



顾静秋,郭旺,朱华吉,郝鹏,吴华瑞. 基于深度学习的动植物新品种精准推荐方法[J]. 中国农业大学学报,2025,30(07):218-229.

GU Jingqiu, GUO Wang, ZHU Huaji, HAO Peng, WU Huarui. An accurate recommendation method for new animal and plant varieties based on deep learning [J]. *Journal of China Agricultural University*, 2025, 30(07): 218-229.

DOI: 10.11841/j.issn.1007-4333.2025.07.19

基于深度学习的动植物新品种精准推荐方法

顾静秋^{1,2,3} 郭旺^{1,2,3*} 朱华吉^{1,2,3} 郝鹏^{1,2,3} 吴华瑞^{1,2,3}

(1. 北京市农林科学院信息技术研究中心, 北京 100097;

2. 国家农业信息化工程技术研究中心, 北京 100097;

3. 农业农村部数字乡村技术重点实验室, 北京 100097)

摘要 针对我国动植物新品种资源信息服务存在的信息不对称、更新不及时、新品种资源难以落地等问题。提出一种结合深度神经网络和注意力机制的两步推荐算法。该算法首先基于动植物品种资源全文数据,构建动植物品种词汇库,搭建Siamese BERT(Bidirectional Encoder Representations from Transformers)网络,通过BERT获取文本上下文关联语义特征,运用双向最大匹配算法获取品种词典特征向量,融合语义及词典特征,通过相似度计算训练回归模型,实现动植物新品种与历史品种相似判断。接下来融合用户静态属性特征、用户行为特征、动植物品种资源特征等品种推荐影响因子,综合分析农业领域用户的行业差异、习惯差异,提出面向动植物品种推荐的用户特征及品种特征表征方式,基于CNN(Convolutional Neural Network)构建用户兴趣网络模型,并引入注意力机制来实现用户与品种的精准匹配推荐。在互联网实时监控新品种资源,实现全自动化的动植物新品种自动推荐。性能测试与效果验证结果表明,该算法在准确率和F1值评价指标上分别达到84.1%和0.832,相比基于协同过滤、矩阵分解等传统推荐算法,能更精准实现动植物新品种的推荐。

关键词 动植物品种资源; 推荐; 深度神经网络; 两步推荐; 兴趣模型

中图分类号 TP399

文章编号 1007-4333(2025)07-0218-12

文献标志码 A

An accurate recommendation method for new animal and plant varieties based on deep learning

GU Jingqiu^{1,2,3}, GUO Wang^{1,2,3*}, ZHU Huaji^{1,2,3}, HAO Peng^{1,2,3}, WU Huarui^{1,2,3}

(1. Research Center of Information Technology, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China;

2. National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China;

3. Key Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing 100097, China)

Abstract In response to the problems of information asymmetry, untimely delayed updates and difficulty in implementing information services for new animal and plant species resources in China, a two-step recommendation algorithm that combines deep neural networks and attention mechanisms is proposed in this study. Firstly, based on the full-text data of animal and plant species resources, the algorithm constructs a vocabulary library of animal and plant species, and builds a Siamese BERT (Bidirectional Encoder Representations from Transformers) network. BERT was used to obtain text context-related semantic features. The bidirectional maximum matching algorithm is used to obtain variety dictionary feature vectors. Integrates semantic and dictionary features, a regression model was trained through similarity calculation to

收稿日期: 2025-01-06

基金项目: 科技创新2030重大项目子课题(2022ZD0115705-05)

第一作者: 顾静秋(ORCID:0009-0005-5492-473X), 副研究员, 主要从事农业智能系统、大数据挖掘研究, E-mail: gujq@nercita.org.cn

通讯作者: 郭旺(ORCID:0000-0001-9344-0854), 助理研究员, 主要从事农业智能系统、文本数据挖掘研究, E-mail: guow@nercita.org.cn

achieve the similarity judgment between new animal and plant varieties and historical varieties. Next, by integrating factors influencing variety recommendation, such as user static attribute characteristics, user behavior characteristics, and animal and plant variety resource characteristics, the industry differences and habits of users in the agricultural field are comprehensively analyzed. A user feature and variety feature representation method is proposed for animal and plant variety recommendation, and a user interest network model is constructed based on Convolutional Neural Network (CNN) and the attention mechanism is introduced to achieve accurate matching and recommendation between users and varieties. By monitoring new variety resources in real time on the internet, fully automatic recommendation of new animal and plant varieties is realized. The experimental performance testing and effectiveness verification results show that the algorithm achieves the accuracy and F1 score evaluation indicators of 84.1% and 0.832, respectively. Compared with the traditional recommendation algorithms based on collaborative filtering and matrix decomposition, the algorithm proposed in this study can more accurately recommend new varieties of animals and plants.

Keywords animal and plant variety resources; recommend; deep neural network; two-step recommendation; interest model

动植物品种资源是农业生产的基础,对于提高农产品的产量和质量,促进农业的可持续发展具有重要作用。随着科技的发展和农业生产方式的转变,新品种资源的研发、推广以及相关技术的更新都产生了大量的信息^[1]。尽管信息数量庞大,但其中的有效信息却往往被淹没在海量的数据中,及时了解并推广应用新的动植物品种资源对于提升农业生产效益具有重要意义。目前,我国动植物新品种资源信息服务存在以下突出问题:1)冷启动问题。以国家农作物品种审定委员会数据库为例,2021—2023年新审定的3 207个农作物品种中,约42%的品种因缺乏历史推广数据而难以匹配适宜区域。2)信息更新延迟。对15个省级农业信息平台的调研显示,品种信息更新周期平均为6.8个月,滞后于品种审定公告时间4.3个月。3)信息不对称。中国农技推广平台用户调查表明,68.5%的农户反映难以获取适合本地区的品种信息,而科研机构研发的品种中有37%因信息传递不畅未能及时推广。

因此,本研究拟提出一种基于深度学习模型的动植物新品种资源推荐方法,通过综合分析用户画像(基础信息、行为偏好)与动植物品种资源特征(品种特性、地理位置、产业类型等元数据),构建用户-资源关联模型,缓解推荐系统冷启动问题,实现动植物新品种需求识别与精准适配,实现精准、及时的农业知识推送服务。

1 相关技术

1.1 用户画像技术

画像构建技术是指通过收集和分析大量数据,

构建个体或群体的详细特征和行为画像,以便更好地理解 and 预测其需求和行为^[2-4]。主要包括基于用户行为的画像方法,基于用户兴趣偏好的画像方法及基于主题的画像方法^[5-6]。用户行为是指用户为满足特定的信息需求在信息行为中采取的各种动作和表现,岳怡然^[7]将用户在使用数字图书馆的过程中产生的一系列的信息数据进行采集,依据这些数据挖掘用户描述性标签,以便能够更精确地定位用户、识别用户。赵又霖等^[8]提出从用户情感认知、用户行为习惯和用户社交网络3种内容维度构建用户画像。基于用户兴趣偏好的画像方法,是以用户兴趣、偏好等用户数据为基础建立用户画像。石宇等^[9]在结合用户群体认知对资源特征进行识别的基础上,对用户感兴趣资源进行兴趣建模。张玉洁等^[10]通过获取用户兴趣偏好建立用户画像实现个性化的广告推荐。基于主题的画像方法是通过用户对用户关联的各种文本信息进行挖掘分析以建立全面而精准的用户画像。陈泽宇等^[11]使用LDA (Latent Dirichlet Allocation)主题模型为每个查询词分配主题,将查询词和其主题共同放入神经网络模型中学习得到其主题词向量,最后采用随机森林分类算法对用户基本属性进行分类构建用户画像。

总体来看,用户画像是一个新兴的研究领域,国内外的学者们从其内涵、特征、模型构建、技术、应用领域等多个角度进行了研究和探讨。但是目前国内关于用户画像的研究尚未形成完整、成熟的理论和体系,用户画像模型构建流程、画像结果准确性的判断、模型评估研究力度仍有不足。从应用领域来看,画像技术在电子商务、图书馆、旅游以

及医疗行业等得到了应用,在其他领域的研究应用尚处于初级阶段。

1.2 资源推荐技术

资源推荐技术近年来得到了快速发展,其核心目标是根据用户的需求和偏好,从大量的资源中筛选出最符合用户需求的资源^[12]。主要的方法有基于协同过滤的推荐方法、基于内容的过滤方法、混合推荐方法、基于深度学习的推荐方法、基于知识图谱的推荐方法。周洋涛等^[13]提出一种基于静态与动态学习需求感知的知识点推荐方法,通过静态感知与动态感知相结合的方式建模复杂知识关联下的用户学习行为,以实现细粒度的个性化知识点推荐。耿杰等^[14]融合注意力机制和指针生成网络,构建了具有双层注意力机制的可解释用户重复消费推荐方法,实现了基于用户重购行为的产品推荐。龙恩等^[15]设计了一种顾及用户画像的遥感信息智能推荐方法,构建融时间、空间、载荷、分辨率、产品级别5项核心元素的可扩展主题用户画像模型,结合各元素权重及兴趣特征值,实现遥感数据有序智能推荐。赵继春等^[16]提出基于LDA标签生成的领域知识图谱构建方法,并构建涉农学习资源知识图谱,设计研发了基于知识图谱语义相似度的协同过滤推荐系统,为用户提供个性化、定制化的学习资源。陈玲洪等^[17]将知识图谱和读者画像技术应用用于图书推荐,构建了读者-图书知识图谱,针对性地改善数据稀疏和冷启动问题。提高了图书推荐的精准度。侯琛等^[18]系统介绍了知识图谱作为农业信息推荐的辅助信息,用于减少冗余推荐,提高推荐结果的精准度的应用现状。刘畅等^[19]将知识图谱构建作为目标,通过结合自注意力机制的动态用户兴趣建模、分层嵌入技术表示和图神经网络多跳推理机制,逐层筛选和聚合复杂关系信息,实现蔬菜生产知识服务的精准推荐。Gopi等^[20]提出了一种新的基于多模态机器学习的作物推荐和产量预测(MMML-CRYP)技术,采用具有内核极限学习机(KELM)技术的平衡优化器(EO)对作物进行有效推荐。Musanas等^[21]基于主要农作物数据集及其关键生长参数训练神经网络模型,进行农作物推荐。

当前推荐技术的发展趋势包括深度学习与神经网络的应用增加、跨领域知识融合、解释性推荐和增强学习的应用,这些趋势提升了推荐系统的精

度和个性化。然而,在特定的应用领域仍存在一些不足,例如数据质量和覆盖范围的限制、模型的适应性和泛化能力不足、对农户画像数据考虑不足,存在知识与用户关联信息缺失、知识更新不灵活、知识匹配不精准等问题。

2 动植物新品种资源数据两步推荐方法

方法基于深度学习框架构建品种特征识别模型获取新品种资源与历史资源的特征相关性,依据用户基础信息,行为信息,动植物新品种资源品种特性、地理位置,产业类型等元数据信息及关联关系构建用户兴趣模型,训练神经网络学习用户与品种资源关注关系。通过品种相似推荐,历史品种推荐用户两步推荐为用户提供动植物新品种资源。

步骤1:进行品种相似推荐。收集动植物品种的文本信息,包括描述、标签、属性等。使用预训练的BERT模型对品种的文本信息进行编码,得到品种的文本表示。构建基于Siamese BERT的品种文本相似度模型,该模型采用共享参数的孪生网络架构,分别对2个品种的文本表示进行编码,通过BERT模型提取[CLS]标记的语义特征向量,最终直接计算2个特征向量之间的余弦相似度作为输出得分。所有历史品种的特征向量均存储在专用向量库中,当新增品种时,通过该模型生成其特征向量并更新至向量库,而模型本身仅在BERT基座或领域词典更新时才需要重新训练。这种设计既保证了相似度计算的准确性,又实现了高效的特征检索。

步骤2:基于用户静态属性数据与行为数据提取用户基本属性特征、用户行为特征。采集动植物品种资源,提取品种资源属性信息。基于用户与品种资源特征数据构建用户-品种关联矩阵,其中行表示用户,列表示品种,每个元素表示用户对该品种的喜好程度或关联程度,构建神经网络模型,模型的输入包括用户静态属性、用户行为特征和品种属性特征,输出为用户与品种的匹配分数。

最后构建知识推荐服务。当新品种出现时,首先利用步骤1的模型获取与新品种相似的历史品种,然后利用步骤2的模型,输入与新品种相似的历史品种信息,获取应该推荐的用户。通过主动推送的形式,将新品种推送至符合用户需求的用户,提供个性化精准知识服务。具体推荐流程见图1。

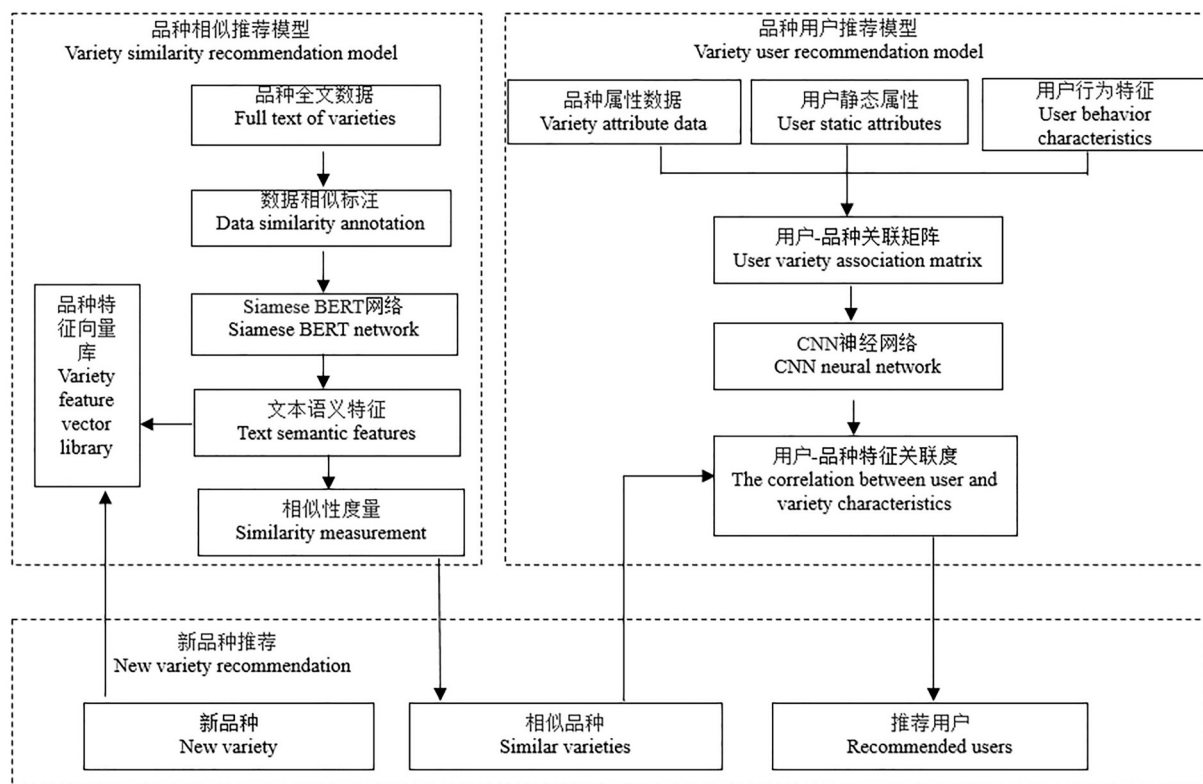


图1 推荐方法流程图

Fig. 1 Recommended method flowchart

2.1 相似品种推荐

相似品种推荐流程见图2。新出现的动植物新品种,在推荐系统的训练模型中并不存在,会产生冷启动问题。为此,设计通过新品种与已有历史品种的相似度进行匹配,将新品种推荐转化为向关注相似历史品种的用户推荐。传统基于词频统计的方法 tf-idf (Term frequency-inverse document frequency)、N-gram 会生成高维稀疏向量,仅依赖字面匹配,忽略语义信息。而BERT作为一种深层语义信息表示的预训练语言模型,用于文本匹配的任务有较好的效果,因此使用Siamese BERT网络进行品种相似匹配推荐。

2.1.1 数据采集预处理

收集动植物新品种的资源数据,对数据进行清洗,去除重复数据、格式统一化,对文本数据进行预处理,分词、去除停用词、词干提取等。对于每个动植物新品种,生成正样本对和负样本对,自身与自身的样本对被标注为相似,记为1,随机选择其他不同品种的样本对,将它们标注为不相似,记为0,确保模型在学习时不会将不同品种的数据误认为相

似。每个文本数据经过预处理后,形成独立的tokenized序列,其中包含了CLS和SEP特殊token。

2.1.2 品种特征表示

Siamese网络由2个相同结构的BERT子网络组成,共享参数,通过BERT网络分别对2个文本特征进行向量化表示,每个子网络输入1个品种。本文使用BERT-base版本,向量的维数为768维。段嵌入(Segment embeddings)包含2个可学习的向量EA和EB,分别用于区分输入文本中的2个句子,将EA分配给第1个句子的所有token,EB分配给第2个句子的所有token。位置向量给出了每个token在句子中的位置的向量表示,同一个位置的token具有相同的向量表示。输入向量提交给BERT的核心模块TRM(Transformer encoder),经过自注意力模块及残差操作,再对向量进行线性转换,完成TRM层的工作,通过多层TRM的叠加处理,最终,对于每个输入品种文本序列,提取CLS token对应的向量作为该品种资源的特征向量。

由于农业动植物新品种包含丰富的领域词汇信息,因此通过构建动植物品种资源词汇库,获取

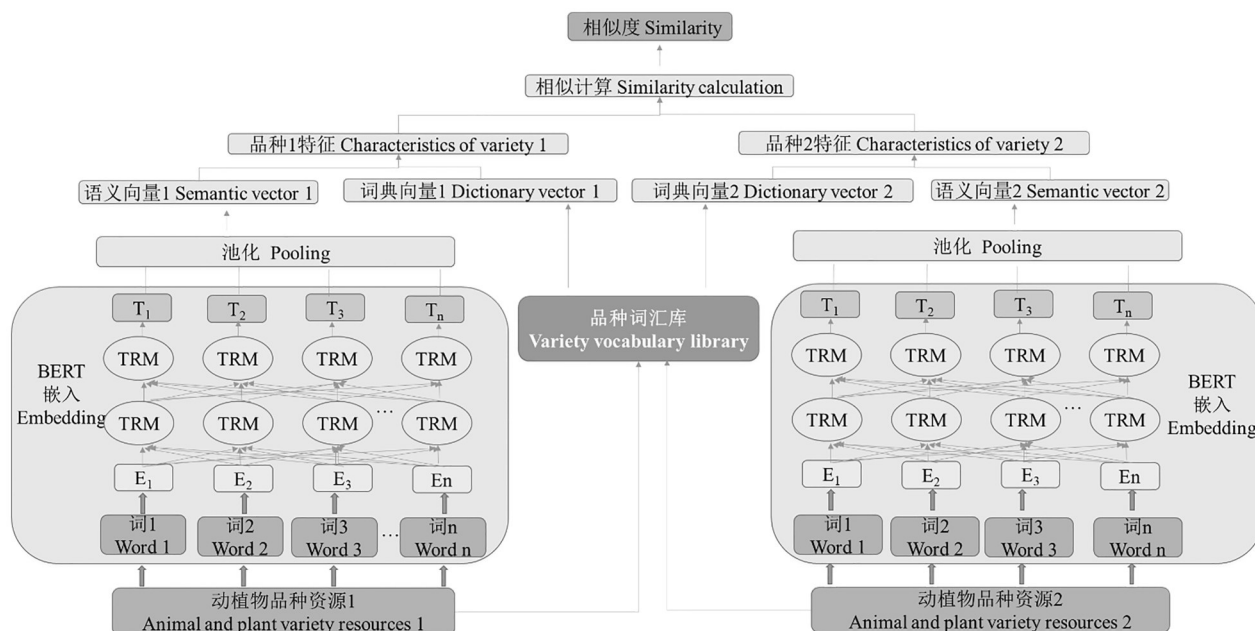


图2 相似品种推荐流程

Fig. 2 Recommendation process for similar varieties

品种信息的词典特征向量,与BERT网络获取的特征向量融合,更全面的表征品种信息资源。具体过程如下:

1)首先基于全部品种数据构建领域词典,抽取整理动植物品种领域的相关词汇,包括动植物的名称、分类、特征描述等信息,筛选出现频率 ≥ 5 次的动植物专业术语(如“抗倒伏”“条锈病抗性”等)构建词汇库,使用哈希表进行存储。

2)使用双向最大匹配算法进行实体识别并向量化。首先执行正向最大匹配,算法从文本起始位置开始扫描,初始化当前字符位置为0。在每一步处理中,算法截取从当前位置开始的最长10字符词组,查询预构建的动植物品种资源词汇库。若未匹配成功,则逐步缩短词组长(每次减1字符)继续查询,直至匹配成功或缩减至单字符。匹配成功后,将该词组标记为实体,并将扫描位置更新至该实体末尾继续后续处理。接着执行逆向最大匹配,算法从文本末尾开始反向扫描,采用与正向匹配相同的处理逻辑,区别在于扫描方向和词组截取顺序改为从右向左。逆向匹配能有效捕捉正向匹配可能遗漏的特殊词组组合。最后,将正向匹配和逆向匹配结果进行对比、合并、去重,得到最终的实体识别结果。本研究仅保留与词汇库匹配的实体词,通过 Word2Vec(Word embedding via neural networks)

进行向量化,得到整体文本的词典特征向量。将该特征向量与BERT层获取的特征向量拼接,作为最终的品种资源特征向量。

2.1.3 模型训练

计算2个品种特征向量之间的相似度得分,使用余弦相似度度量2个向量之间的相似性,公式如下:

$$\text{Similarity}(h_1, h_2) = \frac{h_1 \cdot h_2}{\|h_1\| \cdot \|h_2\|} \quad (1)$$

式中: $\text{Similarity}(h_1, h_2)$ 为2个品种特征向量之间的相似度得分; h_1 为第1个品种特征向量; h_2 为第2个品种特征向量。

采用对比损失函数(Contrastive Loss)进行模型训练,使得相似品种的特征表示在向量空间中距离更近,不相似品种的特征表示距离更远。对比损失函数为:

$$L = \frac{1}{2N} \sum_{i=1}^N [Y_i \cdot (1 - s_i)^2 + (1 - Y_i) \cdot \max(0, s_i - \alpha)^2] \quad (2)$$

式中: L 为对比损失值; N 为训练样本的数量; Y_i 为相似性标签(1表示相似,0表示不相似); s_i 为2篇文章的相似度,通过文章向量的余弦相似度计算; α 为相似度阈值,通过验证集上的性能表现, α 值为0.5。

使用随机梯度下降SGD(Stochastic Gradient

Descent)进行参数更新。通过不断地迭代训练, Siamese BERT 网络学习到将相似样本对的特征向量距离拉近,而将不相似样本对的特征向量距离推远的能力,从而实现动植物品种资源相似匹配的任务。

2.2 用户品种兴趣建模

用户品种兴趣模型架构见图3。在农业动植物新品种推荐过程中,用户特征数据包含用户基础静态特征、用户行为特征,品种特征数据包括品种的名称、特性等属性。对数据特征的全方位提取,以及构建用户品种关注矩阵将能直接影响兴趣模型

的准确度。以往深度兴趣预测模型还没有针对不同类型农业用户,根据用户的地域、行业等多样性兴趣差异,以及用户对品种数据的关注行为构建符合实际情况的基础数据体系。本研究提出方法在兴趣数据采集预处理过程中,充分考虑农业知识推荐的领域场景复杂、用户兴趣多变、需求表达不明确特点,综合用户静态属性特征、用户行为特征、品种资源特征等影响用户知识推荐的主要因素,基于多层 CNN(Convolutional Neural Network) 网络架构构建用户品种兴趣模型。

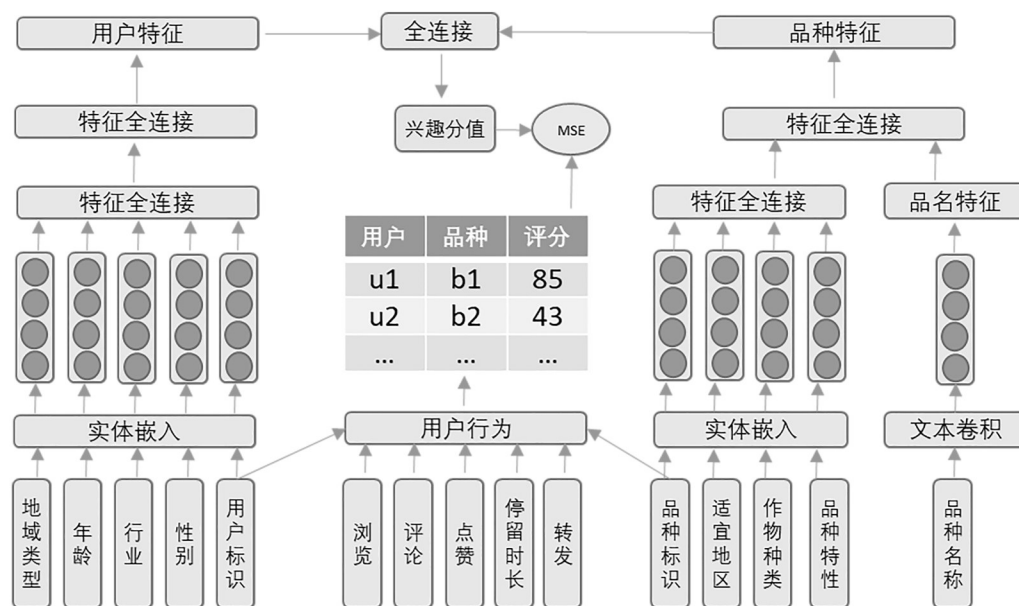


图3 用户品种兴趣模型

Fig. 3 User variety interest model

首先构建完整的数据指标体系,作为模型的输入指标。其中用户静态属性数据包括用户在中国农技推广平台注册的静态信息,包括姓名、性别、年龄、地域、身份类型、行业。其中行业又划分为种植业、畜牧业、林业、渔业、副业,每类行业可具体再划分为不同产业,如种植业又可以细分为粮食、油料、棉花、麻、豆等。身份类型包括小农户、家庭农场、专业大户、农民专业合作社、社会化服务组织、龙头企业,农业科研人员、农技推广人员、农业管理人员。行为信息通过用户在农技推广平台的所有操作产生,包括了用户对品种资源的浏览、点赞、评论,以及停留时长。品种资源数据包括品种的名称、标识、对应的作物品类、品种特性、适用区域、产量,其中品种特性包括特异性、适应性、一致性、稳定性、

地区性、时间性、局限性,适用区域包括东北、西北、华北等地。

在数据预处理环节,用户标识及品种标识数据不变,其他特征数据转为数字,例如,年龄数据分段转成8个连续数字0~7,身份数据转为数字0~8。对于文本类数据首先创建文本到数字的字典,然后将文本中的描述转成数字的列表。为了构建用户品种关联矩阵,需要采集用户对动植物品种的评分数据,但公开的数据集及互联网系统没有符合需求的格式化数据,因此基于农技推广平台的用户对品种资源数据的访问行为进行间接评分。通过用户的浏览、点赞等行为赋予不同的分值,最终获取特定用户对某个品种资源的评分数据。最终将原始用户特征、品种特征,用户品种关联特征合并如表1格式。

表1 用户特征组合
Table 1 User feature combination

用户ID User ID	用户姓名 User name	身份 Identity	身份说明 Identity description	品种名 Variety name	评分 Score
1	王明义	1	小农户	(番茄)红金宝001	76
2	李婉华	2	家庭农场	(黄瓜)玉秀2号	45
3	张晓静	6	农业科研人员	(辣椒)红源10号	90

用户特征提取首先是通过嵌入层获得维度为16的特征向量,例如性别、年龄、行业分别为 $\mathbf{h}\{\text{gender}\}$ 、 $\mathbf{h}\{\text{age}\}$ 、 $\mathbf{h}\{\text{industry}\}$,输入第1层全连接层,输出维度为128维的特征向量,再输入第2层全连接层,最终输出特征为256维的用户特征。通过多层全连接层,模型将低维用户特征进行组合和转换,学习到更加丰富的高阶特征表示。

品种特征提取首先对品种名称之外的其他属性特征通过嵌入层获取维度为16的特征向量,之后输入第1层全连接层,输出维度为128维的特征向量。品种名称通过文本卷积的方式获取特征,网络的第1层是词嵌入层,由每一个单词的嵌入向量组成的嵌入矩阵;下一层使用多个不同尺寸(窗口大小)的卷积核在嵌入矩阵上做卷积,窗口大小定为每次滑动2、3、4和5个单词;第3层网络是最大池化层,得到一个长向量,并在每一次训练迭代中,随机将某些神经元的输出置为零,通过这种dropout做正则化,最终得到了品种名称的特征向量。将品种名称特征与品种其他特征通过多个隐藏层变换,最终通过1个全连接层连接成为256维品种特征向量。

由于用户特征与动植物品种中的特征重要程度不同,如用户特征中的行业信息,品种特征中的作物品类。因此加入注意力机制,使模型更加关注用户和物品之间的重要特征,从而提高模型的性能和效果。

首先,计算用户特征和物品特征之间的注意力权重,使用点乘方法来计算注意力权重,具体公式如下:

$$\text{Attention}(\mathbf{u}_e, \mathbf{b}_e) = \text{softmax}(\mathbf{u}_e^T \cdot \mathbf{W}_a \cdot \mathbf{b}_e) \tag{3}$$

式中: $\mathbf{u}_e$ 为用户特征向量; $\mathbf{b}_e$ 为动植物品种特征向量; $\mathbf{W}_a$ 为学习到的注意力权重矩阵;softmax函数用于将注意力权重归一化。将用户特征和物品特征按照注意力权重进行加权求和,得到加权的特征表

示 F ,计算公式如下:

$$F = \text{Attention}(\mathbf{u}_e, \mathbf{b}_e) \cdot [\mathbf{u}_e; \mathbf{b}_e] \tag{4}$$

将加权的特征表示送入全连接层进行处理,预测评分。全连接层设置2个隐藏层,全连接层的输出表示为:

$$\text{output} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{5}$$

式中: $\mathbf{x}$ 为输入向量; $\mathbf{W}$ 为连接输入和隐藏层的权重矩阵; $\mathbf{b}$ 为偏置向量;模型的输出是用户对物品的评分的预测值。将预测值与真实评分进行比较,并使用均方误差(MSE)作为损失函数来优化模型。MSE优化损失函数的定义如下:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \tag{6}$$

式中: N 为样本数量; y_i 为真实评分; $\hat{y}_i$ 为模型预测的评分; i 为样本序号。使用Adam优化器来最小化MSE损失函数,根据梯度的1阶矩估计和2阶矩估计进行自适应地调整学习率。

3 两步推荐方法验证与结果分析

3.1 试验数据集

研究数据使用中国农技推广信息服务平台的农业用户注册数据、用户在平台上发布的问题、浏览的动植物主导品种数据以及农技知识数据。试验数据集专门包含冷启动测试集(占总数15%),由2023年新审定的320个品种构成,其特性为:1)无历史用户行为数据;2)审定时间<3个月;3)平台累计曝光量<50次。该部分数据用于验证模型对新品种的推荐效果,其余85%数据用于常规性能测试。数据集包括用户数据10 000条,用户问题数据2 000 000条,浏览的主导品种数据200 000条,浏览文章摘要数据200 000条,用户针对问题的浏览、回复、点赞视同关注本问题。平台数据审核过程中,人工对数据进行有效性验证并标注,数据集数据为

已标注数据。将数据以 7:1.5:1.5 的比例分为训练集、验证集和测试集。在分词环节,采用 HanLP (Han Language Processing) 工具包进行分词操作。本研究使用 HanLP 2.1.0 版本,并基于农业动植物

领域的专业特点,对其内置词典进行了扩充,将动植物品种领域的专业词汇纳入其中,以提高分词在该领域的准确性。用户行为数据、品种数据和用户品种关联数据形式分别如表 2、表 3 和表 4 所示。

表 2 用户行为数据

Table 2 User behavior data

用户 ID User ID	品种 ID Variety ID	行为类型 Behavior type	行为时间 Operating time	停留时长/s Duration of stay
1	101	浏览	2023-07-01 10:20:30	120
2	102	点赞	2023-07-01 10:25:30	0
3	106	评论	2023-07-02 14:15:10	60

表 3 品种数据

Table 3 Variety data

品种 ID Variety ID	名称 Variety name	分类 Variety classification	描述 Variety description	适用区域 Suitable region	产量/ (kg/hm ²) Yield
1	京科 968	粮食作物	高产、高抗玉米螟	东华北地区	11 250
2	连梗 11 号	粮食作物	株型紧凑,穗型较大,群体整齐度较好,熟期转色好,抗倒性较强	江苏省淮北地区	9 750
3	毛粉 802	蔬菜作物	果实光滑、美观、脐小、肉厚,不易裂果	陕西、山东、湖北、天津、北京、河北、安徽、辽宁、吉林、黑龙江	75 000

表 4 用户品种关联数据

Table 4 User variety associated data

用户 ID User ID	品种 ID Variety ID	评分 Score
1	101	80
2	102	90
3	106	75

表 4 中的用户-品种评分通过多维度行为数据加权计算获得。浏览行为按停留时长比例赋分(每 60 s 记为 1 分,上限 10 分),点赞固定 20 分,评论基础分 30 分并按字数附加分值(每字 0.5 分,上限 10 分),复合行为累计总分上限 100 分。所有原始评分经过 Z-score 标准化处理线性映射至[0,100]。

3.1 试验参数

在本试验中,使用了 Siamese BERT 网络进行品种相似推荐,并使用多层 CNN 网络进行用户品种兴趣建模。Siamese BERT 模型采用 BERT-base 版本,包含 12 层 Transformer Encoder,每层的隐藏

层维度为 768 维。输入数据经过分词和预处理,生成包含 CLS 和 SEP 特殊 token 的 tokenized 序列,利用 BERT 网络提取 CLS token 对应的向量作为品种资源的特征向量。农业动植物新品种的领域词汇通过构建词汇库,获取词典特征向量,并与 BERT 网络获取的特征向量融合。相似度使用余弦相似度度量,损失函数采用对比损失函数(Contrastive Loss),margin 超参数设定为 0.5,优化器使用随机梯度下降(SGD),学习率为 2e-5。用户品种兴趣模型采用多层 CNN 网络,结合用户静态属性特征、行为特征和品种属性特征。用户和品种的文本特征通过嵌入层转化为 16 维特征向量,并输入全连接层输出 128 维特征向量。品种名称的文本特征经过词嵌入层,使用窗口大小为 2、3、4 和 5 的卷积核进行卷积,并通过最大池化层和随机 dropout 正则化,最终输出品种特征向量。综合用户和品种的特征向量,结合注意力机制,预测用户对品种的评分,优化器使用 Adam 优化器,损失函数为均方误差(MSE)。参数设置见表 5。

表5 参数设置

Table 5 Parameter settings

参数名称 Parameter	Siamese BERT 模型 Siamese BERT model	用户品种兴趣模型 User variety interest model
模型架构	Siamese BERT 网络	多层 CNN 网络
BERT 版本	BERT-base(12 层 Transformer Encoder, 隐藏层维度 768 维)	—
输入数据	CLS 和 SEP 特殊 token 的 tokenized 序列	用户静态属性特征、行为特征、品种属性特征的嵌入表示
特征表示	BERT 提取 CLS token 对应的向量	用户和品种的文本特征通过嵌入层转化为 16 维特征向量
领域词汇特征融合	构建动植物品种资源词汇库, 获取词典特征向量, 并与 BERT 网络获取的特征向量融合	—
相似度度量	余弦相似度	—
损失函数	对比损失函数(Contrastive Loss), margin=0.5	均方误差(MSE)
优化器	随机梯度下降(SGD), 学习率=2e-5	Adam 优化器
嵌入层	—	用户和品种文本特征转化为 16 维特征向量
卷积层	—	使用窗口大小为 2、3、4、5 的卷积核
池化层	—	最大池化层
正则化	—	随机 dropout

3.2 试验结果分析

系统采用精确率 P 、召回率 R 、F1 值、平均绝对误差、均方误差作为用户兴趣分类模型评价指标, 精确率表示模型分类为正例的兴趣标签中分类正确的占比。召回率表示应该被正确分类的, 模型正确分类的比例。F1 值是兼顾精确率与召回率的综合指标。平均绝对误差(MAE)预测评分与实际评分的平均绝对误差, 均方误差(MSE)预测评分与实际评分的均方误差。

精确率 P 、召回率 R 、F1 值指标计算公式如下:

$$P=\frac{TP}{(TP+FP)} \tag{7}$$

$$R=\frac{TP}{(TP+FN)} \tag{8}$$

$$F1=\frac{2\times P\times R}{(P+R)} \tag{9}$$

式中: TP 为分类正确的样本数量; FP 为误分为正类的样本数量; FN 为误分为负类的样本数量。MAE 与 MSE 计算公式如下:

$$MAE=\frac{1}{N}\sum_{i=1}^N|y_i-\hat{y}_i| \tag{10}$$

$$MSE=\frac{1}{N}\sum_{i=1}^N(y_i-\hat{y}_i)^2 \tag{11}$$

式中: N 为样本数量; y_i 为真实评分; $\hat{y}_i$ 为模型预测的评分。

试验结果的详细数据见表 6。可知, 本研究提出的两步推荐算法在各项评价指标上均优于其他对比算法。本研究提出的两步推荐算法在各项指标上均表现出色, 特别是在准确率、召回率和 F1 值方面表现尤为突出, 这表明本研究方法在推荐动植物新品种资源方面具有较高的推荐精度和用户满意度。基于协同过滤的推荐算法利用用户行为数据进行推荐, 在处理冷启动问题时表现较差。此外, 该算法未能充分利用动植物品种的文本特征, 导致推荐精度较低。基于矩阵分解的推荐算法通过分解用户-品种评分矩阵来进行推荐, 能够捕捉用户和品种的潜在特征, 但在处理稀疏数据时效果有限, 且无法处理新品种的冷启动问题。基于深度学习的单步推荐算法结合了因子分解和深度神经网络的优点, 能够较好地处理高维稀疏数据, 推荐效果较好。但在处理动植物品种的特定领域特征时, 效果仍不如本研究提出的两步推荐算法。

表 6 对比试验

Table 6 Comparative experiment

模型 Model	精确率/% Accuracy	准确率/% Precision	召回率/% Recall	F1 值 F1 score	平均绝对误差 Mean absolute error	均方误差 Mean squared error
本研究两步推荐算法	85.2	84.1	81.4	0.832	0.128	0.256
基于协同过滤的推荐算法	74.2	72.5	70.1	0.721	0.157	0.314
基于矩阵分解的推荐算法	76.8	74.6	72.9	0.748	0.148	0.296
基于深度学习的单步推荐算法	80.1	78.2	76.7	0.784	0.138	0.278

此外,为验证模型对冷启动问题的处理能力,设计 2 组对比试验:1)新品种冷启动测试。从试验数据中提取 2023 年新审定的 320 个零行为数据品种作为测试集,其首推准确率达到 68.5%(传统协同过滤方法为 42.3%),且通过 Siamese BERT 的特征映射,AUC(Area Under ROC Curve)仅比常规品种下降 5.7%(传统方法下降 15.2%)。2)新用户模拟试验。随机屏蔽 500 名用户的历史行为,仅基于静态属性推荐,前 3 次交互后推荐准确率提升至 72.1%,显著高于基线方法的 53.4%。这表明本研究提出的两步推荐机制能有效利用品种语义关联和用户属性特征缓解冷启动问题,尤其在新品种推荐场景表现突出,但新用户冷启动的初始准确率(61.2%)仍存在优化空间,可能与农业用户的地域特异性相关。

3.3 消融试验

分别针对以下 4 个模块进行了消融试验:

1)BERT 文本编码。用于提取动植物新品种文本信息的上下文语义特征。消融试验中,去除 BERT,改用传统的 TF-IDF 特征表示。

2)Siamese BERT 网络。用于计算动植物品种之间的相似性。消融试验中,去除孪生网络,改用

单向 BERT 模型计算文本相似度。

3)CNN 兴趣网络模型。用于用户画像与动植物品种资源特征的匹配。消融试验中,去除 CNN 网络,改用简单的全连接网络进行推荐。

4)注意力机制。用于提升模型对用户行为和品种特征的关注度。消融试验中,去除注意力机制,改用平均池化来对输入特征进行聚合。

试验中使用准确率、召回率、F1 值及 AUC 值、 P 值评估各个模块的贡献。其中,AUC 值为随机选取的正样本(相似品种对)得分高于负样本(非相似对)的概率, P 值用于评估模型性能差异的统计显著性。试验结果见表 7。可知,当去除 BERT 文本编码模块,替换为传统的 TF-IDF 特征表示时,准确率下降了 7.5%,F1 值也大幅降低, P 值为 0.002。这表明 BERT 模型在捕捉动植物新品种文本信息中的上下文语义特征方面起到了关键作用。去除 Siamese BERT 网络后,准确率下降至 85.4%,召回率下降至 83.0%, P 值为 0.003。相比完整模型,Siamese 网络在计算品种相似性上更为精准,特别是对于较难分类的样本,孪生网络可以更好地区分相似品种与不相似品种,其影响程度仅次于 BERT。在去除了 CNN 兴趣网络模型后,模型性能仅有小幅下降,

表 7 消融试验

Table 7 Ablation experiment

模型配置 Model configuration	准确率/% Precision	召回率/% Recall	F1 值 F1 score	AUC 值 AUC value	P 值 P value
完整模型	89.6	87.3	0.884	0.921	
去除 BERT 文本编码	82.1	80.5	0.813	0.865	0.002
去除 Siamese BERT	85.4	83.0	0.842	0.889	0.003
去除 CNN 兴趣网络	86.8	84.7	0.857	0.902	0.012
去除注意力机制	87.3	85.2	0.862	0.910	0.021

但准确率仍然有2.8%的差距。这表明CNN在捕获用户兴趣特征与品种特征之间的深层次关联上发挥了重要作用,其影响程度次于Siamese网络。全连接网络的替代方案虽然能有效进行推荐,但无法像CNN充分提取局部特征。当去除了注意力机制时,模型的性能也有所下降,特别是F1值降低了2.2%。这表明注意力机制帮助模型更好地识别出哪些特征更值得关注,尤其是在用户兴趣匹配方面,但总体影响程度较弱。通过以上消融试验,验证了各个模块对动植物新品种资源精准推荐模型的重要性,BERT文本编码与Siamese BERT网络是模型的核心模块,显著提升了品种相似度的计算精度。CNN兴趣网络在捕捉用户兴趣特征与品种特征之间的深层次关联方面发挥了重要作用。注意力机制虽然对模型性能的影响相对较小,但仍能够提升用户与品种的精准匹配能力。

本研究提出的两步推荐算法结合了Siamese BERT网络的深层语义表示能力和多层CNN网络的用户兴趣建模能力,能够充分利用动植物品种的文本特征和用户行为数据,在新品种的冷启动问题上表现尤为突出。此外,通过综合用户静态属性特征、行为特征和品种属性特征,本研究方法能够更加精准地预测用户对品种的兴趣,提高推荐的个性化和准确性。

4 结束语

本研究提出了一种基于Siamese BERT网络和多层CNN网络的动植物新品种资源两步推荐算法。该算法通过Siamese BERT进行品种相似性推荐,并利用多层CNN网络对用户的品种兴趣进行建模。结果表明,本研究算法在AUC、准确率、召回率、F1值等各项评价指标上均显著优于现有主流推荐算法,尤其是在处理新品种冷启动问题和用户兴趣建模方面展现了卓越的性能。相较于传统算法,本研究方法能够更精准地捕捉新品种的特性,并有效应对用户兴趣动态变化的问题,提升了推荐的准确性与个性化程度。特别是在冷启动情境下,该算法对新品种的推荐能力得到了显著增强,体现出较强的实用性。

未来的工作将继续致力于优化模型结构,探索更加复杂的深度学习模型,如基于图神经网络(GNN)或强化学习的推荐框架,未来将探索图神经

网络在品种关联推理中的深度应用。通过构建[品种-性状-环境]三元关系图谱,利用GNN的消息传递机制挖掘隐性关联规则(如抗病基因相似性、地域适应性传导等),实现跨品种的知识迁移推荐。此外,针对用户兴趣的动态变化,进一步考虑时间序列模型的结合,以提升对用户长期偏好的捕捉能力。同时,可以尝试引入多模态数据,融合文本、图像、视频等多种信息,以实现更全面和精细的推荐效果。通过这些方向的深入研究和改进,期望能进一步提升动植物新品种推荐系统的效果与应用价值。

参考文献 References

- [1] 李红军, 孟晨霞, 张红伟, 宋伟彬, 刘兴旺. 基于多源数据的中美作物种质资源管理的比较及启示[J]. 中国农业大学学报, 2024, 29(10): 161-171
Li H J, Meng C X, Zhang H W, Song W B, Liu X W. Comparison and implication of crop germplasm resource management in China and the United States based on multi-source data [J]. *Journal of China Agricultural University*, 2024, 29(10): 161-171 (in Chinese)
- [2] Wu F, Lyu F, Ren J, Yang P, Qian K, Gao S J, Zhang Y X. Characterizing internet card user portraits for efficient churn prediction model design [J]. *IEEE Transactions on Mobile Computing*, 2024, 23(2): 1735-1752
- [3] Wu H R, Liu C, Zhao C J. Personalized agricultural knowledge services: A framework for privacy-protected user portraits and efficient recommendation [J]. *The Journal of Supercomputing*, 2024, 80(5): 6336-6355
- [4] Leng H Y, Shao J X, Zhang Z, Qian Y R, Ma M N, Li Z C. Research on dual-channel user portrait construction method based on DPCNN-BiGRU and attention mechanism [J]. *Journal of Intelligent & Fuzzy Systems*, 2023, 45(2): 2579-2591
- [5] 黄玉民, 赵婵婵. 基于改进主题模型方法的三级短视频用户画像的研究[J]. 计算机科学, 2024, 51(S1): 698-704
Huang Y M, Zhao C C. Study on three-level short video user portrait based on improved topic model method [J]. *Computer Science*, 2024, 51(S1): 698-704 (in Chinese)
- [6] 王世奇, 刘智锋, 王继民. 学者画像研究综述[J]. 图书情报工作, 2022, 66(20): 73-81
Wang S Q, Liu Z F, Wang J M. A review of scholar profiling research [J]. *Library and Information Service*, 2022, 66(20): 73-81 (in Chinese)
- [7] 岳怡然. 面向知识服务的用户画像构建与应用研究[D]. 北京: 中国农业科学院, 2021
Yue Y R. Research on the construction and application of user profiles oriented to knowledge service [D]. Beijing: Chinese Academy of Agricultural Sciences, 2021 (in Chinese)
- [8] 赵又霖, 程丽洁, 庞烁, 徐竟楠, 陆颖隼. 社会感知驱动下的用户画像时空数据研究进展[J]. 情报科学, 2022, 40(9): 185-192
Zhao Y L, Cheng L J, Pang S, Xu J N, Lu Y J. A review on spatiotemporal data of user profile driven by social sensing [J]. *Information Science*, 2022, 40(9): 185-192 (in Chinese)
- [9] 石宇, 胡昌平, 时颖惠. 个性化推荐中基于认知的用户兴趣建模研究[J]. 情报科学, 2019, 37(6): 37-41

- Shi Y, Hu C P, Shi Y H. User profiles modeling based on cognition in personalized recommendation[J]. *Information Science*, 2019, 37(6): 37-41 (in Chinese)
- [10] 张玉洁, 董政, 孟祥武. 个性化广告推荐系统及其应用研究[J]. *计算机学报*, 2021, 44(3): 531-563
- Zhang Y J, Dong Z, Meng X W. Research on personalized advertising recommendation systems and their applications[J]. *Chinese Journal of Computers*, 2021, 44(3): 531-563 (in Chinese)
- [11] 陈泽宇, 黄勃. 改进词向量模型的用户画像研究[J]. *计算机工程与应用*, 2020, 56(1): 180-184
- Chen Z Y, Huang B. Research on user portrait of improved word vector model[J]. *Computer Engineering and Applications*, 2020, 56(1): 180-184 (in Chinese)
- [12] Song C X, Dong H Y. Application of intelligent recommendation for agricultural information: A systematic literature review [J]. *IEEE Access*, 2021, 9: 153616-153632
- [13] 周洋涛, 李青山, 褚华, 李佳楠, 高明彪, 卫彪彪. 基于静态与动态学习需求感知的知识点推荐方法[J]. *软件学报*, 2024, 35(9): 4425-4447
- Zhou Y T, Li Q S, Chu H, Li J N, Gao M B, Wei B B. Knowledge point recommendation method based on static and dynamic learning demand perception[J]. *Journal of Software*, 2024, 35(9): 4425-4447 (in Chinese)
- [14] 耿杰, 刘春丽, 魏雪梅, 程明月, 袁昆, 李洋, 刘业政. 基于用户重购行为的产品推荐方法[J]. *计算机研究与发展*, 2023, 60(8): 1795-1807
- Geng J, Liu C L, Wei X M, Cheng M Y, Yuan K, Li Y, Liu Y Z. Product recommendation method based on user repurchase behavior[J]. *Journal of Computer Research and Development*, 2023, 60(8): 1795-1807 (in Chinese)
- [15] 龙恩, 吕守业, 岑鹏瑞, 杨宇科, 韦二龙, 白龙. 顾及用户画像的多源遥感信息智能推荐方法[J]. *测绘学报*, 2023, 52(2): 297-306
- Long E, Lü S Y, Cen P R, Yang Y K, Wei E L, Bai L. An intelligent recommendation method of multi-source remote sensing information considering user portrait[J]. *Acta Geodaetica et Cartographica Sinica*, 2023, 52(2): 297-306 (in Chinese)
- [16] 赵继春, 孙素芬, 郭建鑫, 钟瑶, 王洪彪, 王敏, 秦莹. 农业在线学习资源知识图谱构建与推荐技术研究[J]. *计算机应用与软件*, 2022, 39(8): 69-75
- Zhao J C, Sun S F, Guo J X, Zhong Y, Wang H B, Wang M, Qin Y. Research on the construction of knowledge graph and recommendation technology of agricultural online learning resources [J]. *Computer Applications and Software*, 2022, 39(8): 69-75 (in Chinese)
- [17] 陈玲洪, 潘晓华. 基于知识图谱和读者画像的图书推荐研究[J]. *数据分析与知识发现*, 2023, 7(12): 164-171
- Chen L H, Pan X H. Recommending books based on knowledge graph and reader profiling[J]. *Data Analysis and Knowledge Discovery*, 2023, 7(12): 164-171 (in Chinese)
- [18] 侯琛, 牛培宇. 农业知识图谱技术研究现状与展望[J]. *农业机械学报*, 2024, 55(6): 1-17
- Hou C, Niu P Y. Review of research status and prospects of agricultural knowledge graphs [J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2024, 55(6): 1-17 (in Chinese)
- [19] 刘畅, 王春山, 缪祎晟, 朱华吉, 郭旺, 吴华瑞. 一种面向蔬菜生产知识服务的主动推荐方法[J]. *河北农业大学学报*, 2024, 47(6): 127-135
- Liu C, Wang C S, Miao Y S, Zhu H J, Guo W, Wu H R. An active recommendation method for vegetable production knowledge services[J]. *Journal of Hebei Agricultural University*, 2024, 47(6): 127-135 (in Chinese)
- [20] Gopi P S S, Karthikeyan M. Multimodal machine learning based crop recommendation and yield prediction model[J]. *Intelligent Automation & Soft Computing*, 2023, 36(1): 313-326
- [21] Musanase C, Vodacek A, Hanyurwimfura D, Uwitonze A, Kabandana I. Data-driven analysis and machine learning-based crop and fertilizer recommendation system for revolutionizing farming practices [J]. *Agriculture*, 2023, 13(11): 2141

责任编辑: 申青苗



第一作者简介: 顾静秋, 博士, 副研究员。1995—1999年在燕山大学计算机科学与技术专业学习, 获工学学士学位; 1999年8月至今在北京市农林科学院信息技术研究中心工作。工作期间, 2002—2005年在北京理工大学计算机应用技术专业学习, 获工学硕士学位, 2010—2018年在北京交通大学计算机科学与技术专业学习, 获工学博士学位。主要从事农业智能系统、大数据挖掘研究。作为执行主持或技术骨干先后参加了国家和省部级重点科研项目30余项。2007年入选北京市科技新星培养计划, 2009年获得计算机信息系统集成项目经理资质。参与的“农业智能系统平台(PAID)研究与应用”项目获北京市科学技术一等奖。在农业信息化发展的关键技术研究、应用、产品化和推广模式方面做出了创新性工作。



通讯作者简介: 郭旺, 硕士, 助理研究员。2009年9月—2013年6月在黑龙江八一农垦大学电子信息工程专业学习, 获工学学士学位; 2013年9月—2016年1月在北京科技大学信息与通信工程专业学习, 获工学硕士学位; 2016年8月—2022年7月在派得伟业公司担任开发工程师, 负责技术研发工作; 2022年8月至今在北京市农林科学院信息技术研究中心工作, 主要从事农业智能系统、大数据挖掘研究。工作期间, 作为执行主持或技术骨干先后参加了国家和省部级重点科研项目30余项, 其中主持国家科技支撑计划子课题1项, 国家重点研发计划子课题1项。发表论文10余篇, 参编著作1部, 取得软件著作权登记60余个, 获得科技奖励4项。