation

This paper uses the Conceptual similarity before and after semantic extension to evaluate the impact of the introduction of extended search terms on the precision.

Definition 1 mainly introduce the relationship between the initial retrieval concept words and the superordinate concepts obtained after semantic query extension Is -a and member-of., and the degree of conformity between these extended concepts and the initial retrieval concept in the retrieval semantic theme is called the concept similarity before and after extended retrieval.

Conceptual similarity reflects the degree of semantic similarity of two concepts in ontology, that is, semantic consistency in retrieval of two concepts in retrieval. As the initial concept of retrieval words c0 stands for the hyponym or ci stands for meronym. Although there are some differences in concept similarity, but if from the point of user retrieval intention, these words are on the original concept semantic specialized retrieval. Therefore, retrieval results which include these extended concepts are conform to the requirements of the concept of the retrieval results, and will not reduce the retrieval accuracy. For example, the initial retrieval concept is "Melocanna", and the text has the content of "pear bamboo". Since the concept of "pear bamboo" is hyponym of "Melocanna", although there are differences in conceptual semantics between them, from the perspective of users' retrieval intention, this result is completely consistent with the requirements.

Therefore, the method of conceptual similarity calculation based on semantic query extension is as follows:

$$SR(c_i, c_o) = \begin{cases} \text{length}(c_o, c_i), & \text{otherwise} \\ \frac{1}{2^{5+S}} & , c_i \in K \end{cases}$$  \hspace{1cm} (1)
\[ K = \{ k | k = c_o \lor k \} \]

Where, \( \text{length}(c_o, c_i) \) is the shortest path from node \( c_o \) to node \( c_i \), \( S \) is the depth of ontology tree, and \( k \) is a hyponym or meronym of \( c_o \).

2.3. Query extension retrieval method

In order to solve the limits that the traditional information retrieval technology relies on keyword matching effectively, if the matching fails, empty records will be returned. Users need to further change the query conditions and try again. The system does not have the function of automatic reasoning. By calculating the similarity of concepts in the domain ontology of bamboo and rattan, this paper extends the user's retrieval concepts semantically, mainly realizing the expansion of the semantic relationship of query words (the equivalence relation or the hyponymy). After query extension, the corresponding equivalent semantic expressions, upper semantic expression and lower semantic expression are obtained. In the concrete implementation, it is expressed in the form of regular expression, which provides convenience for ontology reasoning in the next step.

Semantic retrieval after query expansion is mainly to retrieve concepts and their relationships at the semantic level. The key to its realization is reasoning between concepts. Jena inference machine is used in this paper for ontology inference. Within, Jena provides three ways: (1) adopting the reason according to add regularly (including RDFSReasoner, OWLReasoner, etc.), which includes common reasoning functions; (2) users can customize inference rules as required; (3) registration of third-party inference engines. In this paper, a custom rule inference engine is used in the experiment, and its trigger mechanism is a forward chain engine.

3. Experimental results and analysis

In this paper, Volume 9 of Flora of China is used as experimental data set, which contains 37 genera, 516 species of bamboo, 516 documents and 530,000 words. Through experiments, this paper analyzes the influence of the semantic retrieval model based on semantic query extension proposed in this paper on the improvement of retrieval recall rate in the field of bamboo and rattan.

First, select the "leaf", "rhizome", "rod", "flower" and "bamboo Tuo" five concepts for the primitive retrieval concept. These five retrieval concepts were selected because of the five iconic retrieval concepts frequently used in the rapid retrieval of bamboos in the Retrieval Table of Chinese Bamboos Subfamily Genera [5]. By calculating the concept similarity of the original retrieval concept and extending semantic query, the extended retrieval concept is obtained in the end, as shown in table 2.

To this field of bamboo & rattan semantic retrieval model, the semantic query expansion model retrieval effect is compared before and after the experiment of "leaf", "rhizome" and "pole", "flower" and "bamboo Tuo" five concept semantic extension before and after the experiment, the experimental results as shown in table 3.

| NO. | Retrieve the concept | Semantic extension | Retrieve instance |
|-----|---------------------|-------------------|------------------|
| 1   | leaf                | Leaf sheath, ear, tongue, leaf blade, petiole | lanceolate       |
| 2   | subterranean stem   | subterranean stem, False whip, bamboo whip | solid core       |
| 3   | pole                | Stem, stem base, internodes, handle Tuo, section, section, internodes tip | light            |
| 4   | flower              | Stamen, pistil, lemma, lemma, scale coat, ovary, stigma, anther | yellow           |
| 5   | Bamboo Tuo          | Bamboo Tuo, stalk Tuo, Tuo sheath, Tuo ear, | Unapparent       |
Tuo leaf, Tuo tongue, rod Tuo, wool

Table 3. Retrieval results before and after semantic query expansion.

| NO. | Semantic query extension before | Semantic query expanded | Total number of relevant results |
|-----|---------------------------------|-------------------------|--------------------------------|
|     | Total search results | Correct retrieval number | Total search results | Correct retrieval number |
| 1   | 875               | 712                    | 917               | 763                    | 878               |
| 2   | 103               | 87                     | 132               | 108                    | 127               |
| 3   | 47                | 36                     | 50                | 38                     | 43                |
| 4   | 631               | 489                    | 657               | 507                    | 608               |
| 5   | 289               | 231                    | 320               | 254                    | 304               |

In this paper, three evaluation indexes (accuracy, recall rate and F value) are selected as evaluation indexes to evaluate the influence of semantic query expansion on the performance of the model. The calculation formula is as follows:

\[
\text{Precision Rate: } P = \frac{\text{Correct Retrieval Number}}{\text{Total Search Results}} \times 100\% \tag{2}
\]

\[
\text{Recalling Rate: } R = \frac{\text{Correct Retrieval Number}}{\text{Total Number of Relevant Results}} \times 100\% \tag{3}
\]

\[
\text{F Value: } \quad F = \frac{2 \times P \times R}{P + R} \times 100\% \tag{4}
\]

The retrieval result data before and after semantic query expansion in table 2 were analyzed and calculated. The accuracy, recall rate and F value of model retrieval effect from the before semantic query extension and the after semantic query extension were obtained respectively, as shown in table 4. As can be seen from the data in table 4, the retrieval accuracy of the model was not significantly affected before and after semantic query expansion, but the recall rate and F value were both improved to some extent. Because, in this paper, under the condition of reasonable ontology construction in bamboo & rattan domain, the retrieval method based on semantic extension of concept similarity can improve the retrieval performance of the model by expanding the retrieval scope and increasing the number of correct retrieval results without affecting the accuracy of model retrieval (to make sure the user’s search intention remains unchanged to the greatest extent).

Table 4. Impact of semantic query extension on model performance.

| NO. | Semantic query extension before | Semantic query expanded | F Value |
|-----|---------------------------------|-------------------------|---------|
|     | Precision Rate | Recalling Rate | F Value | Precision Rate | Recalling Rate | F Value |
| 1   | 81.37%           | 81.09%             | 81.23% | 83.21%       | 86.90%        | 85.01% |
| 2   | 84.47%           | 68.50%             | 75.65% | 81.82%       | 85.04%        | 83.40% |
| 3   | 76.60%           | 83.72%             | 80.00% | 76.00%       | 88.37%        | 81.72% |
| 4   | 77.50%           | 80.43%             | 78.93% | 77.17%       | 83.39%        | 80.16% |
| 5   | 79.93%           | 75.99%             | 77.91% | 79.38%       | 83.55%        | 81.41% |

4. Conclusion

On the basis of ontology semantic expression of bamboo & rattan domain knowledge, this paper studies the measurement of semantic relevance of text data in bamboo & rattan domain information, and proposes a semantic information retrieval model based on correlation. According to the semantic correlation in concepts of bamboo & rattan domain, the retrieval model measures the importance of search terms in bamboo and rattan domain ontology, and expands the scope and depth of retrieval. The experimental results show that the proposed semantic query expansion based on concept similarity has
little effect on the retrieval accuracy of the model, but the recall rate and F value have been greatly improved.

Acknowledgments
The findings and the opinions are partially supported by projects of the major science and technology projects of Yunnan province, No.2018ZJ001. This work is supported by Engineering and Technology Research Center for Yunnan Agricultural Big Data and Yunnan Agricultural University, China.

References
[1] Zhang Xinpeng. Development Prospects of Bamboo and Rattan in the world[J]. World Forestry Research, 2003. 16(1):26-30.
[2] Qiu Erfa, Hong Wei, Zheng Yushang. Review on Diversity and Utilization of Bamboo in China[J]. JOURNAL OF BAMBOO RESEARCH, 2001, 20(2):11-14.
[3] Chinese flora editorial committee. Flora of China (volume 9, volume 1) [M] Beijing: Sciences Press, 1996.
[4] Jiang Zehui. The world's bamboo rattan(English edition) [M]. Beijing: China Forestry Publishing House., 2007.
[5] Yi Tongpei, Ma Lisha, Shi Junyi. Chinese bamboo subfamily genus search[M]. Science Press, 2009.
[6] Zhang Junlin. Research on Information Retrieval System Based on Language Model[D]. Doctoral Dissertation of Institute of Software, Chinese Academy of Sciences, 2004.
[7] Manning C. D. write. Yuan Chunfa translate. Basic Statistics of Natural Language Processing[M]. Publishing House of Electronics Industry, 2005.
[8] Zhen Tianqiao. Research of the Mixed Language Model Based on Statistical and Latent Semantic Analysis[D]. Master's Thesis of Harbin Institute of technology, 2007.
[9] Li Zheng. Research on Chinese Web Page Information Retrieval Based on Statistical Language Model[D]. Master's Thesis of Central China Normal University, 2012.
[10] (Chile) Ricardo Baeza-Yates write, Huang Xuanjing translate. Modern information retrieval (2nd edition) [M]. China Machine Press, 2012.
[11] Gao Shan. Research on Query Wxtensoin and Related Technology in Information Retrieval [D]. Master's Thesis of Central China Normal University, 2008.
[12] Wang Qi. Semantic retrieval technology and application of ancient fresco [D]. Doctoral Dissertation of Zhejiang University, 2011.
[13] Han Xin. Research on resource semantic retrieval in bioinformatics based on ontology [D]. Master's Thesis of North University of China, 2012.