doi:10.6041/j.issn.1000-1298.2021.S0.021

基于知识图谱的 Android 端农技智能问答系统研究

张博凯　李　想
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 针对目前普及程度较高的以电话直接咨询、集中技术培训和专家现场指导为主的农业信息服务,受时空和人力限制,存在及时性和便捷性欠缺的问题,研究开发 Android 端农技智能问答机器人 APP,为农民提供信息服务。利用爬虫工具采集互联网平台上的海量农技问答数据,经过预处理后形成语料。对语料特征进行自动标注后训练 CRF 模型识别农技命名实体。并根据词频和信息熵计算命名实体的评价指数,构建"农作物-病虫害-农药"三元组知识库。将知识库导入 Neo4j 建立农技知识图谱。在 Android 端集成命名实体识别和知识图谱查询推荐算法,解决用户问题的关键词识别和查询结果的择优推荐问题。所设计问答系统为农技问答提供了一种智能解决方案,具有较高的自动化程度和应用价值。

关键词: 农业信息服务; 智能问答系统; 知识图谱; 命名实体识别; 条件随机场
中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2021)S0-0164-08

Design of Agricultural Question Answering System Based on Knowledge Graph

ZHANG Bokai　LI Xiang
(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: At present, the popular agricultural information services are mainly telephone direct consultation, centralized technical training and expert on-site guidance in China. Due to the limitation of time and space and manpower, there is a lack of timeliness and convenience. Through the research and development of Android agricultural technology intelligent question answering robot APP, agricultural information service can be provided for farmers. Crawlers were used to collect a large number of agricultural technology Q&A data on Internet platforms, which were preprocessed to form a corpus. The CRF model was trained to recognize the agricultural technology named entity after automatically labeling the corpus features. According to word frequency and information entropy, the evaluation index of named entity was calculated to construct the triple knowledge base of "crops, pests and pesticides". The knowledge base was imported into Neo4j to establish the agricultural technology knowledge map. The algorithm of named entity recognition and knowledge map query recommendation was integrated in Android to solve the problem of keyword recognition and query result recommendation. This question answering system can provide a intelligent solution for agricultural technology Q&A, which had a high degree of automation and application value.

Key words: agricultural information service; intelligent question answering system; knowledge graph; named entities recognition; conditional random fields

0　引言

我国是农业大国,虽然农业是一个传统行业,但及时获取信息的需求却随着时代的发展显得越来越迫切,尤其是新品种、新技术的推广以及市场信息、病虫害防治方法等[1]。然而熟悉农业生产活动、掌握信息资源的从业人员和专家的数量远不能满足需要。其次,受限于农村网络覆盖率低和稳定性较差,

收稿日期: 2021-07-10　修回日期: 2021-09-21
基金项目: 国家自然科学基金项目(61601471)
作者简介: 张博凯(1997—),男,硕士生,主要从事农业自然语言处理和图像识别研究,E-mail: im.bokaizhang@gmail.com
通信作者: 李想(1983—),男,副教授,博士,主要从事农业大数据挖掘和实时复杂事件处理研究,E-mail: cqlixiang@cau.edu.cn

如今广泛开展的农业信息服务方式仍以电话直接咨询、集中技术培训和专家现场指导为主，依托信息技术的线上服务很少[2]。而 5G 技术的应用，使得农业信息传递更加高效便捷，以及带来的海量农业信息资源，给农业的信息化建设带来了新的发展机遇[3]，构建一个及时传输、使用简单的基于海量数据的智慧农业系统成为可能。

农业生产的主体具有生物多样性和变异性，再加上我国幅员辽阔，复杂多样的地理条件和迥异的区域气候对农业生产的影响，导致了农业数据本身具有关系复杂、增长率高等特征[4]。因此一般的关系型数据库难以实现农业知识数据存储。

知识图谱（Knowledge graph）由 Google 公司提出，最初目的是为了优化搜索引擎的返回结果，增强用户搜索质量及体验。从本质上来讲，知识图谱就是一种揭示命名实体之间关系的语义网络，可以对现实世界中的客观事物以及它们之间的相互关系进行形式化描述[5]。

知识图谱技术很快被发掘了其在深度问答方面的强大潜力。以往的问答系统主要采用关键字匹配的方法完成答案检索，而应用知识图谱技术的问答系统允许用户进行口头表达式的提问，能够充分理解用户的提问并通过推导给出准确的答案[6]。尤其是在中文问答中存在的表达方式的多样化、不符合规范语法等问题，知识图谱搜索技术拥有更大的优势[7]。其次，在专业领域（如农业），问答系统所需的庞大专业知识储备导致的过度占用存储空间和降低查询效率的问题可以被知识图谱解决，据研究在规模相同的数据中，运用知识图谱可以节省 10% 的存储空间，查询效率提高近 30 倍[8]。

随着知识图谱技术的不断发展，应用知识图谱的中文问答系统已经成为目前研究热点。其研究内容主要包括命名实体识别（Named entity recognition，NER）、知识图谱构建和问答系统研发。

命名实体识别属于自然语言处理（Natural language processing，NLP）的一个分支。该技术的目的是定位文本中的命名实体并标记为提前定义的类别，本质上承担的是信息提取任务[9-10]。命名实体识别的方法目前主要有 3 种，即基于规则和词典的方法、基于统计的方法、混合规则和统计的方法。

最早出现的基于规则和词典进行命名实体识别的方法，需要专业领域的专家编写规则模板，主要采用模式匹配和字符串匹配等手段，该方法很大程度上依赖人工构建的知识库和词典[11-13]。基于统计的方法利用语料进行训练，主要分为 4 类：支持向量机（Support vector machine，SVM）、隐马尔可夫模型（Hidden Markov model，HMM）、最大熵隐马尔科夫模型（Maximum entropy Markov model，MEMM）和条件随机场（Conditional random fields，CRF）。SVM 分类器不能使用实体的上下文特征进行识别或训练[14]。而隐马尔可夫模型识别英文地名、机构名以及人名的效果较好[15]。后来出现的最大熵隐马尔科夫模型，不仅结构紧凑，而且在不同的专业领域都可以使用。最大熵模型的优点在于其能灵活地选择和运用特征[16]。但由于最大熵隐马尔科夫模型只能达到局部最优解，不能达到全局最优解，该模型也存在标记偏置的问题。之后的条件随机场模型不仅具有隐马尔可夫模型和最大熵模型的优点，也解决了上面提到的标记偏置问题[17-18]。

知识图谱的构建方法也一直是研究热点，文献[19]使用社会网络分析法构建知识图谱网络，清晰地反映了个体之间的联系，但却忽略了每个个体自身的属性。复杂网络（Complex network）是知识图谱构建的一个方向，其构建方法是首先对数据构建网络模型，然后运用各种数据挖掘方法分析网络模型，进而对模型结构可视化[20-23]。但这些构建知识图谱方法的数据保存在传统的关系型数据库中，仍然存在数据存储冗余、处理效率低下等问题。Neo4j 是一个高性能的非关系型图数据库，以图数据形式存储数据，因此以 Neo4j 构建的知识图谱具有存储空间小、查询速度快等优点，可以使用各种图的遍历算法进行数据查询[24-25]。

在问答系统研发方面，研究最广泛的就是专家系统（Expert system）[26-28]。从应用角度看，目前的农业专家系统或问答系统，普遍存在的问题就是局限于某一类农作物或是农事生产的某个阶段，不能很好地覆盖较多类农作物和农事生产的全过程。从使用友好的角度看，专家系统本身具有较强的技术性，对使用者本身的科学素养具有比较高的要求，不利于在农业生产者中使用推广。

本文应用条件随机场进行农技命名实体识别，基于 Neo4j 构建农技知识图谱，实现聊天式的 Android 端农技智能问答系统。

1　数据获取和预处理

我国农村网民数量迅猛增长。在这种情况下，旨在服务农民，定位于农民和农业专家之间桥梁的线上问答平台层出不穷。活跃用户数量较多、影响较广的有农管家、天天学农等。这些问答平台积累了大量的农技问答数据，不仅覆盖作物种类较全面，而且回答的权威性也有一定保障。

借助爬虫工具获取互联网农技问答平台上的问答数据。根据观察平均每条问题的回答数量稳定在3条,故设定采集字段为:问题、回答1、回答2、回答3。采集数据示例如表1所示。

表1 采集数据示例

Tab.1 Sample data collection

提问	回答1	回答2	回答3
各位老师,请问这是病毒病还是其它生理性病害?	生理性裂痕果	裂果肥水不均衡,氮肥过量,磷、钾肥不足,开花坐果时缺硼、钙元素,激素浓度使用不当等原因造成	裂果,高温,干旱,强光,后期雨水过多温湿度失衡,应用激素过高,缺钙,加强苗期管理,可喷0.3%氯化钙
请问老师,这是苹果什么病,现在用什么药?	炭疽病危害,建议用42.8%氟菌肟菌酯1 500倍液喷雾,10%苯醚甲环唑1 500倍液喷雾,25%吡唑醚菌酯1 500倍液喷雾,75%肟菌戊唑醇10~15克/亩喷雾	炭疽病侵染导致。建议叶面喷施戊唑醇肟菌酯或者咪鲜胺或者苯甲嘧菌酯或者啶氧菌酯防治	炭疽病,可以用苯咪甲环唑咪鲜胺或氟硅唑或吡唑醚菌酯
各位老师帮忙看一下是什么情况?说是打了叶面肥就绿下,一个星期以后又黄了!新发的叶子是绿的	根系吸收不好缺素症,冲施枯草芽孢杆菌养根系生根,叶面喷施氨基酸加微量元素缓解下,7到10天喷施一次	考虑土壤环境差,根系发育不良导致,建议随水冲施腐殖酸+枯草芽孢杆菌,促进根系生长发育,叶面喷施碧护+磷酸二氢钾调节生长	土壤酸化板结,透气性差,根系吸收障碍,缺铁、营养不足引起的生理性黄叶,冲施生物菌肥或者腐植酸肥,促进根系发育,叶面喷施磷酸二氢钾加氨基酸叶面肥加禾丰铁加芸苔素调节

在进行命名实体识别前,需要对问答数据进行简单的格式整理,并利用NLPIR-ICTCLAS汉语分词系统对问答数据进行分词,形成语料数据。

在问答数据中,有的提问者会使用问答平台提供的表情符号(Emoji),而在文本数据中,这些Emoji会变成被"[]"包裹的文本,如:"[握手]"。而在训练CRF模型时,可能会用到"[]"作为语料中标注实体的符号,因此需要将采集的问答文本数据中的"[]"替换为"()",例如:"谁知道草莓籽上发芽了,属于什么原因?谢谢![握手]"改为"谁知道草莓籽上发芽了,属于什么原因?谢谢!(握手)"。

本文使用的分词工具NLPIR-ICTCLAS系统的分词效果是在原文的基础上将词语以空格分割并使用"/"标记,例如"欢迎/v 使用/v NLPIR/x ICTCLAS/x 汉语/nz 分词/v 系统/n。/wj",而问答数据中同样也存在"/",为了防止混淆,需要将"/"替换为其表达的实际含义,如:"由土壤中残留病菌传染导致建议用25%甲霜·霜霉威125~187克/亩喷雾"改为"由土壤中残留病菌传染导致建议用25%甲霜·霜霉威125~187克每亩喷雾"。

最后,训练CRF模型需要提供基本的训练语料,因此本文使用NLPIR-ICTCLAS汉语分词系统作为分词工具,问答数据经过分词后被切割为语料。该工具主要的功能有中英文混合分词、关键词提取、新词识别与自适应分词和导入用户专业词典等。该工具的新词识别等功能基于交叉信息熵算法,而农业中的专业名词的逆文本频率指数(Inverse document frequency,IDF)并不能很好地反映自身的关键性,因此对农业专业名词的区分识别率并不是很高。所以本文只使用了该工具的分词功能,即基于该工具的通用词典进行切割语料,切割农业问答数据后效果如表2所示。

表2 预处理结果

Tab.2 Pretreatment result

原数据	分词后
注意螨类危害,用哒螨灵,螺螨酯,阿维炔螨特,乙螨唑,联苯肼酯等喷雾防治。	注意/v 螨/w 类/n 危害/vn,/wd 用/v 哒/x 螨/w 灵/a,/wd 螺/ng 螨/w 酯/ng,/wd 阿/b 维/b 炔/w 螨/w 特/d,/wd 乙/m 螨/w 唑/x,/wd 联/vg 苯/n 肼/w 酯/ng 等/udeng 喷/v 雾/n 防治/vn。/wj
炭疽病侵染导致,建议叶面喷施戊唑醇肟菌酯或者咪鲜胺或者苯甲嘧菌酯或者啶氧菌酯防治。	炭疽/n 病/n 侵/vg 染/v 导致/v,/wd 建议/v 叶面/n 喷/v 施/v 戊/Mg 唑/x 醇/ag 肟/w 菌/n 酯/ng 或者/c 咪/ng 鲜/a 胺/n 或者/c 苯/n 甲/n 嘧/x 菌/n 酯/ng 或者/c 啶/x 氧/n 菌/n 酯/ng 防治/v。/wj
褐斑病,已经到后期造成落叶了!果树上目前完好的叶片在用药后还会继续落叶,属于正常情况。一般使用三唑类药剂防治:戊唑醇、氟硅唑、腈菌唑等,建议再配些营养,比如氨基酸或者磷酸二氢钾等!	褐斑病/n,/wd 已经/d 到/v 后期/f 造成/v 落叶/n 了/y!/wt 果树/n 上/n 目前/t 完好/a 的/ude1 叶片/n 在/p 用/p 药/n 后/f 还/d 会/v 继续/v 落叶/n,/wd 属于/v 正常/a 情况/n。/wj 一般/ad 使用/v 三/m 唑/x 类/q 药剂/n 防治/v:/wm 戊/Mg 唑/x 醇/ag,/wn 氟/n 硅/n 唑/x,/wn 腈/w 菌/n 唑/x 等/udeng,/wd 建议/n 再/d 配/v 些/q 营养/n,/wd 比如/v 氨基酸/n 或者/c 磷酸/n 二/m 氢/n 钾/n 等/v!/wt

2　命名实体识别

CRF + +采用了特征模板，因此可以自动生成特征函数，具体特征需要给定。本文使用 5 个特征对语料进行标注，分别为界定词、词性、左右指界词、偏旁部首和数量词[29]。

界定词：在采集的农技问答数据中，涵盖的农作物中基本上都含有“瓜、菜、果”，病虫害中含有“病、虫、蛾”，农药则含有“酯、唑、醇、胺、腈”等。上述各词基本上都能很好地指示农作物、病虫害和农药等实体，但由于提问者会有“专家给我看看这是啥毛病”、“请问这是什么虫子”等表述，因此设定跳过提问文本，从回答开始查找界定词。

词性：通过预处理后，语料都被标记了词性。经过统计，农作物、病虫害和农药的词性都为名词，且病虫害和农药有可能是复合名词，包含动词、形容词等。

指界词：农技命名实体常常与某些词伴随出现，出现在实体左边称为左指界词，右边称为右指界词。常见指界词如表 3 所示。

表 3　常见指界词

Tab. 3　Common boundary words

类型	左指界词	右指界词
农作物	这种、看看、老师们	怎么、咋、什么
病虫害	句首、预防、注意	句末、导致
农药	使用、施、喷	调节、防治、缓解

偏旁部首：农作物命名实体常包含“豆、艹、木”等部首，例如“豌豆、草莓、辣椒”等。病虫害中大部分都以“病”作为结尾，因此界定词就可以很好地标注。除此之外，虫害可以以“虫”作为偏旁部首特征进行识别。农药中大部分为有机物，据研究主要包括“有机磷酸酯类、拟除虫菊酯类、氨基甲酸酯类”等[30]。因此可选择“酉、月、口”等作为农药的偏旁部首特征。

数量词：数量词常常出现在农药实体附近，用来指示农药的使用量，如“42.8% 氟菌肟菌酯 1 500 倍液”。另一种情况是农药实体中出现量词：“甲、乙、双、三等”，这是由于化学命名造成的，具体如“氟虫双酰胺、磷酸二氢钾、苯醚甲环唑”等。因此，数量词也是一个很重要的农药实体特征。

通过数据预处理和标注特征，得到了语料特征序列以及输出特征序列，基于这些序列对 CRF 模型进行训练，多次迭代，优化各特征的权重，完成训练后可通过输入文本数据，输出农技命名实体的识别结果。

模型完成训练后，可用于识别文本中的农技命名实体。能够将标注好的命名实体整理生成用户词典。为了提高识别的效率，本文将适当部分的问答数据输入模型进行识别，将生成的用户词典导入 NLPIR - ICTCLAS 汉语分词系统进行快速识别大部分的问答数据，具体识别效果如表 4 所示。

表 4　命名实体识别结果

Tab. 4　Named entity recognition result

原数据	识别结果
根腐病危害。建议选用内吸性杀菌剂乙蒜素 + 甲基硫菌灵、溴菌腈等喷施防治。	根腐病/n_disease 危害/vn。/wj 建议选用/n_new内/f 吸/v 性/ng 杀菌剂/n_medicine 乙蒜素/n _ medicine + /m 甲基硫菌灵/n _ medicine 、/wn 溴菌腈/n_medicine 等/v 喷施防治/n_new。
注意防治炭疽病侵染，用阿砣，苯醚甲环唑，咪鲜胺锰盐，腐霉利，氟硅唑，啶氧菌酯，苯甲溴菌腈，吡唑嘧菌酯等药剂交替使用喷雾防治。	注意/v 防治/vn 炭疽病/n_disease 侵染/n_new，用/p 阿砣/n_medicine，苯醚甲环唑/n_medicine，咪鲜胺锰盐/n_medicine，腐霉利/n_medicine，氟硅唑/n_medicine，啶氧菌酯/n_medicine，苯甲溴菌腈/n_medicine，吡唑嘧菌酯/n_medicine 等/udeng 药剂/n 交替使用/n_new 喷雾/n_new 防治/vn。/wj

3　知识图谱构建

本文核心是构建农技知识图谱，采用 Neo4j 图数据库构建。采集的问答数据经过预处理、命名实体识别后，构建“农作物-病虫害-农药”命名实体三元组，形成农技知识库，并将知识库导入 Neo4j 构建知识图谱。为农民提供农事解决方案时，即涉及到农药等命名实体的推荐算法设计时，使用评价指数(Power index，PI)来辅助农药等命名实体的推荐。

Neo4j 是一个非关系型的图数据库，其基本结构是由节点(node)和边(relationship)构成，节点和边除了自身的名称还可以添加属性值。除了其知识图谱在空间效率和查询速度方面的优势，由于其数据添加只需要创建节点和建立关系，因此在面对使用爬虫工具采集农业问答的更新速度快的问题时，Neo4j 不仅可以迅速根据新数据添加数据库内容，甚至还可以设计新的节点和关系的类别，并和原有数据便捷地建立联系。综上所述，在本文的应用场景下，Neo4j 的表现远远超过传统的关系型数据库。

构建知识图谱首先需要准备数据。基于 Python 的文本处理程序提取标记后的文本数据中的标签及关系。标签为农作物(n_crop)、病虫害(n_disease)、农药(n_medicine)。将两实体位于同一句话中或处于上下文滑动窗口中视为实体间存在关联。上下文

滑动窗口指的是处于同一个问答关系数据组内,最终构建农技节点三元组“农作物-病虫害-农药”。节点间的关系分别为“感染”、“适用”。

在 Neo4j 使用 Cypher 将所需数据导入库内。同时建立节点并创建索引,防止建立重复节点导致数据结构混乱等问题。然后需要导入节点间关系,建立节点间的边。最终形成所需的农技知识图谱,如图 1 所示。

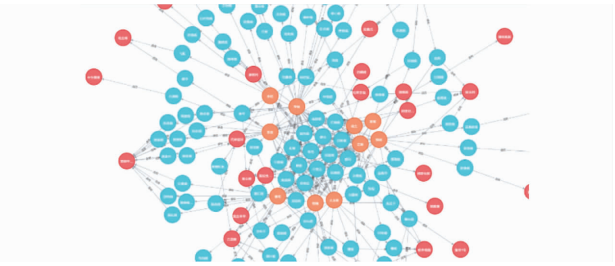

图 1 农技知识图谱

Fig. 1 Agricultural knowledge graph

4 智能问答系统搭建

4.1 Android 端设计

基于知识图谱的农技智能问答系统承载平台选用 Android。利用 Android Studio 搭建 APP 框架,调用 ChatKit 工具包实现聊天功能,最后使用 Neo4j Driver 实现与知识图谱服务器之间的通信。

ChatKit 是一个开源的 UI 聊天组件,由 LeanCloud 开发。它的底层聊天服务是基于同一个开发者的 IM 即时通信服务 LeanMessage 开发。本文的聊天功能主要使用 ChatKit 实现。

在使用 Git 将 ChatKit 导入项目中后,新建一个继承 Application 的类并在 Application. onCreate 中初始化,随后需要在 AndroidMainfest. xml 配置该类,语句为:

```
<application android:name=".MessagesActivity" >
</application>
```

在聊天界面中,不仅要展示聊天内容,还有展示聊天双方的头像等信息,出于保证聊天功能的通用性和扩展性,ChatKit 提供了获取聊天用户信息的接口,需要提供用户信息的获取方式。为此新建了继承接口 LCChatProfileProvider 的类,提供聊天用户的信息。

打开即时通讯功能,进入聊天界面有两种方式,一是通过指定另一个参与者的 clientId 的方式,开启一对一的聊天;另一种是指定一个已经存在的 AVIMConversation id,建立多人聊天群。本文的聊天双方为用户和机器人,因此只有一对一聊天的需求。在这种方式下,通过调用 intent. putExtra (LCIMConstants. PEER_ID, "peermemberId") 来传递参与者的 clientId。

聊天界面的布局和事务实现,在 ChatKit 中是由 LCIMConversationFragment 和 LCIMConversationActivity 承担。

由于本文需要与知识图谱服务器进行 HTTP 访问,如果在 UI 主进程内进行这项任务,由于网络延迟等问题将有可能造成应用卡在某个界面而无法操作的问题,即“假死”问题。为了避免“假死”,建立了专门用于与服务器通信的异步消息处理子线程,在主进程中定义消息分发对象 Handler,负责把当前的查询消息发送给子线程,并定义用于储存消息的对象 Message,当子线程把查询结果放入 Message 传回主进程后,系统将会执行 Handler 中定义的界面刷新等操作。

在查询子线程内,使用高性能的二进制 bolt 协议,查询语句与上文同样使用 Cypher 语句,这里以查询某作物感染某病虫害需要使用什么农药为例:MATCH (m:Crop) -[:感染]-(n:Disease) + WHERE m. crop = {crop} + RETURN n. disease AS disease。最终实现效果如图 2 所示。

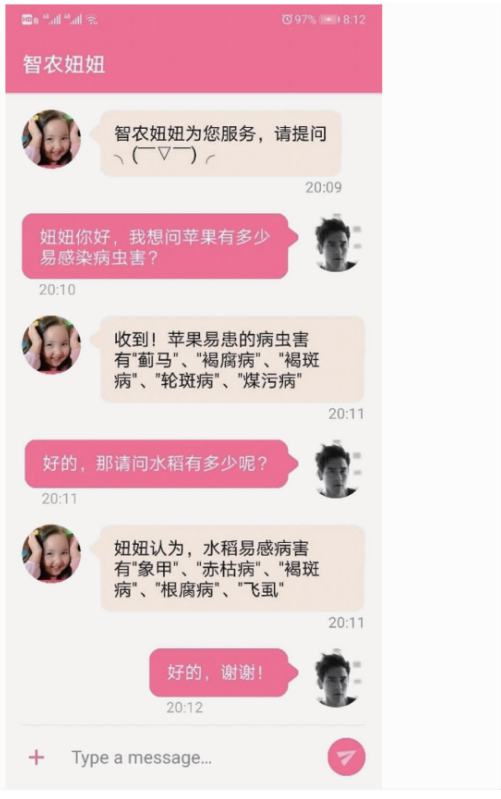

图 2 问答系统界面

Fig. 2 Q&A system interface

4.2 关键词提取与实体推荐算法

在智能问答中,最关键的 2 个问题就是用户问题中关键词的提取和回答中实体的推荐。这 2 个部分都是基于本文之前讲述的技术实现。

由于需要在知识图谱中根据节点进行搜索，进而建立知识链条推导出答案。因此需要对用户提出的问题中的关键词进行提取。经过训练 CRF 模型得到的用户词典导入 NLPIR－ICTCLAS 汉语分词系统后，已经具备了较强的命名实体识别能力，在用户输入问题后，通过基于词典的分词，能够较准确地提取问题中的关键命名实体。但由于 Neo4j 是一个基于事务的图数据库，一个事务中所含的 Cypher 语句的查询功能是确定且唯一的，所以需要 APP 根据实体的识别情况判定查询的内容，本文目前已经设定的判定方法如表 5 所示。

表 5　查询判定

Tab. 5　Query decision

实体识别情况	查询语句
Crop	MATCH (m:Crop)－[:感染]－(n:Disease) WHERE m.crop = {crop} RETURN n.disease AS disease
Crop, Disease	MATCH (m:Crop)－[:感染]－(n:Disease)－[:适用]－(o:Medicine) WHERE m.crop = {crop} AND n.disease = {disease} RETURN o.medicine AS medic

表 5 中的 2 条查询分别对应查询某种农作物的易感疾病和查询某种农作物感染某种病虫害后的最佳农药。为了实现最优推荐，这里用到了每个命名实体的评价指数属性。

虽然本文采集的农技问答数据的来源是国内比较专业的农业问答平台，但是仍然不能保证回答数据的准确性，尤其是在农药的推荐方面。专家储备的农技知识不同或知识更新速度过慢，不能了解最新最有效的农药，所处地域的农药经销商不同导致农民能够买到的农药种类有限，农民的经济实力不同导致对农药价格较为敏感，都有可能使问答数据中专家推荐的农药具有较大的不确定性和推广难度，进而导致基于问答数据构建的农技智能问答机器人的推荐农药不准确。

为了克服上述问题，本文提出使用评价指数 PI 来量化某个命名实体在农业生产中的影响力或受欢迎程度，该指数使用词频(Term frequency，TF)和信息熵(Information entropy，IE)计算。

在新闻领域中，一个词语能否代表一段内容或者是否是内容中的关键词，很重要的一个参考指标就是该词语在这段内容中出现的次数或频率。在农业中同样如此，一种农药的有效性和受欢迎程度，可以从其是否被反复提及观察得出。因此使用词频作为评价指数计算的重要参数。在文本分类研究中，有 2 种计算词频的方法，一种是直接统计某个词在文本中出现的次数计算词频 f_{TF}，其计算公式为

$$f_{TF} = \frac{f_i}{F} \tag{1}$$

式中　f_i——某个词在文本中出现的次数

F——文本中所有词数量

但是在文本分类或是关键词提取中，f_{TF} 并不能很好地评价关键词来区别文本，因为在中文文档中经常出现如“我，的，地”等没有实际含义的词，用这些词作为关键词显然是不正确的。为了避免这种情况，研究者们又定义了逆文本频率(IDF)，其计算公式是

$$f_{IDF} = \lg \frac{N}{N_i} \tag{2}$$

可得

$$W = f_{TF} f_{IDF} \tag{3}$$

式中　N——参与统计的文本总数量

N_i——有某个词出现的文本数量

W——该词的最终词频指数

通过式(3)可以看出，这种方法的出发点是某个词如果在所在的文本中的出现次数较高，且在整个参与统计的多个文本内较少出现，那么该词的代表性较高且能明显区分不同的文本。虽然本文也是为了找出具有代表性的命名实体，但是词频统计只是针对每个农作物类别的一个文本文件，没有可以参考的其他文本，其次要推荐的命名实体应具有较高的知名度和推广度，这与逆文本频率的计算方式有悖，因此本文采用第 1 种方法计算实体的词频。

信息熵是香农(SHANNON C E)在 20 世纪 40 年代提出的概念。在热力学中的热熵表示系统中分子的混乱程度，与之类似，信息熵描述了信息的不确定程度，也可以理解为信息出现在文本中的概率。从传播学的角度看，信息出现在文本中的概率越大，信息的传播范围就越广，其价值也就越高。因此信息熵可以作为本文计算评价指数的另一个参数。其计算公式为

$$H(X) = -\sum_{i=1}^{n} P(X_i)\ln P(X_i) \tag{4}$$

式中　X_i——随机变量，代表所有需要计算的信息

$P(X_i)$——输出概率，随机变量 X_i 的输出概率函数

本文通过计算每一个命名实体的信息熵可以量化其信息价值，其中参数输出概率 $P(X_i)$ 的实际意义就是对于一个给定文本，X_i 所代表的词在其中出现的概率。在本文中，该概率其实已经求得，即 TF，因此，在计算中直接代入求解。

最后，本文使用评价指数(PI)综合考虑词频和信息熵，并引入权重系数来改变参数对评价指数的影响程度，方便后期进行调参优化。每个命名实体

的评价指数 I_{PI_i} 的计算公式为

$$I_{\mathrm{PI}_i} = af_{\mathrm{TF}_i} + bH(X_i) \tag{5}$$

式中 a——词频权重系数，调整词频 f_{TF_i} 在 I_{PI_i} 中的占比

b——信息熵权重系数，调整信息熵 $H(X_i)$ 在 I_{PI_i} 中的占比

根据查询情况，APP 会自动排序并选择评价指数最高的实体作为推荐答案，并将其匹配插入到预设好的回答模板中，在聊天窗口中发送给用户。

农技问答系统构建方法还有许多可改进的地方。如在命名实体识别方面，虽然 CRF 模型和特征标注可以提高一定的识别准确率，但是在识别特征辨识率较低的农技命名实体（如农药“银法利”）时，基本上无法识别和标注。可以使用人工辅助编写词典方法，提升这些特征度较低的命名实体的识别率。其次，在使用小规模数据集时，Neo4j 的加载和查询效果很好，但使用大规模数据集时，串行机制和超级节点会影响图数据库的性能发挥。而且经过测试发现，当结果集规模相对较小时，查询速度较快，基本能够达到联机事务处理过程（OLTP）的要求，而当查询得到的结果集规模十分庞大时，消耗的时间相对较长。通过查阅资料，北京大学计算机科学技术研究所数据库团队所研发的基于图数据库理论的开源 RDF 知识图谱数据的存储和查询系统 gStore，在大规模的 Benchmark 数据集测试上，平均性能优于目前的 Virtuoso 和 Apache Jena 等国外同类产品。因此本文中的知识图谱构建可使用 gStore 提高。

5 结束语

提出了一种基于知识图谱的农技智能问答机器人构建方法。该方法使用爬虫工具采集海量问答数据，通过整理格式和数据预处理后形成输入语料，基于农技特征标注训练 CRF 模型，识别农技命名实体建立农技知识库，将知识库内的实体和实体间关系导入 Neo4j 建立农技知识图谱。一定程度上缓解农业生产者的及时获取信息的需求，为实现智慧农业提供有力的技术支持，使农业信息化发展趋于更加智能、高效。

参 考 文 献

[1] 张元宝. 农业专家系统的构建与应用[D]. 兰州：兰州大学，2013.
ZHANG Yuanbao. Construction and application of agricultural expert system[D]. Lanzhou: Lanzhou University, 2013. (in Chinese)

[2] 吴杨，张一博，汪明召. 基于农民需求视角的农业信息服务探究——以武汉市为例[J]. 科技创业月刊，2019，32(10)：83－86.
WU Yang, ZHANG Yibo, WANG Mingzhao. Research on agricultural information service from the perspective of farmers' demand—take Wuhan as an example[J]. Pioneering with Science & Technology Monthly, 2019, 32(10): 83－86. (in Chinese)

[3] 杨琇涵，宿丽丽，王青蓝. 论 5G 时代农业信息化的发展趋势[J]. 农业科技管理，2020，39(2)：11－12，35.
YANG Xiuhan, SU Lili, WANG Qinglan. Discussions on development trends of agricultural informatization in 5G era[J]. Management of Agricultural Science and Technology, 2020, 39(2): 11－12, 35. (in Chinese)

[4] 苑严伟，冀福华，赵博，等. 基于 Solr 的农田数据索引方法与大数据平台构建[J]. 农业机械学报，2019，50(11)：186－192.
YUAN Yanwei, JI Fuhua, ZHAO Bo, et al. Index method of farmland data based on Solr and construction of big data platform[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11): 186－192. (in Chinese)

[5] 袁培森，李润隆，王翀，等. 基于 BERT 的水稻表型知识图谱实体关系抽取研究[J]. 农业机械学报，2021，52(5)：151－158.
YUAN Peisen, LI Runlong, WANG Chong, et al. Entity relationship extraction from rice phenotype knowledge graph based on BERT[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(5): 151－158. (in Chinese)

[6] 张紫璇，陆佳民，姜笑，等. 面向水利信息资源的智能问答系统构建与应用[J]. 计算机与现代化，2020(3)：65－71.
ZHANG Zixuan, LU Jiamin, JIANG Xiao, et al. Construction and application of intelligent question answering system for water conservancy information resources[J]. Computer and Modernization, 2020(3): 65－71. (in Chinese)

[7] 杜泽宇，杨燕，贺樑. 基于中文知识图谱的电商领域问答系统[J]. 计算机应用与软件，2017，34(5)：153－159.
DU Zeyu, YANG Yan, HE Liang. Question answering system of electric business field based on Chinese knowledge map[J]. Computer Applications and Software, 2017, 34(5): 153－159. (in Chinese)

[8] 宫法明，李翛然. 基于 Neo4j 的海量石油领域本体数据存储研究[J]. 计算机科学，2018，45(增刊)：549－554.
GONG Faming, LI Xiaoran. Research on ontology data storage of massive oil field based on Neo4j[J]. Computer Science, 2018, 45(Supp.): 549－554. (in Chinese)

[9] RAU L F. Extracting company names from text [C] // Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications, 1991: 29－32.

[10] GRISHMAN R, SUNDHEIM B. Message understanding conference 6: a brief history [C] // Copenhagen, Denmark: Association for Computational Linguistics, 1996: 466－471.

[11] HANISCH D, FUNDEL K, MEVISSEN H T, et al. ProMiner: rule based protein and gene entity recognition[J]. Bmc Bioinformatics, 2005, 61(1): S14.

[12] POMARES Q A, SIERRA M A, GONZALEZ R R A, et al. Named entity recognition over electronic health records through a combined dictionary based approach[C]// Amsterdam, Netherlands: Elsevier, 2016: 55 - 61.

[13] 王宁,葛瑞芳,苑春法,等.中文金融新闻中公司名的识别[J].中文信息学报,2002,16(2):1 - 6.
WANG Ning, GE Ruifang, YUAN Chunfa, et al. Company name identification in Chinese financial domain[J]. Journal of Chinese Information Processing, 2002, 16(2): 1 - 6. (in Chinese)

[14] 魏芳芳,段青玲,肖晓琰,等.基于支持向量机的中文农业文本分类技术研究[J].农业机械学报,2015,46(增刊):174 - 179.
WEI Fangfang, DUAN Qingling, XIAO Xiaoyan, et al. Classification technique of Chinese agricultural text information based on SVM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(Supp.): 174 - 179. (in Chinese)

[15] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name [J]. Machine Learning Journal Special Issue on Natural Language Learning, 1999, 34(1 - 3): 211 - 231.

[16] 李素建,王厚峰,俞士汶,等.关键词自动标引的最大熵模型应用研究[J].计算机学报,2004,27(9):1192 - 1197.
LI Sujian, WANG Houfeng, YU Shiwen, et al. Research on maximum entropy model for keyword indexing[J]. Chinese Journal of Computers, 2004, 27(9): 1192 - 1197. (in Chinese)

[17] LIAO W. VEERAMACHANENI S. A simple semi-supervised algorithm for named entity recognition [C]// Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing, 2009: 58 - 65.

[18] 郭旭超,唐詹,刁磊,等.基于部首嵌入和注意力机制的病虫害命名实体识别[J].农业机械学报,2020,51(增刊2):342 - 350.
GUO Xuchao, TANG Zhan, DIAO Lei, et al. Recognition of Chinese agricultural diseases and pests named entity with joint radical-embedding and self-attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(Supp.2): 342 - 350. (in Chinese)

[19] 宗乾进.中国极地研究知识图谱——南京大学知识图谱研究组系列论文[J].现代情报,2011,31(5):12 - 15,20.
ZONG Qianjin. Mapping knowledge domains of China polar research—one of series thesis of Nanjing University mapping knowledge domains research group[J]. Journal of Modern Information, 2011, 31(5): 12 - 15, 20. (in Chinese)

[20] 马井刚. 面向复杂网络的可视化分析工具的设计与实现[D].北京:北京邮电大学,2010.
MA Jinggang. The design and implementation of visual analysis tool towards complex networks[D]. Beijing: Beijing University of Posts and Telecommunications, 2010. (in Chinese)

[21] NOY N F, SINTEK M, DECKER S, et al. Creating semantic web contents with protege-2000[J]. IEEE Intelligent Systems, 2001, 16(2): 60 - 71.

[22] 王晰巍,韦雅楠,邢云菲,等.社交网络舆情知识图谱发展动态及趋势研究[J].情报学报,2019,38(12):1329 - 1338.
WANG Xiwei, WEI Yanan, XING Yunfei, et al. Research on the dynamics and trends of the development of public opinion topic maps in social networks[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(12): 1329 - 1338. (in Chinese)

[23] QIANG Yao, KAI Chen, LAN Yao, et al. Scientometric trends and knowledge maps of global health systems research[J]. Health Res. Policy Syst., 2014, 45(5): 12 - 18.

[24] 姜惠娟,郭文龙.基于 Neo4j 的药膳方图数据库设计与优化[J].中央民族大学学报(自然科学版),2019,28(3):48 - 55.
JIANG Huijuan, GUO Wenlong. Design and optimization of medicated diet's diagram database based on Neo4j[J]. Journal of Minzu University of China(Natural Sciences Edition), 2019, 28(3): 48 - 55. (in Chinese)

[25] ZHU Z, ZHOU X, SHAO K. A novel approach based on Neo4j for multi-constrained flexible job shop scheduling problem[J]. Computers & Industrial Engineering, 2019, 130: 671 - 686.

[26] 赵明,董翠翠,董乔雪,等.基于 BIGRU 的番茄病虫害问答系统问句分类研究[J].农业机械学报,2018,49(5):271 - 276.
ZHAO Ming, DONG Cuicui, DONG Qiaoxue, et al. Question classification of tomato pests and diseases question answering system based on BIGRU[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(5): 271 - 276. (in Chinese)

[27] 冯绍振. 测土配方施肥专家系统的研究与实现[D]. 泰安:山东农业大学,2017.
FENG Shaozhen. Research and realization of soil testing and fertilizer recommendation expert system[D]. Taian: Shandong Agricultural University, 2017. (in Chinese)

[28] 陈瑛,陈昂轩,董玉博,等.基于 LSTM 的食品安全自动问答系统方法研究[J].农业机械学报,2019,50(增刊):380 - 384.
CHEN Ying, CHEN Angxuan, DONG Yubo, et al. Methods of food safety question answering system based on LSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(Supp.): 380 - 384. (in Chinese)

[29] 李想,魏小红,贾璐,等.基于条件随机场的农作物病虫害及农药命名实体识别[J].农业机械学报,2017,48(增刊):178 - 185.
LI Xiang, WEI Xiaohong, JIA Lu, et al. Recognition of crops, diseases and pesticides named entities in Chinese based on conditional random fields[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(Supp.): 178 - 185. (in Chinese)

[30] 林海琳,杨辉荣.农药有机合成的进展与展望[J].化工科技,1998(4):17 - 20.
LIN Hailin, YANG Huirong. The development and prospect for organic synthesis of pesticides[J]. Science & Technology in Chemical Industry, 1998(4): 17 - 20. (in Chinese)