Research on Semantic Information Retrieval Model of Bamboo Rattan Domain Based on Semantic Relevance

Lin Peng, Xu-peng Kou, Lin-nan Yang

College of Big Data, Yunnan Agricultural University, Kunming 650201

TCorrespondence to: Lin-nan Yang (lny5400@sina.com)

Abstract: At present, the research of semantic information retrieval technology is mostly in the initial stage, and the research is mostly in exploratory theoretical research. Among them, the construction of domain semantic information retrieval model is still immature. In this paper, bamboo rattan field was selected as the research object, and the key problems in semantic information retrieval technology of bamboo rattan domain were analyzed. Based on the domain ontology of bamboo and rattan, the weight calculation method of bamboo rattan domain terms is defined. This paper introduces semantic relevance degree extension into the domain semantic information retrieval model of bamboo rattan, constructs a domain semantic information retrieval model suitable for bamboo rattan, and realizes the association retrieval and semantic query extension of text information in bamboo rattan domain.

1. Introduction

Bamboo rattan is the combined name of bamboo and rattan plants. It is a large family in the plant kingdom, which plays an important role in the forest resources of the world. At present, a large number of data related to bamboo rattan have been digitized, which lays a foundation[1-5] for rapid identification and semantic retrieval of bamboo rattan. Semantic information retrieval, however, the comprehensive realization of bamboo rattan areas, especially in the semantic information retrieval of textual data, also need to solve the following questions:

(1) The representation and measurement of knowledge in bamboo rattan domain. In the research of bamboo rattan information, there are uncertainties and inaccuracies in text information. How to use artificial intelligence technology to effectively represent the knowledge of bamboo rattan field is the basis for realizing semantic retrieval of bamboo rattan information. (2) The correlation retrieval of text messages in the bamboo fields. It mainly need to achieve the following two levels of association retrieval: According to the semantic information of bamboo rattan text, the identification and retrieval of related bamboo rattan are realized. Taking literature as a bridge, it realizes the retrieval from bamboo rattan to related literature, and then from literature to bamboo rattan. (3) The semantic query expansion problem. At present, semantic information retrieval technology does not take into account the wide range of domain knowledge involved in bamboo and rattan domain knowledge and lacks the ability to query multi-level domain knowledge. At the same time, there are some limitations in using ontology to realize semantic inspection. How to effectively improve the retrieval effect based on the existing ontology technology is the key to the design of semantic information retrieval model in the field of bamboo rattan.

In view of the above problems, based on the construction of bamboo rattan ontology, this paper defines the weight calculation method of bamboo rattan domain terms and introduces the semantic correlation measure. Combined with the semantic characteristics of knowledge in the field of bamboo...
rattan, a semantic information retrieval model suitable for the field of bamboo rattan was proposed, and the classification of bamboo rattan based on text-based data was realized.

2. Weight Calculation of Terms in the Field of Bamboo Rattan

2.1. Bamboo Rattan Domain Term Weight Definition

This paper uses the TF-IDF algorithm for reference to define and calculate the domain term weight. The idea is that: When the species of bamboo and rattan are identified according to the search term set, the domain document set is set as *Flora of China*. If the frequency of a certain domain term appears in the descriptive document of one kind of bamboo rattan, the greater the identification degree of the domain term to the bamboo rattan document, the higher the weight of the domain term will be. At the same time, if the domain term appears less frequently in all other description documents of bamboo rattan in *Flora of China*, the greater the identification degree of the domain term to bamboo rattan document, the higher the weight of the domain term will be.

2.2. Weight Calculation of Bamboo Rattan Domain Terms

According to the TF-IDF algorithm, the weight calculation process of bamboo and rattan domain terms. Therefore, the weight calculation formula of terms in the field of bamboo rattan is as follows:

\[
sweight[i, j] = tf_{i, j} \times idf_j = \frac{n_{i, j}}{\sum_k n_{k, j}} \times \log \frac{N}{n_i}
\]

where \( sweight \) represents the weight of domain terms; \( tf_{i, j} \) represents TF value and is the frequency of domain term \( I \) in document \( j \); \( idf_j \) represents IDF value and is the frequency of reverse documents in all documents of the domain term \( I \); \( n_{i, j} \) represents the number of times the domain term \( I \) appears in document \( j \); \( N \) is the total number of documents in the domain document set; \( n_i \) is the number of documents containing the domain term \( I \).

3. Semantic correlation degree calculation

This paper takes the ninth volume of *Flora of China* as corpus document, in which each kind of bamboo is a document. Therefore, each document belongs to one concept. In the semantic identification of bamboo and rattan species, it is necessary to find the semantic related concept set based on query description (generally multiple word combinations), so as to determine the corresponding document set of the concept set. Among them, the establishment of the semantic correlation between the retrieval word set and concepts is the most core link, that is, the calculation of the semantic correlation between words and concepts in the semantic query extension.

In order to solve this problem, this paper USES the degree of word - concept - document to calculate the semantic relevance between the retrieval word set and the concept. The specific ideas are as follows:

If in the process of querying leaf attributes only in the ontology, the retrieval word set \( q \) is \{ lanceolate, glabrous, ..... , rough \}, then it can be expressed as the retrieval vector \( \bar{q} = \{ w_{1,q}, w_{2,q}, \cdots, w_{n,q} \} \), Where \( n \) represents the number of words to retrieve the word set; The leaf attribute of each bamboo in the domain ontology of bamboo rattan is represented by \( d_j \), so the leaf attribute vector can be represented as \( \bar{d}_j = \{ w_{1,j}, w_{2,j}, \cdots, w_{n,j} \} \).

When the search term appears, write 1 at the component of the corresponding vector; If the search term does not appear, 0 is written at the corresponding component.
Through the similarity between vectors \( \vec{d}_j \) and \( \vec{q} \), the degree of correlation between the leaf attribute \( d_j \) of the ontology and retrieval word set \( q \) can be evaluated. The cosine law between vectors can be used to quantify, that is, the cosine value of the angle between two vectors can be calculated:

\[
\text{sim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} (2)
\]

Where \( Q = (w_{1,q}, w_{2,q}, \cdots, w_{n,q}) \) represents the retrieval vector; \( d_i = (d_{1,i}, d_{2,i}, \cdots, d_{n,i}) \) is the concept vector.

The weight of the bamboo and rattan domain terms themselves varies. Where, the higher the frequency of a certain bamboo and rattan domain term appears in one of the documents, the greater the identification degree of the domain term to the document, and the higher the weight of the domain term; At the same time, if the domain term appears less frequently in all other descriptive documents of bamboo rattan from Flora of China, it means that the greater the identification degree of the domain term for this document, the higher the weight of the domain term will be.

\[
\text{Sim}(Q, D_i) = \frac{\sum_{j=1}^{n} w_{q,j} d_{j}}{\sqrt{\sum_{j=1}^{n} (d_{j})^2 \sum_{j=1}^{n} (w_{q,j})^2}} (3)
\]

Where, \( Q \) represents the retrieval word set, \( Q = (w_{1,q}, w_{2,q}, \cdots, w_{n,q}) \); \( D_i \) represents the concept vector, \( D_i = (d_{1,i}, d_{2,i}, \cdots, d_{n,i}) \); \( d_{j} \) represents the term weight of the \( j \) entry in the concept vector.

4. Experimental results and analysis

This paper makes an experimental analysis on whether the semantic information retrieval model of domain term weight factor is scientific and reasonable, and the influence of the term weight factor on the retrieval results before and after the introduction of domain term weight factor. The ninth volume of Flora of China was used as the experimental data set, including 37 genera, 516 species of bamboo, 516 documents and 530,000 words. The retrieval concepts and retrieval examples in table 1 were used as the original retrieval items in the experiment. Two sets of retrieval results were obtained by opening and closing the weight of domain terms in the experiment. The retrieval results are shown in table 2 and table 3.

| NO. | retrieval concept | Semantic extension | retrieval examples |
|-----|------------------|--------------------|-------------------|
| 1   | leaf             | sheath, blade ear, ligule, blade, petiole | lanceolate |
| 2   | subterranean stem| subterranean stem, pseudorhizome, whip made of bamboo | solid |
| 3   | pole             | Stem handle, culm base, internode, ring, Knot, ntrahode, septa intersegmental, internode shoots | bright |
| 4   | flower           | androecium, pistil, pale, inferior palea, lodicule, ovary, stigma, anther | yellow |
Table 2 the TOP5 semantic relevance value before introducing the domain term weight

| Number | Semantic Relevance Value | (Before Introducing Domain Terminology Weight) |
|--------|--------------------------|-----------------------------------------------|
|        | leaf (lanceolate)       | Underground stem (solid) | pole (bright) | flower (yellow) | Culm (not obvious) |
| 1      | 0.141421356             | 0.316227766             | 0.288675135 | 0.208514414 | 0.223606798 |
| 2      | 0.145864991             | 0.316227766             | 0.29415734  | 0.242535625 | 0.258198889 |
| 3      | 0.146173762             | 0.353553391             | 0.251573256 | 0.258198889 | 0.223606798 |
| 4      | 0.147441956             | 0.333333333             | 0.223606798 | 0.208514414 | 0.208514414 |
| 5      | 0.138675049             | 0.333333333             | 0.204124145 | 0.208514414 | 0.208514414 |

Table 3 the TOP5 semantic relevance value after introducing the domain term weight

| Number | Normalized Value of Semantic Relevance Degree | (After Introducing Domain Terminology Weight) |
|--------|-----------------------------------------------|-----------------------------------------------|
|        | leaf (lanceolate)       | Underground stem (solid) | pole (bright) | flower (yellow) | culm (not obvious) |
| 1      | 0.282842712             | 0.316227766             | 0.288675135 | 0.294883912 | 0.223606798 |
| 2      | 0.291729983             | 0.316227766             | 0.29415734  | 0.32997717  | 0.35418372 |
| 3      | 0.382546028             | 0.353553391             | 0.302546028 | 0.353553391 | 0.365148372 |
| 4      | 0.294883912             | 0.333333333             | 0.387298335 | 0.35418372  | 0.387298335 |
| 5      | 0.138675049             | 0.333333333             | 0.288675135 | 0.294883912 | 0.294883912 |

Table 2 and table 3 are retrieval results of Top5 semantic relevance value before and after the introduction of domain term weight respectively. Retrieval results corresponding to the varieties of bamboo: Melocanna baccifera (No.1), Schizostachyum dumetorum (No.2), Schizostachyum pseudolima (No.3), Schizostachyum funghomii (No.4) and Dendrocalamus tibeticus (No.5). Among them, the difference of relevance value of retrieval results of the same concept is mainly determined by the different weights of retrieval items and domain terms that constitute concept vectors, reflecting the degree to which retrieval items represent different concepts and documents.

Before and after the introduction of the weight of domain terms, the sorting matrix of semantic relevance is $L'$ and $L$, namely:

\[
L' = \begin{bmatrix}
4 & 1 & 4 & 4 & 3 \\
3 & 5 & 3 & 2 & 2 \\
2 & 5 & 1 & 1 & 1 \\
1 & 5 & 2 & 3 & 5 \\
5 & 5 & 5 & 5 & 4
\end{bmatrix}
\]

\[
L = \begin{bmatrix}
4 & 1 & 4 & 3 & 4 \\
3 & 5 & 3 & 5 & 2 \\
1 & 5 & 1 & 2 & 1 \\
2 & 5 & 2 & 1 & 5 \\
5 & 5 & 5 & 4 & 3
\end{bmatrix}
\]

Where the item whose semantic relevance is 0 indicates that the retrieval item is not included in the concept vector. In order to express the correlation degree of 0 more clearly, this paper sets all the items with semantic correlation degree of 0 as the last one, that is, the fifth one.

By contrast, it can be seen that the domain term weight plays an important role in the screening sequence of results. By contrast, $L'$ and $L$ in the concept of "leaf", the term "lanceolate" in Schizostachyum pseudolima and Schizostachyum funghomii domain term weights in the two documents, directly affected the "leaf" semantic relevancy in "lanceolate" and Schizostachyum pseudolima and Schizostachyum funghomii, making Schizostachyum pseudolima and Schizostachyum funghomii raft swaps.

At the same time, the concept of "sheath" key words "not clear" in the Melocanna baccifera and Dendrocalamus tibeticus also have the same problem. However, the concept "flower" is the one that is
most affected by the weight of domain terms. Before and after the weight of domain terms is turned on and off, the semantic relevance ranking of 5 bamboo species changes from (4, 2, 1, 3, 5) to (3, 5, 2, 1, 4), which affects the ranking of all retrieval results.

Compared with the documents corresponding to the five bamboos in Flora of China, it can be found that the retrieval words with significant domain term weight appear more frequently in the corresponding documents, and the corresponding concept description language is shorter, indicating that the weight of domain term can better distinguish the representative degree of the retrieval words.

Therefore, the term weight of bamboo rattan domain proposed in this paper is introduced into the retrieval model of bamboo rattan semantic information, so that the results with a high degree of relevance to the retrieval words can be ranked in front without affecting the retrieval results.

5. Conclusion
Based on the semantic expression of the domain ontology of bamboo rattan, this paper finds that the domain terms of bamboo rattan are closely related to the standard description information of bamboo rattan, proposes the weight calculation method of bamboo rattan domain terms based on ID-IDF, and constructs the semantic information retrieval model of bamboo rattan domain combining with the semantic extension based on correlation. The model describes the relationship between the concept of retrieval term and the species of bamboo rattan from two aspects of the domain term weight and semantic relevance. It can automatically recognize and extract the domain knowledge of bamboo rattan in the text information of bamboo rattan and screen the results based on correlation degree without requiring users to have a higher professional knowledge background, which solves the problem of "semantic gap" between the real retrieval intention of the discriminator and the domain knowledge of bamboo rattan.

Acknowledgments
The findings and the opinions are partially supported by projects of the major science and technology projects of Yunnan province, No.2018ZI001. This work is supported by Engineering and Technology Research Center for Yunnan Agricultural Big Data and Yunnan Agricultural University, China.

References
[1] Zhang Xinping. 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 Zeihui. The world's bamboo rattan(English edition) [M]. Beijing: China Forestry Publishing House, 2007.
[5] Qiu Yanxian. Research on Automatic Extraction of Domain Terms and Classification of Relationships [D]. Master's Thesis of Kunming University of Science and Technology, 2009
[6] SUN Xia, WANG Xiao-feng, DONG Le-hong, WU Jiang. Study on Term Relation Extraction from Domain Text[J]. Computer Science, 2010, 37(2):189-191, 215.
[7] Xv Jianmin. Extension Research of Bayesian Network Information Retrieval Model Based on Term Relation [D]. Doctoral Dissertation of Tianjin University, 2007.
[8] Xue Zheng. Hot Topic Discovery of Text Information Based on Improved TF-IDF [D]. Master's Thesis of Wuhan Res,2009
[9] Wang Qi. Semantic retrieval technology and application of ancient fresco [D]. Doctoral Dissertation of Zhejiang University, 2011.
[10] Han Xin. Research on resource semantic retrieval in bioinformatics based on ontology [D]. Master's Thesis of North University of China, 2012.