An Intelligent Search Engine for Agricultural Disease Prescription

Weijian Ni, Mei Liu, Qingtian Zeng, Tong Liu

An Intelligent Search Engine for Agricultural Disease Prescription

Weijian Ni, Mei Liu, Qingtian Zeng and Tong Liu

Shandong University of Science and TechnologyQingdao, Shandong Province, 266510 P.R.China

niweijian@gmail.com

Abstract. Generic search engines have played a significant role in helping people locate their needed information on the web. However, they don't perform as desired on domain-specific queries. In this paper, we focus on the domain of agriculture and develop a novel search engine specifically for agricultural disease prescription retrieval. In order to improve the performance of search for prescription documents, we exploit the domain-specific characteristics embedded in agricultural disease prescription, and propose a domain-specific query expansion approach as well as a BM25-based structural retrieval function. An intelligent search engine for agricultural disease prescription is then implemented based on the proposed retrieval model. User interfaces of the developed search engine are demonstrated.

Keywords: Search Engine; Agricultural Disease Prescription; QueryExpansion; Information Retrieval

1 Introduction

A document of agricultural disease prescription generally covers multifaceted information about some agricultural disease, including symptom, transmission route, control method and etc. Given an appropriate piece of agricultural disease prescription, farmers are likely to know about the treatment plans, causes and other aspects of the agricultural disease he was encountering. Due to massive amounts of agricultural disease prescriptions on the web, search engine becomes an indispensable tool for farmers when looking for the right treatment plan.

However, providing farmers with agricultural disease prescriptions of practical application is not a trivial task. As an example, a farmer may expect to know about *control methods* of an agricultural disease he encounters during plant cultivation, by typing a keyword query describing the *symptom* to search engines; nevertheless, a document containing extended yet repeated information about the input symptom is more likely to be returned than that containing information about control method. That is partly because most traditional information retrieval models are based on keyword-

* Corresponding author: Qingtian Zeng

match between user's query and candidate documents. However, when searching for agricultural disease prescription documents, user's intent is not only embedded in the literal denotation of the query, but also dependent on the semantics associated with the initial query which is hard to be captured in the keyword-match retrieval schema.

Table 1. An example of agricultural disease prescription document

芹菜病毒病（Celery Virus Disease）
一、症状（I. Symptom） 全株染病。初叶片皱缩，呈现浓、淡绿色斑驳或黄色斑块，表现为明显的黄斑花叶，严重时，全株叶片皱缩不长或黄化、矮缩…… **二、病原（II. Etiology）** 由黄瓜花叶病毒（CMV）和芹菜花叶病毒（CeMV）侵染引起。两种病原引起的叶症状相似。芹菜花叶病毒（CeMV）粒体线形，寄主范围窄，主要侵染菊科、藜科、茄科中几种植物，病毒汁液稀释限点100～1000倍，钝化…… **三、传播途径（III. Transmission route）** CMV和CeMV田间主要通过蚜虫传播，或通过人工操作接触磨擦传毒…… **四、发病条件（IV. Epidemic factor）** 栽培管理条件差，干旱、蚜虫多发病重…… **五、防治方法（V. Control method）** 主要采取防蚜、避蚜措施进行防治。其次是加强水肥管理，提高植株抗病能力，以减轻为害。其他方法参见番茄病毒病……

Compared with plain-text documents, an essential difference of agricultural disease prescription documents is that the passages therein typically play a clear role in describing different aspects of the corresponding agricultural disease, such as symptom and control method. The roles may be either provided explicitly by document authors just as the example shown in Table 1, or implied in the content of text. In other words, agricultural disease prescription document is essentially a type of structured document in which passages with their roles make up of the fields or components. Obviously, the performance of prescription retrieval would be promoted if the document structure embedded in agricultural disease prescriptions was exploited in a principled way.

Based on the above observation, we propose a novel structural retrieval model and further develop an intelligent search engine for agricultural disease prescriptions. Specifically, we propose a field probabilistic model and a field associative model to formulate the structure information embedded in prescription documents, and incorporate the document structure into the prescription retrieval process through domain-specific query expansion and structured BM25 retrieval function. The domain-specific search engine is then implemented using Apache Lucene toolkit [1].

2 Related Work

2.1 Information Retrieval

Information Retrieval (IR) has gained substantial research attentions due to massive amount of data emerging, especially on the web [2]. Traditional IR models such as vector space model, language model, probabilistic model and learning-to-rank [3], have become one of core functions of modern search engines. A key to IR models is to score a document against a user's query through evaluating how the document fits user's information need.

If documents have structure more than plain-text, it would be hard to incorporate the structure information into traditional IR models because they are mostly developed for plain-text documents. Therefore, a number of retrieval models scoring documents through considering structure evidences, referred to as Structural Retrieval, have been proposed. A basic idea of structural retrieval is to combine the scores calculated on each components within the document using traditional score functions [4,5]. More recently, structural retrieval models have found wide applications in job-resume matching [6], question answering [7] and etc. However, to our knowledge, no work exists on adopting structural retrieval model for the structured agricultural disease prescription documents.

2.2 Agriculture-specific Search Engine

According to the statistics, by the end of 2009, there had been more than 30,000 agricultural web sites on the Internet, which cover heterogeneous agricultural information about market trends, agricultural technologies and etc. Since generic search engines usually don't incorporate characteristics of agriculture domain, it is necessary to develop agriculture-specific search engines to help farmer acquiring their interested information from the Internet. Key techniques of agricultural IR have been widely studied recently. As example, Huang [8] proposed an agriculture deep web entry discovery algorithm to acquire agriculture-specific deep web resources; Zhou [9] focused on filtering the agriculture unrelated topic pages based on URL and content during web pages crawling.

Several Chinese agriculture-specific search engines have been put into production, e.g. AgriSou [10], Sounong [11] and Agr365 [12]. There are also numbers of world wide agriculture-specific search engines: AgNIC (Agriculture Network Information Collaborative) [13] is a knowledge discovery system and collaborative platform for agricultural resource interchanging and searching supported by U.S. national agricultural library; Agriscape [14] is an online directory on agriculture related information and provides a simple query interface.

Most of the existing agriculture-specific search engines aim to retrieve various types of agricultural information such as farm price and agricultural news, which is so heterogeneous that embodies little common domain-specific characteristics; while we aim to develop an agriculture-specific search engine especially for one type of agricultural information, i.e., agricultural disease prescription. Furthermore, the agricul-

ture-specific search engine in the paper is developed based on a novel structural retrieval model instead of tradition IR models.

3 Architecture of Prescription Search Engine

Figure 1 shows the architecture diagram of the developed agricultural disease prescription search engine based on structural retrieval. There are mainly two components: structure information model and structural retrieval function. The former aims to derive the structure information embedded in prescription documents that would contribute to retrieval accuracy; the latter then provides users with prescription documents with high relevance w.r.t. users' queries.

In the following section, we will present the proposed structural retrieval model in details. Formally, let $\mathcal{C} = \{D_1, \cdots, D_n\}$be a prescription collection composed of n documents and assume every passages in a prescription document are tagged with one (or possibly a few) field$f$ from a given field set $F = \{f_1, \cdots, f_k\}$.

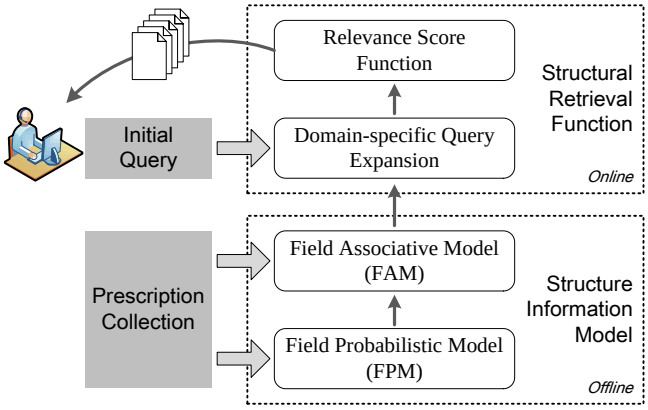

Fig. 1. Architecture diagram of the developed search engine

4 Structural Retrieval Model for Prescription Search

4.1 Structure Information Model

Field Probabilistic Model.

Due to the different semantics of prescription fields, the affinities of a term with different fields are not necessarily always equal. For example, the terms *yellow* and *patches* would give a stronger hint on the *symptom* than other aspects of a disease, so it is necessary to pay emphasis on the *symptom* field when searching for the query term *yellow patches*. Therefore, we propose a Field Probabilistic Model (FPM) to infer field affinity for each term.

A FPM is a multinomial distribution$p(f|w)$ that attempts to capture the probability of a field f when seeing a given term w. By Bayes' rule, the posterior probability

$p(f|w)$ can be computed by combining the prior probability of each field and the conditional term likelihood. That is,

$$p(f|w) = \frac{p(w|f)p(f)}{p(w)} = \frac{p(w|f)p(f)}{\sum_{f' \in F} p(w|f')p(f')}$$

where $p(f)$ is the prior probability of field f observed in prescription documents and $p(w|f)$ is the probability of seeing a term w in a given field f. We estimate the both probabilities by collection statistics of a given prescription collection$\mathcal{C}$.

Field Associative Model.

In agricultural disease prescription documents, there are often potential associations among field terms. Taking agricultural diseases with the symptom *oozing black liquid* as an example, if the control methods containing the terms *quarantine* and *spraying suspension agent* were found to be more effectual than that containing the term *removing weeds* in related prescription documents, the documents with the term *oozing black liquid* in the *symptom* field, as well as the term *quarantine* or *spraying suspension agent* in the *control method* field, would be preferred for the query *oozing black liquid*. Therefore, we propose a Field Associative Model (FAM) to discover these strong inter-term associations.

A FAM is a set of association rules [15] each of which is an implication of the form $R: A \Rightarrow B$, where both the antecedent and consequent are sets of terms. The utility of each rule is measured by two metrics, namely *support* and *confidence*.

Support of a rule is defined as in a given prescription collection the percentage of documents where every terms in the rule are seen. It can be simply computed as the term set probability: $sup(A \Rightarrow B) = p(A \cup B)$. *Confidence* of a rule is the probability of seeing terms in consequent in the prescription set where the terms in antecedent appear, i.e.,$conf(A \Rightarrow B) = p(B|A)$.

Intuitively, high *support* of a rule implies that the rule covers a considerable part of the prescription collection, while high *confidence* implies that the terms in consequent are highly likely to appear if the terms in antecedent are seen. Therefore, in FAM, only the rules whose *support* and *confidence* satisfies minimum thresholds could be considered meaningful. Practically, we view a prescription document with its composed terms as a piece of transaction data, and employ FP-growth [16], a frequent-itemsets mining algorithm with high time and space efficiency, to mine FAM-rule candidate from the given prescription collection.

4.2 Structural Retrieval function

Based on the above structure information models, we propose a prescription retrieval mechanism to provide users with proper prescription documents. In particular, we first propose a domain-specific query expansion approach to capture user's actual

query intent, and then propose a score function that calculate relevance of prescription documents w.r.t. the expanded user's query.

Domain-specific Query Expansion.

Instead of employing traditional query expansion approach in generic IR, we leverage the derived structure information about agricultural disease prescriptions to uncover possible query intents behind query terms. The idea is to view FAM-rules as a prescription knowledge repository composed of informational associations among term sets, and expand the query terms in antecedent of a FAM-rule by the terms in its corresponding consequent.

Formally, given a set of FAM-rules $\mathcal{R} = \{(R_1: A_1 \Rightarrow B_1), \cdots, (R_m: A_m \Rightarrow B_m)\}$and an initial user's query composed of t terms: $Q = \{q_1, \cdots, q_t\}$, the expanded user query takes the following form:

$$Q' = Q \cup Q_E \qquad (1)$$

where Q_E is the set of expanded query terms and

$$Q_E = \{q' \mid \forall R_i \in \mathcal{R}, \exists q_j \in Q, q_j \in A_i \wedge q' \in B_i\} \qquad (1)$$

Note that the terms q in the expanded user query Q' are assigned with different weights w_q according to the expanding confidence. Generally, for the terms $q \in Q, w_q = 1$; whereas for the terms $q \in Q_E$,

$$w_q = \frac{\sum_{R \in \mathcal{R}_q} conf(R)}{|\mathcal{R}_q|} \qquad (1)$$

where $\mathcal{R}_q$ is the set of FAM-rules according to which the term q is expanded and $conf(R)$ is *confidence* of a given FAM-rule R.

As mentioned above, the field affinity of each term varies among prescription fields, which is depicted in FPM, we further structurally expand user's query using FPM to derive the underlying structure of users' search intent.

Formally, given a set of prescription fields $F = \{f_1, \cdots, f_k\}$ and a query $Q' = \{q'_1, \cdots, q'_s\}$ expanded using FAM, the query further expanded using FPM can be written as:

$$Q'' = \begin{array}{c} \\ q'_1 \\ \vdots \\ q'_s \end{array} \begin{array}{c} \begin{array}{ccc} f_1 & \cdots & f_k \end{array} \\ \begin{pmatrix} w_{1,1} & \cdots & w_{1,k} \\ \vdots & \ddots & \vdots \\ w_{s,1} & \cdots & w_{s,k} \end{pmatrix} \end{array} \qquad (1)$$

where each element $w_{i,j}$ in the matrix indicates query intensity of the expanded query term q'_i on the field f_j, and $w_{i,j} = w_{t_i} \cdot p(f_j|t_i)$.

Prescription Score Function.

After expanded using FAM and FPM, user's query will be structurally represented as a matrix. It makes most traditional IR models difficult to be leveraged directly. In the paper, we extend the traditional BM25 model [17] to deal with the structured query in a natural way. The traditional score function of BM25 model is shown as follows:

$$score(D,Q) = \sum_{q \in Q} \frac{(k_1+1) \cdot tf(q,D)}{k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right) + tf(q,D)} \cdot \log \frac{N - df(q) + 0.5}{df(q) + 0.5} \quad (1)$$

where tf and df are the term frequency function and the document frequency function, respectively. k_1 and b are free parameters.

However, for structured documents and queries, it would not be adequate to calculate term frequency by simply counting occurrence of a term in a document. Thus we adapt the score function in BM25 model by revising the term frequency function. The idea is that an occurrence of some query term in a field is weighted by the field weight of that term. Assume that a prescription document is segmented according to prescription fields, i.e., $D = \{D_{f_1}, \cdots, D_{f_k}\}$, and each field of document D_{f_j} is viewed as unstructured text, term frequency of a term q_i' $(1 \leq i \leq s)$in the structured query Q'' is calculated as:

$$tf_{struct}(q_i', D) = \sum_{j=1,\cdots,k} tf(q_i', D) \cdot w_{ij} \quad (1)$$

Then, we substitute Equation 7 into Equation 6 and get the relevance score function for agricultural disease prescription retrieval.

5 Implementation

A search engine for agricultural disease prescription was developed based on the proposed structural retrieval model. The indexing and retrieval function are implemented using Apache Lucene toolkit [1]. To construct prescription database for the search engine, we crawled web pages about agricultural diseases from several agricultural websites. For the pages in each individual website, a set of rules were manually designed to extract the text of prescription and to segment the text according to predefined prescription fields in {*Symptom*, *Etiology*, *Pathogenesis*, *Infection_Way*, *Control_Method*}. In total, there are 7903 prescription documents indexed in the developed search engine.

Figure 2 shows the query interface of the developed search engine. The user can type his query in the top box on the view. Our search engine also supports facet search and each of the facets is a prescription field. Below the query box, the user can find the retrieved prescription documents ranked by the relevance score. Unlike generic search engines that return users with excerpts of relevant webpages, each result

of the developed search engine lists excerpts of every prescription fields and a photo of that disease (if any exist). When the user clicks on the titles of retrieved results, the details of the corresponding agricultural disease prescription will be displayed in a pop-up window as show in Figure 3. All the information about the selected agricultural disease is organized by the prescription fields. A navigation menu on the top-right side of the view is designed to help user browse details of the prescription document. We can roughly say that, through using our system, it would be much easier for farmers to locate their needed information in agricultural disease prescriptions.

Fig. 2. Query interface of prescription search engine

Fig. 3. Detail display interface of prescription search engine

6 Conclusion and Further Work

In this paper, we developed a domain-specific search engine for agricultural disease prescription. To exploit the domain characteristic of agricultural disease prescription, we based the search engine on a novel structural retrieval model. The proposed retrieval model includes modeling structure information embedded in prescription documents, structurally expanding user's query and a structural retrieval function. We constructed a real-world prescription collection and implemented the search engine using Apache Lucene toolkit. The search results are structurally organized in the user interface to facilitate user's information needs.

As a primary effort in domain-specific search engine for agricultural disease prescription, there is still much room for improvement. One direct and effective way to improve retrieval result is to enlarge the indexed prescription collection. In this way, more potential relevant prescription would be included in retrieval results. Besides, instead of BM25, we will leverage other recently proposed IR models, e.g. learning to rank, in this IR task.

Acknowledgments.

This paper is supported partly by National Natural Science Foundation of China(No. 61170079 and 61202152), Excellent Young Scientist Foundation of Shandong Province (No. BS2012DX030), Higher Educational Science and Technology Program of Shandong Province (No. J12LN45), Postdoctoral Science Foundation of China (No. 2012M521363), National Statistical Science Foundation of China (No. 2012LY001), Special Fund for Agro-scientific Research in the Public Interest (No. 201303107), Special Fund for Fast Sharing of Science Paper in Net Era by CSTD (No. 2012107) and Sci. & Tech. Development Fund of Qingdao(13-1-4-153-jch).

References.

1. Apache Lucene (2013). http://lucene.apache.org/
2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction toInformation Retrieval. Cambridge University Press (2008)
3. Tie-Yan Liu. Learning to Rank for Information Retrieval. Springer (2011)
4. Ross Wilkinson. Effective retrieval of structured documents. In Proceedings of SIGIR, pp. 311–317 (1994)
5. Paul Ogilvie, Jamie Callan. Combining document representations for known-itemsearch. In Proceedings of SIGIR, pp. 143–150 (2003)
6. Xing Yi, James Allan, W. Bruce Croft. Matching resumes and jobs based on relevance models. In Proceedings of SIGIR, pp. 809–810 (2007)
7. Le Zhao, Jamie Callan. Effective and Efficient Structured Retrieval. In Proceedingsof CIKM, pp. 1573–1576 (2009)
8. He Huang. Complex Adaptive Agriculture Vertical Search Model and its Implementation. Dissertation: University of Science and Technology of China (2010)

9. Peng Zhou. Research on key techniques of agricultural search engine. MS Thesis:Capital Normal University, China (2009)
10. AgriSou (2013). http://www.agrisou.com/
11. Sounong (2013). http://www.sounong.net/
12. Agr365 (2013). http://so.ag365.com/
13. AgNIC (2013). http://www.agnic.org/
14. Agriscape (2013). http://www.agriscape.com/
15. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques,Third Edition. Massachusetts:Morgan Kaufmann (2011)
16. Jiawei Han, Jian Pei, Yiwen Yin. Mining frequent patterns without candidategeneration. In Proceedings of SIGMOD, pp. 1–12 (2000)
17. Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu. Okapi atTREC-7. In Proceedings of TREC, pp. 199–210 (1998)