一种面向领域文档的结构化检索模型及其在农技处方检索中的应用

刘 彤 倪维健
(山东科技大学信息科学与工程学院 青岛 266590)

摘 要 各种专业领域中的文档往往具有显著的结构化特征,即一篇文档往往是由具有不同表达功能的相对固定的多个文本字段构成,同时这些字段蕴含了相关的领域知识。针对专业文档的结构化和领域化特征,设计了一种面向结构化领域文档的信息检索模型。在该模型中,首先对领域文档集进行挖掘以构建能够反映领域知识的结构化模型,之后以此为基础设计了结构化文档检索算法来为用户查询返回相关的领域文档。选择一类典型的领域文档——农技处方开展了应用研究,利用一份现实的农技处方文档数据集将提出的方法与传统的信息检索方法进行了实验对比分析,并开发了农技处方检索原型系统。

关键词 信息检索,农技处方,查询扩展,结构化检索
中图法分类号 TP391.3 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.10.056

Information Retrieval Model for Domain-specific Structural Documents and its Application in Agricultural Disease Prescription Retrieval

LIU Tong NI Wei-jian
(College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract Different from plain text, professional documents in various domains are mostly a type of structural document which is composed of several roughly fixed textual fields and embeds rich domain knowledge. To incorporate the inherent structure information and domain knowledge, we proposed a novel retrieval model for professional documents based on structural retrieval. In particular, we first derived a domain model from a given professional document collection, and then used it as a basis to design a domain-specific structural retrieval function. We applied the proposed structural retrieval model to agricultural disease prescriptions, i.e., a representative type of professional document in agriculture, and developed a prototype search engine for agricultural disease prescription. The experimental results on a real prescription collection show advantages of the proposed model to conventional information retrieval approaches.

Keywords Information retrieval, Agricultural disease prescription, Query expansion, Structural retrieval

1 引言

在各类现实应用领域中广泛存在着大量的专业文档,与传统意义上的纯文本相比,它们具有显著的特点:一方面,同一领域内的专业文档在内容上具有相关性,蕴含了丰富的领域知识;另一方面,专业文档在形式上体现出了很强的结构化特征,即一篇专业文档往往包含一些相对固定的字段。表1给出了农业领域中的一类专业文档——农技处方文档的一个示例。农技处方是指农业技术人员针对各类动植物病虫害所提供的处理方案,其中通常涵盖了某种农业病虫害的发病症状、发病原因、防治手段等各方面的信息。通过表1可以发现,农技处方文档中包含了症状、病原等多个相对固定的字段,因而可以被认为是一种"结构化文档"。除农技处方外,这种结构化特征在其他领域的专业文献中也普遍存在,比如药品说明文档中通常包含成份、性状、功能主治等字段;菜谱文档中通常包含用料、操作步骤、营养成分等字段;应急预案文档中通常包含指挥体系、预防机制、保障措施等字段;教案文档中通常包含教学目标、教学内容、作业、实验等字段。

这些结构化文档中存在的字段通常在语义上代表了相应的表达主题(症状、病因等),而不同的表达主题在单词分布等方面存在着很大差异;此外,字段之间也存在一定的关联,比如特定的症状(叶片皱缩、黄斑等)可能是由特定的病因(花叶病毒等)所导致,另外也决定了特定的治疗手段(抗毒剂等)。字段之间的差异性和关联性使得面向这类具有结构化特征的专业文档的信息检索任务与传统信息检索任务具有很大的差异。以农技处方文献检索为例,以下是两个典型的应用场景:

到稿日期:2014-05-20 返修日期:2014-07-21 本文受山东省优秀中青年科学家科研奖励基金(BS2012DX030),中国博士后科学基金(2012M521363),全国统计科学研究计划项目(2012LY001),山东省高校科技计划(J12LN45,J14LN33),山东省博士后创新项目专项(201303072)资助。

刘 彤(1982—),女,博士,讲师,主要研究方向为文本分类、数据挖掘;倪维健(1981—),男,博士,副教授,主要研究方向为数据挖掘、机器学习,E-mail:niweijian@gmail.com(通信作者)。

(1)农户在作物种植过程中遇到了某种病虫害,他向搜索引擎描述这种病虫害的症状,希望能够得到针对这种病虫害的有效的治疗方法;

(2)农户为感染了病虫害的作物采取了某种治疗方案,他向搜索引擎描述所使用的药品或措施,希望确定这些方法能否对症治疗所遇到的农业病虫害。

在上述两个应用场景中,如果农户使用了基于传统检索模型的搜索引擎,那么他极有可能无法得到满意的结果,主要原因在于传统检索模型主要采用了基于内容匹配的检索方法。假设当农户使用症状词作为查询词时,返回的检索结果将更倾向于对症状进行更为详细的描述,而非重点介绍与症状词可能对应的病虫害的治疗方法。因此,传统信息检索模型在直接应用于这种结构化领域文档的检索问题时会存在一定局限。

表1 农技处方文档示例

芹菜病毒病
一、症状 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 致病毒原为黄瓜花叶病毒(CMV)、芹菜花叶病毒(CeMV)。CMV 病毒呈颗粒球状,直径 28～30 纳米,病毒汁液稀释终点 1000～10000 倍,钝化温度 60～70℃,10 分钟,体外存活期 3～4 天。芹菜花叶病毒(CeMV)呈粒体线形,病毒汁液稀释限点 100～1000 倍,钝化温度 55～65℃…… 三、传播途径 CMV 和 CeMV 在田间主要通过蚜虫传播,也可通过人工操作接触磨擦传播。 四、发病条件 栽培管理条件差,干旱,蚜虫数量多发病重,夏季高温易发病。 五、防治方法 (1)选用抗病的芹菜品种。(2)加强水肥管理,提高植株抗病力,以减轻危害。春季栽培时,采取早育苗,简易覆盖或棚室栽培,以提早收获,避开蚜虫及高温等易发病的因素。高温干旱时期应搭棚遮阴。(3)定植时剔除病苗。(4)从苗期起及时防治蚜虫。(5)发病初期喷洒 1.5%植病灵乳油 1000 倍液、20%病毒 A 可湿性粉剂 500 倍液、0.5%抗毒剂 1 号水剂……

本文以领域文档中蕴含的这种结构化特征给传统信息检索模型带来的挑战为出发点,设计了一种新的结构化检索模型,以更为有效地解决结构化领域文档检索问题。具体而言,本文主要开展了如下两方面的工作:

(1)对同一领域内的结构化文档集进行挖掘,量化分析领域文档中文本与字段、字段与字段之间的关系,构建相应的领域模型,使之能够为后续的文档检索过程提供有益的线索;

(2)设计了能够有机融合领域知识的结构化检索方法,在评估结构化文档相关性时除了考虑文档与查询在字面上的匹配关系,还充分考虑了文档和查询中潜在的领域知识,以提升领域查询的检索效果。

本文第2节对相关工作进行了简要介绍,并概括了工作的创新点;第3节给出了所提出的结构化领域检索模型的整体框架,并详细介绍了其中的各个组成部分;第4节介绍了以农技处方文档为应用案例的实验研究和原型系统开发;最后对本文工作进行了总结和展望。

2 相关工作

信息检索模型作为搜索引擎的核心技术之一,一直是学术界和产业界关注的热点。经典的信息检索模型包括语言模型、向量空间模型、概率模型等,这些模型的一个特点是检索对象主要以纯文本为主,一般难以直接处理检索对象中除文字之外的其他特征。为此,近年来研究者们对传统信息检索模型进行了多种扩展,以更好地处理具有各种复杂结构的检索对象,其中与本文工作密切相关的工作主要包括如下两方面:

(1)结构化检索

结构化检索(Structural Retrieval)主要指针对具有结构化特征的文档的信息检索方法。一般而言,传统的信息检索模型可以通过适当扩展来更为有效地处理结构化文档。BM25F[1]和 BM25E[2]是经典的 BM25 模型的两个结构化扩展版本。BM25F 模型与 BM25 模型的区别在于并不在文档层面直接计算词频,而是首先在文档中的各个字段上计算词频,之后将各个字段上的词频进行加权合并作为整个结构化文档的词频;BM25E 模型对 BM25F 模型进行了进一步改进,在各个文档字段上计算词频时考虑了相关字段的影响。经典的语言模型也可以进行结构化扩展,其基本思想是在各个字段上分别构建语言模型,之后进行加权合并,代表工作包括层次语言模型(Hierarchical Language Model, HLM)[3]和混合字段语言模型(Mixture of Field Language Models, MFLM)[4]。HLM 和 MFLM 的主要问题在于模型合并权重是固定的,无法反映查询词与字段之间不同的相关度。为此,J. Kim 等提出了面向半结构数据的概率检索模型(Probabilistic Retrieval Model for Semi-structured Data, PRM-S)[5],该模型为每个查询词估计了字段后验概率,并将其作为字段语言模型合并权重;J. Kim 等进一步提出了字段相关度的概念,并设计了多种估计方法[6],使得模型合并权重能够融合文档集、相关反馈等多种知识。

由于结构化文档普遍存在于各种 Web 应用中,结构化检索技术具有广泛的应用范围,其中的一个典型应用是 XML 文档检索,目前 BM25 的各种结构化扩展版本在 XML 文档检索任务中均取得了较好的效果[7,8]。此外,研究者们还将结构化检索技术应用于个人简历、具有主谓宾结构的句子等结构化文本,解决了个人简历与工作岗位自动匹配[9]、自动问答[10]等应用问题。

本文同样以结构化文档上的信息检索方法为研究对象,已有工作大多仅侧重于对检索算法的设计,而本文则从结构化文档的领域属性出发,通过对同领域内结构化文档集的挖掘得到字段依赖性等各种领域知识,并将之融入至结构化检索过程中。

(2)主题模型在信息检索中的应用

主题模型是一种重要的文本挖掘方法,它将文档表示成一个关于主题的多项分布,同时每个主题是一个关于单词的多项分布[11]。由于主题模型能够在主题而非单词层面上描述文档的内容,因此很多研究者将之应用于信息检索过程中,以期得到更好的检索结果。

目前关于主题模型在信息检索中的应用研究主要集中在利用主题模型中提供的主题分布信息对基于语言模型的信息检索方法进行优化,优化手段包括文档语言模型估计、查询扩展和文档扩展。将主题模型用于文档语言模型估计的基本方法是根据主题模型中计算得到的文档主题分布以及主题单词分布,在所有的主题上利用全概率公式计算出文档的单词分布 $p(w|D)$,之后将该分布作为一个平滑项用于估计文档的语言模型[12]。基于主题模型的查询扩展的基本思想是把主

题模型中的一个主题当作一个“文档”，把所有的主题当作一个“文档集”，之后在这个“文档集”上利用相关性模型（Relevance Model）[13]等相关反馈技术对查询词进行扩展[12]。文档扩展的目标是利用指定文档的近邻文档集对该文档进行扩展表示，主题模型在应用于文档扩展时主要利用主题分布计算近邻文档权重 $p(D_j^{NN}|D)$ 和扩展词权重 $p(w|D_j^{NN})$，相比于传统文档扩展方法，基于主题模型的文档扩展方法能够有效避免文档偏置和噪音词对扩展结果的影响[14]。

大量实验结果表明，恰当地应用主题模型能够在一定程度上提升信息检索的效果[12]，然而在已有研究工作中很少涉及主题之间的关联性对信息检索效果的影响。结构化领域文档中的字段与主题模型中的主题类似，但是相比于主题间的关联性，领域知识的存在往往使得字段之间的关联性更为丰富，本文对字段关联性在信息检索过程中的应用进行了探索。

3 结构化领域检索模型

面向结构化领域文档的信息检索模型主要由离线挖掘和在线检索两个部分组成，其中离线挖掘部分的目标是对指定领域内的结构化文档集进行挖掘，得到字段与单词、字段与字段之间的关联关系模型，以此作为用于结构化检索的领域模型；在线检索部分的目标是结合离线挖掘得到的领域知识库，设计面向领域文档的结构化检索算法，为用户查询返回相关的领域文档。

3.1 面向信息检索的结构化领域模型

3.1.1 字段归属模型

字段归属模型的目标是对单词与字段之间的关系进行量化。假设给定结构化领域文档集 $\mathcal{D}=\{D_1,\cdots,D_N\}$，对应的词表 $V=\{w_1,\cdots,w_M\}$，文档集中存在 k 个字段 $F=\{f_1,\cdots,f_K\}$，那么在字段归属模型中为每个单词 $w\in V$，估计如下多项分布：

$$\{p(f_i|w)|i=1,\cdots,K\}$$

根据贝叶斯公式，给定单词的字段后验概率可以通过字段先验概率和单词似然计算得到，即：

$$p(f_k|w)=\frac{(w|f_k)p(f_k)}{p(w)}=\frac{p(w|f_k)p(f_k)}{\sum_{k=1}^{K}p(w|f_k)p(f_k)} \tag{1}$$

利用给定的结构化领域文档集，可以估计得到字段先验概率 $p(f_k)$ 以及单词条件似然 $p(w|f_k)$。字段先验概率可以通过包含指定字段的文档占全体文档的比例进行估计，即：

$$p(f_k)=\frac{\sum_{i=1}^{N}1_{(f_k\in D_i)}}{N} \tag{2}$$

需要注意的是，任意字段的先验概率并不总等于 $1/K$，主要原因是领域内的 K 个字段并不总是同时出现在每一篇领域文档中。比如在农技处方文档中，“病原”字段通常出现在描述作物病害的文档中，而在描述作物虫害的文档中则较少出现。

单词条件似然刻画了在指定字段下单词出现的概率，它可以通过统计文档集中该字段内单词出现的频率得到，计算方式如下：

$$p(w|f_k)=\frac{1}{N}\cdot\sum_{i=1}^{N}(\lambda\cdot tf(\omega,D_i^{f_k})+(1-\lambda)\cdot tf(w,D_i)) \tag{3}$$

其中，D_i 和 $D_i^{f_k}$ 分别是第 i 个文档以及其中第 k 个字段的内容。上式右侧第一项表示文档字段内的单词频率，由于文档中单一字段的内容比较稀疏，因此应用文档内的单词频率作为平滑项（上式右侧第二项）。λ 是平滑项调节权重，通常 $0.5\leqslant\lambda<1$。

3.1.2 字段关联模型

字段关联模型的目标是对字段与字段之间的关系进行量化。字段之间的关联性主要通过字段内的单词来体现，因此首先挖掘单词之间的关联关系，结合之前构建的字段归属模型，即可得到对字段关联性的形式化描述。

单词之间的关联关系具有多种形式，如一对一关系、一对多关系、多对多关系等。传统的单词关联性分析方法主要面向单词对（一对一关系），一般并不针对多对多关系进行分析[15]，而在领域文档中，一对一关系并不能充分地反映领域知识。为此，本文在字段关联模型中刻画了单词之间的多对多关联关系。

具体而言，字段关联模型由一组单词集关联规则构成，每个单词集关联规则形式如下：

$$R:R^{ant}\Rightarrow R^{con}$$

其中，R^{ant} 和 R^{con} 分别表示规则前件和后件，它们是领域词表的任意子集，即 $R^{ant},R^{con}\in 2^V$。对于每个规则，字段关联模型给出了两个量化指标：支持度（support）和置信度（confidence）。

支持度是指在规则前件和后件中所有单词同时出现在一个领域文档中的概率，它可以通过对给定的领域文档集进行统计得到，计算公式如下：

$$sup(R)=p(R^{con}\cup R^{ant})=\frac{\sum_{i=1}^{N}1_{(R^{ant}\cup R^{con}\subseteq D_i)}}{N} \tag{4}$$

一个单词集关联规则的支持度越大，说明它能够覆盖越多的领域文档，从而对于后续的文档检索过程能够产生越大的影响。因此，只有支持度高于某指定阈值的规则才被纳入字段关联模型中。

置信度是指规则后件中的所有单词出现在一个出现了规则前件中所有单词的领域文档中的概率，它同样可以通过统计方式得到，计算公式如下：

$$conf(R)=p(R^{con}|R^{ant})=\frac{\sum_{i=1}^{N}1_{(R^{ant}\cup R^{con}\subseteq D_i)}}{\sum_{i=1}^{N}1_{R^{ant}\subseteq D_i}} \tag{5}$$

一个单词集关联规则的置信度越大，表示前件单词集有更大的能力“预测”后件单词集的出现，即后件单词集和前件单词集之间具有更强的关联关系。因此，将置信度高于某阈值的规则纳入字段关联模型中，以反映领域单词集之间的关联性。

在领域文档集中，大量的领域知识可以通过单词集关联规则来体现。以农技处方文档为例，假设芦荟出现了黄化、矮化等现象，那么一种常见的处理方式是追肥，而不宜过多浇水。这种领域知识可以体现为如下单词集关联规则：

$$R_1:\{芦荟,黄化,矮化\}=\{追肥\}$$

$$R_2:\{芦荟,黄化,矮化\}=\{浇水\}$$

其中，R_1 的支持度和置信度将均高于 R_2。

给定某领域文档集，可以使用传统关联规则挖掘算法获得满足最小支持度和置信度阈值的单词集关联规则。为了能够利用传统的关联规则挖掘算法，需要将每个单词看作一个

项，将每篇领域文档看作一个事务，进而可以把给定的领域文档集转换成传统关联规则中的事务集形式，之后应用 FP-growth[16] 等频繁项集挖掘算法得到频繁单词集。最后，根据各个单词集在领域文档集中出现的概率，即可得到每个规则的置信度。通过指定一个最小置信度阈值，在字段关联模型中仅保留置信度较高的规则。

需要特别指出的是，由领域文档集得到的事务数据集具有一个显著的特点：每个事务中项的数量（取决于对应领域文档中包含的不同单词的数量）通常大于数据库领域中常见的事务，这使得频繁项集挖掘算法的效率大为降低。为了解决该问题，在进行事务转换之前首先对领域文档进行预处理，即过滤掉领域文档中的一些单词，缩减事务长度。由于字段关联模型主要反映字段之间的关联关系，而那些在各个字段中均频繁出现的单词很难成为字段关联关系的载体，因此本文计算了字段归属模型中每个单词的字段分布的信息熵：

$$H(w)=-\sum_{k=1}^{K}p(f_k|w)\cdot\log p(f_k|w) \tag{6}$$

如果一个单词的信息熵越大，说明它在各个字段中出现的概率越均等，从而具有比较弱的字段属性。通过指定一个最大信息熵阈值，在领域文档中过滤掉高于该阈值的单词，进而实现缩减事务长度的目标。这种缩减方法一方面能够保证频繁项集挖掘算法的效率，另一方面使得挖掘得到的规则侧重于反映字段之间单词集而非字段内部单词集之间的关联关系。

3.2 结构化领域文档在线检索方法

在得到结构化领域模型之后，本文进一步设计了结构化检索算法，该算法的主要特点是能够将前期通过挖掘领域文档集得到的结构化领域模型有机融合至检索过程，进而优化检索效果。结构化检索算法具体由结构化查询扩展算法和结构化相关性算法两部分组成。

3.2.1 结构化查询扩展算法

在结构化文档检索任务中，用户查询依然采用无结构的形式，但通常隐含着结构化的查询意图。比如，用户输入查询为“芦荟长得矮如何施肥”，那么其查询意图可以表示为：“作物：芦荟；症状：长得矮；治疗方法：施肥”。为此，我们对用户领域查询进行了结构化扩展，使之能够更好地与结构化文档进行匹配。结构化查询扩展主要包含两个阶段：首先利用字段关联模型对用户在多个字段上的查询意图进行推理和扩展，之后利用字段归属模型获得用户在各个字段上的查询意图权重。

基于字段关联模型的查询扩展的基本思想是使用单词集关联规则中的后件单词对前件查询词进行扩展，具体流程如算法 1 所示。在查询扩展算法中，首先初始化一个用于保存每个规则匹配查询词数量的哈希表（行 1－3）；之后，在字段关联模型中查找规则前件能够匹配查询词的规则集（行 4－9），并根据查询词匹配数量将它们进行合并，得到规则前件匹配了所有查询词的规则集 $\mathcal{R}_Q$（行 10）；再后，把 $\mathcal{R}_Q$ 中每个规则的前件中的单词作为一个扩展查询词，其权重由该规则的证据度决定（行 12－14）；最后，对原始查询词和扩展查询词的权重进行归一化处理（行 16－19），返回扩展之后的查询 Q'。

算法 1 基于字段关联模型的查询扩展算法

输入：字段关联模型：$\mathcal{R}=\{R_1:R_1^{ant}\Rightarrow R_1^{con},\cdots,R_L:R_L^{ant}\Rightarrow R_L^{con}\}$

用户查询：$Q=\{q_1,\cdots,q_U\}$

输出：扩展后的用户查询及查询词权重：$Q'=\{(q_1',w_1'),\cdots,(q_s',w_s')\}$

1. foreach $R\in\mathcal{R}$ do
2. $C\{R\}=0$;
3. endfor
4. foreach $q\in Q$ do
5. $\mathcal{R}_q=\{R|(R\in\mathcal{R})\wedge(q\in R^{ant})\}$;
6. foreach $R\in\mathcal{R}_q$ do
7. $C\{R\}++$;
8. endfor
9. endfor
10. $\mathcal{R}_Q=\{R|(R\in\mathcal{R}_q)\wedge(|R^{ant}|=C\{R\})\wedge(q\in Q)\}$;
11. $E=\emptyset$; norm=0;
12. foreach $R\in\mathcal{R}_Q$ do
13. $EU=\{(t,w)|(t\in R^{con})\wedge(w=att(R)/|R^{con}|)\}$;
14. norm+=w;
15. endfor
16. foreach $(t,w)\in E$ do
17. $w/=2\cdot norm$;
18. endfor
19. $Q'=\{(q_1,1/2U),\cdots,(q_U,1/2U)\}\cup E$;

算法 1 的时间复杂度主要由行 4－9 决定，行 5 中在字段关联模型中查找前件匹配查询词的所有规则是其中最为耗时的操作。为了能够提高查找效率，我们设计了一种倒排索引来存储字段关联模型中的所有规则。在该倒排索引中，以每个单词作为索引项，每个单词对应的倒排列表中记录了前件中包含该单词的所有规则。由于可以为所有单词构建查找效率为 O(1) 的索引表，所以算法 1 整体的时间复杂度为 $O(|Q|\cdot|\mathcal{R}_Q|)$，即算法效率由用户查询长度和被匹配到的规则的数量共同决定。

在得到扩展的查询词及其权重后，我们利用字段归属模型对用户查询意图在各个字段上的分布进行了进一步的量化。假设扩展后的用户查询为 $Q'=\{(q_1',w_1')\cdots,(q_s',w_s')\}$，可以得到如下矩阵形式的结构化用户查询：

$$Q'=\begin{array}{c} \\ q_1' \\ \\ w_s' \end{array}\begin{array}{c} \begin{array}{ccc} f_1 & \cdots & f_K \end{array} \\ \begin{bmatrix} w_{1,1} & \cdots & w_{1,K} \\ \vdots & \ddots & \vdots \\ w_{s,1} & \cdots & w_{s,K} \end{bmatrix} \end{array} \tag{7}$$

在用户查询矩阵中的每个元素 $w_{i,j}$ 表示查询词 q_i' 在第 k 个字段上的查询强度，其计算方式如下：

$$w_{i,j}=w_i'\cdot p(f_j|q_i') \tag{8}$$

在式(8)中，$\sum_{j=1}^{K}w_{i,j}=w_i'$。因此可以认为，本文利用字段归属模型把查询词权重在各个领域字段上进行了分配。

3.2.2 结构化相关性算法

用户查询经过结构化扩展之后，体现为一个式(7)所示的查询矩阵的形式，而传统的信息检索算法在计算文档相关性时均将查询处理为单词空间中的向量形式，这使得它们无法被直接应用于结构化领域文档检索任务中。在研究过程中，为了重点突出领域模型和查询扩展对结构化文档检索效果提升的作用，对经典的相关性算法进行了简单改造，而不再设计复杂的全新算法。

具体而言，本文对 BM25 这一被大量实践所证明的简单

有效的相关性算法进行结构化改造。经典的 BM25 相关性计算公式如下：

$$rel(Q,D)=\sum_{q\in Q}\frac{(k_1+1)\cdot tf(q,D)}{k_1\cdot(1-b+b\cdot\frac{|D|}{avgdl})+tf(q,D)}\cdot \log\frac{N-df(q)+0.5}{df(q)+0.5} \quad (9)$$

在式(9)中主要有两个影响相关性的因素：词频 $tf(q,D)$ 和文档频率 $df(q)$，在传统的计算方法中，两者均没有考虑文档内部的结构。我们重点对词频的计算方式进行了结构化改造，其基本思想是在各个字段上分别计算查询词的词频，并根据用户查询在各个字段的权重对之进行合并。

假设给定由 K 个字段组成的领域文档 $D=\{D^{f1},\cdots,D^{fK}\}$，文档中的每个字段由纯文本构成，对于式(7)所示的用户查询矩阵中的查询词 $q_i'(1\leqslant i\leqslant S)$，其在文档 D 中的词频计算方式如下：

$$tf^{struct}(q_i',D)=\sum_{i=1}^{K}tf(q_i',D^{fj})\cdot w_{ij} \quad (10)$$

其中，$tf(q_i',D^{fj})$ 为查询词 q_i' 在文档第 j 个字段中的词频，w_{ij} 对应着式(7)查询矩阵中的权重。

传统结构化检索方法中的一个基本方法是在各个字段上分别计算查询与文档的相关性得分，之后在各个字段上将相关性得分进行加权合并。相比于这种传统方法，提出的结构化相关性算法的主要优势在于：

(1) 参数数量少：我们仅对词频的计算方法进行了改造，保留了传统的相关性公式，使得参数的数量(k_1，b)与传统方法保持一致，而且在调参方法上可以从传统方法中得到借鉴；但是如果在各个字段上分别应用传统的相关性公式进行计算，那么参数的数量将会增加到原有公式的 K 倍。

(2) 保留了相关性原有的物理意义：保留传统的相关性公式，使得已有的针对传统相关性算法的理论分析成果可以直接应用到本文的结构化文档相关性上；由于文档字段在文本长度、单词分布等各方面上存在很多差异，使得在各个字段上计算得到的相关性之间并不具有可比性，直接将它们进行线性合并，使得相关性原有的物理意义不再存在。

4 实验结果与分析

4.1 实验数据集

本文选取农技处方检索作为典型案例对结构化领域文档检索方法进行了应用。在领域文档集方面，我们手工选择了多个农业技术类网站，通过编写定向网络爬虫抓取了一系列与农技处方相关的网页。抓取得到的农技处方文档中，只有一部分文档被作者注明了“症状”、“病原”等字段(见表 1)，而其它文档并没有这种显式的标记。因而，将具有标注信息的农技处方文档作为训练集，应用机器学习方法对未标记文档进行自动字段标注，具体方法可参见文献[17]。最终用于本文实验的领域文档集中共包含 10712 篇农技处方文档，所有文档中共包含“作物品名”、“分布危害”、“症状形态”、“病原病因”、“预防治疗”等 5 个字段。

在用户查询方面，获取了某商业搜索引擎提供的查询日志，从中手工提取了 58 个与农技病虫害检索有关的查询(为了便于展示结构化检索效果，我们对部分查询词进行了改动)。表 2 展示了实验中使用的部分查询词。对于每个查询词，在农技处方文档集中为之人工标注了若干相关的文档。为简单起见，在实验中没有对相关性进行分级，检索得到的文档仅被分为相关和不相关两类。

表 2 查询词示例

序号	查询词
1	杜鹃花叶子变干脱落
2	小竹子好容易枯萎
3	四季风 TY1 号、TY2 号防控番茄病毒病
4	芦荟长得矮如何施肥
5	冬季室内植物为何枯萎
6	小麦类似黄色蚊子的害虫使用什么药剂
7	栀子花叶子发黄腐烂
8	梨树叶子上有铁锈
9	仙客来花骨朵未开就枯萎了是怎么回事
10	水稻叶子褐色斑点缺素

4.2 检索结果对比

分别选择了经典的信息检索方法——基于 KL 散度的语言模型(KLLM)和语言模型的一种结构化检索版本——层次语言模型(HLM)与本文提出的结构化领域文档检索方法进行对比。从实验数据集中随机抽样了一个验证集用于为这两个基准方法选择较优的参数。在检索结果评价指标方面，实验中使用两类常见的信息检索评估指标——MAP (Mean Average Precision)和 Pr@n (Precision at n，n=1,5,10)。

表 3 列出了提出的结构化领域文档检索方法(Domain-specific Structural Retrieval)和基准方法的检索效果。可以发现，本文方法在大多数指标上(MAP、Pr@1、Pr@5)取得了比基准方法更高的值，特别是在 Pr@1 指标上相比于基准方法获得了最大的提升(提升率分别为 9%和 10%)。其主要原因在于通过利用字段关联模型对用户查询进行扩展，有助于使相关的文档获得更好的得分，而且一些包含了被反向扩展的查询词的文档的得分会受到相应的惩罚，这些因素均有助于将相关的文档排至检索结果的顶端。此外，还可以发现结构化检索方法(HLM)比传统信息检索方法(KL-LM)取得了更好的检索结果，特别在 Pr@10 指标获得了最高值，这说明结构化检索方法能够综合考虑多个字段的影响，保证相关文档均得到较高的评分。

表 3 检索结果对比

	MAP	Pr@1	Pr@5	Pr@10
KLLM	0.3131	0.5041	0.4660	0.3576
HLM	0.3230	0.4988	0.4703	**0.3792**
DSR	**0.3319**	**0.5490**	**0.4871**	0.3741

4.3 各检索模块影响力分析

为了能够更好地验证本文提出的模型中各个组成部分对检索结果的影响，将各个组成部分变为可选项，得到了一系列模块化的领域文档检索方法，如表 4 所列。在表 4 中，结构化领域文档检索模型中的领域模型、查询扩展算法、相关性算法被独立成 3 个模块，其中领域模型模块有两种选择(使用、不使用)，查询扩展算法有 3 种选择(不使用、使用基于关联规则的扩展方法[18]、使用基于字段关联模型的扩展方法)，相关性算法有两种选择(经典 BM25 公式、本文提出的结构化 BM25 公式)。可以发现，表 4 中的 DSR-null-null 方法即是经典的 BM25 检索算法，DSR-DM-FA 即是本文提出的领域文档检索方法。

表 4　模块化检索方法

	领域模型	查询扩展	相关性算法
DSR-null-null	无	无	BM25
DSR-null-AR	无	基于关联规则	BM25
DSR-DM-null	有	无	$BM25^{struct}$
DSR-DM-FA	有	基于字段关联模型	$BM25^{struct}$

图 1 展示了表 4 中的 4 种检索方法的评估结果，从中可以得出以下结论：

(1)DSR-DM-null 取得了比 DSR-null-null 更优的检索效果，这说明相比于经典的 BM25 算法，利用字段归属模型改造后的 BM25 相关性计算公式更加适合结构化文档的检索任务。

(2)DSR-DM-FA 取得了比 DSR-DM-null 显著更优的检索效果，这说明利用字段关联模型进行查询扩展是提升结构化文档检索效果的主要因素。

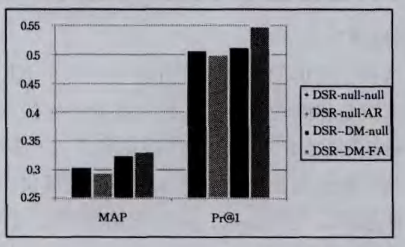

图 1　模块化检索方法结果对比

DSR-null-AR 与 DSR-null-null 相比，前者取得的效果提升并不显著，甚至有所降低，这说明单纯使用关联规则进行查询扩展并不能很好地提升检索效果，其主要原因在于基于关联规则的查询扩展方法倾向于把一些通用词作为扩展词，而这些词一般不会增加文档匹配的精确性。这也在另一方面验证了本文所提出的综合运用字段归属模型和字段关联模型的查询扩展方法相比于单纯使用关联规则更加适用于信息检索中的查询扩展任务。

4.4　检索效率分析

本文提出的结构化领域文档检索方法包括在线和离线两部分。对于在线部分，查询扩展算法(算法 1)具有线性时间复杂度，这与各类经典的查询扩展算法具有相似的复杂度；相关性算法本质上是经典 BM25 公式的一种变形，它与经典的 BM25 算法也具有相同的时间复杂度。因此，本文方法中相比于传统信息检索方法更为耗时的是离线部分，故我们主要对字段关联模型离线挖掘的效率进行了实验分析。

使用 4.1 节描述的农技处方文档集挖掘单词集关联规则将文档进行事务数据转换，并过滤掉一些重复、无效事务后，事务数据集中共包含 9992 个事务，8518 个项，每个事务平均包含 152 个项。该实验使用 FP-growth 算法挖掘频繁单词集，硬件上使用了 Xeon E5640 CPU。

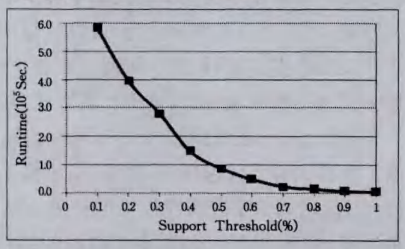

图 2　离线挖掘时间变化趋势图

图 2 描述了字段关联模型挖掘时间随单词集关联规则最小支持度的变化曲线。可以发现，随着支持度阈值的减小，字段关联模型挖掘时间呈指数级上升趋势。其主要原因是挖掘时间主要由频繁单词集挖掘算法决定，而该算法通常具有指数级的时间复杂度。根据图中曲线，如果把支持度阈值设为 0.1%，那么挖掘过程通常可以在一周内完成，这对于离线挖掘任务来说基本是可以接受的。

4.5　原型系统开发

基于提出的检索模型，开发了一个农技处方检索引擎。图 3 展示了该原型系统的主要界面。我们基于开源搜索引擎 Indri 开发了该原型系统中的农技处方文档索引和 BM25 相关性算法模块。用户可以向该系统中输入自然语言查询，同时该系统也支持分面搜索(facet search)，其中每个分面是农技处方中的一个字段。系统的检索结果主要由排序后的相关的农技处方文档组成，与传统搜索引擎直接为用户返回每个文档的摘要不同，该系统为相关农技处方文档的每个字段生成了摘要展示给用户，并且列出了相关的农作物病虫害图片。通过这些手段，该系统在一定程度上提升了农技检索的易用性。

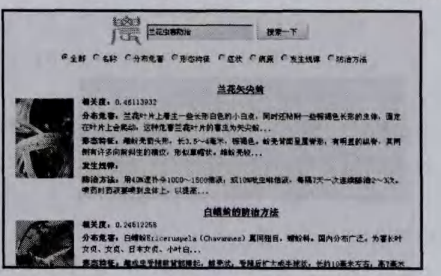

图 3　原型系统界面

结束语　本文提出了一种面向结构化领域文档的信息检索模型。相比于传统的信息检索方法和结构化检索方法，本文方法侧重于对领域文档集中所蕴含的领域知识的挖掘，通过构建结构化领域模型为领域文档检索提供有益的支持；同时通过设计结构化用户查询算法，以及对经典 BM25 相关性公式进行改造，在检索过程中实现了对领域模型的有效利用。通过在农技处方文档检索任务上的实验和分析，证实了本文所提出的结构化领域检索框架是完整、有效的。

本文工作尚存在一定改进空间，主要包括：(1)受困于公开的结构化检索数据集的获取难度，前期仅在农技处方文档上开展了实验研究，后续我们将通过各种手段获得更多的结构化检索数据集，进一步丰富实验结果；(2)结构化领域模型蕴含了丰富的领域知识，如何结合专家经验对之进行合理解释，并将之应用于其它的领域文本挖掘任务，是一个有意义的研究课题；(3)目前针对结构化相关性算法的研究还相对简单，设计更为精细的相关性算法对于提升领域检索的效果将有着重要的作用。

参 考 文 献

[1] Robertson S, Zaragoza H, Taylor M. Simple BM25 Extension to Multiple Weighted Fields[C]// Proceedings of the 13th ACM CIKM. Washington DC, USA, 2004: 42-49

[2] Lu W, Robertson S, MacFarlane A. Field-Weighted XML Retrieval Based on BM25[C]// Proceedings of the 5th Workshop of INEX. Germany, 2006: 161-171

[3] Ogilvie P, Callan J. Hierarchical language models for XML component retrieval[C]// Proceedings of the 4th Workshop of INEX. Germany, 2005: 224-237

(下转第 286 页)

An Li-ping, Chen Zeng-qiang, Yuan Zhu-zhi. Multi attribute decision analysis based on rough set theory [J]. Control and Decision, 2005, 20(3): 294-298

[7] 马峻，吉晓民. 利用粗糙集理论实现工艺决策的冲突消解[J]. 计算机辅助设计与图形学报，2005，17(3)：600-604
Ma Jun, Ji Xiao-min. Implementation of Conflict Resolution for Process Decision Based on Rough Theory [J]. Journal of Computer Aided Design & Computer Graphics, 2005, 17(3): 600-604

[8] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001
Wang Guo-yin. Rough set theory and knowledge acquisition [M]. Xi'an: Xi'an Jiaotong University Press, 2001

[9] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001
Zhang Wen-xiu, Wu Wei-zhi, Liang Ji-ye, et al. The rough set theory and method [M]. Beijing: Science Press, 2001

[10] Hu X, Cercone N. Learning in relational databases: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338

[11] Swiniarski R W, Skowron A. Rough set methods in feature selection and recognition[J]. Pattern Recognition Letters, 2003, 24(6): 833-849

[12] Felix R, Ushio T. Rough sets-based machine learning using a binary discernibility matrix[C] //Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials, 1999(IPMM'99). IEEE, 1999: 299-305

[13] 杨萍，李济生，黄永宣. 一种基于二进制区分矩阵的属性约简算法[J]. 信息与控制，2009，38(1)：70-74
Yang Ping, Li Ji-sheng, Huang Yong-xuan. A attribute reduction algorithm based on binary Discernibility Matrix [J]. Information and Control, 2009, 38(1): 70-74

[14] 张颖淳，苏伯洪，曹娟. 基于粗糙集的属性约简在数据挖掘中的应用研究[J]. 计算机科学，2013，40(8)：223-226
Zhang Ying-chun, Su Bo-hong, Cao Juan. Study on application of Attributive Reduction Based on Rough set in Data mining [J]. Computer Science, 2013, 40(8): 223-226

[15] 常犁云，王国胤，吴渝. 一种基于 Rough Set 理论的属性约简及规则提取方法[J]. 软件学报，1999，10(11)：1206-1211
Chang Li-yun, Wang Guo-yin, Wu Yu. A Method of Attribute Reduction and Rule Extraction Based on Rough Set Theory[J]. Journal of Software, 1999, 10(11): 1206-1211

[16] 鄂旭，邵良杉，张毅智，等. 一种基于粗糙集理论的规则提取方法[J]. 计算机科学，2011，38(1)：232-235
E Xu, Shao Liang-shan, Zhang Yi-zhi, et al. Method of Rule Extraction Based on Rough Set Theory [J]. Computer Science, 2011, 38(1): 232-235

[17] 张利，卢秀颖，吴华玉，等. 基于粗糙集的启发式值约简的改进算法[J]. 仪器仪表学报，2009(1)：82-85
Zhang Li, Lu Xiu-ying, Wu Hua-yu, et al. Improved heuristic algorithm used in attribute value reduction of rough set [J]. Chinese Journal of Scientific Instrument, 2009(1): 82-85

[18] Suresh B V, Viswanath P. Rough-fuzzy weighted k-nearest leader classifier for large data sets[J]. Pattern Recognition, 2009, 42(9): 1719-1731

[19] Astudillo C S A, Oommen B J. On achieving semi-supervised pattern recognition by utilizing tree-based SOMs[J]. Pattern Recognition, 2013, 46(1): 293-304

[20] 任靖，李春平. 最小距离分类器的改进算法-加权最小距离分类器[J]. 计算机应用，2005，25(5)：992-994
Ren Jing, Li Chun-ping. Improved minimum distance classifier-weighted minimum distance classifier [J]. Computer Applications, 2005, 25(5): 992-994

(上接第 280 页)

[4] Ogilvie P, Callan J. Combining document representations for known-item search[C]// Proceedings of the 26th ACM SIGIR. Toronto, Canada, 2003: 143-150

[5] Kim J, Xue X, Croft W B. A Probabilistic Retrieval Model for Semistructured Data[C]// Proceedings of the 31th ECIR. Toulouse, France, 2009: 228-239

[6] Kim J, Croft W B. A Field Relevance Model for Structured Document Retrieval[C]// Proceedings of the 34th ECIR. Barcelona, Spain, 2012: 97-108

[7] Itakura K Y, Clarke C L. A framework for BM25F-based XML retrieval[C]// Proceedings of the 33rd ACM SIGIR. Geneva, Switzerland, 2010: 843-844

[8] 刘德喜，万常选，刘喜平，等. 基于结点权重模型的 XML 片段检索策略[J]. 计算机学报，2013，36(8)：1729-1744
Liu, De-xi, Wan Chang-xuan, Liu Xi-ping, et al. A Snipet Retrieval Strategy Based on Element Weighting Model[J]. Chinese Journal of Computers, 2013, 36(8): 1729-1744

[9] Yi X, Allan J, Croft W B. Matching resumes and jobs based on relevance models[C]// Proceedings of the 30th ACM SIGIR. Amsterdam, 2007: 809-810

[10] Zhao L, Callan J. Effective and Efficient Structured Retrieval [C]// Proceedings of the 18th ACM CIKM. Hong Kong, China, 2009: 1573-1576

[11] Blei D M, Ng A Y, Jordan M I. Latent Dirichletallocation[J]. Journal of Machine Learning Research, 2003, 3(4/5): 993-1022

[12] Yi X, Allan J. A Comparative Study of Utilizing Topic Models for Information Retrieval[C]// Proceedings of the 31th ECIR. Toulouse, France, 2009: 29-41

[13] Lavrenko V, Croft W B. Relevance-based language models[C]// Proceedings of the 24th ACM SIGIR. New Orleans, Louisiana, USA, 2001: 120-127

[14] Ganguly D, Leveling J, Jones G J F. An LDA-smoothed relevance model for document expansion: a case study for spoken document retrieval[C]// Proceedings of the 36th SIGIR. Dublin, Ireland, 2013: 1057-1060

[15] Bai J, Song D, Bruza P, et al. Query Expansion Using Term Relationships in Language Models for Information Retrieval[C]// Proceedings of the 14th CIKM. Bremen, Germany, 2005: 688-695

[16] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]// Proceedings of SIGMOD. Dallas, Texas, USA, 2000: 1-12

[17] Liang Y, Liu T, Ni W. Augmented Vector Space Model for Passage Intention Classification in Chinese Agricultural Prescription Documents[J]. Journal of Computational Information Systems, 2014, 10(1): 101-108

[18] Songa M, Song I-Y, Hu X, et al. Integration of association rules and ontologies for semantic query expansion[J]. Data & Knowledge Engineering, 2007, 63(1): 63-75