Enhancing Keyword-Based Botanical Information Retrieval with Information Extraction

Xiaoya Tang
School of Library and Information Management, Emporia State University
1200 Commercial St., Campus Box 4025
Emporia, KS 66801
1-620-341-5071

xtang@emporia.edu

ABSTRACT
Keyword-based retrieval matches search terms and documents via term co-occurrence. Such an approach does not allow matching based on the specific plant characteristic descriptions that are often used in botanical text retrieval. This study applies information extraction techniques to automatically extract plant characteristic information from text and allows users to search using such information in combination with keywords. An evaluation experiment was conducted using actual users. The results indicate that this approach enhances task-based retrieval performance.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods.*

General Terms
Design, Experimentation, Performance

Keywords
Information Retrieval, Botanical Text Retrieval, Information Extraction

1. INTRODUCTION
Documents in a domain-specific collection usually contain very similar terms with similar frequencies, though each document might describe a different topic or object. For such collections, keywords or phrases are not very effective in identifying particular documents.

Botanical text collections contain documents that describe various plant characteristics that are critical to plant species identification. Species identification usually requires very specific and accurate information. However, most retrieval systems on botanical text collections are keyword-based and do not work effectively with such specific retrieval. Furthermore, keyword-based retrieval systems are not effective for matching user query terms and document terms as the two use very different vocabularies [4]. Efforts have been made to extract useful information from specialized collections such as biomedical literature [3] to create databases. Information extraction techniques have also been applied to retrieval results as a post-retrieval filter [1]. In this study, information techniques are used to enhance keyword-based retrieval by automatically extracting plant characteristic information from texts and allowing users to search with both extracted information and keywords. A retrieval system was implemented and evaluated by real users on a sub-set of the Flora of North America (FNA) collection.

SIGIR'08, July 20-24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. METHOD
2.1 Extraction of Plant Characteristic Information
This study adapted and enhanced an information extraction system to extract the plant characteristic information from each document. The system WHISK [2] was originally created to extract rental information from short newspaper advertisements. It uses machine learning techniques and learns how to extract information based on manually created examples. Changes were made to generalize the system so that it can work with botanical texts that have longer sentences, broader vocabulary, and a greater variation in structure among different documents.

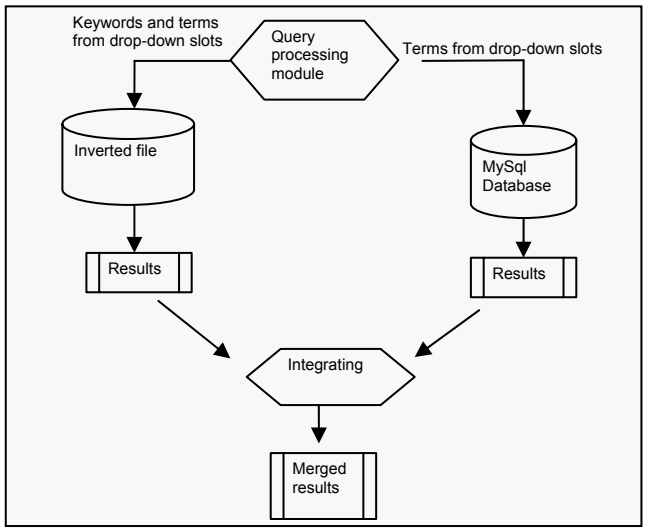
The enhanced system was then used to extract a set of plant characteristics from the original documents that are most often used in user search queries, including information regarding plant leaf shape, leaf margin, leaf color, leaf length and width, leaf arrangement, leaf apex, leaf base, and fruit/net shape. Plant distribution information, plant genus and species names were also extracted.

2.2 Retrieval System Implementation
An information retrieval system SEARFA (SEARch Flora Advanced system) was implemented to allow users to search using both extracted information and keywords. The search interface included a search form to allow the use of the extracted information in search. It also included a search box to allow users to search using keywords. The search form was designed to be as simple as possible so that any impact on retrieval performance introduced by the form itself was minimized. As users often used an attribute-value like pattern in their search queries [4], the search form included a list of plant attributes. To describe the value of each attribute, users could select extracted information from a drop-down menu. The search terms from the drop-down menus were also searched as keywords in order to mitigate against possible errors in extraction. The search results from different parts were merged, as shown in Figure-1.

Different weights were assigned to keywords and extracted information. Keywords were weighted using the inverse document frequency weighting algorithm $tf \times idf$, whereas extracted information was assigned higher weights as it is very

specific and expresses much more accurate meaning than keywords, and is therefore considered to have a greater value in discriminating documents. The results of the two parts were merged and the weights of the same document in the two result sets were summed to obtain its final weight.

Figure-1. Merging of the Results from the two Parts



3. EVALUATION

An experiment was conducted to evaluate the system by real users to accomplish a set of tasks that were created to best simulate practical use of the text collection. A retrieval system SEARF (SEARch Flora system) that only provides keyword search was used as a benchmark. Twenty four biological science students with certain level of botanical knowledge were recruited and randomly assigned to two groups. They were asked to search the collection of documents using one of two systems to find the exact documents that describe the species of a set of 8 plant specimens after a short training session to help them get acquainted with the system.

In this study, the retrieval performance was measured by a task-oriented approach focusing on the utility of information and subjects' efforts. The following four measures were used to evaluate the search performance: NTH (Number of Tasks accomplished per Hour), which is the number of tasks that were successfully completed by a subject in an hour; SSR (Search Success Rate), which is the ratio of the number of successful searches to the total number of searches of a subject; NST (Number of Searches per Task), which is the average number of searches a subject conducted for a task, whether it was successfully completed or not; and NSAT (Number of Searches per Accomplished Task), which is the average number of searches a subject conducted for a task that was successfully completed. The results are shown in Table-1.

The experiment results indicate that subjects in the SEARFA group have better performance regarding all the measures, which means that on average the automatically extracted information does enhance retrieval performance.

Table-1. Means of Performance of Subjects in the Two Groups

Groups	NTH	SSR	NST	NSAT
SEARFA	8.0775	0.2100	6.045	4.74
SEARF	3.5975	0.0533	11.938	8.43

One possible reason for the better performance of the SEARFA system was that the extracted information provided more specific document representations and allowed more accurate matching between user queries and document content. As a simple example, when a user typed in the search query "leaf ovate" in the keyword search, the results documents would be those that contain these two words. However, these documents might describe ovate seeds and oblong leaves, which are not what the users want. But with extracted information, the users would search by defining "leaf-shape: ovate" in a query which would match a document with extracted information "leaf-shape: ovate".

The extracted information also helped by providing the connection between users' vocabulary and the vocabulary used in the collection. As the collection contained many botanical terminologies, it was more difficult for users to find the appropriate search terms by themselves. It was relatively easier for them to select terms from a given list.

4. CONCLUSIONS AND FUTURE WORK

The study indicates that information extraction is a promising technique for enhancing botanical text retrieval. However, information searching is a complicated process which involves many factors from the system's side as well as from the user's side. One future direction of this study will be the investigation of how the extracted information impacts search performance of users having different levels of domain knowledge.

5. REFERENCES

[1] Bear, J., Israel, D., Petit, J., and Martin, D. Using information extraction to improve document retrieval. In E. M. Voorhees, D. K. Harman (eds.), *The Sixth Text REtrieval Conference (TREC-6)*, 367-377. NIST Special Publication 500-240, 1998.

[2] Soderland, S. Learning information extraction rules for semi-structured and free text. *Machine Learning, Feb. 1999,34, 1-3,* 233-272.

[3] Subramaniam, L. V., Mukherjea, S., Kankar, P., Srivastava, B., Batra, V.S., Kamesam, P.V., and Kothari, R. Information extraction from biomedical literature: methodology, evaluation and an application. In *Proceedings Of The Twelfth International Conference On Information And Knowledge Management*, November 2003, 410 – 417.

[4] Tang, Xiaoya, & Heidorn, P. Bryan. (2007). The Loss of domain knowledge in user search queries: A Query log analysis of a botanical retrieval system. In *Proceedings of Annual Meeting of the American Society for Information Science and Technology,* October 19-24, 2007, Milwaukee, Wisconsin.