Enhancing Plant Species Retrieval in Flora Through Language Model Integration

De-Kai Kao[‡], Chih-Kai Yang[‡], Chien-Hsing Chen[§]

‡ Department of Forestry, National Pingtung University of Science and Technology, Pingtung, Taiwan
§ Department of Biomechatronics Engineering, National Pingtung University of Science and Technology, Pingtung, Taiwan

Corresponding author: De-Kai Kao (block58697@gmail.com)

Abstract

Traditionally, textual data storage and retrieval systems were designed primarily for human reading, mainly relying on paper records. However, as information technology has advanced, computerized searches have become common. However, Boolean logic-based data retrieval systems often struggle to handle data's diversity and richness effectively. These systems rely on strict matching rules, which can lead to either too few or too many results. For example, when searching for plant species descriptions, a query like "circle" AND "ellipse" may exclude relevant records that describe these traits using slightly different terms (e.g., "round" or "oval"). Conversely, broader queries like "oblong" may return an overwhelming number of irrelevant results. This rigidity limits the system's ability to adapt to the nuanced and varied ways users describe data. With the advent of advanced semantic models such as SBERT (Sentence-Bidirectional Encoder Representations from Transformers) (Reimers and Gurevych 2019), we can now delve deeper into the semantic relationships within textual data. Unlike general-purpose large language models, SBERT is specifically designed for efficient semantic similarity computation.

In plant taxonomy, records in Flora, such as *Flora of Taiwan* or *Flora of China*, play a crucial role in understanding plant diversity in specific regions. These records provide critical information on plant growth environments, morphological characteristics, and economic values.

Our research aims to enhance the efficiency of retrieving plant data using language models. Specifically, we transform textual descriptions from Flora and user queries into vector representations (Fig. 2) and calculate their cosine similarity to determine the relevance between user inputs and species records. Cosine similarity, a metric commonly used in text mining and information retrieval, quantifies the similarity between two vectors by measuring the cosine of the angle between them. The similarity score ranges from -1 (completely dissimilar) to 1 (identical), where higher scores indicate greater similarity. By applying this method, we can provide users with ranked scores of plant species related to their queries (Fig. 1). This approach not only streamlines data

retrieval but also introduces new perspectives for botanical research and data management, fostering a more efficient exploration of plant diversity.

Our results demonstrate the potential of language models to facilitate biodiversity research and data management, especially in retrieving plant taxonomy information. Our approach provides a novel tool for future biodiversity data analysis and retrieval, thereby contributing to the progress of biodiversity conservation.

Keywords

species identification, cosine similarity, semantic retrieval

Presenting author

De-Kai Kao

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084 URL: https://arxiv.org/pdf/1908.10084

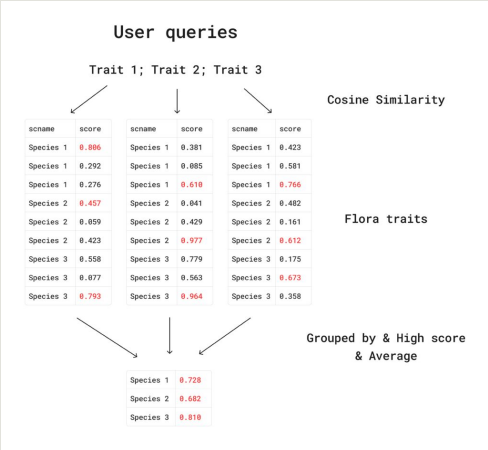

Figure 1.

Cosine similarity and aggregated scoring for Flora trait queries.

The calculation process provides a visual representation of the cosine similarity scores between user queries and Flora traits. In the middle section, each row represents a specific trait of a plant species, while columns correspond to the user's query traits (Trait 1, Trait 2, Trait 3). The cosine similarity score measures how closely a trait from the user's query aligns with traits in the Flora dataset.

Red numbers highlight the highest similarity score for each species across all its traits, representing the trait most relevant to the user's query. At the bottom, the aggregated scores show the average of these highest scores, providing an overall similarity score for each species and ranking their relevance to the user's query.

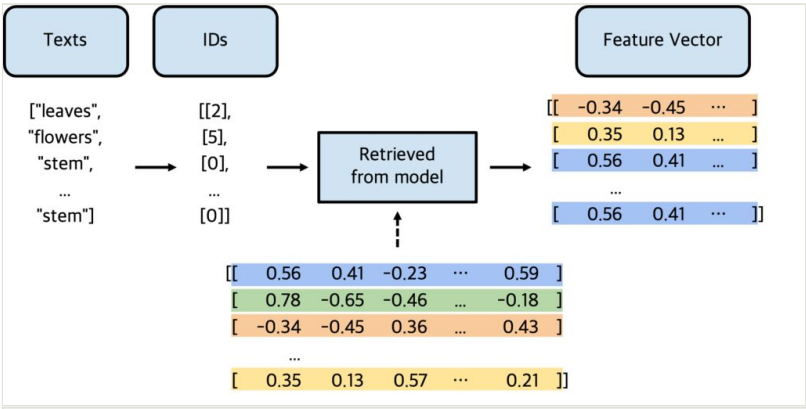

Figure 2.

Embedding lookup retrieved from the language model.

This workflow illustrates the process of transforming textual data into vector representations using a pre-trained language model. The leftmost column contains the original textual inputs, including both species descriptions and user queries. These texts are associated with unique IDs (middle column) for reference. The retrieved vector representations (rightmost column) are numerical embeddings generated by the language model. Each row represents a unique vector, which captures the semantic meaning of the corresponding text. The numbers within each vector represent the values of individual dimensions in the vector space. These values are used for calculating the cosine similarity between the vectors.