基于 Citespace 知识图谱的农业知识语义智能检索方法

陈素清[1] 武 浩[1] 王超凡[1]
CHEN Suqing WU Hao WANG Chaofan

摘 要 当前对农业知识语义的智能检索方法往往难以精准捕捉知识的深层语义，导致检索结果与目标集之间存在较大偏差，使得检索推荐列表中目标信息的平均倒数排名偏低，一定程度上影响了用户对农业知识的获取效率。为此，提出一种基于 Citespace 知识图谱的农业知识语义智能检索方法。通过对农业知识词向量化表示，将其转化为向量空间中的数值表示。利用 Citespace 软件进行关系抽取和图谱展示，建立农业知识图谱。在此基础上，对农业知识图谱中的特征向量进行扩展，以丰富其语义维度。引入余弦相似度算法来计算检索信息与扩展后的农业知识图谱中特征向量之间的相似度。基于相似度值计算结果，筛选出与检索信息相似度最高的农业知识文本。根据相似度排序生成一个检索文本列表，从而实现农业知识语义智能检索。实验结果表明，所提出的方法在检索推荐列表中的平均倒数排名保持在 0.90 以上，且交并比超过 0.95，表明所提出的方法在农业知识语义智能检索领域具有良好的应用前景和显著优势。

关键词 Citespace 知识图谱；农业知识语义；智能检索；词向量化表示；关系抽取；相似度算法

doi：10.3969/j.issn.1672-9528.2024.08.049

0 引言

随着信息技术的迅猛发展，特别是大数据、云计算和人工智能技术的深度融合，知识的获取、组织与利用方式正经历着深刻的变革。农业作为国民经济的基础产业，其知识的积累、传承与创新对于推动农业现代化、提升农业综合竞争力具有至关重要的作用。语义智能检索，通过利用自然语言处理技术、机器学习算法和知识图谱等技术手段，对文本数据进行深度解析和语义理解，实现了精准、高效的知识检索。相比于传统的基于关键词的检索方法，语义智能检索更能准确理解用户的查询意图，深入挖掘文本中的深层含义，从而为用户提供更加准确、全面的检索结果。目前，语义检索方式已被广泛应用，并且吸引了众多专家与学者在此领域展开深入研究。

例如，文献 [1] 提出了基于词向量扩展语义检索方案，阐述了语义检索关键技术，通过对知识文本词向量扩展，处理知识文本之间的关系，构建具有关联关系的文本集合，从中抽选出检索结果。文献 [2] 提出了基于语义特征挖掘的语义检索方法，利用语义关联挖掘技术对知识文本进行关联分析，生成检索列表。文献 [3] 提出了基于自然语言理解的电力调度文本检索方法，基于正则表达式，构建电力调度文本解析模型，抽取关键信息并构建知识图谱；提出相似度计算方法，通过计算与知识图谱实体的相似距离，实现电力文本信息检索。文献 [4] 提出了基于人工智能技术的海量计算机信息检索方法，基于反向语义推理获取检索本体，计算语义关联度，使用加权法生成检索算法，实现海量计算机信息的精准检索。由于农业知识的复杂性和多样性，用户在检索时往往需要表达更加精确和具体的需求。传统的知识检索方法在处理农业知识时，常常局限于简单的关键词匹配和文本搜索。这种方法虽然简单易行，但在面对复杂多样的农业知识时，忽视了知识背后的深层语义关系，仅仅根据字面上的关键词进行匹配，导致检索结果往往只能涵盖部分相关的知识，而无法全面准确地反映用户的真实需求，影响了用户对农业知识的获取效率。为了解决传统方法存在的弊端，本研究提出一种基于 Citespace 知识图谱的农业知识语义智能检索方法。通过这一方法，能够更深入地理解农业知识的深层语义和关联性，为用户提供更加准和全面的检索结果。

1 农业知识词向量化表示

在农业知识语义智能检索过程中，传统的文本处理方法通常将词汇视为离散的、不可分割的符号，这种方法难以捕捉词汇之间的语义关系。而通过将词汇转化为向量表示，可以在连续的向量空间中表达词汇之间的相似性。这种向量化

1. 集宁师范学院 内蒙古乌兰察布 012000

[基金项目] 集宁师范学院科学研究项目“基于知识图谱的智慧农业信息资源推荐系统研究”（jsky202254）；集宁师范学院横向课题“基于知识图谱的智能检索服务系统研究”（hx202217）

表示方式不仅能够有效捕捉词汇间的语义联系，还提供了一个统一的度量标准。这意味着，即使检索关键词与文档中的词汇不完全相同，只要它们在语义上相近，系统也仍然能够给予较高的相似度评分，从而将相关文档准确地检索出来。这种表示方式有助于提高检索效率和准确性，并适应深度学习模型的需求。

为了满足农业知识语义检索需求，采用CBoW模型[5-6]（连续词袋模型）来对农业知识文本中的词汇进行向量化表示。CBoW模型通过农业知识文本的上下文来不断预测中心词，从而确保词向量能够蕴含丰富的上下文语义信息。模型的训练目标是最大化预测正确概率，这有助于获取根据表达能力的词向量。

CBoW模型是一个典型的三层神经网络，其结构分为输入、隐藏和输出三层[7]。在对农业知识文本进行词向量化表示之前，需要先对文本进行预处理，包括清洗和分词。清洗过程旨在去除文本中的噪声数据，如无效字符、HTML标签等，以保证分词的准确性。分词之后，将文本转化为一系列词汇的序列。根据分词和文本结构中的词汇，构建出一个词典，并为每个词分配一个唯一的索引。这个词典将作为CBoW模型的输入参考，用于将词汇转化为向量形式，从而便于神经网络进行学习和预测。假设农业知识文本中第 j 个词为 x_t，连续词袋模型截取与该词相邻的上下文词汇作为输入层输入向量，输入层接收目标词前后一定范围内的上下文词汇的词向量，利用独热编码对截取的词汇编码成具有多维特征的向量集，转换为初始的词向量[8]。将编码后的词向量输入到隐藏层，将所有输入的词向量进行累加或取平均，假设编码上下文词汇的数量为 n，词向量的维度为 v，词典的大小为 R，每个上下文词汇经过One-Hot编码后，与嵌入矩阵相乘，得到对应的词向量，其公式为：

$$h=\frac{1}{n}W\left(\sum_{j=1}x_{j,*}v\right)^{R} \tag{1}$$

式中：h 表示连续词袋模型隐藏层输出向量；$x_{j,*}$ 表示独热编码后的初始词向量[9]。将得到的词向量发送到输出层，该层是一个softmax层，用于预测目标词，输出向量为目标词，利用损失函数softmax训练模型。训练时，模型会遍历语料库中的每个词，将其作为目标词，并选择其前后的上下文词汇作为输入[10]。模型会计算输入上下文词汇的词向量与输出目标词之间的概率分布，并通过反向传播算法和梯度下降优化方法来更新词向量的值，其用公式表示为：

$$y=\frac{\exp\left(u\left(h\right)\right)}{\int\exp u\left(h\right)\mathrm{d}h} \tag{2}$$

式中：y 表示词向量为目标词的概率；$u(h)$ 表示词向量的得分。

训练完成后，模型中的参数即为每个词汇的词向量表示。这些词向量成功捕捉了词语之间的语义关系，因此在向量空间中，语义相近的词汇会聚集在相近的位置。完成上述步骤后，农业知识文本中的词汇得到了有效的向量化表示，这为后续利用Citespace构建知识图谱奠定了基础。

2 构建农业Citespace知识图谱

农业知识通常涉及大量实体和它们之间复杂的相互关系。通过将词汇向量化，能够在连续空间中表达词汇之间的语义关系，这是构建有效知识图谱的基础。知识图谱通过图形化表示来展现实体之间的关系，使得复杂的语义关系变得直观易懂。Citespace软件能够将这些关系以图例的形式展示，Citespace知识图谱的构成包括节点和边，其中每个节点代表一个“实体”，而每条边则代表实体与实体之间的“关系”。这种数据结构本质上构成了一个语义网络，其目的在于从“关系”的角度出发来深入分析问题。

在构建知识图谱的过程中，首先将向量化表示的农业知识文本导入到Citespace软件中；然后，采用监督学习方法从文本中抽取关系，特别是以谓词为中心的事实三元组，从而从大规模非结构化的自然语言文本中提取出结构化信息。具体来说，将未标注的句子、头实体、关系、尾实体等输入到监督学习模型中，让模型根据之前学习到的知识自动判断这些实体之间的关系，并给出相应的关系标签。这样，就能够得到一个包含丰富实体和关系的农业知识图谱，为后续的语义检索和分析提供有力支持[11]。

以下给出相应的关系标签：

$$f\left(r,t\right)=-\left\|r\odot h-t\right\| \tag{3}$$

式中：$f(r,t)$ 表示抽取的农业知识实体关系；r 表示农业知识本文中的头实体向量；t 表示农业知识本文中的尾实体向量；h 表示农业知识文本关系向量；⊙表示向量之间求哈达玛积的操作；$\|\cdot\|$ 表示对向量取范数操作。将抽取的关系在Citespace软件中进行图谱展示，将抽取的关系添加到三元组表中，将三元组表导入到Citespace知识图谱中，实现以图例的形式将抽取的关系表示：

$$D=\left\langle L,V\right\rangle \tag{4}$$

式中：D 表示农业知识图谱；L 表示知识图谱的节点，即三元组表中的实体向量；V 表示知识图谱的边，即三元组表中的关系向量。

3 语义检索

在农业知识语义智能检索过程中，农业知识图谱的构建旨在形成一个结构化的语义网络，能够清晰地表示出农业领域中的实体（如作物、病害、防治方法等）以及这些实体之间的关系。为了提高检索的准确性和效率，本研究对农业知识图谱中的特征向量进行扩展。这样做能够更全面地捕捉和

表达文本的含义，尤其是那些依赖于特定上下文或背景的信息。接下来，引入余弦相似度算法来计算检索信息与扩展后的农业知识图谱中特征向量之间的相似度。余弦相似度是一种常用的向量相似度计算方法，它通过测量两个向量之间的夹角余弦值来评估它们的相似程度。在农业知识语义检索中，这种方法能够准确度量检索信息与知识图谱中实体或关系的语义接近度。基于计算出的余弦相似度值，筛选出与检索信息相似度最高的农业知识文本。最终，系统会根据相似度排序生成一个检索文本列表，从而实现农业知识语义智能检索。上述步骤能够显著提高检索的准确性和效率，为用户提供更加精准和有用的农业知识信息。

例如，如果原有的农业知识图谱中特征词向量是数值的有序列表 (freq1, freq2, ⋯, freqm)，那么扩展后的词向量可以表示为 ((S, freq1), (S, freq2), ⋯, (S, freqm))，其中 S 代表当前句子的 ID。在其基础上，引入余弦相似度算法对检索信息与知识图谱中的特征向量相似度分析。余弦相似度可以表征两个文本向量（通常是由文本中的词汇、n-gram 或 TF-IDF 权重等特征构成的向量）的相似程度，如果两个文本向量的夹角越小，那么它们的内容就越相似。假设检索信息为向量 1，农业知识图谱中的特征向量为向量 2，则两个向量的余弦相似度计算为：

$$\cos S = \frac{\|\alpha\| \times \|\beta\|}{\alpha \cdot \beta} \tag{5}$$

式中：$\cos S$ 为检索信息与扩展后的农业知识图谱中的特征向量余弦相似度；α 为用户输入的检索信息；β 表示扩展后的农业知识图谱中的特征向量 [12]。余弦相似度的其值域被严格限定在 [-1, 1] 这一范围内。当余弦相似度的值越接近 1 时，意味着这两个向量在方向上的夹角越小，而它们的相似度越高。具体来说，如果两个向量的余弦相似度接近 1，那么可以认为这两个向量在内容上具有很高的相似性，或者它们所代表的文本、图像或其他数据类型在某种意义上是相近的。相反，当余弦相似度的值越接近 -1 时，表示这两个向量在方向上的夹角越接近 180°，即它们是反向的，这意味着两个向量所代表的信息在很大程度上是不相似甚至是相反的。而当余弦相似度的值为 0 时，表示这两个向量是正交的。在几何意义上，正交意味着两个向量相互垂直，没有共同的方向分量。因此，输出余弦相似度值为前 5 或者前 10 的农业知识文本，生成检索文本列表，实现农业知识语义智能检索。

4 实验论证

4.1 实验数据与指标

为了验证基于 Citespace 知识图谱的农业知识语义智能检索方法的性能，在本次实验中，选择了两个数据集来进行测试，这两个数据集分别为 IKHFAASG 和 IFUAOUYFA，它们分别包括了大约 1.25 GB 和 1.24 GB 的农业知识信息。利用本文设计的农业知识语义智能检索方法，对这两个数据集进行了实验。实验环境为：Ubuntu20.5620 操作系统，Intel（R）Core i8 CPU，256 GB 内存。

对于检索效果的评价，本次实验选择了平均倒数排名（MRR）和交并比作为评价指标。平均倒数排名是信息检索领域里一个关键的度量标准，它主要关注检索到的相关条目是否位于用户容易注意到的位置，并特别强调结果列表的排序重要性。MRR 是对于每个查询，第一个结果正确的逆位置（即位置 r 的逆是 $1/r$）的平均值。若目标答案未出现在任何候选答案中，则 MRR 值将被设定为 0；而当候选答案列表中仅包含一个目标答案时，MRR 值则为该答案在列表中的逆位置。MRR 值计算公式为：

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}\frac{1}{E_{et}} \tag{6}$$

式中：MRR 表示检索结果在候选答案列表中平均倒数排名；N 表示检索次数；i 表示候选答案列表中文本数量；E_{et} 表示目标答案在候选答案列表中所处位置。MRR 值越大，则表示检索精度越高。

交并比（IoU）是一种广泛用于评估目标检索精度的标准。它主要用于衡量检索框与实际目标框之间的重叠程度。当 IoU 的值为 0 时，意味着检索框没有覆盖到任何目标区域，即检索完全失败；而当 IoU 的值为 1 时，表示检索框与实际目标框完全重合，意味着检索完全成功且没有误差。因此，IoU 的值越接近 1，说明检索的精度越高，即预测的目标位置与实际目标位置越接近，其计算公式为：

$$\mathrm{IoU} = \|S \cup A\| / \|S \cap A\| \tag{7}$$

式中：IoU 表示检索集与目标集的交并比；S 表示检索框面积；A 表示目标框面积。

4.2 实验结果与讨论

为了验证检索方法的有效性，实验过程中使用了数据集 IKHFAASG 对检索方法进行初步检验，并利用数据集 IFUAOUYFA 对检索方法的交并比（IoU）进行专门评估。为了确保实验结果具有说服力和可对比性，选择了文献 [2] 中提出的基于语义特征挖掘的图书馆文献资源智能检索方法，以及文献 [3] 中介绍的基于自然语言理解的电力调度文本检索方法作为对比方法。图 1 展示了三种方法在数据集上的 MRR 对比。图 2 给出了它们对应的交并比（IoU）值。通过对比分析这两种评价指标，可以全面评估三种不同检索方法的性能表现。在平均倒数排名（MRR）方面，较高的值意味着检索系统能够更快地返回与用户查询最相关的结果。而在

交并比（IoU）方面，它衡量了检索结果框与真实目标框的重叠程度，值越接近 1 表示检索精度越高。

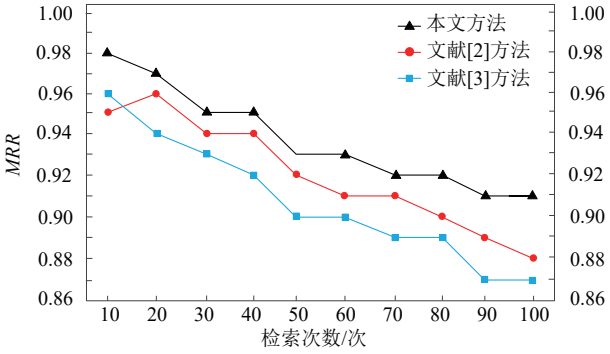

图 1 不同检索方法 MRR 对比

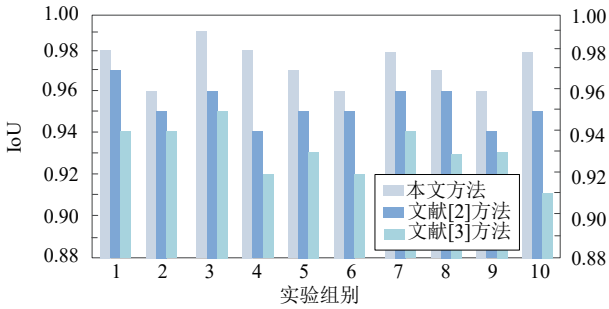

图 2 不同检索方法交并比对比

对比图 1 和图 2 中的数据可知，本文提出的方法在目标检索推荐列表中的平均倒数排名（MRR）为 0.94，这表示当检索列表中包含 10 个农业知识文本时，本文方法检索到的目标知识能够排在第二位。这一结果显著高于文献 [2] 和文献 [3] 中的对比方法。另外，本文方法的检索集与目标集的交并比（IoU）始终保持在 0.95 以上，这说明检索结果与目标知识在内容上基本重合，检索精度远高于主流方法。实验结果表明，本文方法能够实现对农业知识语义的智能精准检索，并且在与两种对比方法的比较中展现出了显著的优势。本文方法的优势在于构建农业知识图谱。通过这一步骤，为后续的检索工作提供了一个结构化的知识库和丰富的上下文信息。利用 Citespace 软件，能够清晰地展示农业领域中的实体、它们之间的关系以及这些关系之间的语义联系，从而构建出一个结构化的语义网络。这个网络不仅广泛涵盖了农业知识，还能够揭示知识间的内在联系和层次结构。基于这个图谱，通过扩展关键词向量和采用相似度算法匹配等策略，显著提升了检索的准确性和效率。

5 结语

本研究针对当前农业知识语义智能检索方法难以精准捕捉深层语义的局限性，提出了一种基于 Citespace 知识图谱的创新方法。本研究的创新点主要体现在以下几个方面：一是将词向量化表示与 Citespace 知识图谱相结合，实现了对农业知识深层语义的精准捕捉；二是对特征向量进行扩展，丰富了语义维度，提高了检索的准确性和全面性；三是引入余弦相似度算法，实现了对检索结果与用户需求之间相似度的精确计算。这些创新点共同构成了本研究的核心竞争力，为农业知识语义智能检索领域的发展提供了新的思路和方法。

参考文献：

[1] 杨曦宇 . 基于词向量扩展的语义信息检索研究综述及应用展望 [J]. 林业科技情报 ,2024,56(1):212-215.

[2] 陈彦海 . 基于语义特征挖掘的图书馆文献资源智能检索方法 [J]. 信息与电脑 (理论版),2024,36(2):125-127.

[3] 张小韬 , 季小龙 . 基于自然语言理解的电力调度文本检索方法研究及应用 [J]. 黑龙江电力 ,2023,45(5):466-470.

[4] 魏凡其 . 基于人工智能技术的海量计算机信息检索方法设计 [J]. 电子技术与软件工程 ,2022(20):216-219.

[5] 尹梦岩 , 王梦霞 . 智能检索环境下语义分词调整策略的研究 [J]. 河南科技 ,2022,41(20):155-158.

[6] 颜小平 , 严长春 , 马顺 , 等 . 智能检索系统中生成语义分词的原理及调整策略 [J]. 中国发明与专利 ,2022,19(9):42-51.

[7] 陆柳杏 , 吴丹 . 非物质文化遗产领域汉藏双语本体的语义检索策略研究 [J]. 图书情报工作 ,2022,66(13):15-24.

[8] 吴旭东 , 黄文波 . 智能化升级检索系统中语义检索策略及实例分析 [J]. 中国发明与专利 ,2022,19(5):74-79.

[9] 林宗英 , 林民山 . 基于语义相似度的数字文献推广信息智能检索算法 [J]. 齐齐哈尔大学学报 (自然科学版), 2022, 38(1): 33-38.

[10] 倪子健 , 李文强 , 唐忠 . 基于网络表示学习的本体语义挖掘与功能语义检索方法 [J]. 工程设计学报 , 2021, 28(5): 539-547.

[11] 张云中 , 祝蕊 . 面向知识问答系统的图情学术领域知识图谱构建：多源数据整合视角 [J]. 情报科学 ,2021, 39(5): 115-123.

[12] 蒋红健 . 高校数字档案资源智能语义检索技术策略研究 [J]. 兰台世界 ,2020(12):57-60.

【作者简介】

陈素清（1982—），女，蒙古族，内蒙古呼和浩特人，硕士，副教授，研究方向：教育信息化、教育大数据、计算机应用。

（收稿日期：2024-05-28）