

doi:10.6041/j.issn.1000-1298.2023.03.031

基于 CDSSM 的作物病害处方推荐方法

张领先¹ 赵聃桐¹ 丁俊琦¹ 乔岩²

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 北京市植物保护站, 北京 100029)

摘要: 作物病害诊断积累了大量电子处方数据,对电子处方数据二次利用,实现作物病害处方智能推荐是植保领域重要的研究内容。对此,本文构建基于 CDSSM 的作物病害处方推荐模型,实现多种类作物病害的诊断和处方推荐。基于病害标准知识库对作物病害处方数据进行筛选,并进行数据扩充,同时结合领域知识构建标准处方库;构建基于 CDSSM 的作物处方推荐模型,根据文本特征生成语义向量,计算语义向量的余弦距离,结合标准处方库完成融合地区、时间、作物种类、生长期等多个因素的处方精准推荐。从病害诊断、处方推荐、针对番茄病害处方推荐和不同输入对处方推荐的影响 4 方面展开结果分析,并与基于 DSSM、DSSM-LSTM、Cosine、Jaccard、BM25 的模型结果进行对比分析;结合实际应用需求设计并构建面向移动终端的作物病害处方推荐应用“处方宝”。结果表明,基于 CDSSM 的作物病害处方推荐模型病害诊断正确率为 71%,处方推荐准确率为 82%,优于其他 5 种作物病害处方推荐模型;针对番茄病害处方推荐准确率更高。本文构建的基于 CDSSM 的作物处方推荐模型可以满足实际应用需求,还能够进行病害种类的扩充,可以作为作物病害处方推荐的高效辅助工具。

关键词: 作物病害处方; 智能推荐; 语义相似度; 深度语义匹配模型

中图分类号: TP391 文献标识码: A 文章编号: 1000-1298(2023)03-0308-10 OSID: 

Recommendation Method of Crop Disease Prescription Based on CDSSM

ZHANG Lingxian¹ ZHAO Dantong¹ DING Junqi¹ QIAO Yan²

(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

2. Beijing Plant Protection Station, Beijing 100029, China)

Abstract: Crop disease diagnosis has accumulated a large number of electronic prescription data. It is an important practical problem that how to make secondary use of electronic prescription data to realize intelligent recommendation of crop disease prescription in the field of plant protection. A CDSSM-based crop disease prescription recommendation method was constructed to realize the diagnosis and prescription recommendation of multiple crop diseases. Based on the disease standard knowledge base, the crop disease prescription data were screened and expanded, and the standard prescription database was constructed combining with the domain knowledge. The CDSSM-based crop prescription recommendation model was constructed, semantic vector was generated according to text features, Cosine distances of semantic vectors were calculated, and prescription recommendation was completed with standard prescription database. The results were analyzed from four aspects of disease diagnosis, prescription recommendation, tomato disease prescription recommendation and influence of different inputs on prescription recommendation. The results were compared with models based on DSSM, DSSM-LSTM, Cosine, Jaccard and BM25. Combined with the actual application requirements, the mobile terminal oriented crop disease prescription recommendation application “Prescriptionist” was designed and constructed. The results showed that the accuracy of disease diagnosis of CDSSM was 71%, the accuracy of prescription recommendation was 82%, which were better than that of the other five crop disease prescription recommendation models. The recommendation accuracy of tomato disease prescription was higher. The CDSSM-based crop prescription recommendation model constructed can meet the practical application requirements, and also expand the disease types, which can be used as an efficient auxiliary tool for crop disease prescription recommendation.

Key words: crop disease prescription; intelligent recommendation; semantic similarity; deep semantic matching model

0 引言

病害防治一直是农业生产中不可忽视的关键环节之一。作物病害不仅造成农作物产量严重损失,也给生产者带来经济损失,严重威胁农业生产安全^[1-2]。

处方开具是作物病害防治中重要的一部分^[3]。然而,传统的作物病害处方以人工开具为主,这对于工作人员的专业素质要求较高,且成本高、效率低;随着计算机技术的发展,专家系统问世,大大提升了开具作物病害处方的效率,但是专家系统后期维护成本高,且往往只针对一种或几种作物,普适性不足^[4]。

目前,随着农业信息化建设的推进,以电子病历为载体的农作物病害处方数据大量积累,给作物病害防治提供了新的可能^[5]。挖掘现有处方数据中的有效信息,开展处方推荐相关研究,将有效缓解当前作物病害防治的困境。而在生物医学领域,通过挖掘处方数据中的有效信息进行处方推荐已取得诸多成果^[6-9]。通过大数据分析获得问诊对象的发病特征,结合多个对象之间的特征相似度以及处方之间的相似度,推测出问诊对象可能需要的处方并生成推荐列表,实现智能化开具处方^[10-12]。

处方数据以文本为主,因此挖掘处方数据离不开文本特征分析。近年来,深度学习在文本特征分析领域取得了突破性的进展,结合卷积神经网络(Convolutional neural network, CNN)的深度语义匹配模型(Deep structured semantic model, DSSM)——CDSSM^[13]作为目前最有效、应用最广的文本特征计算方法之一,在广告推荐、检索结果推荐等场景中取得了许多成果^[14-15]。CDSSM 的神经网络架构能够分析文本的多层次语义,将文本表示为包含语义特征的向量,进而依据不同处方之间文本的语义匹配度实现处方推荐。与传统的文本相似度计算方法相比,CDSSM 的文本特征学习和表达能力更优。传统的文本相似度计算方法局限于表面文字相似度,计算过程较为机械;在数据量较大、表示较为复杂的情况下,CDSSM 通过语义匹配相比传统方法能够实现更为精准的推荐。

针对作物病害处方开具困难,处方数据大量积累但缺少有效利用的问题,本文提出基于 CDSSM 的作物病害处方推荐方法,通过计算处方文本的语义特征,实现基于文本相似度的作物病害处方推荐。

1 处方库与处方相似度模型构建

本文作物病害处方推荐分为数据预处理、模型构建和面向应用场景的处方推荐3部分。

数据预处理经过数据规范化、有效数据筛选等,通过数据集划分和数据扩充得到可行的数据集,并在数据清洗的基础上构建作物病害标准处方库,为后续处方推荐提供数据支持;模型构建部分,基于 CDSSM 构建包含输入层、表示层和匹配层的处方推荐模型;面向应用场景的处方推荐部分,通过作物病害处方推荐模型计算作物病害描述与标准处方库的标准描述间的语义相似度,将得分由高到低排序,选取得分最高的标准描述,在标准处方库中匹配标准描述对应的诊断,将该诊断作为作物病害诊断结果,对应治疗方案为待诊断作物的最佳治疗方案,最终完成处方推荐(图1)。

1.1 数据获取

本文所使用的数据为北京市植物保护站植物医生2017年10月30日—2021年7月13日的诊断记录,共计10万余条数据。由北京市植物保护站主导,开发出包括问诊农户、植物诊所、作物和症状、诊断结果和防治建议的标准化处方。

本文数据内容可分为结构化数据和非结构化数据两部分,其中结构化数据有所属区县、所属乡镇、所属村庄、开具时间、作物、发育阶段、受害部位、发生面积(亩)、发生比重、主要症状、田间分布症状、诊断结果共12部分;非结构化数据有问诊记录、开具农药名称、开具农药数量3部分。由农户提供作物病害信息,植物医生凭借个人从业经验,并依据基于CABI plantwise网站编写的《病害明白纸》^[16]等专业文献,结合病害发生机理,按照传统病害诊疗方式进行诊断,最终开具作物病害处方(图2)。

1.2 数据预处理

本文数据预处理分为数据规范化、有效数据筛选、数据集划分和数据扩充。通过数据预处理获得本文模型可用的数据集,以及本文模型进行处方推荐时所依赖的标准处方库和领域词典。根据原始处方数据具体内容以及作物病害领域知识,迭代性地进行数据清洗和有效数据选择,最终将经过优化的作物病害处方作为标准处方库,作物病害信息为唯一值,与诊断结果和治疗建议一一对应,为后续进行处方推荐提供数据支持(图3)。

1.2.1 数据规范化

本文研究对象为作物病害,故选择病害诊断结果的诊断记录作为原始数据集,共42726条。由于采用的数据直接取自植保站,问诊信息由“植物医生”诊断时手动输入生成,带有口语化特点,规范性不足。本文通过数据清洗,将不规范诊断结果,如霜霉病、黄瓜角斑病、靶斑病修正为标准结果白菜霜霉病、黄瓜细菌性角斑病、黄瓜靶斑病。整理前病害类

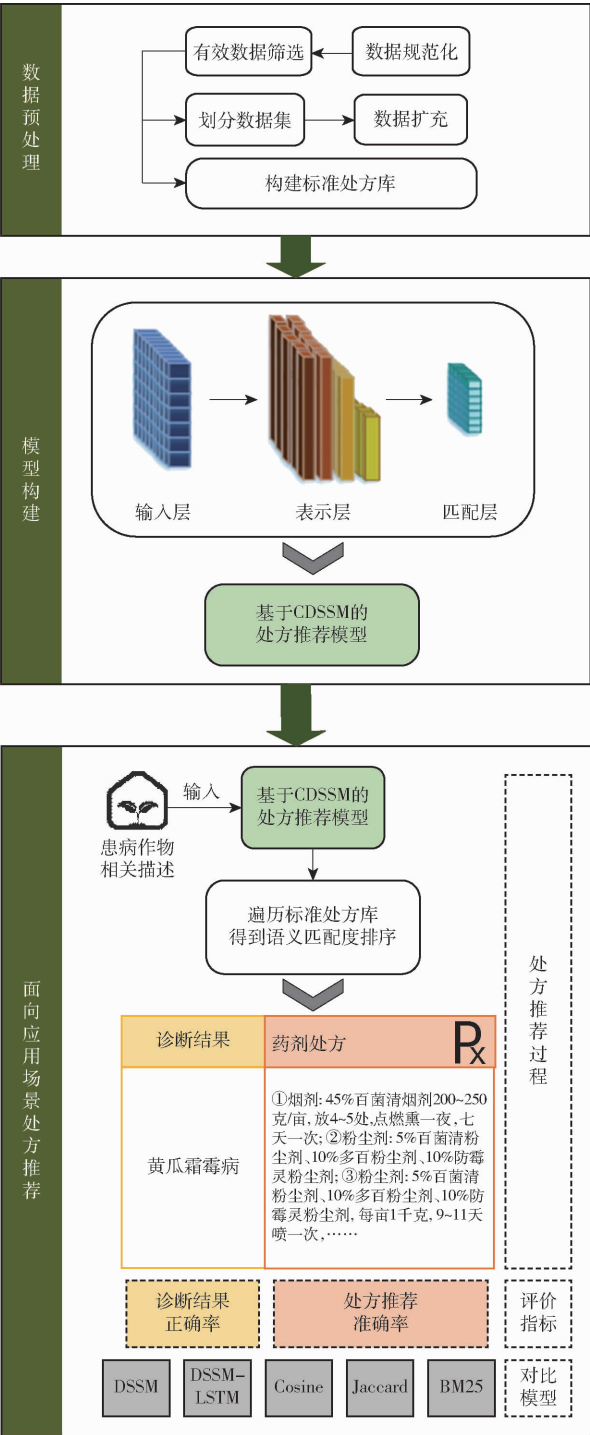


图1 作物病害处方推荐流程图

Fig. 1 Crop disease prescription recommendation process

型为54种,整理后为32种。

1.2.2 有效数据筛选

传统的作物病害诊断知识具有经验性、因果性的特点,植物医生水平参差不齐,诊断内容均为手动输入,导致部分处方诊断结果有误、开具农药名称不准确、个别信息异常。

目前,国内外已构建许多可公开使用的农作物病害知识库,各个知识库的特点和内容存在差异。本文选择基于CABI plantwise 网站编写的病害明白

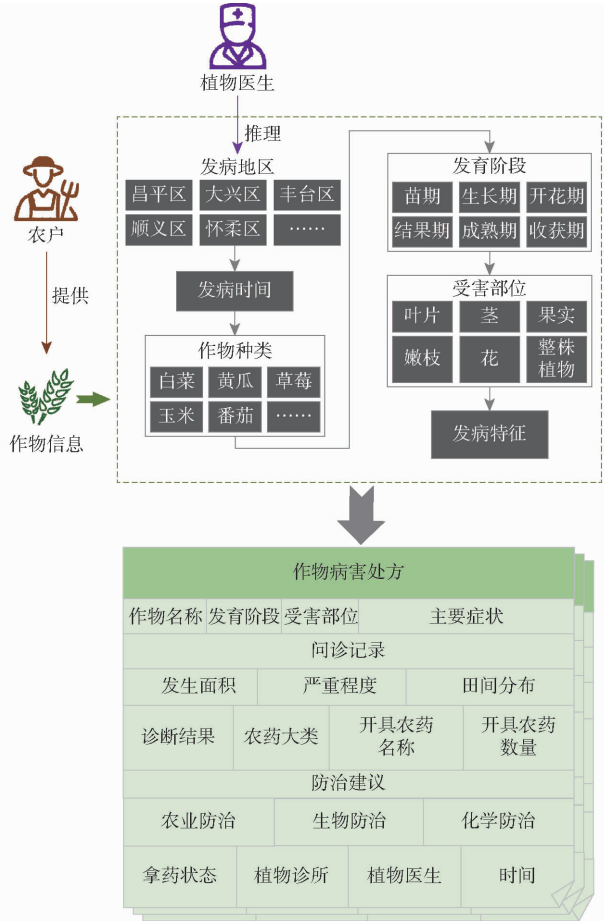


图2 作物病害处方开具过程

Fig. 2 Course of crop disease prescription

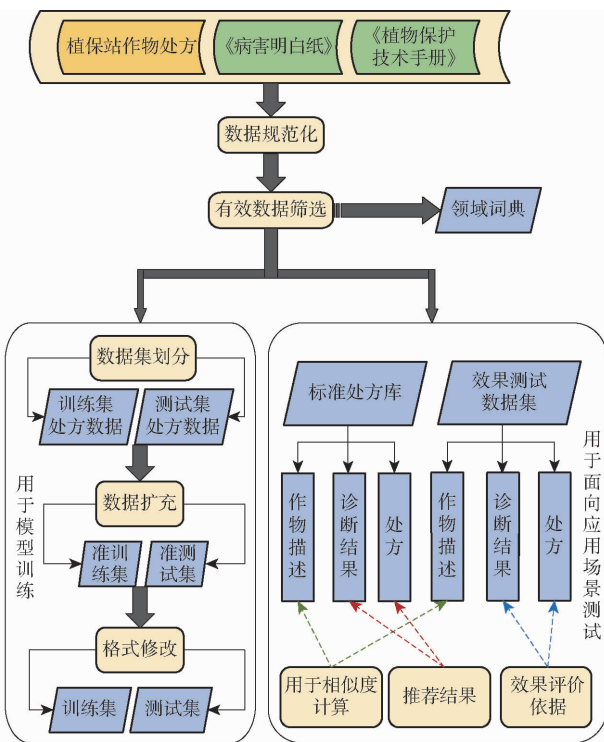


图3 数据预处理流程

Fig. 3 Data preprocessing process

纸知识库,结合《农业植物病理学》^[17],无论在知识数据支持还是实用性上,都能便捷准确地提供作物

发病特征、发生原因、防治方法、农药分类等信息。借助知识库在数据预处理时剔除上述失误造成的无效数据,从而保证数据集质量。筛选完成后得到 20 911 条可用的处方记录。处方数据主要包括白菜、草莓、黄瓜、番茄、辣椒、西瓜 6 种作物,涵盖的病害种类共 32 种,其中根结线虫的发病作物无特定种类,对应 972 条处方数据。与黄瓜相关的病害种类最多,共 10 种,与番茄相关的处方数据量最多,共 9 430 条,其次是黄瓜(6 753 条),处方数据量在一定程度上反映了北京市作物种植比例(表 1)。

表 1 数据集数量统计
Tab.1 Data set quantity statistics

作物	病害	处方 数量/条	作物	病害	处方 数量/条
白菜	白菜黑腐病	216	番茄	番茄病毒病	2 725
	白菜霜霉病	167		番茄灰霉病	1 761
	白菜炭疽病	136		番茄溃疡病	139
	总计	519		番茄青枯病	367
草莓	草莓白粉病	697	番茄	番茄晚疫病	3 128
	草莓灰霉病	127		番茄叶霉病	642
	总计	824		番茄早疫病	668
白菜、西瓜、黄瓜	根结线虫	972	总计		9 430
黄瓜	黄瓜靶斑病	70	辣椒	辣椒病毒病	227
	黄瓜白粉病	1 717		辣椒疮痂病	127
	黄瓜猝倒病	86		辣椒炭疽病	165
	黄瓜灰霉病	462		辣椒疫病	313
	黄瓜枯萎病	101	总计		832
	黄瓜立枯病	219	西瓜	西瓜白粉病	189
	黄瓜霜霉病	2 370		西瓜枯萎病	211
	黄瓜炭疽病	138		西瓜蔓枯病	190
	黄瓜细菌性角斑病	1 477		西瓜炭疽病	736
	黄瓜疫病	113		西瓜疫病	255
	总计	6 753	总计		1 581

1.2.3 数据集划分

采用分层抽样,按照 7:3划分 14 618 条和 6 293 条数据用于生成训练集和测试集。

一般而言,一条完整的作物病害处方应当包含问诊农户信息、受害作物信息、诊疗处方。简单起见,本文暂时仅关注受害作物信息和诊疗处方。其中受害作物信息是作物病害诊断的重要参考因素,是进行作物病害诊断和诊疗处方推荐的核心判断依据。由此,本文以受害作物信息为主筛选标准处方集。

1.2.4 数据扩充

由于作物病害种类多,数据筛选后部分病害对应的数据量较少,直接使用将导致模型过度依赖部分属性,降低模型性能,本文采用数据增强对数据集

进行扩充。

本文使用 EDA(Easy data augmentation)^[18] 分别对训练集数据和测试集数据按照同一比例进行扩充。其中包括 4 种方法,分别为同义词替换(Synonyms replace, SR)、随机插入(Randomly insert, RI)、随机交换(Randomly swap, RS)和随机删除(Randomly delete, RD)。设句长为 l 个单词,替换比例为 α ,同义词替换基于 Python 的 Synonyms 模块从句子中随机选取 $l \times \alpha$ 个不属于停用词集的单词,并随机选择其同义词进行替换;随机插入将 $l \times \alpha$ 个单词的同义词插入到句子中任意位置;随机交换将句子中 $l \times \alpha$ 个单词随机交换;随机删除剔除句子中 $l \times \alpha$ 个单词。本文设置替换比例 $\alpha = 0.1$ 。

数据扩充后基本实现数据平均分布。图 4 中横坐标轴数字代表不同病害类型,1 代表白菜黑腐病,2 代表白菜霜霉病,3 代表白菜炭疽病,4 代表草莓白粉病,5 代表草莓灰霉病,6 代表番茄病毒病,7 代表番茄灰霉病,8 代表番茄溃疡病,9 代表番茄青枯病,10 代表番茄晚疫病,11 代表番茄叶霉病,12 代表番茄早疫病,13 代表根结线虫,14 代表黄瓜靶斑病,15 代表黄瓜白粉病,16 代表黄瓜猝倒病,17 代表黄瓜灰霉病,18 代表黄瓜枯萎病,19 代表黄瓜立枯病,20 代表黄瓜霜霉病,21 代表黄瓜炭疽病,22 代表黄瓜细菌性角斑病,23 代表黄瓜疫病,24 代表辣椒病毒病,25 代表辣椒疮痂病,26 代表辣椒炭疽病,27 代表辣椒疫病,28 代表西瓜白粉病,29 代表西瓜枯萎病,30 代表西瓜蔓枯病,31 代表西瓜炭疽病,32 代表西瓜疫病。

通过随机匹配的方式,将扩充后的处方数据进一步处理,最终得到可以输入本文处方推荐模型的训练集数据 53 367 条,测试集数据 22 109 条,共计 75 476 条。

1.2.5 作物病害标准处方库构建

本文以筛选完成的有效数据为基础,进行标准处方库的构建。标准处方库应体现原始数据集中作物发病规律特征,并保证处方数据准确无误,以辅助本文在实际应用环节进行作物病害处方推荐。本文结合作物病害领域知识从 CABI plantwise 网站、《植物保护技术手册》等现有网站文献和标准处方集,整理得到标准处方库。

1.3 基于 CDSSM 的作物病害处方相似度计算模型构建

1.3.1 深度语义匹配模型

本文以作物病害处方的精准推荐为目标,需要在标准处方库中进行匹配,如何进行文本的正确匹配,即如何正确分析文本语义十分重要。传统的匹

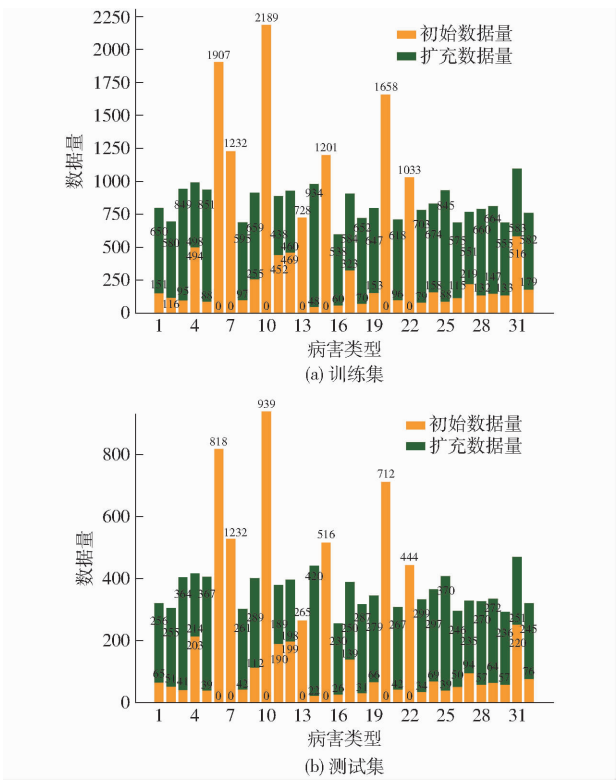


图4 数据集扩充
Fig.4 Data expansion

配方法没有考虑语义信息,大多从词、句式、语法结构出发,依赖于特定规则,硬性匹配效果并不理想^[19]。

结合卷积神经网络(Convolutional neural network, CNN)的深度语义匹配模型(Deep structured semantic model, DSSM)CDSSM,也称为卷积潜在语义模型(Convolutional latent semantic model, CLSM),其核心思想是将一对文本映射到共同维度的语义空间中,通过余弦相似度计算一对文本的语义相似度,从而训练得到隐含语义模型。DSSM模型^[20]最初以词袋模型(Bag of words, BoW)为文本映射方式,但词袋模型不考虑文本的上下文关系,只考虑单词所占权重,即单词出现的频率,所以本文将词袋模型改为词向量映射,模型包括输入层、表示层和输出层(图5)。

在输入层完成文本数据到词向量的转换,在表示层通过CNN结构将词向量转换为语义向量,表示处方内容概念的多个层次,最终在匹配层计算语义向量间的余弦相似度,作为处方相似度的评价依据。

将CDSSM应用于农业领域的作物病害处方推荐,基于词向量映射、卷积神经网络的语义表示和语义向量匹配,能够对现有的作物处方数据展开有效的二次利用,根据历史数据文本特征体现的作物发病规律,融合地区、时间、作物种类、生长期等多个因素进行处方精准推荐。同时,由于数据的多样性,实

现的处方推荐面向多种作物和病害,有效降低了病害相关研究的模型训练成本。

1.3.2 输入层

输入层将作物病害数据映射到向量空间,并将其作为表示层的输入(图6)。将预处理后的处方数据进行分词,生成词向量,再将其按序组合生成句向量,保持了上下文语义关系。

本文设置词汇表大小为3 000,最大序列长度为60。

1.3.3 CNN表示层

经过输入层处理后生成的句向量为 x ,经过一维卷积层和最大池化层处理后生成维度为512的输出向量 O_2 。用 w_i 表示第 i 层的权值矩阵, b_i 表示第 i 层的偏差项。则第一隐藏层向量 l_1 、第 i 个隐藏层向量 l_i 、输出向量 y 可分别表示为

$$l_1 = w_1 x \tag{1}$$

$$l_i = f(w_i l_{i-1} + b_i) \quad (i = 2, 3, \dots, n-1) \tag{2}$$

$$y = f(w_n l_{n-1} + b_n) \tag{3}$$

用tanh作为激活函数。最终输出一个128维的低维语义向量(图7)。

本文设置学习率衰减为 1×10^{-4} ,批量大小为32,设置参数input_keep_prob为0.8。

1.3.4 匹配层

匹配层通过余弦相似度实现对两个语义向量相似度的计算。对于作物病害处方 p_1 和 p_2 ,其对应的语义向量分别为 y_1 和 y_2 ,作物病害处方 p_1 和 p_2 相关性计算公式为

$$R(p_1, p_2) = \text{cosine}(y_1, y_2) = \frac{y_1^T y_2}{\|y_1\| \|y_2\|} \tag{4}$$

1.4 硬件、软件环境和评价指标

实验软件环境为Python 3.6和TensorFlow 1.15.0,操作系统为Windows 10(64位),硬件环境为AMD Radeon(TM) R7 350(2 048 MB)显卡、Intel(R) Core(TM) i7-6700 CPU @ 3.40 GHz。

在实际农业生产中,农户及植物医生对作物病害的描述以及开具处方的多样性非常高,农药的使用及组合十分灵活,即使是同种作物同一时期的同种病害的描述,也可能由于农户个人偏好的不同有较大差异,且对于同种病害,不同植物医生开具的处方也可能有所不同。这使得预测完全一致的作物病害处方有较大困难,因此本文采用top-n推荐方式,提升作物病害处方推荐能力。

在处方推荐效果评价中,本文模拟实际应用场景,将作物病害描述作为输入,以标准处方库为依据,获得诊断结果及处方,对诊断结果的正确率(Accuracy)和推荐处方准确率(Precision)进行计

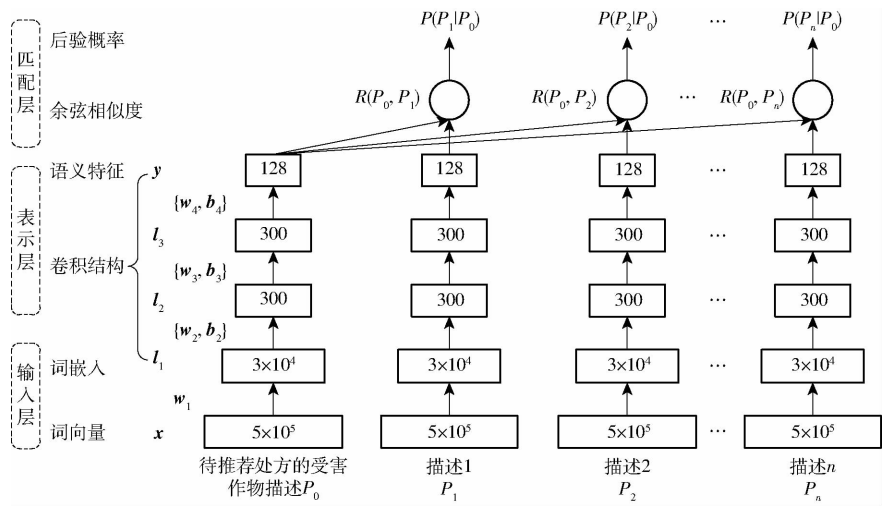


图 5 CDSSM 模型结构
Fig. 5 CDSSM model structure

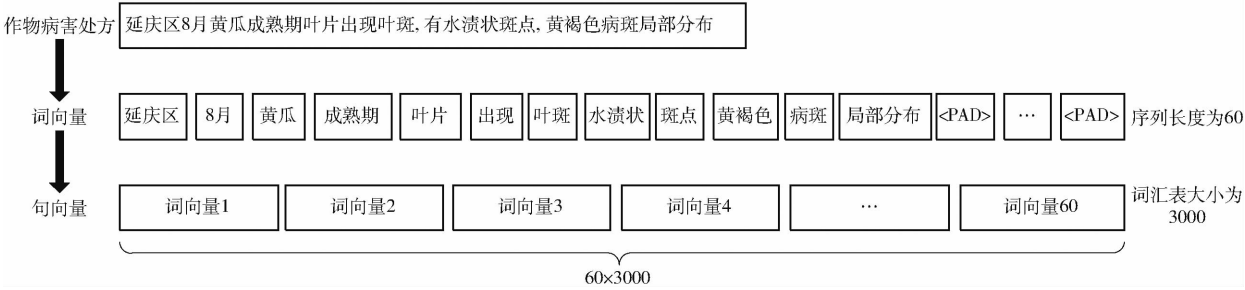


图 6 CDSSM 输入示例
Fig. 6 Input sample of CDSSM

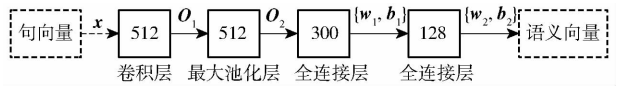


图 7 CNN 表示层
Fig. 7 Presentation layer of CNN

算,作为处方推荐效果的评价指标。其中, n_p 表示推荐准确的处方数量, n_t 表示推荐的处方总数目。本文设置处方相似度阈值为 0.8,选取 top - 5 推荐方式。

2 实验结果与分析

2.1 作物病害诊断

目前文本相似度计算方法可分为表面文本相似度计算和语义相似度计算。表面文本相似度计算直接针对原始文本,作用于字符串序列或字符组合,其算法原理简单、易于实现,是研究历史最长的一类文本相似度算法。本文选取 3 种表面文本相似度计算方法与所研究的基于深度学习的语义相似度计算模型进行对比,分别为 Cosine^[21]、BM25^[22]、Jaccard^[23]。另选取 DSSM、DSSM - LSTM 两种与 CDSSM 结构类似的深度学习语义相似度计算模型作为对照。结合实验对比结果,选择结合领域词典的结巴分词,基于 CDSSM 的处方推荐模型进行对比

分析。

基于 Cosine、BM25、Jaccard 和 DSSM、DSSM - LSTM、CDSSM 6 种方法的作物病害处方推荐模型的正确率分别为 33%、24%、21% 和 40%、38%、71%。其中基于 CDSSM 的模型处方推荐正确率更高,这说明语义相似度计算更有助于正确分析文本信息,深度学习在作物病害处方推荐方面具有研究和应用价值。CDSSM 使用了 CNN 结构,诊断准确率略高于 DSSM - LSTM(图 8)。

对作物病害诊断结果正确率分析可知,数据量对诊断结果有一定影响,作物病害对应的数据量越大,其诊断效果越好。基于深度学习的作物病害推荐模型效果总体好于基于表面文本相似度的模型,说明基于深度学习的模型对语义进行计算后,对于处方文本相关度计算优于表面文本相似度。对于白菜黑腐病、辣椒炭疽病等,所有模型的诊断正确率均较好,这与数据量大有关,也与其患病特征辨识度高、关键词集中有关。对于辣椒病毒病等病害出现了基于深度学习的模型正确率低于基于表面文本的模型的情况,这是由于该病害测试数据较短,且存在易混淆描述,导致基于语义相似度的深度学习模型无法发挥计算优势,造成误判,这也是本文模型的不

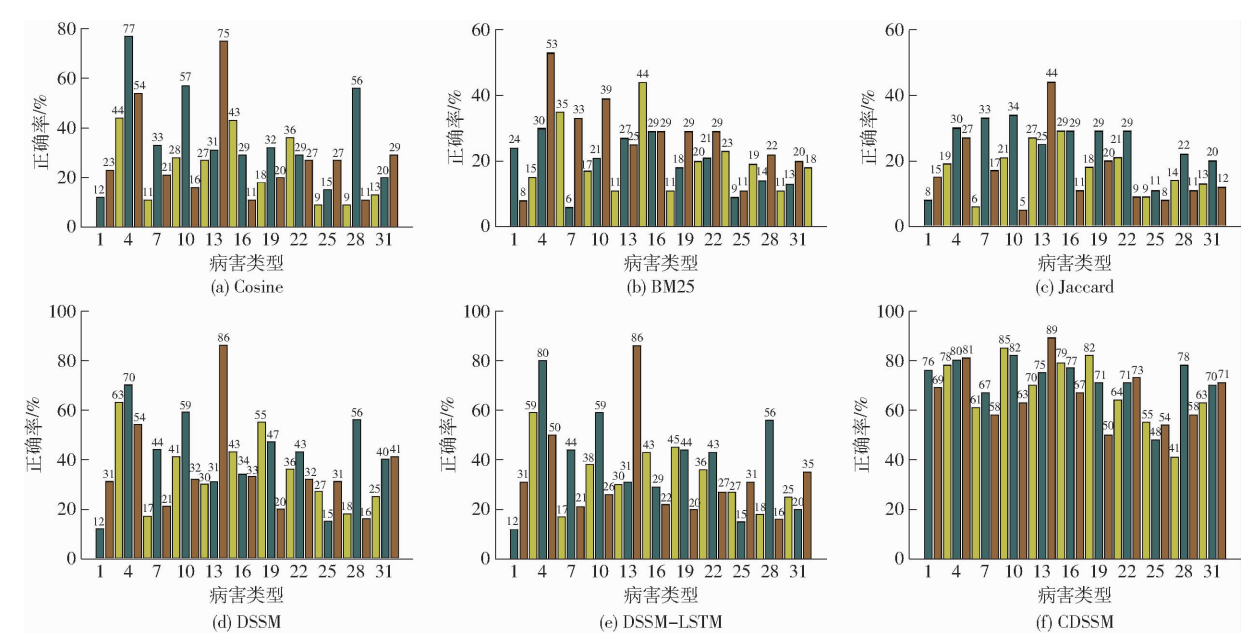


图 8 不同模型对不同病害的诊断正确率

Fig. 8 Accuracy of different models for different diseases

足之处,若扩大数据量,同时改进模型训练方式,将有所改善。

2.2 处方推荐

本文将模型推荐的处方与标准处方进行余弦相似度计算,对模型处方推荐准确率评价。Cosine、BM25、Jaccard、DSSM、DSSM-LSTM、CDSSM 模型准确率分别为:63%、49%、41%、70%、66%、82%,基于深度学习的作物病害推荐模型准确率总体好于基于表面文本相似度的模型。对于白菜霜霉病、番茄溃疡病等病害出现诊断结果正确率不高,处方相似度高的情况,这是由于该病害对应的处方中药剂种类较多,诊疗方法丰富,易出现与其他病害对应处方内容相似的情况。对于番茄叶霉病等病害,在基于表面文本相似度计算的方法中,出现处方相似度较低,诊断正确率较高的情况,这是由于该病害不同发病时期对应不同药剂处方,基于表面文本相似度的计算方法只能正确判断病害种类,无法判断病害所处发生时期。

总体来看基于深度学习的模型表现优于传统方法,说明语义分析能够更准确地计算药剂处方与患病特征之间的关联,从而实现处方精准推荐(图 9)。

在推荐处方准确率评价中,CDSSM 明显优于 DSSM-LSTM,所以基于 CDSSM 的作物病害处方推荐模型表现更好。

2.3 番茄病害处方推荐

为研究处方推荐模型对于单一作物的处方推荐效果,本文选取种植数量、问诊记录最多的番茄作为研究对象,对番茄的 7 种病害(番茄病毒病、番茄灰

霉病、番茄溃疡病、番茄青枯病、番茄晚疫病、番茄叶霉病、番茄早疫病)展开实验分析。同时对不同的 top-n 推荐方式进行对比。根据全部病害的实证结果分析可知,本文使用的基于深度学习的语义匹配模型相比传统的文本相似度计算方法具有更强的文本分析能力,因此选择基于深度学习的语义匹配模型(DSSM、DSSM-LSTM、CDSSM)作为文本分析工具,番茄病害处方推荐实证分析实验结果如表 2 所示。

总体来看,基于 CDSSM 的处方推荐模型在所有的番茄病害中,诊断结果正确率以及处方推荐准确率都更高。在 top-5 推荐策略下,对于大多数番茄病害的诊断正确率以及处方推荐准确率都能达到 95%,对于每一种病害处方推荐表现都基本能够满足应用需求。相比其他番茄病害,3 种模型对于番茄病毒病的诊断和处方推荐都表现较差,这是由于番茄病毒病的病原种类较多,不同病原对应的发病部位不同,所以作物受害的症状较为多样^[24],而训练集与测试集中的番茄病毒病处方数据没有涵盖所有病原,所以模型训练时对番茄病毒病的症状语义分析能力有所欠缺,无法准确分析数据集中没有涵盖的病原所导致的番茄病毒病发病症状。如果丰富番茄病毒病的处方数据,将有效提升模型的处方推荐效果。

由于处方的多样性,本文使用了 top-n 推荐策略。根据表 2 可以看出,随着 top-n 的选择范围扩大,处方推荐效果有所提升,但整体的准确率变化幅度不大,说明模型对于作物受害描述的语义分析能力是可靠的,能够正确将语义相似度较高的推荐结

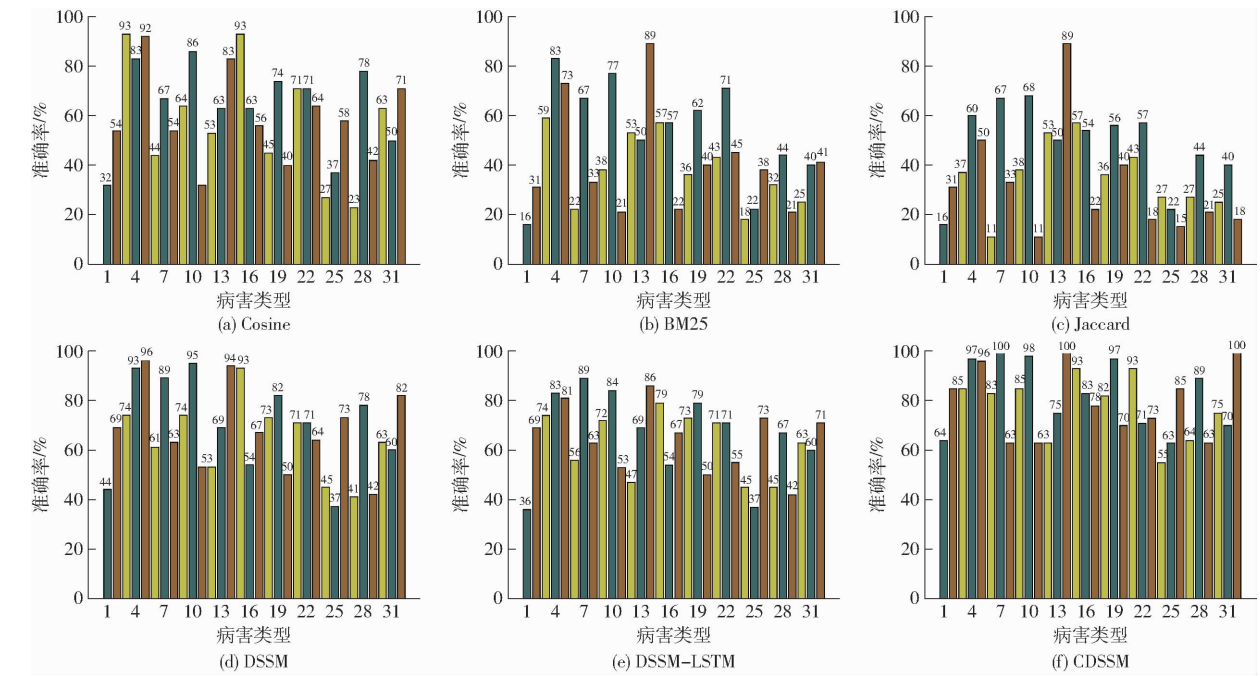


图9 不同模型对不同病害的处方推荐准确率

Fig.9 Precision of different models for different diseases

表 2 番茄病害处方推荐结果

Tab.2 Disease prescription recommended results of tomato

模型	番茄病害名称	top - 5		top - 4		top - 3		top - 2		top - 1	
		正确率	准确率	正确率	准确率	正确率	准确率	正确率	准确率	正确率	准确率
DSSM	番茄病毒病	20	60	20	50	20	50	20	50	10	40
	番茄灰霉病	95	70	95	70	95	70	70	70	70	70
	番茄溃疡病	80	80	80	80	80	80	70	70	30	30
	番茄青枯病	95	95	95	95	95	95	95	95	60	60
	番茄晚疫病	95	95	70	95	70	95	70	95	70	95
	番茄叶霉病	50	70	30	70	30	70	20	30	20	20
	番茄早疫病	40	80	40	80	40	80	40	40	40	20
	平均值	60	80	50	80	50	80	50	70	40	50
DSSM - LSTM	番茄病毒病	30	60	30	60	20	50	10	40	10	30
	番茄灰霉病	70	70	30	70	30	70	30	70	10	10
	番茄溃疡病	20	20	20	20	10	10	10	10	10	10
	番茄青枯病	40	40	40	40	40	40	30	30	30	20
	番茄晚疫病	95	95	95	95	30	30	20	30	20	10
	番茄叶霉病	20	20	20	20	20	20	20	20	20	20
	番茄早疫病	95	80	95	80	95	80	80	60	60	50
	平均值	50	60	50	60	30	40	20	30	10	20
CDSSM	番茄病毒病	70	60	60	60	60	60	60	60	50	50
	番茄灰霉病	95	95	95	95	95	95	70	70	70	60
	番茄溃疡病	95	95	80	80	80	80	80	80	80	80
	番茄青枯病	95	95	95	95	95	95	95	95	60	60
	番茄晚疫病	95	95	95	95	95	95	95	95	70	70
	番茄叶霉病	95	95	95	95	95	95	95	95	95	95
	番茄早疫病	95	95	95	95	95	95	95	95	95	95
	平均值	90	90	90	90	80	90	70	80	70	70

果排在前边。为了满足实际应用需求,并控制推荐结果的长度,本文选择 top - 5 作为处方推荐策略。

2.4 不同输入对方推荐的影响

根据病三角原理^[25],本文方法在实际场景进行作物病害诊断时,输入内容可分为环境、作物、症状

3 部分,其中环境包括发病地区、发病时间,作物包括作物种类、发育阶段、受害部位,症状为发病特征。番茄是北京市种植较广泛且种植量最多的作物之一,番茄晚疫病为番茄发病率较高的病害之一,本文以番茄晚疫病为例,不同输入对方推荐的影响如表 3 所示。

表 3 不同输入对方推荐的影响

Tab.3 Result of different inputs on prescription recommendation			
模型	输入	正确率/%	准确率/%
DSSM	症状	16	39
	环境 + 症状	20	43
	作物 + 症状	41	77
	环境 + 作物 + 症状	61	89
DSSM - LSTM	症状	18	43
	环境 + 症状	23	70
	作物 + 症状	34	77
	环境 + 作物 + 症状	59	84
CDSSM	症状	41	57
	环境 + 症状	45	77
	作物 + 症状	61	84
	环境 + 作物 + 症状	82	93

由于基于语义相似度计算的处方推荐模型明显优于基于表面文本相似度的方法,本文选择了基于语义相似度的推荐方法进行不同输入的对比研究。发现症状描述对于模型的诊断及处方推荐影响显著,而环境信息和作物信息影响较弱。这表明在进行处方推荐时需要着重描述作物病害症状,可以提高处方推荐准确率(表 3)。

同时仅仅输入症状信息的处方推荐效果不如输入完整信息时,准确率有所降低,说明环境信息和作物信息对于处方推荐有一定影响,在实际应用中输入的信息越全面,模型处方推荐准确率越高。在输入完整信息时,使用了 CNN 结构的 CDSSM 模型推荐的处方准确率最高,也体现了 CNN 结构对于数据特征提取的优势(表 3)。

3 移动终端构建与应用

3.1 设计目标

为实现快速准确地为受害作物开具处方,及时控制作物病情,减轻农户种植负担,本文设计并实现了基于 CDSSM 的面向移动终端的作物病害处方推荐系统“处方宝”,为农户提供快捷高效的作物病害处方推荐服务。

3.2 系统实现

作物病害处方推荐系统的主要用户为广大农户,需要具有便捷、即时的特点,因此采用

Django 与微信小程序结合的方式进行移动终端的开发。

农户通过输入受害作物的相关描述,实现即时便捷的作物病害处方推荐;同时可以查看并下载个人问诊历史记录,进一步降低受害作物问诊成本,并提升个人病害防治水平(图 10)。



图 10 “处方宝”功能界面

Fig. 10 Function interface of Prescriptionist

4 结论

(1)基于 CDSSM 的作物病害处方推荐模型,能够更好地计算处方语义相似度,病害诊断和处方推荐更加精准。病害诊断正确率为 71%,高于 DSSM (40%)、DSSM - LSTM (38%)、Cosine (33%)、BM25 (24%)、Jaccard (21%);处方推荐准确率为 82%,高于 DSSM (70%)、DSSM - LSTM (66%)、Cosine (63%)、BM25 (49%)、Jaccard (41%)。

(2)对一种作物进行病害处方推荐时,模型推荐准确率更高。使用 top - 5 推荐策略时,对番茄病害的整体诊断正确率为 90%,处方推荐准确率为 90%,其中番茄灰霉病、番茄溃疡病、番茄青枯病、番茄晚疫病、番茄叶霉病和番茄早疫病的诊断正确率和处方推荐准确率为 100%。

(3)本文提出的作物病害处方推荐方法不局限于特定的作物种类和病害类型,便于进行病害种类的扩充,可以充分二次利用历史数据,并降低作物病害相关研究的模型训练成本。

参 考 文 献

- [1] 刘万才,卓富彦,李天娇,等.“十三五”期间我国粮食作物植保贡献率研究报告[J]. 中国植保导刊, 2021, 41(4): 33 – 36,51.
LIU Wancai, ZHUO Fuyan, LI Tianjiao, et al. Study on the contribution of plant protection to main crops in China during the 13th Five-Year Plan period[J]. China Plant Protection, 2021, 41(4): 33 – 36,51. (in Chinese)
- [2] 翟肇裕,曹益飞,徐焕良,等.农作物病虫害识别关键技术研究综述[J]. 农业机械学报,2021,52(7):1 – 18.
ZHAI Zhaoyu, CAO Yifei, XU Huanliang, et al. Review of key techniques for crop disease and pest detection[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(7): 1 – 18. (in Chinese)
- [3] 邱荣洲,陈韶萍,赵健,等.农作物病虫害防治农药速查系统构建与应用[J]. 农药学学报, 2022, 24(3):630 – 636.
QIU Rongzhou, CHEN Shaoping, ZHAO Jian, et al. Construction and application of pesticide fast searching system for crop diseases and pests control[J]. Chinese Journal of Pesticide Science, 2022, 24(3): 630 – 636. (in Chinese)
- [4] 张彬露. 农业专家系统概述与优化研究[J]. 农村经济与科技,2020,31(15):341 – 342,354.
ZHANG Binlu. Overview and optimization of agricultural expert system[J]. Rural Economy and Science-Technology, 2020, 31(15): 341 – 342,354. (in Chinese)
- [5] 田娜,杨晓文,单东林,等. 我国数字农业现状与展望[J]. 中国农机化学报, 2019, 40(4): 210 – 213.
TIAN Na, YANG Xiaowen, SHAN Donglin, et al. Status and prospect of digital agriculture in China[J]. Journal of Chinese Agricultural Mechanization, 2019, 40(4): 210 – 213. (in Chinese)
- [6] ACOSTA J N, DORR F, GOICOCHEA M T, et al. Acute headache diagnosis in the emergency department accuracy and safety of an artificial intelligence system[J]. Neurology, 2019, 92: 5 – 10.
- [7] ARAUJO J M M, PEIXOTO Z M A. A new proposal for automatic identification of multiple soybean diseases[J]. Computers and Electronics in Agriculture,2019, 167: 105060.
- [8] 臧靖,陈宁,李锦,等. 我院门急诊处方前置审核系统的设计与运行实践[J]. 中国药学杂志, 2019,54(5):420 – 423.
ZANG Jing, CHEN Ning, LI Jin, et al. Design and operation practice of up-front prescription review/intervention system for outpatients in our hospital[J]. Chinese Pharmaceutical Journal, 2019, 54(5): 420 – 423. (in Chinese)
- [9] RUAN C, WU Y, YANG Y, et al. Semantic-aware graph convolutional networks for clinical auxiliary diagnosis and treatment of traditional chinese medicine[J]. IEEE Access, 2021, 9: 8797 – 8807.
- [10] ZENG X, JIA Z, HE Z, et al. Measure clinical drug-drug similarity using electronic medical records[J]. International Journal of Medical Informatics, 2019, 124: 97 – 103.
- [11] 赵明,董翠翠,黄乔雪,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, HUANG 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)
- [12] 程铭,熊蜀峰,李霏,等. 基于混合注意力机制的脑卒中康复方案推荐[J]. 武汉大学学报(理学版), 2021, 67(6):569 – 577.
CHENG Ming, XIONG Shufeng, LI Fei, et al. Rehabilitation program recommendation for stroke based on hybrid attention mechanism[J]. Journal of Wuhan University (Natural Science Edition), 2021, 67(6): 569 – 577. (in Chinese)
- [13] SHEN Y, HE X, GAO J, et al. A latent semantic model with convolutional-pooling structure for information retrieval[C] // Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, 2014:101 – 110.
- [14] PONTES E L, HUET S, LINHARES A, et al. Predicting the semantic textual similarity with siamese CNN and LSTM[J]. arXiv preprint arXiv:1810.10641, 2018.
- [15] XU J, HE X, LI H. Deep learning for matching in search and recommendation[C] // International ACM Sigir Conference. ACM, 2018:1365 – 1368.
- [16] 张熠. CABI 及其与中国的合作[J]. 世界农业, 1999(11): 20 – 21.
ZHANG Yi. CABI and its cooperation with China[J]. World Agriculture, 1999(11): 20 – 21. (in Chinese)
- [17] 董金皋. 农业植物病理学[M]. 北京: 中国农业出版社, 2015: 408 – 415.
- [18] WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv preprint arXiv:1901.11196, 2019.
- [19] 王春柳,杨永辉,邓霏,等. 文本相似度计算方法研究综述[J]. 情报科学, 2019, 37(3):158 – 168.
WANG Chunliu, YANG Yonghui, DENG Fei, et al. A review of text similarity approaches[J]. Information Science, 2019, 37(3): 158 – 168. (in Chinese)
- [20] HUANG P, HE X, GAO J, et al. Learning deep structured semantic models for web search using clickthrough data[C] // Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, 2013:2333 – 2338.
- [21] GUNAWAN D, SEMBIRING C, BUDIMAN M. The implementation of cosine similarity to calculate text relevance between two documents[C] //2nd International Conference on Computing and Applied Informatics. IOP Publishing, 2018:978.
- [22] LAMSIYAH S, MAHDAOUY EI A, OUATIK EI ALAOUIS, et al. Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, BM25 model, and maximal marginal relevance criterion[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 2:1 – 18.
- [23] NIWATTANAKUL S, SINGTHONGCHAI J, NAENUDORN E, et al. Using of Jaccard coefficient for keywords similarity[C] // Proceedings of the International Multiconference of Engineers and Computer Scientists, IMECS, 2013.
- [24] 李英梅,刘晨,王周平,等. 番茄病毒病的症状识别特征与防治策略[J]. 现代农业科技, 2020(12):143 – 145.
LI Yingmei, LIU Chen, WANG Zhouping, et al. Recognition characteristics and control measures of tomato virus disease[J]. Modern Agricultural Science and Technology, 2020(12): 143 – 145. (in Chinese)
- [25] POKHREL A. Role of individual components of disease triangle in disease development; a review[J]. Journal of Plant Pathology & Microbiology, 2021, 12(9):1 – 7.