Research and Construction of Yunnan Plant Vertical Retrieval System

Lin Peng, Zhi-run Ma, Xin Zhang

College of Big Data, Yunnan Agricultural University, Kunming 650201, China
TCorrespondence to: Lin Peng(penglin2286351@163.com)

Abstract. The plant richness of Yunnan Province ranks first in the country, and scientific and rational protection and utilization of Yunnan plant resources are of great significance for maintaining China's biodiversity. In view of the current distribution patterns and physiological characteristics of plants in Yunnan Province, as well as related geography, climate and wild habitats, this paper constructs a set of vertical search system for Yunnan plants through the research of vertical search architecture, network data collection, semantic analysis of network information and the pattern of network information retrieval. The system can effectively solve the technical obstacles of Yunnan plant network resource utilization, and provide new technical support and theory for the macro-level and deep research of plants in Yunnan Province.

1. Introduction

Yunnan is an important gene pool of species in China and is known as the “plant kingdom”. According to statistics, there are 274 families, 2076 genera and 17,000 species of higher plants in Yunnan Province, accounting for 62.9% of the total number of higher plants in the country. Due to geological history changes, Quaternary glacial effects and vertical changes in the mountains, many plants in Yunnan have more ecological genetic types and stronger ecological adaptability [1].

A large number of studies have shown that the distribution of plants is closely related to changes in multiple real-time environmental factors such as longitude, altitude, atmospheric circulation, topography, temperature, and precipitation [2-6]. A comprehensive survey of the surrounding environment of plants, through the study of plant diversity and its distribution patterns, genesis, zoning and “hot spots” in the region, is the focus of plant research, conservation and utilization. However, the current distribution patterns and physiological characteristics of plants in Yunnan, as well as related geography, climate and wild habitats are very difficult to obtain, which greatly restricts the comprehensive large-scale and deep-level systematic research of Yunnan plants. Specifically, there are mainly the following problems:

1. Because Yunnan plant field knowledge involves many professions, various information resources are distributed in different databases and professional websites, and various resource retrieval system methods are different. Researchers need to master digital resource system retrieval technologies with different interfaces, and spend a lot of time and energy to browse, retrieve and summarize various types of information, resulting in a low level of comprehensive utilization of these information resources, restricting the large-scale, comprehensive and systematic research of Yunnan plants.

2. In the process of scientific research on Yunnan plants, not only static information such as plant shape description, image samples, distribution areas, scientific literature, etc., but also dynamic information such as soil data, water resources data, climate data, and remote sensing data are needed. Real-time information such as market price, planting area, production, and sales volume may also be required for economic plants. This information is distributed in different systems (databases),
independent of each other, intricate and intertwined, and there is a potential relationship. However, at present, both the general retrieval system and the agricultural vertical retrieval system at home and abroad do not take into account the interrelationship between the static, dynamic and real-time information in the plant field, and the real-time requirements of the retrieval results are not high, greatly it restricts the retrieval effect of the retrieval system.

(3) Because Yunnan plant domain information exists in different databases and professional websites, this information exists in different data formats such as TXT, HTML, XML, RTF, PDF, PSZ/PS, and in different languages such as Chinese, English, Latin, etc. However, most of the current retrieval tools can’t provide information retrieval of heterogeneous data; At the same time, these information from the Internet and professional databases can only achieve simple keyword-based retrieval, without semantic association ability. This leads to the fact that the user sometimes has to change the query word because the word does not match, and often only the query word appears in the document before it can be retrieved. As a result, there are frequent cases where documents related to user query requests cannot be retrieved because of different words.

(4) For those who work in Yunnan plant research, most of them use the general search engine to obtain relevant information on the Internet. When a user uses a general search engine for information retrieval, a large number of data sets containing duplicate information and spam information are often obtained, and the user cannot quickly and accurately locate the demand information. Moreover, the acquisition of Yunnan plant epitaxial information is still difficult to break through, which greatly restricts the large-scale and systematic research of Yunnan plants. In particular, the processing capability of real-time data on the network is not ideal, and the data fusion and semantic analysis of real-time network data need to be further studied.

(5) In view of the above problems, this paper introduces the vertical search engine technology, establishes a vertical search system for Yunnan plants, collects real-time plant network data in Yunnan, precise query, in-depth analysis, and provides Yunnan plant-related literature, pictures, videos and ecological environment information for researchers and related personnel.

2. Design Ideas of Yunnan Plant Vertical Retrieval System
At present, domain vertical search system generally has the problems of poor timeliness of information, difficulty in semantic interoperability and low search efficiency. This paper will solve the problem of information timeliness by establishing a federation architecture of vertical search system for Yunnan plants and combining it with creating an open field real-time perception and topic collection mechanism for Yunnan plants; By studying the topic clustering of network texts, it solves the problem of network data semantic interoperability; By studying the problem of Yunnan plant network information fusion, it creates a cloud-oriented plant search model to solve the problem of low efficiency.
3. Federated Architecture Construction of Distributed Vertical Retrieval System

The federal architecture of the distributed vertical search system can effectively support solving the problem of “specific and wide compatibility”, implicit association mining and real-time retrieval response in the vertical search system of Yunnan plant field. This paper adopts the design concept of “division and cooperation”. The whole search system consists of several vertical search engine nodes (unit nodes) and a federal central node for different topics. As shown in Fig. 2, each unit node is a meta-search engine for a specific topic, and can independently retrieve information of a specific topic independently; Each unit node can also accept the user retrieval request by the central node, decompose the request into different topic tasks, and distribute them to the corresponding different unit nodes to complete together. By adopting such a "division and collaboration" mechanism, it is desirable to solve the problem of compatibility between the "special" of the search results of the vertical search system and the "wide" of the collected data. At the same time, the central node performs in-depth processing such as fusion, clustering and association analysis on the unit node search information under different topics to solve the implicit association mining problem of multi-source data. At the same time, in the federated search system, the real-time perception mechanism of the unit node to the network information and the linkage data collection mechanism of the federated search system are established. When a node in the federated system perceives certain network information, it triggers nodes of other related topics to perform linkage data collection and data analysis, thereby solving the real-time response problem of the federated vertical search system.
4. Semantic analysis method for network text topic classification

The biggest challenge of network text topic classification is the feature matrix sparsity problem. The existing network text topic classification research only considers the influence of synonym and near-synonym on feature sparsity in traditional texts, and does not consider the influence of features such as less vocabulary and less semantics on the sparseness of features in network texts. Aiming at the existing research problems, the project intends to use the network text topic classification method based on concept and strong feature library to solve the feature sparse problem of network text space vector model. The overall framework of the method is shown in Figure 8, which is mainly composed of three parts. The first is to generate the concept of network text, the second is to generate strong features of the network text domain, the third is to generate a network text space vector model, train the classifier, and realize the classification of network text topics.

Figure 3. Thematic classification framework based on conceptual features and strong feature libraries.

(1) Generate network text concept features. Firstly, the establishment of the Yunnan plant domain ontology library based on the shape characteristics of Yunnan plants, the terminology and concepts are mainly derived from the "Yunnan Flora", with reference to "Chinese Flora"; Then, using the concept and the structural relationship in Yunnan plant ontology, the similarity of different feature words in
Yunnan plant ontology is calculated. Finally, the semantic connection is established between synonyms and near-synonyms, and the conceptual features of the network text are generated by replacing the feature words in the network text with concepts.

(2) Generate strong features in the network text domain. It is proposed to adopt a domain-oriented web crawler to capture massive domain network text data from the Internet as a domain subject data source. Based on the LDA theme model, statistical modelling is performed to obtain a strong feature vocabulary with high degree of classification. In order to reflect the importance of strong feature words in the subject classification, and to reduce the influence of fewer feature words in the network text on the sparse feature matrix, a new strong feature weighting method is proposed. This method defines the domain contribution of strong feature words based on information gain; Then, based on the contribution of strong feature fields, a strong feature weighting method is proposed.

(3) Generate network text space feature vectors, train classifiers, and implement network text topic classification. It is proposed to use mutual information for text feature selection. After selecting the first M features with the largest mutual information, the network text feature features based on the Yunnan plant ontology generation and the domain strong features generated based on the domain strong feature database are added to jointly generate the network text space feature vector. Finally, the spatial feature vector of the network text consists of two parts, one is the ordinary feature word and the concept feature, and is weighted by TF-IDF; The second is a strong feature word, which is weighted by a strong feature weighting method based on domain contribution. Based on the network text space vector, the polynomial Bayesian classifier is initially proposed as the network text topic classifier to realize the network text topic classification.

5. Construction of Yunnan Plant Vertical Search Prototype System

Combined with the vertical search prototype system architecture and functional requirements for Yunnan plants, a vertical search engine system for Yunnan plants was constructed. The basic framework of the system is shown in Figure 4.
The system can implement traditional document retrieval and retrieval of network-oriented information. In the document-level index, we index all information related to the document, such as the body, keywords, burst words, information ID corresponding to the document, time, location, source, importance of the document in the information, and so on. In the information retrieval system, users can submit two different types of queries: 1) Query with only time or place; 2) A query containing query keywords. For the first type of query, it mainly supports the user to browse the information needs that occur at a specific time or at a specific place. For the second type of query, it mainly supports the user's need to understand the information related to the query word. Of course, for the second type of query, the user can also specify the time and place. In the information level index, the list of documents corresponding to each information will be saved. Through this list, you can quickly find out which information documents belong to this.

6. Conclusion
This paper introduces the vertical search engine technology, establishes a vertical search system for Yunnan plants, collects, accurately queries and in-depth analyses the network data of Yunnan plants in real time, and provides Yunnan plants (including text, pictures, videos and ecological environment, etc.) data information, effectively solves the technical obstacles in the utilization of digital resources in Yunnan plant networks, and comprehensively enhances the real-time response capability and overall use effect of Yunnan plant network digital resources.

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] Zai-fu Xu. Approach on some characteristics of Yunnan plants under the evolution from ancient to modern[J]. GUIHAIA, 2003, 23(4): 294-297.
[2] ZHU H, MA Y, YAN L, et al. The relationship between geography and climate in the generic-level patterns of Chinese seed plants[J]. Acta Phytotaxonomica Sinica, 2007, 45(2): 134-166.
[3] Zhi-jian Yu. Discussion on the role of climate in the transformation of yunnan into a plant kingdom[J]. AGRICULTURE AND TECHNOLOGY, 2014, 34(1): 178.
[4] Wen-hong Ma, Yuan-he Yang and Jin-sheng He. Biomass of temperate grassland in Inner Mongolia and its relationship with environmental factors[J]. Science in China(Series C): Life Science, 2008, 38(1): 84-92.
[5] Jian-meng Feng. Spatial patterns of species diversity of seed plants in China and their climatic explanation[J]. Biodiversity Science, 2008, 16(5): 470-476.
[6] Jian-meng Feng and Xv Dongcheng. The large-scale distribution pattern of seed plant species richness in China and its relationship with geographical factors [J]. Ecology and Environmental Science, 2009, 18(1): 249-254.
[7] Xiang-qin Liu. Research and Design on Key Technologies of Vertical Search Engine Oriented Soybean Theme[D]. Northeast Agricultural University, 2013.
[8] Xiao-rong Yang. Research and application of key technologies of distributed agricultural science and technology information sharing[D]. Doctoral dissertation of the Chinese academy of agricultural sciences, 2011.
[9] Jing Li. Ontology theory and its application in agricultural document retrieval system -- a case study of floriculture ontology modeling[D]. Doctoral dissertation of graduate school of Chinese academy of sciences, 2004.
[10] Xiao-rui Li. The Present Situation of Archaeology in Yunan Province[J]. Relics From South, 2016, 1: 166-170.