基于改进 VSM 的农业实用技术自动问答系统研究

罗长寿,张峻峰,孙素芬,魏清凤　(北京农林科学院农业科技信息研究所,北京 100097)

摘要　针对农业实用技术需求及传统向量空间模型在农业特定领域应用中存在的问题,从特征词规范化、文档特征词专业性权值修正、查询特征词分布权值修正、系统效率优化 4 方面对其改进,并进行农业技术自动问答系统开发。结果表明,改进后的计算模型提高了自动问答的准确率和召回率,且系统检索性能也得到明显改善。

关键词　向量空间模型;农业;自动问答系统
中图分类号　S126　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)28 - 13948 - 03

Study on the Question-Answer System of Agricultural Practical Technology Based on Improved VSM

LUO Chang-shou et al　(Agricultural Science Information Institute, Beijing Academy of Agriculture and Forestry science, Beijing, 100097)

Abstract　According to the existed problems of agricultural practical technology requirement and traditional VSM when they applied to the agricultural Field, the paper solved the problems by characteristic term standardization, the weight of specialized word revision, the inquiry of characteristic words distribution revision, and system efficiency optimizer, and developed a Question-Answer system based on the improved VSM. The results showed that the improved computer model raised theaccuracy , recall rate of and the system retrieval performance.

Key words　VSM; Agriculture; Question-Answer System

在近年中央 1 号文件的号召及支持下,农业信息资源激增,解决"最后一公里"问题的信息服务备受关注。基于自动问答模式的解决方案,以简单的自然语言提问方式提供准确的答案信息成为了目前的研究热点,它与根据关键词检索并返回一堆相关信息的传统搜索引擎(如 google、Baidu)有着根本的区别[1-2]。农业生产离不开实用技术。由于农业特定领域的自动问答系统不多见,而在一般开放域的问答系统中寻求专业技术帮助,回答满意率较低,因此研究开发农业实用技术自动问答系统显得十分必要。

自动问答的经典检索模型——向量空间模型(VSM)在文本信息处理领域中一直占据着非常重要的地位[3-4]。但由于其对组成文档的各特征词之间关系做了相互独立的正交假设,而在农业技术信息服务应用中,这一假设难以满足,使得模型计算度量与实际结果有较大的误差。笔者研究分析 VSM 在农业技术服务应用中存在的具体问题,通过特征词规范化,考虑文档特征词专业性和查询特征项分布对权重的影响来修正向量空间模型,并采用效率优化策略进行实用技术智能问答系统开发,以提高系统的查准率和检索效率,方便高效地为涉农用户提供实用技术信息服务。

1　传统向量空间模型及应用分析

1.1　传统向量空间模型基本原理　传统向量空间模型(VSM)提出,在若干特征词的向量空间下,将查询语句与文档按照系特征词的维度分别量化,然后计算两向量间夹角余弦以得到查询与文档的相似度,并按其排序,优先检索相似度大的文档[5]。实践中,用特征词表示文本特征项,用 TF - IDF(词频—逆文档词频)进行特征项权重计算,用特征项权重构建向量,用向量夹角余弦进行文本相似性度量,据相似度值进行倒排文档索引。

若用户提问与答案文档包含 n 个特征词,则在该特征词空间下:

(1)任一文档 $d_i \in D$(D 为文档集合),可描述为由若干权值构成的 n 维向量:

$$d_i = (W_1^{\ d}, W_2^{\ d}, \cdots, W_n^{\ d})$$

$$W_i^{\ d} = TF \times IDF = (f_i/c_i) \times \log(N/n_i)\ (1 \leqslant i \leqslant n) \qquad (1)$$

其中,分量 $w_i^{\ d}$ 为第 i 个特征词 T_i 在文档 d_i 中的所有权重,由词频 TF 与反词频 IDF 的积得来。f_i 为特征词 T_i 在文档 d_i 中出现的次数,c_i 为文档 d_i 的总词数,为避免长文档包含特征项的次数多,TF 用 f_i/c_i 对其进行归一化。N 为文档集合 D 的总文档数,n_i 为 D 中含有特征词 T_i 的文档总数。

(2)任一提问查询语句 q_j,可描述为若干查询特征词串的 n 维向量[6]:

$$q_j = (W_1^{\ q}, W_2^{\ q}, \cdots W_n^{\ q})$$

$$W_j^{\ q} = \begin{cases} 1 & T_j \in q_j \\ 0 & T_j \notin q_j \end{cases} \qquad 1 \leqslant j \leqslant n \qquad (2)$$

其中,分量 $W_j^{\ q}$ 为第 j 个特征词 T_j 在提问问题 q_j 中的权重。

(3)提问查询语句 q_j 与文档 d_i 的相似度可以利用各自转化形成向量在 n 维空间的相对位置来决定,较常采用相似度计算指标是两个向量的夹角余弦函数:

$$\mathrm{sim}(d_i, q_j) = \frac{d_i \times q_j}{|d_i| \times |q_j|} = \frac{\sum_{i=1}^{t} W_{ij} \times W_{iq}}{\sqrt{\sum_{i=1}^{t} W_{ij}^{\ 2} \times \sqrt{\sum_{i=1}^{t} W_{iq}^{\ 2}}}} \qquad (3)$$

1.2　VSM 在农业特定领域的应用及改进分析　VSM 成功地将非结构化的文本信息表示成向量形式,为文本信息处理和操作奠定了数学计算基础[7]。它所采用的权值计算方法——TF/IDF,综合考虑了词在单个文档的出现频率(TF)和该词在整个文档集中的分辨能力(IDF),是目前信息检索领域广泛采用且效果较好的检索模型[8-9]。但在该特定领域的应用中有其不足,存在的问题主要有:

农业词语的多种表达在基于词形匹配的检索中,漏检现象严重。VSM 采用词形匹配加权处理模式,缺乏对词同义不

基金项目　北京市农林科学院青年基金(2007030610);北京农委农业科技项目"农业实用技术信息供求智能对接系统应用推广"(20080801)。

作者简介　罗长寿(1974 -),男,山西大同人,博士,副研究员,从事农业信息技术研究。

收稿日期　2009-06-08

同形情况的考虑。如常用氮肥碳酸氢铵,被俗称为“碳铵”,又表达成“NH_4HCO_3”,在查询其有关施用方法时,其他表达方式则因词形不匹配而无法被检索到。因此,可对提问语句和答案文档的特征词进行规范化统一描述,在同形词的基础上计算权值(详请),以避免同义不同形信息的漏检。

缺乏对特征词专业重要性考虑,削弱了该类词对主题的贡献力。VSM 侧重于考虑词频因子来衡量特征项权重,在文档中出现频率高,文档集中出现频率低的词,往往被赋予较高的权重[10]。而本应用部分情况相反,如某文档围绕“桃树砧木的选择”的展开介绍中,重要专业词“砧木”常通过其他方式表达而出现频率较低。而文档集中关于果林花卉嫁接技术介绍又不免再次涉及该词而逆文档词频率较高,从而该词综合权值较低。因此在该领域应用中,需进一步考虑词专业性因子对特征词权重影响予以修正。

缺乏对查询特征词分布重要性考虑,使对相关文档的识别能力较弱。VSM 认为同一特征项表达文档的能力完全相同[6],因而特征词出现在标题或正文中权值相同,无法分辨出前者对主题更为重要。在提问分析形成多词查询时,对于查询语句而言,特征词全部分布于标题的文档比分布于正文的文档更为相关,因为后者常常是阐述另一主题时对查询特征词有所涉及而已。因此,需考虑查询特征词分布因子对相似度值影响予以修正。

相似度计算量大,影响检索的速度。VSM 基于统计学方法进行大量特征词权值计算,并在稀疏矩阵中构建高维度特征向量进行相似度计算,检索代价高,因此需采用一定的忧化策略以提高系统工作效率。

2　农业实用技术自动问答系统设计

2.1　系统框架　自动问答功能主要通过问题理解,信息检索及答案抽取三大模块实现(图 1)。问题理解,主要对提问语句进行分析以理解用户提问意图。信息检索,包括对技术文档进行提取特征词的预处理,建立特征词权值的索引,以及答案候选文档集选择。答案抽取,在上述环节的基础上实现提问语句与候选文档的相似度计算及结果输出。

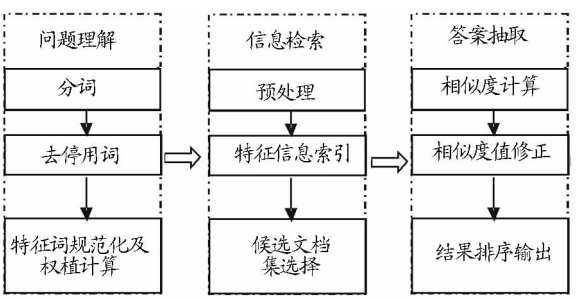

图 1　系统框架

Fig. 1　The framework for the Question – Answer System

2.2　用户提问理解　对用户给出的自然语言问题进行分词,停用词过滤,提取能表达提问语句信息的查询特征词。并对其进行规范化及权值计算,形成表达提问意图的查询向量。

(1) 分词及去停用词。分词的算法已比较成熟,较大程度影响分词效果的是所基于词库的词收录状况。参考《农业主题词表》、《农业叙词表》,收录关于种植、养殖等技术术语、专业学名物名及其相应非正式叙词,在基于专业词库的分词引擎上进行分词。停用词表收录标点符号以及在文档中权值较小的词。分词引擎先根据用户所提的问题如“请问西瓜怎么样栽培?”可切分成“请问”、“西瓜、“怎么样”、“栽培”、“?”,利用停用词表,过滤分析出名词、形容词、动词等与主题相关的特征词集合,如“西瓜”、“栽培”。

(2)特征词规范化。根据《农业叙词表》的等同关系项,选取词条,构建同义词表。将分词、去停用词后的结果集在同义词表看有无与之对应的主题词。若有,则使用规范的主题词作为该文档的特征词条,若没有,则用原词作为特征词条。

2.3　文档信息检索　对技术文档进行同上分词、停用词过滤,以及特征项规范化的预处理,利用改进的专业词重要性权值算法计算每个特征项的权值,形成构建文档向量的文档特征词权值索引表。并以查询向量为条件,选择候选答案文档集。

文档特征词权值计算:在农业特定领域的自动问答系统中,专业词的所传达的信息不容忽视。如“葡萄冬季修剪方法”,专业词“葡萄”、“修剪”为专业词,比一般词“冬季”、“方法”在表达主题上更显重要。将组成文档的特征词分为专业词和一般词两大类,对实用技术文档特征项权值进行修正,特征词 T_i 在文档 d_i 中的权重为:

$$W_i^d = m \times TF_i \times IDF_i \tag{4}$$

m 为词类加权系数,表示专业词或非专业词对文档主题贡献力大小。

候选结果集选择:考察实际技术文档,标题和正文中不包含查询特征词的文档必然与用户提问的相关性非常小,在分析时应剔除这类文档,以减少不必要的系统计算。具体可以利用查询特征词构成找查询条件,选择标题或正文中含有特征项的文档为候选文档集,在限定范围内计算相似度,以节约系统开销。

2.4　答案抽取　在候选结果集范围内,对提问语句与文档的相似度进行量化与修正,并按照修正的相似度值对结果进行排序输出。

相似度计算及基于特征词分布权重的修正:据式(3)计算查询向量与文档向量的相似度值,然后根据查询特征词分布状况进行相似度值修正。标题是反映文档主题的关键。对于查询语句而言,标题全部包含其特征词比标题部分包含更为相关,而后者又比标题不包含而正文全部包含更为相关。在答案候选集中,将分布情况分三种,即全部分布于标题、部分分布于标题、不分布于标题,针对不同的情况分别赋以不同的权值:

$$Msim(d_i,q_j) = \begin{cases} \lambda_1 sim(d_i,q_j)\, T_i \epsilon title \\ \lambda_2\, sim(d_i,q_j)\, T_i \text{ 部分 } \epsilon title\, T_i \text{ 部分 } \epsilon content \\ \lambda \cdots\cdots 3 sim(d_i,q_j)\, T_i \neq title \end{cases} \tag{5}$$

λ_i 为词分布加权系数,表示不同查询特征词分布情况对相似度贡献度大小。

结果集排序输出:对候选答案集中的文档进行排序,并取出大于给定阈值的若干答案。将与问题相关度权值最大的答案内容信息直接显示在主界面,剩下的在“问题相关”栏

目中列表显示。系统经检索在阈值之上无答案时,则给以提示消息。

2.5 系统效率优化 VSM 模型的改进增加了计算的复杂度,同时由于其权值及相似度值计算量非常之大,若不采用优化的设计模式,系统运行效率非常慢,实际应用性差。因此,在设计中,以尽量让后台承担更多的计算任务的思想分配系统资源。技术文档预处环节,采用全文索引,以及各文档特征词权值、文档向量、文档模(即$|d_i|$)提前计算存储,以减少与前台提问向量即时计算的工作量。相似度计算环节,选择标题或正文包含提问关键词的文档集范围内进行计算,以减少与提问无关的无用计算。同时,由于查询向量对候选文档集中的各文档向量都是相同的,因此式(3)中是否除以$|d_i|$,对下一步排序没有影响,运算时作为公共因子消去。

3 应用实验及结果分析

系统选用"北京农业数字信息资源中心"[11]中经过经过精心审核、组织的各种实用技术信息作为自动问答文档库,涉及种植技术、养殖技术、加工保鲜、综合技术四大类,共18 370条。为测评本研究对 VSM 改进的实际效果,系统从准确率、召回率及查询效率3个方面进行测试。其中准确率定义为正确答案文档与所有被检索到的文档之比,召回率定义为检索到的相关文档与库中所有的相关文档之比。查询效率以点击提交到系统给出答案的系统计时来衡量。准确率和召回率分两类对象进行测试,即由内部项目组人员和外部应用农户根据4大类各提5问。两类人员20次提问测试结果平均值见表1。

表1 农业技术自动问答实验结果

Table1 The experiment results of Question-Answer System

模型 Model	准确率//% Accuracy rate		召回率 //% Recall rate		查询效率//s Query efficiency
	内	外	内	外	
VSM	70.27	55.42	78.48	70.67	0.42 376
改进 VSM ImprovedVSM	87.31	85.56	88.25	84.68	0.00 145

以上数据表明,改进 VSM 在计算准确率、召回率及查询效率方面较传统 VSM 均有所提高,说明针对农业应用特点的改进是有效的。同时发现,在准确率、召回率指标的测试中,项目组内部人员的测试结果高于外部实际应用农户。分析其原因,项目组内部人员对农业实用技术资源库比较熟悉,提问内容和库中现有文档较接近,因此准确率、召回率较高。外部基层农户则完全按照自己的需求和语言方式提问,进行自动问答时,因部分地方俗语在短期内未被系统词表收录导致用户提问理解不充分,以及系统答案文档覆盖面不够而一定程度影响了检索结果。因此也说明,随着系统词表的进一步完善,技术答案文档库的规模扩大,改进成效将会更明显。

4 结束语

笔者对 VSM 在农业技术自动问答特定领域的应用进行了改进研究,并相应开发了农业技术自动问答系统。研究的主要特点有:

(1)对特征词进行规范化,从而进行基于同词形的相似度计算,保证检索具有较好召回率的同时,较以往同义词扩展的方式,降低了特征词维数,对解决 VSM 稀疏矩阵问题有积极作用。

(2)查询主题是多查询特征词综合表达的结果,因此,本研究考虑多查询词分布的权值修正,对相关文档的识别能力更强,较一般研究只孤立考虑单个词位置权重修正的检索效果更佳。

(3)利用候选集选择、文档特征词权值预计算和去提问分母的方式,减轻了前台计算的工作量,也克服了因模型改进而增加算法计算复杂度的问题,使得系统效率较大提高。

从实验结果看出,改进方法取得了一定的成效。但由于农业基层应用的特殊性及中文语言处理的复杂性,检索效果还有待提高。今后将结合农业应用特点,进一步完善专业词库,并融合语义、句法相似度计算方法进行研究,以期为解答农业生产疑难提供更好的服务。

参考文献

[1] 刘里,曾庆田. 自动问答系统研究综述[J]. 山东科技大学学报:自然科学版,2007,24(4):73-76.

[2] 安玉璞. 自然语言问答系统的设计与实现[D]. 哈尔滨:哈尔滨工业大学,2003.

[3] 白曦,吕晓枫,孙吉贵. 基于加权向量空间模型的网络搜索[J]. 计算机应用研究,2007,24(2):51-56.

[4] 苏小虎,杨思春. 基于改进 VSM 的中文问答系统研究[J]. 情报理论与实践,2008,31(4):624-627.

[5] 梁斌. 走进搜索引擎[M]. 北京:电子工业出版社,2007:201-212.

[6] 李玉鑑,操卫平,周兰珍. 结构化向量空间模型及其在 Web 信息检索中的应用[J]. 北京工业大学学报,2008,34(4):441-444.

[7] 苏新宁. 信息检索理论与技术[M]. 北京:科学技术文献出版社,2004:32-35.

[8] 林永民,吕震宇,赵爽,等. 向量空间模型中特征加权的研究[J]. 情报杂志,2008(3):5-10.

[9] 郭庆琳,李艳梅,唐琦. 基于 VSM 的文本相似度计算的研究[J]. 计算机应用研究,2008,25(11):3256-3258.

[10] 苏小虎. VSM 的权重改进对文档相似度的影响研究[J]. 电脑知识与技术,2008(10):135-137.

[11] 北京农业数字资源中心[EB/OL]. http://www.agridata.gov.cn/bjdrc_web20060413/NewFront/index.aspx.

(上接第 13928 页)

学校能否进行正常教学的关键。现代技术在教学管理上的应用越来越广泛,四川农业大学开发出的排课系统在教学管理工作中发挥了重要的作用,极大地提高了工作效率,有效地利用了现有的教学资源,提高了教务管理的效率。

参考文献

[1] 任克强,赵光甫,张国萍. 高校排课问题的模型与算法[J]. 计算机与现代化,2007(10):80-82.

[2] 周小海. 学分制排课模式研究[J]. 重庆教育学院学报,2007,20(4):93-95.

[3] 李益生. 高校课表编排应注意的几个问题[J]. 辽宁医学院学报:社会科学版,2007,5(2):61-63.

[4] 寿锦雄. 基于充分利用教学资源的排课法探讨[J]. 中等职业教育,2006(24):17-18.