DOI:10.13718/j.cnki.xsxb.2020.09.008

面向农业科研办公的垂直搜索引擎研究与设计①

李 昀[1], 邓 颖[2,3,4], 吴华瑞[2,3,4]

1. 北京市农林科学院,北京 100097; 2. 国家农业信息化工程技术研究中心,北京 100097;
3. 北京市农业信息技术研究中心,北京 100097; 4. 农业农村部农业信息技术重点实验室,北京 100097

摘要:在农业科研办公过程中,科研人员进行信息检索的频率高,信息需求精度高,但传统的综合性搜索引擎检索农业实用技术、政策法规、专题数据等方向性比较强的农业信息,通常返回结果数据量庞大、主旨范围宽泛,导致内容不精准、搜索面太广,筛选结果专业性不足;且现阶段主流的农业领域的垂直搜索引擎的搜索策略主要建立在传统的文本检索上,在自身领域数据量有限的情况下,搜索结果查全率不高,且搜索结果没有排序依据(大多仅仅按信息发生时间为排序依据). 本文对农业互联网信息搜索引擎进行了研究,通过对各级农业管理部门网站、农业科研院所网站、农业新闻网站、农业商业网站等数据源的模块进行定位,通过爬虫进行数据更新检测与定时抓取,从数据源上有效减少不相关信息;基于数百个互联网数据源农业相关模块的信息抽取,采用 word2vec 和本文提出的基于文本特征表达的 doc2vec,分别创建农业词向量、文档向量空间,用来应对搜索关键词为无序词组和有序语句的搜索场景,确保垂直搜索的智能和返回结果的准确. 经过实验验证,本文提出的 doc2vec+tf-idf 搜索算法能够在有序搜索中达到较高的准确率,结合 word2vec 进行的无序搜索,有针对地进行语义搜索,可以进一步提高搜索引擎的查准率,满足日益增长的对农业领域信息搜索的高效高质的需求.

关 键 词:农业信息搜索引擎;语义相似度;word2vec;doc2vec;tf-idf;文本智能搜索

中图分类号:S126 **文献标志码**:A **文章编号**:1000-5471(2020)09-0043-08

伴随农业信息化的快速发展,农业科研协同办公平台中,用户对科研信息的需求量和信息准确度越来越高,且变化的增幅越来越大. 然而面对巨大的网络信息资源,用户在信息搜索时会查出很多与目标信息无关的网页[1]. 同百度、谷歌等通用搜索引擎相比,聚焦农业信息的垂直搜索引擎[2-3]能为农业科研工作者提供更专业性的搜索结果. 国外的农业垂直搜索引擎已经取得了一定的成果[4],如 Agriscape Search,WEBAgriSearch 等. 我国的农业垂直搜索引擎出现相对较晚,自 2007 年首个农业搜索引擎上线以来,目前国内农业搜索引擎主要有农搜网、搜农网等,仍然处在发展时期,存在一些不完善的地方,且尚无专注农业科研的搜索引擎. 首先搜索结果中仍包含了大量的无效信息[5],搜索准确率和用户满意度较低;其次搜索结果过于模式化,搜索结果都按照规定的分类模块显示,而忽略了搜索的关键词是否与预设的分类有关联;农业领域信息缺乏,目前存在的几个主流农业搜索引擎关注点大多在农产品市场价格方面,而如研究热点、重大成果、实用技术、政策法规、领域热点等相关的信息非常稀少. 构建智能化的农业科研办公平台是推动农业科研现代化、信息化发展的重要手段. 本文在传统的农业垂直搜索引擎基础上,保证数据源的精确性,结合语义关联分析查询机制,提供对农业信息的精确及时的检索查询,为农业科研办公的智能化、信息化提供有力技术支撑. 在农业科研办公平台中,小部分数据来自于科研单位办公过程产生的以及手动输入的,主要数据来源于外部互联网数据接入和抓取,在不考虑合作数据对接共享的情况下,如何高效获取平台外的信息成为亟待解决的问题,而垂直搜索引擎是解决这一问题的工具.

① 收稿日期:2020-08-07
基金项目:2020 年度农业农村部农业信息技术重点实验室建设项目(PT2020-03).
作者简介:李 昀(1969-),硕士,高级工程师,主要从事信息化管理应用研究.
通信作者:吴华瑞,博士,研究员.

1 系统设计

现在传统的农业领域搜索引擎对数据来源定位不明确，从质量不高的数据源中获取大量无效信息，导致返回给用户的搜索结果包含很多干扰信息，用户不得不自行对结果的有效性进行二次判断.

面向农业科研办公的垂直搜索引擎由数据收集模块(数据资源池、爬虫模块，信息源数据监测模块)，智能搜索模块(语义分析、智能分类、信息转发模块)组成，以实现数据粗采-筛选-精确搜索的一体化过程(图1).

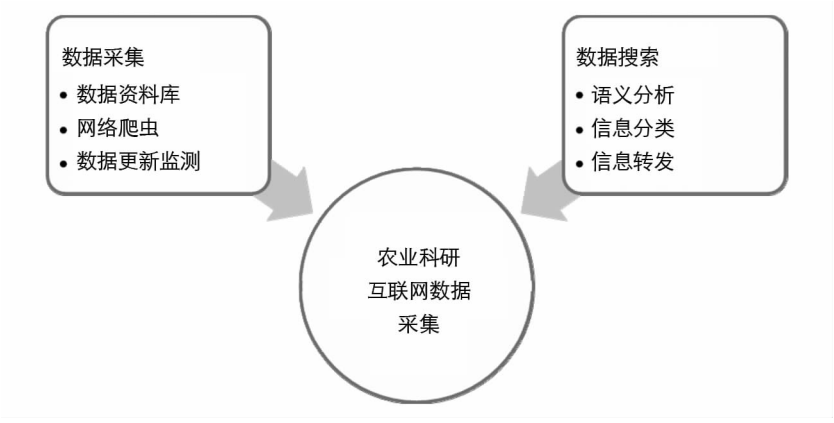

图1 农业科研互联网信息垂直搜索引擎框架设计

1.1 数据采集模块

采用 WebMagic+Phantom JS+Redis 的网页数据抓取框架，对数据源资料库中的网站模块进行更新监测和信息抓取.

本系统采用 WebMagic 爬虫框架[6-8]完成信息获取的基本工作，通过内置的定时任务执行器对录入的指定网站进行广度优先的网页数据遍历. 同时，结合 Phantom JS[9-11]的网页动态渲染技术，获取 html 页面行和经过 JavaScript 渲染的数据源信息，通过 Redis[12-15]缓存框架对爬虫获取的数据进行缓存计算，在缓存中对新获取数据的网页地址、正文标题及内容的 MD5[16-17]加密值与数据库中数据相同项的 MD5 加密结果进行查重比较，仅将 url 不重复且 MD5 对比结果不同的数据新增存入数据库，将 url 重复但 MD5 结果不同的数据进行更新. 开发可视化数据配置界面，对数据获取需要的各个配置项进行定义，并提供实时检测功能，对正在配置的数据源进行实时检测，及时发现配置上的问题.

系统集成了大量全面且定向精准的数据源：农业农村部、省厅级农业管理部门、农业科研单位、院校的官方网站、综合性及农业专业性资讯网站. 同时对数据源网站进行了面向网站子模块的筛选：人工遍历以上数据源的各个子板块直至底层板块(定义详情页的上一级菜单为底层板块)，通过以往人工采集数据的经验指定包含农业领域相关内容多、更新速度快的底层板块，对其更新内容进行监测和爬取. 通过对数据源全面核对以及严格把关，确保了系统抓取数据的全面性和精确性，获取的数据丰富而不冗余.

对本地化的数据实时监控，可视化各个数据源的数据更新情况及有效数据量，帮助及时发现网站改版、地址变更等异常，提醒对相应数据源的重新配置，辅助系统维护，对保持系统的数据质量起到监督预警的作用.

1.2 信息搜索模块

当前大多数的农业信息搜索引擎的检索方法都是进行基于全文检索的关键词模糊查询，搜索过程相对简单，但是所得到的结果只有包含搜索的关键词的信息，相关度仅仅是根据词频来判断，并且无法判断词之间的先后顺序、间隔距离等条件，这样的结果往往不全面且信息相关度不准确.

本系统的智能搜索模块按功能分为搜索和转发两大子模块. 本系统将神经网络应用到了搜索功能中，通过计算语义相似度的方式匹配搜索结果，增加返回信息语义范围的同时按照相关度进行排序，使用户能更容易获得与之查询内容接近的信息内容；在搜索结果界面提供信息发布功能，通过调用各个农业信息服务云平台的 RESTful Api，具有信息发布权限的用户可以将对应的信息发布至各个平台的对应分类目标板块中去，从精准度和方便程度上提高了用户的使用体验.

1.2.1　搜索方法设计

传统的搜索引擎主要使用关键字匹配，利用全文检索技术对爬虫数据建立索引，并对索引进行关键词的模糊查询，然后根据 PageRank[18-19]，Hyperlink-Induced Topic Search(HITS)[20-23]等面向链接的算法对查询结果进行排序[24]. 不同于水平搜索引擎的全领域信息搜索，在农业科研信息垂直搜索引擎中，被检索数据范围比水平搜索引擎少，且对返回结果的精确度要求高，使用面向链接的算法进行搜索会返回很多广告、站点导航等无效页面信息，对搜索结果产生干扰. 本文提出采用基于语义相似度[25-26]的搜索策略，抛开网页之间的链接关系，只考虑搜索内容和返回结果之间的语义关联程度.

本策略将搜索分为语义搜索和非语义搜索 2 类. 非语义搜索即搜索内容仅由不含语义的单词或者词组组成，如“我”“和”“并且”等，这些词汇在停用词列表中，在文本分词时已经从语料中去除；如果用户特意搜索此类单词，本文将使用传统的全文检索模糊查询方法，直接从数据库中进行匹配. 包含停用词表之外单词的搜索内容定义为语义搜索，该类搜索采用语义相似度搜索法进行匹配.

越专业的领域，其专业词汇量越是有限，而专业词汇对语义影响的权重值越高，在进行语义分析工作之前，从农业领域中总结出其专业词汇形成高优先级词典，在分次和关键词提取时，高优先级词典中的单词凭借其具有的高权值会优先被分词算法提取出来.

然后是将采集的大量农业科研互联网数据文档通过 pyhanlp 进行分词、去停用词处理，形成清洁可用的训练语料库. 通过 word2vec 模型对语料库进行词向量空间构建，形成词向量模型.

(1) 语义分析模型

word2vec 是 Mikolov 等提出的语言模型[27]，通过 CBOW 模型和 Skip-gram 模型实现对语料库中所有单词的词向量[28] 的计算与表示.

在搜索过程中，被检索语句 ω 中的每个词 ω_i 都可以用训练好的 word2vec 模型计算表示出其在空间 S 中的向量坐标：

$$\boldsymbol{S}(\omega_i) = (v_{i1}, v_{i2}, \cdots, v_{in}) \tag{1}$$

其中 n 为空间维度，v_{in} 为单词 ω_i 在空间 S 中各个维度上的权值，则文档中所有词的向量求均值可以用来表示该文档在的向量空间中的坐标：

$$\boldsymbol{S}(d) = \frac{\boldsymbol{S}(\omega_1) + \boldsymbol{S}(\omega_2) + \cdots + \boldsymbol{S}(\omega_k)}{k} = \left(\frac{v_{11} + v_{12} + \cdots + v_{1n}}{k}, \frac{v_{21} + v_{22} + \cdots + v_{2n}}{k}, \cdots, \frac{v_{k1} + v_{k2} + \cdots + v_{kn}}{k}\right) \tag{2}$$

其中 k 为文档 d 分词结果中单词个数. 以此为依据计算文档向量 $\boldsymbol{S}(d)$ 与搜索内容 t 在空间 $\boldsymbol{S}$ 中的向量表示 $\boldsymbol{S}(t)$ 的余弦相似度[29-30]，即为所求的搜索排序依据：

$$similarity = \cos(\theta) = \frac{\boldsymbol{S}(t)\boldsymbol{S}(d)}{|\boldsymbol{S}(t)||\boldsymbol{S}(d)|} = \frac{\sum_{i=1}^{n} t_i d_i}{\sqrt{\sum_{i=1}^{n}(t_i)^2} \times \sqrt{\sum_{i=1}^{n}(d_i)^2}} = \frac{\sum_{i=1}^{n} t_i (v_{i1} + v_{i2} + \cdots + v_{in})}{\sqrt{\sum_{i=1}^{n}(t_i)^2} \times \sqrt{\sum_{i=1}^{n}(v_{i1} + v_{i2} + \cdots + v_{in})^2}} \tag{3}$$

其中 t_i 表示单词 t 在向量空间中第 i 个维度的权值，$v_{ij}(j = 1,2,\cdots,n)$ 为文档中第 i 个单词在向量空间中第 j 个维度上的权值. 从公式可知，如此表示文档向量仅考虑了文档中单词以及他们的词频，并没有将词语排列顺序、前后间隔距离等因素考虑在内. 因此本文针对有序查询语句，采用 Doc2vec 模型进行语义相似度的计算. Mikolov 在 2014 年提出了 doc2vec，对自己先前提出的 Word2vec 进行了改进[31].

doc2Vec 同样具有 2 种模型，分别为：Distributed-Memory(DM 模型) 和 Distributed-Bag-of-Words (DBOW 模型). 如图 2，DM 模型跟 Word2vec 的 CBOW(Continuous Bag-of-Word Model)相似，在已知上下文和文档向量时，计算目标词出现的概率，DBOW 模型则是跟 Word2vec 的 Skip-gram 相似，已知文档向量时，计算文档中出现随机词组的概率.

与word2vec类似，通过doc2vec模型计算出各个文档的空间向量，与搜索语句的空间向量直接进行余弦值计算就可以得出对应的语义相似度. 不同的是，这样的计算结果受文档中词语先后顺序，词与词间隔的远近等因素影响. 而doc2vec相对于word2vec，在模型的输入层增加了段落向量(Paragraph vector)，它类似于词向量，用于表示一个段落的向量空间特征，CBOW模型中的训练过程中，每次仅截取文本中的一部分词进行训练，但是忽略了上下文中的其他词，如此训练出的句子的空间向量，只是文本中各个词的向量均值，忽略了词序问题. 虽然doc2vec的训练也是通过固定滑窗大小截取上下文的一部分词来训练，但是他在同一个段落的滑窗移动的过程中共享段落向量，即一段文本通过滑动滑窗会进行若干次训练，但每次训练输入的Paragraph vector也不会因训练词组的改变而改变. 这样段落向量更能表达这段文本的主旨，也更能满足面向农业科研办公的互联网信息垂直搜索引擎对其理解能力、智能程度的需求.

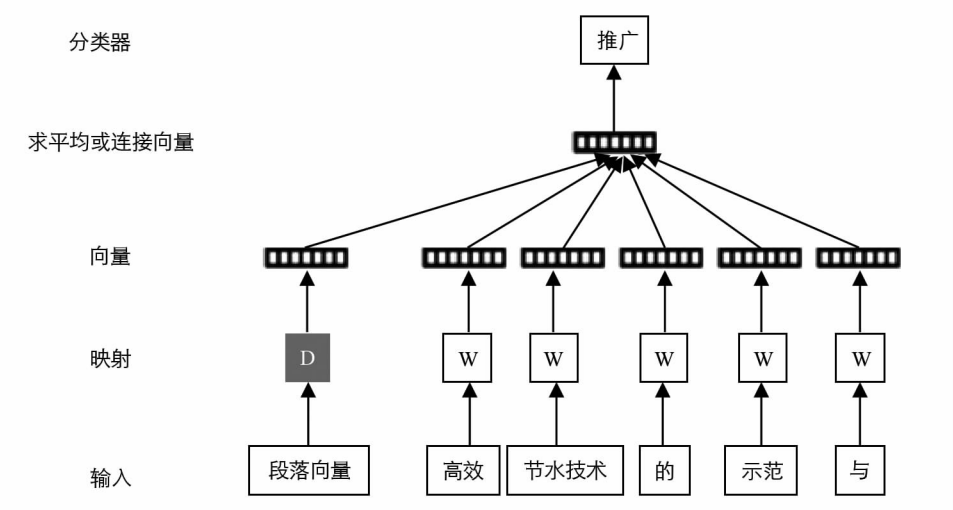

图2 DM模型

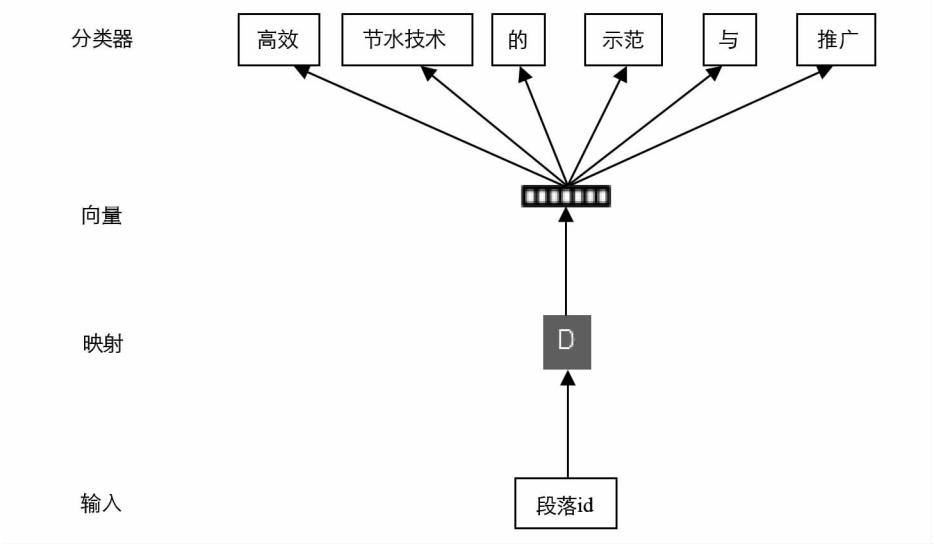

图3 DBOW模型

(2)基于TF-IDF的算法改进

虽然word2vec和doc2vec模型都通过对文档全文分词后的词组进行计算，但是在文档长度较大，包含单词数量较多的情况下，文档的特征比单独的词更复杂，因此文档向量模型的训练所需训练文本数量大，在训练样本数量不够充足的情况下会出现文档向量模型表达不准确的问题，并且训练所需的计算量也高于词向量模型. 现在从事农业领域自然语言研究的人员有限，几乎没有相关的开源数据样本，在纯人工收集的前提下，样本数量的增大将会造成前期样本收集的时间成本增高，对研究实验和生产应用都造成极大的阻碍，并且对全文进行向量空间表达，尤其是长文本的表达时，由于其单词数过多，文本特征数就有可能

越多，在这种情况下，即使训练样本充足，对模型进行训练时也容易产生过拟合现象[32]. 类似卷积神经网络[33]，在每次卷积计算之后通常会连接一次池化计算，通过局部特征的提取进行降维，不但减少模型训练和计算的复杂度，也降低了模型的过拟合程度. 本文提出在农业互联网信息垂直搜索引擎检索模块中通过文本的特征提取，即通过 tf-idf 计算得出文档的关键词，仅用关键词组成的语料来训练 doc2vec 向量模型，在减少文档中参与语义相关度计算的单词量的同时又尽量不影响被检索文档的语义，以此实现在训练样本数量受限的情况下，增加模型的可靠性.

TF-IDF 是计算单词对于其所在文本重要程度的一种统计方法[34-35]. *TF* 表示词频(Term-Frequency)，*IDF* 表示逆文本频率(Inverse-Document-Frequency). *TF* 是词条(t)在任意文档(d)中出现的频率. *IDF* 则是反应 t 在整个语料库中被包含的文档数 n，n 和 *IDF* 呈反比.

$$IDF(x) = \log \frac{N+1}{N(x)+1} + 1 \tag{4}$$

$$TF\text{-}IDF(x) = TF(x) * IDF(x) \tag{5}$$

通过计算每个文档中各个单词的 TF-IDF 值并排序选出文档重要度最高的 k 个词作为保留词，去掉其他词语，并对由保留词构成的文档进行训练文档向量模型训练得到各个文档的空间向量表示结果. 最后通过计算搜索内容和被检索文本的向量余弦相似度作为语义相似度判断依据，并筛选删除相似度低于设定阈值的结果，然后相似度从大到小的顺序对被检索文档进行排序作为查询结果发送给用户.

2　实验与结果

2.1　评价方法

查全率和查准率是评价搜索引擎性能的主要考核标准[36]，模型运算速度和模型训练速度则是作为模型优化程度的重要评价因素. 查全率(recall)又称作召回率，是衡量搜索引擎返回结果与用户查询内容相关度的能力，如公式(6)；查准率(precision)即为精度，用于衡量搜索引擎去除不相关搜索结果能力，如公式(7).

$$\text{recall} = \frac{C}{T} \tag{6}$$

$$\text{precision} = \frac{C}{R} \tag{7}$$

式中 R 表示搜索结果的信息数量，C 表示搜索结果中相关信息的数量，T 表示整个文档中的相关信息数(搜索出和未搜索出的相关信息的总和).

precision 和 recall 互相影响，若 precision 高但 recall 低，查到的信息总量就少，反之，查到的有用信息所返回的信息基数就高，对用户产生的干扰也就越强. 因此，存在一种新的评估指标，F1 测试值：

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

由于垂直搜索引擎数据来自互联网爬虫工具，信息量巨大，无法统计互联网当中相关信息的具体数量，因此将查准率作为评价搜索引擎的重要指标.

2.2　实验设计

搭建信息爬虫系统，以 24 个高校网站、49 个科研机构网站、33 个农业管理部门官方网站、23 个媒体网站、225 个农业商业网站、农业信息网站为数据源，抓取近 10 万条近期数据，人工定义 500 词的农业领域专业词典，从总样本中筛选出 38 749 条与词典内单词相关的数据共 63 M. 标记 10 个单词，10 个无序三元词组以及 10 个完整句(分词后 5 个有序单词以上)的相关文档.

本实验在 Intel i5-4590 CPU，16GB 内存主机上进行. 在 Python 3.5.4 环境下，安装 MySQLdb，PyMySQL 进行数据库操作，安装 pyhanlp，jpypel 等工具进行分词和词向量处理.

实验步骤如下：

1) 连接 mysql 数据库，提取人工筛选出的语料样本；

2) 通过 pyhanlp 中的 crf 分词算法，在基础自然语言词典的基础上，引入农业词典库，对 38 749 条语料数据进行逐条分句、分词；

3) 采用 pyhanlp 自带停用词库进行停用词过滤并存入 txt 文档；

4) 人工检查过滤后生成的文档，对未过滤而应当停用的单词进行记录并添加至停用词列表；

5) 用更新后的停用词文档再次进行过滤，并通过 tf-idf 算法进行关键词提取，提取个数为全文词数的 8%；

6) 通过 pyhanlp 集成的 word2vec 训练生成词空间向量模型；

7) 用生成的 word2vec 模型训练 doc2vec 模型(记录计算时间)；

8) 文档向量表示：

① 用 word2vec 模型表示每条数据中每个分词的词向量，并求平均值；

② 用 doc2vec 模型表示每条数据全文分词的文档向量；

③ 用 doc2vec 模型表示每条数据关键词组的文档向量；

9) 相似度计算：用人工标记所用的 50 组无序单词/词组以及 50 个有序整句与步骤 8 中产生的 a～c3 种向量模型文档向量进行相似度计算，统计并计算相似度阈值取 0.1～1 时，关键词个数 8%时的搜索结果的查准率.

2.3 结果与分析

实验影响因子有：8-①,8-②,8-③ 3 种向量模型生成方法、搜索语句有/无序状、相似度筛选阈值.

从图 4 可以看出，在搜索内容为有序整句的时候，3 种文档模型的查准率在阈值增加的过程中多呈现先快速增后缓慢减少的趋势. 其中 word2vec 模型增加幅度最大，在相似度阈值为 0.4 的时候打到最大查准率；doc2vec 模型增幅最小，在相似度阈值为 0.6 的时候达到峰值，其峰值与 word2vec 模型的峰值相比略高，但是几乎持平；doc2vec 与 tf-idf 关键词提取相结合的文本向量模型增幅居中，在相似度阈值 0.5 时达到峰值，其峰值高于另外 2 个文本向量模型. 可以判断本文提出的方法在整句搜索时，有效文本相似度筛选阈值取值 0.5 时达到最高性能，即表明其在数量有限的农业语料库优先中表现最为优异，另外 doc2vec 模型在收到训练样本数量限制的情况下性能受到较大的负影响.

从图 5 可知，3 种模型的查准率在有序和无序搜索时均呈现先增后减的趋势，无序搜索 3 种模型的查准率随相似度阈值变化的幅度更为平缓，且均在阈值接近 0.5 时达到峰值，峰值情况下，各向量模型的查准率从大到小依次为 word2vec,doc2vec+tf-idf,doc2vec. 本文提出的 doc2vec+tf-idf 方法在相似度阈值大于 0.65 时，具有更高的查准率，但 word2vec 模型的查准率峰值更高.

以农业科研领域的文本数据作为语料，采用领域内的自定义分词词典和停用词表，使用本文提出的 doc2vec 与 tf-idf 提取关键词结合的文本向量模型，通过有序文本搜索时查准率更高；而基于 word2vec 均值的文本向量模型在对无序词组进行搜索时有更好的表现. 针对不同的搜索场景动态选取对应的高性能搜索方法将进一步提高搜索引擎的性能.

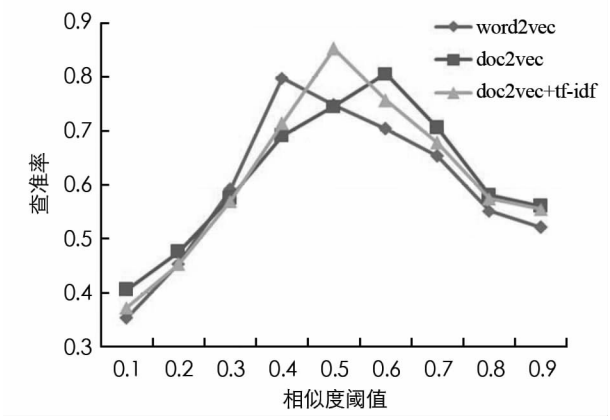

图 4 3 种模型在有序搜索内容条件下的相似度阈值与查准率关系

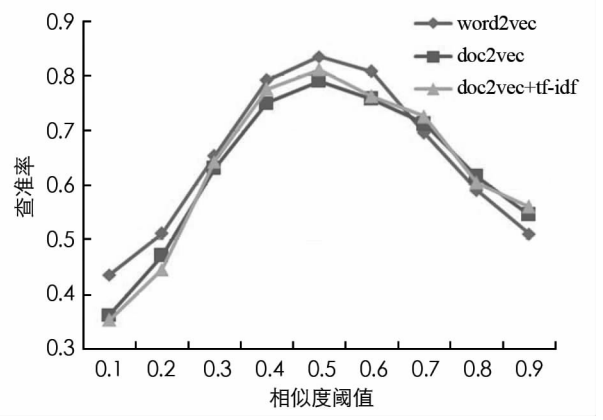

图 5 3 种模型在无序搜索内容条件下的相似度阈值与查准率关系

3 结 论

本文通过人工精确定位数据源、爬虫系统自动抓取海量农业信息互联网信息，通过 doc2vec 与 tf-idf 结合的神经网络算法进行语义相似度匹配搜索. 实验结果，证明本方法在有序文本搜索时具有较高的准确性，但 word2vec 在无序离散的词汇组合搜索时则有更高的查准率. 因此，针对不同的文本搜索场景选用不

同的搜索方法将进一步提高农业科研协同办公平台信息搜索引擎的性能. 在下一步的研究工作中，将对用户搜索内容的有序和无序分类进行判断，以决策针对性的搜索方法，达到更高的准确率；对文本进行预分类，同时判断用户搜索内容的类型，在同一分类下进行搜索，通过缩小搜索范围，降低搜索运算的时间、提升搜索效率.

参考文献：

[1] 李广丽，刘觉夫. 垂直搜索引擎系统的研究与实现 [J]. 情报杂志，2009，28(10)：144-147，169.

[2] 肖冬梅. 垂直搜索引擎研究 [J]. 图书馆学研究，2003(2)：87-89.

[3] 许翰林，王　瑞，王佳丽，等. 基于 Lucene 的新闻垂直搜索引擎设计与实现 [J]. 电脑编程技巧与维护，2018(2)：50-52.

[4] 彭玉容，杨　捧，高　媛. 农业搜索引擎的发展现状及关键技术研究 [J]. 安徽农业科学，2010，38(20)：10971-10972，10977.

[5] 王晓琴，李书琴，景　旭，等. 基于 Nutch 的农业垂直搜索引擎研究 [J]. 计算机工程与设计，2014，35(6)：2239-2243.

[6] 武婷婷. 一种基于 WebMagic 和 Mahout 的信息搜集与推荐系统 [J]. 软件导刊，2016，15(10)：1-3.

[7] 吕太之，毕家钦. 基于 Hadoop 平台的岗位分析和推荐系统的构建 [J]. 河北软件职业技术学院学报，2017，19(4)：1-4.

[8] 张婷婷，刘　凯，王伟军. 科研人员 Web 数据自动抓取模式及其开源解决方案 [J]. 信息资源管理学报，2015，5(2)：21-27.

[9] 李佳欣，潘　伟. PhantomJS 在 Web 自动化测试中的应用 [J]. 计算机光盘软件与应用，2013(18)：76-77，80.

[10] 胡　越，张源伟，雷　军. 自定规则的 AJAX 网页信息采集功能的设计 [J]. 物联网技术，2016，6(9)：86-87.

[11] 李　浩. 基于评论的博客搜索引擎的设计与实现 [D]. 重庆：重庆大学，2016.

[12] ZHU J，HU B，SHAO H. Research of Lightweight Vector Geographic Data Management Based on Main Memory Database Redis [J]. Journal of Geo-Information Science，2014，16(2)：165-172.

[13] GAO X B，FANG X M. High-Performance Distributed Cache Architecture Based on Redis[M]//Lecture Notes in Electrical Engineering. Berlin，Heidelberg：Springer Berlin Heidelberg，2013：105-111.

[14] ROEHM D，PAVEL R S，BARROS K，et al. Distributed Database Kriging for Adaptive Sampling (D2KAS) [J]. Computer Physics Communications，2015，192：138-147.

[15] BALIS B，BUBAK M，HAREZLAK D，et al. Towards an Operational Database for Real-time Environmental Monitoring and Early Warning Systems [J]. Procedia Computer Science，2017，108：2250-2259.

[16] RIVEST R. The MD5 Message-Digest Algorithm[R]. RFC Editor，1992.

[17] SZYDLO M，YIN Y L. Collision-Resistant Usage of MD5 and SHA-1 via Message Preprocessing [J]. Topics in Cryptology - CT-RSA 2006，2006：99-114. DOI：10. 1007/11605805_7.

[18] HAVELIWALA T H. Topic-sensitive Pagerank：a Context-sensitive Ranking Algorithm for Web Search [J]. IEEE Transactions on Knowledge and Data Engineering，2003，15(4)：784-796.

[19] LANGVILLE A N，MEYER C D. Google's PageRank and Beyond [J]. Mathematical Intelligencer，2011，30(1)：68-69.

[20] LORIGO L，KLEINBERG J，EATON R，et al. A Graph-Based Approach towards Discerning Inherent Structures in a Digital Library of Formal Mathematics [J]. Mathematical Knowledge Management，2004：220-235. DOI：10. 1007/978-3-540-27818-4_16.

[21] NOMURA S，OYAMA S，HAYAMIZU T，et al. Analysis and Improvement of HITS Algorithm for Detecting Web Communities [J]. Systems and Computers in Japan，2004，35(13)：32-42.

[22] CHAKRABARTI S，DOM B E，GIBSON D，et al. Topic Distillation and Spectral Filtering [J]. Artificial Intelligence Review，1999，13(5-6)：409-435.

[23] ARASU A，CHO J，GARCIA-MOLINA H，et al. Searching the Web [J]. ACM Transactions on Internet Technology (TOIT)，2001，1(1)：2-43.

[24] 吴莉霞. 浅谈搜索引擎优化策略 [J]. 电脑知识与技术，2014，10(15)：3662-3664.

[25] 赵　谦，荆　琪，李爱萍，等. 一种基于语义与句法结构的短文本相似度计算方法 [J]. 计算机工程与科学，2018，40(7)：1287-1294.

[26] 冯高磊，高嵩峰. 基于向量空间模型结合语义的文本相似度算法 [J]. 现代电子技术，2018，41(11)：157-161.

[27] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient Estimation of Word Representations in Vector Space [EB/OL]. 2013：arXiv：1301. 3781[cs. CL]. https：//arxiv. org/abs/1301. 3781.

[28] 黄承慧，印　鉴，侯　昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报，2011，

34(5):856-864.
[29] 朱命冬,徐立新,申德荣,等. 面向不确定文本数据的余弦相似性查询方法[J]. 计算机科学与探索,2018,12(1):49-64.
[30] HINTON G E. Learning Distributed Representations of Concepts[C]//In Proceedings of the Eighth Annual Conference of the Cognitive Science Society,1986,Amherst MA:Lawrence Erlbaum Associates,c1986:1-12.
[31] LE Q V,MIKOLOV T. Distributed Representations of Sentences and Documents[EB/OL]. 2014:arXiv:1405.4053[cs.CL]. https://arxiv.org/abs/1405.4053.
[32] 覃光华,丁晶,陈彬兵. 预防过拟合现象的人工神经网络训练策略及其应用[J]. 长江科学院院报,2002,19(3):59-61.
[33] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[C]//19th International Conference on Neural Information Processing Systems,November 12-15,2012,Doha,Qatar:Springer,c2012:1097-1105.
[34] KARDARAS D K,KAPERONIS S,BARBOUNAKI S,et al. An Approach to Modelling User Interests Using TF-IDF and Fuzzy Sets Qualitative Comparative Analysis[J]. Artificial Intelligence Applications and Innovations,2018:606-615. DOI:10.1007/978-3-319-92007-8_51.
[35] DHAR A,DASH N S,ROY K. Application of TF-IDF Feature for Categorizing Documents of Online Bangla Web Text Corpus[J]. Intelligent Engineering Informatics,2018:51-59. DOI:10.1007/978-981-10-7566-7_6.
[36] 风元杰,刘正春,王坚毅. 搜索引擎主要性能评价指标体系研究[J]. 情报学报,2004(1):63-68.

On Design of Vertical Search Engine toward Agricultural Scientific Research Office

LI Yun[1], DENG Ying[2,3,4], WU Hua-rui[2,3,4]

1. *Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China;*
2. *National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China;*
3. *Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China;*
4. *Key Laboratory of Agri-informatics, Ministry of Agriculture, Beijing 100097, China*

Abstract: The disadvantage of using traditional comprehensive search engines in Agricultural area is that they returns too many results which are not accurate enough to match the requirement of the agricultural scientific research office due to its non-limited search coverage and using improperly semantic association algorithms. In this article, an Agricultural Web-Info Gathering system monitors have been mentioned, updated information been gathered and accumulated from specific modules of series of agricultural websites such as official websites of national and local agriculture management departments, official websites of agricultural college or research institutes, agriculture magazines websites, and agriculture commercial websites. Specification of data resource reduces non-related data, efficiently limited the search range. The search engine utilized word2vec and text feature based doc2vec models and took data of agriculture oriented websites as text corpus to build word vector space and document vector space to deal with non-ordered words set search and ordered sentence or paragraph search, in order to ensure the search result to be accurate as well as intelligent. According to the result of experiment it is proved that this system with doc2vec +tf-idf search algorithm has higher accuracy in sequential search for agricultural information. With the high performance of word2vec algorithm in nonsequential search, dynamically choosing corresponding algorithm for sequential/nonsequential search could further improve the accuracy of the search engine, and satisfied high quality data resource requirement of Agricultural information.

Key words: agricultural vertical search engine; semantic similarity; word2vec; doc2vec; tf-idf; context based search

责任编辑 王新娟