(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120123455 A
(43)申请公布日 2025.06.10

(21)申请号 202510197098.2

(22)申请日 2025.02.21

(71)申请人 金陵科技学院
地址 211169 江苏省南京市江宁区弘景大道99号

(72)发明人 陈文君 周陈新 洛汤姆 朱明宇 封宇乾 苑惠丽 陈晖

(74)专利代理机构 南京苏高专利商标事务所(普通合伙) 32204
专利代理师 颜盈静

(51)Int.Cl.
G06F 16/31(2019.01)
G06F 16/332(2025.01)
G06F 18/23213(2023.01)
G06F 40/216(2020.01)
G06F 40/279(2020.01)

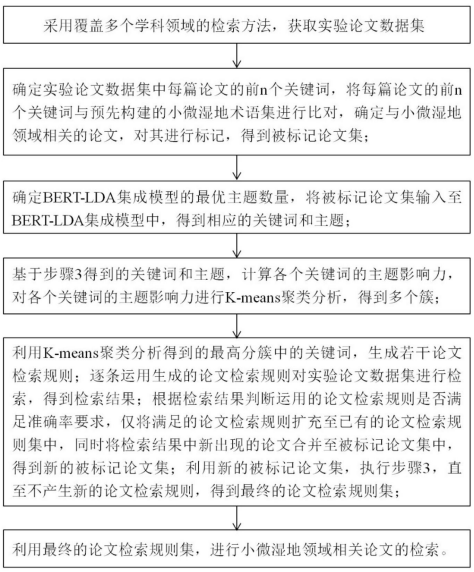
权利要求书2页 说明书9页 附图4页

(54)发明名称

一种面向小微湿地异质性名称的文献检索规则构建方法

(57)摘要

本发明公开了一种面向小微湿地异质性名称的文献检索规则构建方法,包括:获取实验论文数据集;将实验论文数据集与小微湿地术语集进行比对,确定被标记论文集;确定BERT-LDA集成模型的最优主题数量,将被标记论文集输入至BERT-LDA集成模型中,得到关键词和主题;计算各个关键词的主题影响力,对各个关键词的主题影响力进行K-means聚类分析;以此生成若干论文检索规则;逐条运用论文检索规则对实验论文数据集进行检索,根据检索结果判断论文检索规则是否满足准确率要求,仅将满足的论文检索规则扩充至已有的论文检索规则集中,同时更新被标记论文集;重复执行上述步骤,直至不产生新的论文检索规则;利用论文检索规则集,进行小微湿地领域相关论文的检索。



CN 120123455 A

1.一种面向小微湿地异质性名称的文献检索规则构建方法,其特征在于:包括以下步骤:

步骤1:采用覆盖多个学科领域的检索方法,获取实验论文数据集;

步骤2:确定实验论文数据集中每篇论文的前n个关键词,将每篇论文的前n个关键词与预先构建的小微湿地术语集进行比对,确定与小微湿地领域相关的论文,对其进行标记,得到被标记论文集;

步骤3:确定BERT-LDA集成模型的最优主题数量,将被标记论文集输入至BERT-LDA集成模型中,得到相应的关键词和主题;

步骤4:基于步骤3得到的关键词和主题,计算各个关键词的主题影响力,对各个关键词的主题影响力进行K-means聚类分析,得到多个簇;

步骤5:利用K-means聚类分析得到的最高分簇中的关键词,生成若干论文检索规则;逐条运用生成的论文检索规则对实验论文数据集进行检索,得到检索结果;根据检索结果判断运用的论文检索规则是否满足准确率要求,仅将满足的论文检索规则扩充至已有的论文检索规则集中,同时将检索结果中新出现的论文合并至被标记论文集中,得到新的被标记论文集;利用新的被标记论文集,执行步骤3,直至不产生新的论文检索规则,得到最终的论文检索规则集;

步骤6:利用最终的论文检索规则集,进行小微湿地领域相关论文的检索。

2.根据权利要求1所述的一种面向小微湿地异质性名称的文献检索规则构建方法,其特征在于:在所述BERT-LDA集成模型中执行以下操作:

采用全连接层将LDA模型提取到的主题概率分布矩阵和BERT模型到的语义特征矩阵以深度学习的方式串联融合按照一定比例进行融合,得到融合后的高维特征向量;

对融合后的高维特征向量依次进行标准化处理和降维操作,得到相应的关键词和主题。

3.根据权利要求1所述的一种面向小微湿地异质性名称的文献检索规则构建方法,其特征在于:所述的确定BERT-LDA集成模型的最优主题数量,具体操作包括:

确定BERT-LDA集成模型的主题数量取值范围;

以主题数量作为聚类数量,按照下式逐一计算不同主题数量下的silhouette值:

$$silhouette=\frac{(b-a)}{\max(a,b)}$$

其中,b表示每个样本点的平均最近聚类距离,a表示平均聚类质心聚类;

取最高silhouette值的主题数量作为BERT-LDA集成模型的最优主题数量。

4.根据权利要求1所述的一种面向小微湿地异质性名称的文献检索规则构建方法,其特征在于:步骤4中,基于步骤3得到的关键词和主题,按照下式,计算各个关键词的主题影响力:

$$TI(w|t)=k\cdot p(w|t)+(1-k)\cdot\left[Sim(\mathrm{BERT}(w),TopicEmbedding(t))\cdot\frac{p(w|t)}{p(w)}\right]$$

式中,TI(w|t)表示关键词w在主题t中的生成概率,p(w)为关键词w在整个实验论文数据集中的全局概率,BERT(w)表示关键词w的BERT嵌入向量,TopicEmbedding(t)表示主题t

的嵌入向量,它是通过关键词的BERT嵌入向量均值计算的,k为平衡参数,用于控制LDA模型和BERT模型之间的融合比例,p(w|t)表示主题-词分布,Sim(•)表示余弦相似度函数。

一种面向小微湿地异质性名称的文献检索规则构建方法

技术领域

[0001] 本发明涉及自然语言处理技术领域,特别是涉及一种面向小微湿地异质性名称的文献检索规则构建方法。

背景技术

[0002] 术语多样性和命名异质性是特定领域文献检索和信息获取面临的一项挑战,对其深入探讨有助于系统的掌握领域知识,提升理论研究与社会实践的效率。小微湿地是地理、生态环境等交叉学科的新兴概念,发挥着多种生态系统服务功能,但有着术语多样性、命名异质性的特征。开展小微湿地术语多样性研究,并以此为切入,系统构建领域学术知识,对提升小微湿地的保护管理水平具有重要意义。

[0003] 目前,各类小微湿地的具体名称(例如池塘、水塘、沟渠、溪流、春沼、秋洼、泉、人工湿地等)主要通过人工检索归纳而来。然而,人工检索的覆盖面有限,可能遗漏某些特定名称,导致小微湿地检索规则不全面。

发明内容

[0004] 发明目的:为解决现有技术存在的检索覆盖面有限的问题,本发明提出了一种面向小微湿地异质性名称的文献检索规则构建方法。

[0005] 技术方案:一种面向小微湿地异质性名称的文献检索规则构建方法,包括以下步骤:

[0006] 步骤1:采用覆盖多个学科领域的检索方法,获取实验论文数据集;

[0007] 步骤2:确定实验论文数据集中每篇论文的前n个关键词,将每篇论文的前n个关键词与预先构建的小微湿地术语集进行比对,确定与小微湿地领域相关的论文,对其进行标记,得到被标记论文集;

[0008] 步骤3:确定BERT-LDA集成模型的最优主题数量,将被标记论文集输入至BERT-LDA集成模型中,得到相应的关键词和主题;

[0009] 步骤4:基于步骤3得到的关键词和主题,计算各个关键词的主题影响力,对各个关键词的主题影响力进行K-means聚类分析,得到多个簇;

[0010] 步骤5:利用K-means聚类分析得到的最高分簇中的关键词,生成若干论文检索规则;逐条运用生成的论文检索规则对实验论文数据集进行检索,得到检索结果;根据检索结果判断运用的论文检索规则是否满足准确率要求,仅将满足的论文检索规则扩充至已有的论文检索规则集中,同时将检索结果中新出现的论文合并至被标记论文集中,得到新的被标记论文集;利用新的被标记论文集,执行步骤3,直至不产生新的论文检索规则,得到最终的论文检索规则集;

[0011] 步骤6:利用最终的论文检索规则集,进行小微湿地领域相关论文的检索。

[0012] 进一步的,在所述BERT-LDA集成模型中执行以下操作:

[0013] 采用全连接层将LDA模型提取到的主题概率分布矩阵和BERT模型到的语义特征矩

阵以深度学习的方式串联融合按照一定比例进行融合,得到融合后的高维特征向量;

[0014] 对融合后的高维特征向量依次进行标准化处理和降维操作,得到相应的关键词和主题。

[0015] 进一步的,所述的确定BERT-LDA集成模型的最优主题数量,具体操作包括:

[0016] 确定BERT-LDA集成模型的主题数量取值范围;

[0017] 以主题数量作为聚类数量,按照下式逐一计算不同主题数量下的silhouette值:

[0018] $$silhouette=\frac{(b-a)}{\max(a,b)}$$

[0019] 其中,b表示每个样本点的平均最近聚类距离,a表示平均聚类质心聚类;

[0020] 取最高silhouette值的主题数量作为BERT-LDA集成模型的最优主题数量。

[0021] 进一步的,步骤4中,基于步骤3得到的关键词和主题,按照下式,计算各个关键词的主题影响力:

[0022] $$TI(w|t)=k\cdot p(w|t)+(1-k)\cdot\left[Sim(\mathrm{BERT}(w),TopicEmbedding(t))\cdot\frac{p(w|t)}{p(w)}\right]$$

[0023] 式中,TI(w|t)表示关键词w在主题t中的生成概率,p(w)为关键词w在整个实验论文数据集中的全局概率,BERT(w)表示关键词w的BERT嵌入向量,TopicEmbedding(t)表示主题t的嵌入向量,它是通过关键词的BERT嵌入向量均值计算的,k为平衡参数,用于控制LDA模型和BERT模型之间的融合比例,p(w|t)表示主题-词分布,Sim(·)表示余弦相似度函数。

[0024] 有益效果:与现有技术相比,本发明具有以下优点:

[0025] (1)本发明方法结合了BERT模型的长文本语义理解优势与LDA模型的主题可解释性,从知识工程角度,凝聚小微湿地目前分散化的领域知识,促进其保育管理,实现小微湿地异质性名称的自动化挖掘,扩展小微湿地文献检索规则,并为文本主题挖掘的可解释性研究提供新的场景;

[0026] (2)本发明方法通过目标词向量,能够充分结合上下文语义信息,弥补LDA主题模型的劣势,训练出更优的主题向量,得到具有更好细粒度和聚类精准度的关键主题识别效果。

附图说明

[0027] 图1为本发明提出的一种面向小微湿地异质性名称的文献检索规则构建方法的流程图;

[0028] 图2为本发明提出的一种面向小微湿地异质性名称的文献检索规则构建方法的流程图;

[0029] 图3为不同主题数量下,LDA模型的Cv值和Perp值;

[0030] 图4为不同主题数量下,BERT-LDA集成模型的silhouette值;

[0031] 图5为BERT-LDA集成模型首轮与迭代至稳定之后的可视化图谱,其中,图5中的(a)为BERT-LDA集成模型首轮的可视化图谱,图5中的(b)为BERT-LDA集成模型迭代至稳定之后的可视化图谱;

[0032] 图6为论文检索规则。

具体实施方式

[0033] 现结合附图和实施例进一步阐述本实施例的技术方案。

[0034] 如图1所示,本实施例提出了一种面向小微湿地异质性名称的文献检索规则构建方法,其主要包括以下步骤:

[0035] 步骤1:基于预先构建的小微湿地术语集,在实验论文数据集中标记小微湿地领域的相关论文;如图2所示,具体操作包括:

[0036] 实验论文数据集是来源于万方数据库的期刊论文,检索时间跨度为2012年1月1日至2022年12月31日,检索时间为2023年12月。为保证广泛全面的获取小微湿地相关论文,采用了一种囊括众多学科领域的检索方法,具体以“湿地”、“水体”、“水域”三者之一作为关键词,与《普通高等学校本科专业目录(2023年版)》中的所有二级专业名称进行“AND”连接(检索规则示例如表1所示)。以中文期刊文献为研究资料是因为,我国从寒带到热带的地理环境包括多样化的湿地类型,并且中文语境包含异质化的小微湿地具体名称及其使用方式。选取上述三个关键词是因为它们是水科学研究领域的基本术语,可作为小微湿地相关研究的前置条件;学科分类包括哲学、历史学、文学、艺术学、教育学、法学、经济学、管理学、理学、工学、农学、医学12个大类,下设合计93个一级学科分类,共703个二级专业名称。此外,若二级专业名称中包含“与”、“以及”等连接词,需额外对其进行分词和补词,然后执行“OR”查询。例如,将“自然地理与资源环境”拆分为“自然地理”和“资源环境”,将“海洋资源与环境”拆分并补全为“海洋资源”和“海洋环境”(表1)。为控制实验数据量并确保相关性,每次查询依据主题相关度排序,选取前15条记录,最终得到4606篇论文。提取论文的文本内容,接着进行去重、清洗、分词处理,最后采用TF-IDF算法得到每篇论文中,按重要性排序的前n个关键词。

[0037] 表1检索规则示例

[0038]

学科大类	一级学科分类	二级专业名称	检索表达式

[0039]

理学	地理科学类	地理科学	(主题:("湿地" OR "水体" OR "水域") AND 主题:("地理科学") AND 出版时间:[2012-01-01 TO 2022-12-31]
		自然地理与资源环境(拆分为:自然地理 AND 资源环境)	(主题:("湿地" OR "水体" OR "水域") AND 主题:("自然地理与资源环境" OR "自然地理" OR "资源环境")) AND 出版时间:[2012-01-01 TO 2022-12-31]
		海洋资源与环境(拆分为:海洋资源 AND 海洋环境)	(主题:("湿地" OR "水体" OR "水域") AND 主题:("海洋资源与环境" OR "海洋资源" OR "海洋环境")) AND 出版时间:[2012-01-01 TO 2022-12-31]

[0040] 预先构建的小微湿地术语集是采用专家访谈与人工判读法,从小微湿地领域的综述论文中汇总代表性术语、命名构建得到的,作为后续第一轮论文检索与标记的依据。上述综述论文来源于主流学术会议、期刊数据库。

[0041] 将实验论文数据集中每篇论文的前n个关键词与预先构建的小微湿地术语集比

对,若存在交集,则标记此篇论文属于小微湿地研究领域,以此得到被标记论文集。

[0042] 步骤2:搭建BERT-LDA集成模型。如图2所示,在本实施例中,利用LDA模型,提取被标记的论文的主题概率分布矩阵,利用BERT模型,提取被标记的论文的语义特征矩阵;采用Python语言Keras 2.1.0开发包中的全连接层(Dense)将语义特征矩阵与主题概率分布矩阵以深度学习的方式串联融合,得到融合后的高维特征向量,采用填充或截断法对高维特征向量进行标准化,确保所有向量长度一致,并使用umap 0.1.1开发包中的UMAP算法模块,将特征向量降至2维,得到具有较高主题影响力的关键词。

[0043] 主题概率分布矩阵为每个文档在不同主题上的分布情况提供信息,可视为文档的语义“标签”,是理解长文本主题结构与内容、揭示潜在知识的基础。由于主题概率特征不包含词汇之间的上下文关系,LDA模型在处理长文本中复杂的词汇依赖关系和语境信息时存在局限,难以理解词义的多样性和深层语义信息。为此,本实施例引入BERT模型,它采用一种可微调的双向Transformer编码器,而非传统循环神经网络,实现自注意力(self-attention)叠加,即词向量能够更精准的反映不同语境下的含义,从而有效地捕捉上下文语义信息。

[0044] 步骤3:确定BERT-LDA集成模型的最优主题数量;具体操作包括:

[0045] 首先,确定BERT-LDA集成模型的主题数量取值范围;

[0046] 然后,由于BERT-LDA集成模型并非基于概率,主题数量为聚类数量,逐一计算不同主题数量下的silhouette值,确定最优主题数量。

[0047] 轮廓系数(silhouette coefficient)是解释和验证数据聚类一致性的典型指标,计算公式如下:

[0048] $$silhouette=\frac{(b-a)}{\max(a,b)}$$

[0049] 其中,b表示每个样本点的平均最近聚类距离;a表示平均聚类质心聚类,silhouette的取值范围为0~1,该值越接近1,两个聚类样本点彼此越相近。

[0050] 在本实施例中,采用sklearn 1.4.2开发包中的silhouette_score模块计算silhouette。

[0051] 步骤4:将被标记论文集输入至BERT-LDA集成模型中,得到相应的关键词和最优主题数量的主题个数。

[0052] 步骤5:基于步骤4得到的关键词和主题,计算各个关键词的主题影响力,对各个关键词的主题影响力进行K-means聚类分析,得到多个簇;具体操作包括:

[0053] 将模型最优主题数量输入LDA模型,获取全局词频p(w)和每个主题-词分布p(w|t),计算每个关键词相对于整篇论文的主题影响力。为了兼顾局部主题分布与全局词频特征,增强关键词的可解释性,采用线性组合的方式计算LDA模型中每个关键词相对于各个主题的主题影响力。计算公式如下:

[0054] $$TI(w|t)=k\cdot p(w|t)+(1-k)\cdot\frac{p(w|t)}{p(w)}$$

[0055] 式中,TI(w|t)表示关键词w在主题t中的生成概率,p(w)为关键词w在整个实验论文数据集中的全局概率,k为平衡参数,取值为0-1。若关键词在主题内出现概率较高(即p(w

|t)高),而全局出现概率较低(即p(w)低),其主题影响力得分将大幅提升。

[0056] 将BERT模型输出的BERT特征通过余弦相似度与其融合,主题影响力公式可改写为:

[0057] $$TI(w|t)=k\cdot p(w|t)+(1-k)\cdot\left[Sim(\mathrm{BERT}(w),TopicEmbedding(t))\cdot\frac{p(w|t)}{p(w)}\right]$$

[0058] 式中,BERT(w)表示关键词w的BERT嵌入向量,TopicEmbedding)t_表示主题t的嵌入向量,它是通过关键词w的BERT嵌入均值计算的,k控制LDA模型和BERT模型之间的融合比例。为保证均衡取k=0.5,定量计算出BERT-LDA模型提取关键词的主题影响力。

[0059] 对各个关键词的主题影响力进行K-means聚类分析,并采用Gephi 0.10工具构建可视化图谱,以外围大节点和颜色相同、内嵌文字的小节点分别表示主题、关键词,两者之间的连线表示主题影响力的方式直观展示出BERT高维特征融合LDA特征后的结果。

[0060] 步骤6:对K-means聚类得到的最高分簇中的关键词,采用排列组合中的组合操作,连接2个至多个具有高影响力的关键词,生成若干论文检索规则;逐条运用生成的论文检索规则对实验论文数据集进行检索,得到检索结果;根据检索结果判断运用的论文检索规则是否满足准确率要求,仅将满足的论文检索规则扩充至已有的论文检索规则集中,同时将检索结果中新出现的论文合并至被标记论文集中,得到新的被标记论文集;利用新的被标记论文集,执行步骤3,直至不产生新的论文检索规则,即模型与数据均至稳定状态,得到最终的论文检索规则集。

[0061] 若某篇论文的前n个关键词包含该条检索规则的所有关键词时,则初步判定这篇论文属于目标领域。根据检索结果判断运用的论文检索规则是否满足准确率要求,具体操作包括:基于领域专家的人工判读,分析所有初步判定的论文,得到该条检索规则的准确率。基于预先设定的准确率阈值,将高于准确率阈值的新规则扩充至已有的论文检索规则。

[0062] 步骤7:利用最终的论文检索规则集,进行小微湿地领域相关论文的检索。

[0063] 现结合具体案例对实施例提出的构建方法达到的技术效果做进一步说明。

[0064] 表2为采用实施例中步骤1提到的步骤构建的小微湿地术语集。这些术语印证了《小微湿地保护与管理规范(GB/T42481-2023)》中“小微湿地”、“小微湿地群”、“生态保育”、“生态重建”、“自然恢复”等专业概念,并且形成了具象化的命名。将实验论文数据集中每篇论文的前50个TF-IDF关键词与小微湿地术语集比对,若存在交集,则标记此篇论文属于小微湿地研究领域,由此得到298篇领域相关论文。

[0065] 表2小微湿地术语集

[0066]

序号	术语	序号	术语	序号	术语	序号	术语
1	小微湿地	7	小型脆弱水体	13	池塘	19	泉
2	临时湿地	8	被忽视的淡水栖息地	14	水塘	20	溪
3	微型湿地	9	非河漫滩湿地	15	多水塘	21	春沼
4	孤立湿地	10	非洪泛平原湿地	16	井	22	秋洼
5	小型水体	11	人工湿地	17	堰	23	湿地镶嵌体
6	小水体	12	极地湿地	18	沟		

[0067] 为了说明本实施例提出的BERT-LDA集成模型的模型性能优于LDA模型,进行以下比较。

[0068] 设定LDA模型和BERT-LDA集成模型的主题数量取值范围一样,均为2-30;

[0069] 逐一计算不同主题数量下的Perp值和Cv值,确定LDA模型的最优主题数量。

[0070] 模型困惑度(Perplexity)是表示模型泛化能力的代表性指标,计算公式如下:

[0071] $$Perp=\exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\}$$

[0072] 其中,logp(w_d)表示单个文档d的对数概率,M表示文档总数,对所有文档的对数概率求和,得到整个文档集合的对数似然值,即整体拟合程度;$\sum_{d=1}^{M}N_d$表示整个文档集合的总词数。困惑度Perp取值>0,结果越小,模型泛化能力越强。

[0073] 其中,主题一致性是评估主题模型性能的重要指标,它通过度量高频词的语义关联性判断生成主题的质量,计算公式如下:

[0074] $$\begin{cases}Cv=\dfrac{1}{|P|}\sum\limits_{(w_i,w_j)\in P}\mathrm{Sim}(w_i,w_j)\\ \mathrm{Sim}(w_i,w_j)=\dfrac{\mathbf{v}_{W_i}\cdot\mathbf{v}_{W_j}}{\|\mathbf{v}_{W_i}\|\|\mathbf{v}_{W_i}\|}\end{cases}$$

[0075] 式中,w_i,w_j表示主题内的两个高频词,P表示主题内词对的集合,Sim表示计算词对之间的相似度、量化语义相关性的余弦相似度算法。$\mathbf{v}_{W_i}$和$\mathbf{v}_{W_j}$是w_i和w_j的词向量表示,点积·计算词向量的相似度,分母||v||是词向量的L2范数,用来归一化。Cv是该主题的整体一致性得分,通过对所有词对的相似性得分取平均而得,取值范围为0-1,得分越高,主题越紧凑,模型效果越好。

[0076] 由于BERT-LDA集成模型并非基于概率,无法直接得到Perp值,因此通过计算多种聚类簇数的silhouette值,确定最优主题数量。

[0077] 最终,通过比较两种模型最优主题数量的Cv值来评价其优劣

[0078] 采用gensim 4.3.2开发包中的Model模块计算Perp,以及CoherenceModel模块计算Cv;采用sklearn 1.4.2开发包中的silhouette_score模块计算silhouette。

[0079] 图3、图4分别展示了不同主题数量下,LDA模型的Cv值和Perp值,以及BERT-LDA集成模型的silhouette值。表3列出了两种模型在首轮,以及迭代至稳定后,采用最佳主题数所得到的Cv值。总体而言,BERT-LDA集成模型的Cv值在首轮,以及迭代至稳定后,相比LDA模型分别提高了2.96%、3.63%,因此集成模型具有更好的主题建模能力。对于LDA模型,首轮和迭代至稳定后的Perp值均呈现出,随着主题数量的增加,先小幅下降,接着保持平稳,然后显著上升的趋势(图3)。小幅下降与保持平稳的拐点大约出现在主题数量3至5,而保持平稳与显著上升的拐点大约出现在17或18。据此,该模型最优主题数量的有效范围初步确定在3至18。进一步,当主题数量为4时,模型首轮和迭代至稳定后的Cv值均达到最大。综合两种指标,确定LDA模型首轮和迭代至稳定后的最佳主题数量均为4,其Cv值分别为0.4838、0.5187(表3);而相比于其他主体个数,Perp值仅有微小提升(图3),进一步说明4个最佳主

题数量的可靠性。

[0080] 表3两种模型评价对比

[0081]

模型	首轮Cv值	迭代至稳定的Cv值
LDA	0.4838	0.5187
BERT-LDA	0.5134	0.5541

[0082] 对于BERT-LDA集成模型,为保证均衡,设定两种模型的融合比例k为0.5。首轮和迭代至稳定后,融合特征分别在聚类数量4和3时取得最高的silhouette值,因此集成模型最优主题数量分别为4和3。其中,首轮运行时,最优主题数量和LDA模型相同,再次说明其可靠性。将此结果输入LDA模型获取p(w)和p(w|t),再求解出集成模型首轮、迭代至稳定后的Cv值,分别为0.5134、0.5541(表3)。

[0083] 图5展示了BERT-LDA集成模型首轮和迭代至稳定后的可视化图谱,图5中,外围大节点和颜色相同、内嵌文字的小节点分别表示主题、关键词;两者之间的连线表示主题影响力,越粗影响力越大,反之越小;红圈表示聚类后的高影响力关键词。相比以往高维特征向量关键词提取中存在的不可解释性问题,本实施例融合BERT模型生成的768维语义特征向量,以及LDA模型生成的4维主题概率分布向量,然后进行降维、聚类处理,得到主题、关键词及其主题影响力的可视化图谱,直观呈现出从大量期刊论文中挖掘提炼的小微湿地领域知识。

[0084] 从模型首轮运行(图5中的(a))到迭代至稳定状态(图5中的(b)),论文的主题数量由4个减少为3个,关键词数量由87个减少至72个,主题影响力均值由5.49×10^{-3}提升至5.96×10^{-3},表明关键词的聚合性有所增强。此外,不同颜色节点的交叉更少出现,表明关键词的语义空间分布特征更为明确。例如,“景观”和“生态”这两个关键词,由首轮从属于两个主题,转变为稳定后的属于同一主题。由此可见,经过迭代,BERT-LDA集成模型降低了关键词的特征维度,并提升了它们的语义聚合效果。进一步分析,模型首轮运行所挖掘出的4个主题可以概括为“城市湿地公园”、“湿地污染净化”、“池塘水产养殖”、“湿地面积变化”;而迭代至稳定后,“湿地面积变化”类别消失,其中的关键词被分配到另外三个主题。推测原因在于,相比于其他主题更为聚焦的研究内容,“湿地面积变化”中的关键词(例如“面积”、“类型”、“变化”等),大多是从侧面反映小微湿地的多种特征,在实际研究中起着辅助作用,因而多轮迭代后被归入其他类别。

[0085] 如图5所示,在“城市湿地公园”主题中,“湿地”、“城市”、“公园”、“生态”、“景观”是模型迭代至稳定后具有高主题影响力的关键词。在“湿地污染净化”主题中,“人工湿地”、“植物”、“微生物”、“去除”、“污水”、“系统”是模型迭代至稳定后具有高主题影响力的关键词。其中,“人工湿地”在迭代前后的主题影响力始终最高(>15×10^{-3});而“植物”、“微生物”的主题影响力分别由首轮的11.6×10^{-3}上升至12.9×10^{-3}、7.1×10^{-3}上升至8.3×10^{-3}。这一发现与前人研究相互印证,人工湿地是文献中单一名称占比最多的小微湿地类型(约15%),在不同气候和地理环境中均有分布。在“池塘水产养殖”主题中,“池塘”、“养殖”、“水产”、“水体”是模型迭代至稳定后具有高主题影响力的关键词。

[0086] 图6展示了高于70%准确率阈值的112条论文检索规则,它们分别由“湿地”“植物”、“去除”、“污水”、“研究”、“系统”、“微生物”、“城市”、“景观”、“公园”、“生态”这11个关

键词中的部分所组成。在这些检索规则中,109条属于“湿地污染净化”主题,其余3条属于“城市湿地公园”主题。BERT-LDA集成模型首轮迭代后,得到63条检索规则,新增61个论文标记;第二轮迭代后,未得到新规则,但新增5个论文标记,并进入稳定状态,表明迭代运行方法,使得小微湿地领域知识得到逐步挖掘。

[0087] 分析图6可知,关键词“湿地”、“植物”、“去除”、“污水”、“研究”、“系统”中任意3个组合得到的论文检索规则准确率为70%至100%。这表明囊括多学科领域的实验数据集中,有关小微湿地污水净化的论文占绝大多数,包括去污效果、去污机制等主题;相对而言,有关湿地植物和湿地系统的论文则有所减少,推测因为它们主要为去污、净化等主题提供辅助,而非主要研究对象。

[0088] 关键词“微生物”与上述前6个关键词组合后,新增66个论文标记,并且与关键词“植物”所在论文检索规则的检索结果高度相似,表明小微湿地领域中,植物与微生物之间存在密切联系。

[0089] 对于“城市湿地公园”主题中的规则,只有关键词“湿地”、“城市”、“景观”、“公园”,和“湿地”、“城市”、“景观”、“生态”,以及“湿地”、“城市”、“景观”、“公园”、“生态”这三种组合的正确率超过阈值,且低于80%。推测原因在于,部分城市湿地公园的研究侧重于其更为综合的社会服务价值,而非生态环境的单一方面,例如公园道路的设计布局是否便捷,以及周边居民对公园的整体满意度。这类研究并未归入小微湿地的研究范畴,使得该主题下高正确率的检索规则相对较少。

[0090] 由于篇幅限制,表4列举了新增的典型标记论文及其关联检索规则。分析可知,通过本实施例方法挖掘的论文检索规则,也就是高主题影响力的关键词组合,与“湿地”、“水体”、“水域”作“AND”连接之后,能够在脱离小微湿地具体命名术语的情况下,有效识别相关学术论文,实现对《小微湿地保护与管理规范(GB/T42481-2023)》中专业概念,以及专家判读综述论文所得术语集(表2)的有效延升。例如,论文2可由图6中的规则84、107检索得到,论文8可由规则30、65、66、68、73、98、101、102、105检索得到。以上规则并不包含小微湿地已有的、多样化的术语和命名,但是通过迭代式的论文挖掘、训练,自动识别出领域相关论文,辅助汇聚分散化的领域知识。

[0091] 表4典型标记论文及其检索规则

[0092]

序号	论文	规则序号
1	周克成, 梁敏, 饶利军. 人工湿地污水处理技术及其	1-34、36-39、40-77、

[0093]

	发展应用[J]. 皮革制作与环保科技, 2022, 3(15): 32-34.	79-99、101、102、103、105、106、107
2	李晓婷. 水解酸化+中温 UASB+生物接触氧化+人工湿地工艺处理规模化猪场废水的工程实践研究[J]. 水处理技术, 2013, 39(5): 128-130,134.	84、107
3	董辉. IBR 生物反应池+生物湿地组合工艺用于城市污水处理工程[J]. 城市建设, 2013(16).	10、30、32、40、41、65、66、67、68、70、72、73、75、77、80、98、99、100、101、102、103、105
4	王艳春. 安徽合肥四里河生态修复策略研究[J]. 中国园林, 2018, 34(7): 86-90.	12、33、35、38、43、67、71、72、76、77、78、82、99、102、103
5	张爱娣, 郑仰雄, 黄东兵. 5 种湿地植物对含盐生活污水的净化效果及其生理响应[J]. 江苏农业学报, 2020, 36(2): 384-390.	33、66、71、76、99、102
6	罗晶. 城市湿地水生植物治理水污染应用分析[J]. 绿色科技, 2019(18): 83-84.	30、65、66、68、98、73、101、102、105
7	史晓涛, 邱征, 张琼, 等. 城市内河入湖口门多田活水链湿地重构技术研究[J]. 安徽建筑, 2022, 29(6): 59-60.	66、99、102
8	李丹, 郑丙辉, 储昭升, 等. 洱海流域多塘湿地工程综合效果评价[J]. 华东师范大学学报（自然科学版）, 2021(4): 8-16.	30、65、66、68、73、98、101、102、105
9	齐婧含, 朱梦炎, 陆欣鑫. 城市公共资源景观水体浮游植物多样性特征——以丁香公园为例[J]. 国土与自然资源研究, 2020(1): 94-96.	110、111、112
10	曾思妍, 张葳. 城市湿地公园生态景观设计研究[J]. 艺术科技, 2022, 35(10): 210-212.	110、111、112

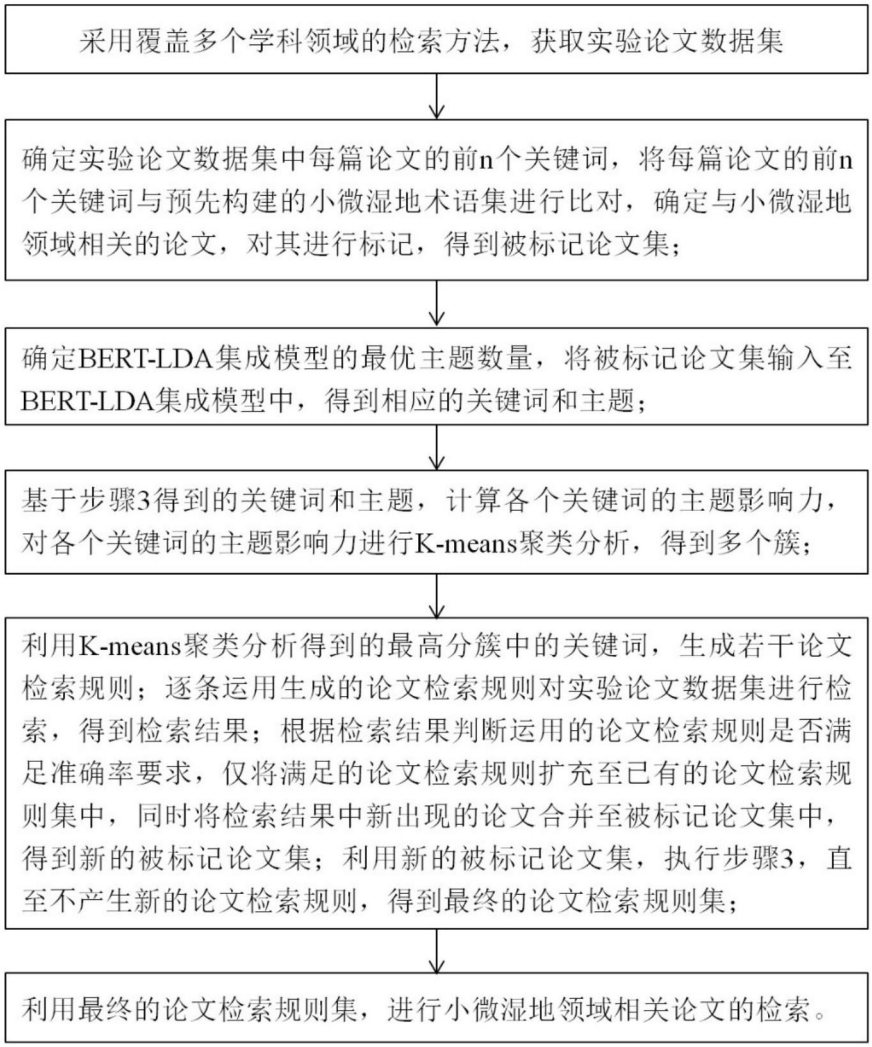

采用覆盖多个学科领域的检索方法，获取实验论文数据集
确定实验论文数据集中每篇论文的前n个关键词，将每篇论文的前n个关键词与预先构建的小微湿地术语集进行比对，确定与小微湿地领域相关的论文，对其进行标记，得到被标记论文集；
确定BERT-LDA集成模型的最优主题数量，将被标记论文集输入至BERT-LDA集成模型中，得到相应的关键词和主题；
基于步骤3得到的关键词和主题，计算各个关键词的主题影响力，对各个关键词的主题影响力进行K-means聚类分析，得到多个簇；
利用K-means聚类分析得到的最高分簇中的关键词，生成若干论文检索规则；逐条运用生成的论文检索规则对实验论文数据集进行检索，得到检索结果；根据检索结果判断运用的论文检索规则是否满足准确率要求，仅将满足的论文检索规则扩充至已有的论文检索规则集中，同时将检索结果中新出现的论文合并至被标记论文集中，得到新的被标记论文集；利用新的被标记论文集，执行步骤3，直至不产生新的论文检索规则，得到最终的论文检索规则集；
利用最终的论文检索规则集，进行小微湿地领域相关论文的检索。

图1

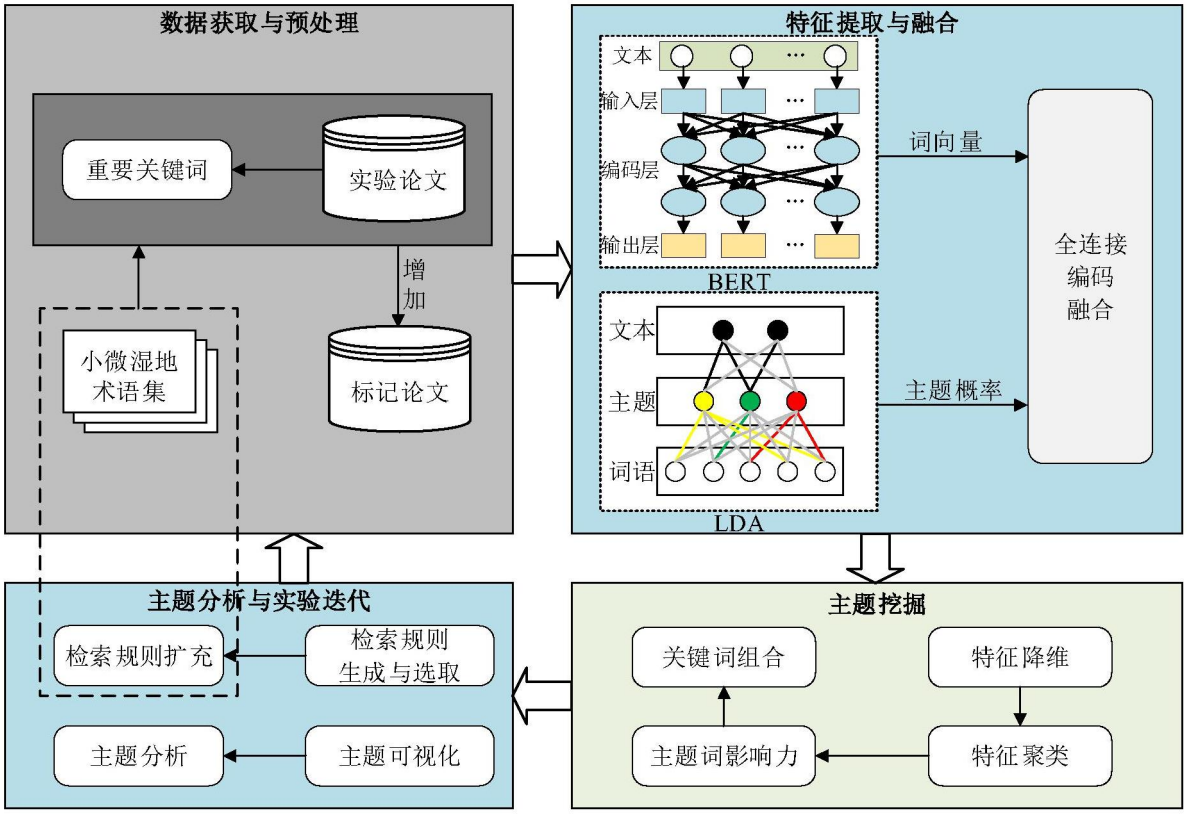

数据获取与预处理
重要关键词
实验论文
增加
小微湿地术语集
标记论文
特征提取与融合
文本
输入层
编码层
输出层
BERT
词向量
全连接编码融合
文本
主题
词语
LDA
主题概率
主题分析与实验迭代
检索规则扩充
检索规则生成与选取
主题分析
主题可视化
主题挖掘
关键词组合
特征降维
主题词影响力
特征聚类

图2

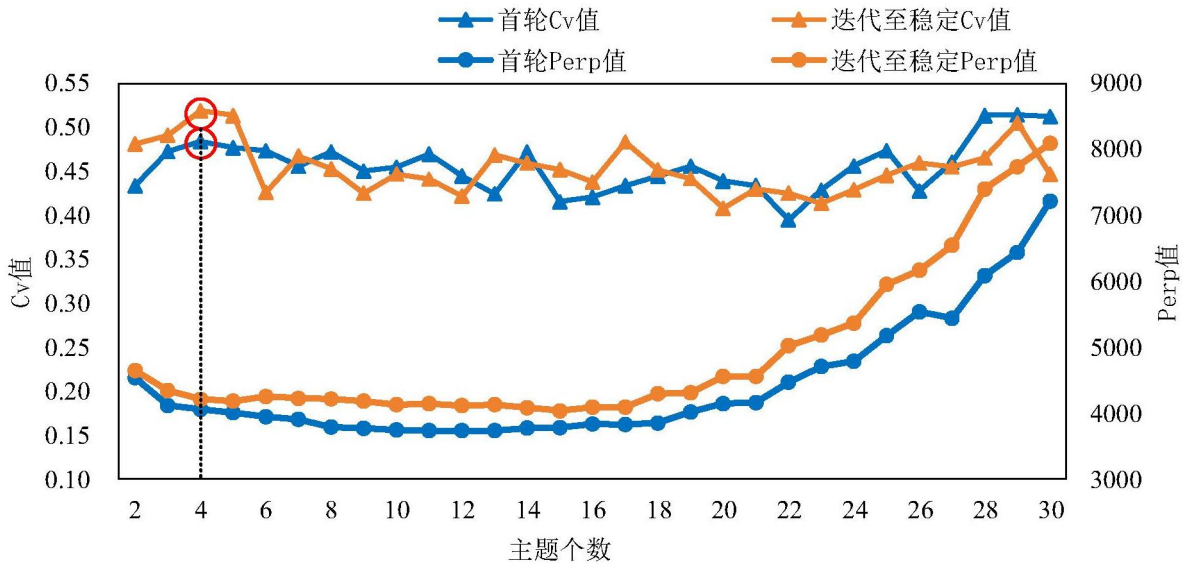

首轮Cv值
迭代至稳定Cv值
首轮Perp值
迭代至稳定Perp值
Cv值
Perp值
主题个数
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
9000
8000
7000
6000
5000
4000
3000
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30

图3

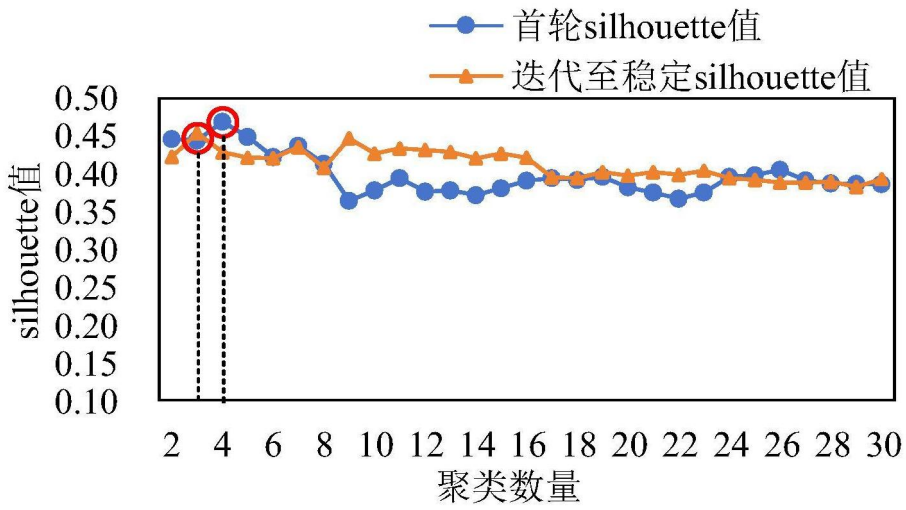

首轮silhouette值
迭代至稳定silhouette值
silhouette值
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
聚类数量

图4

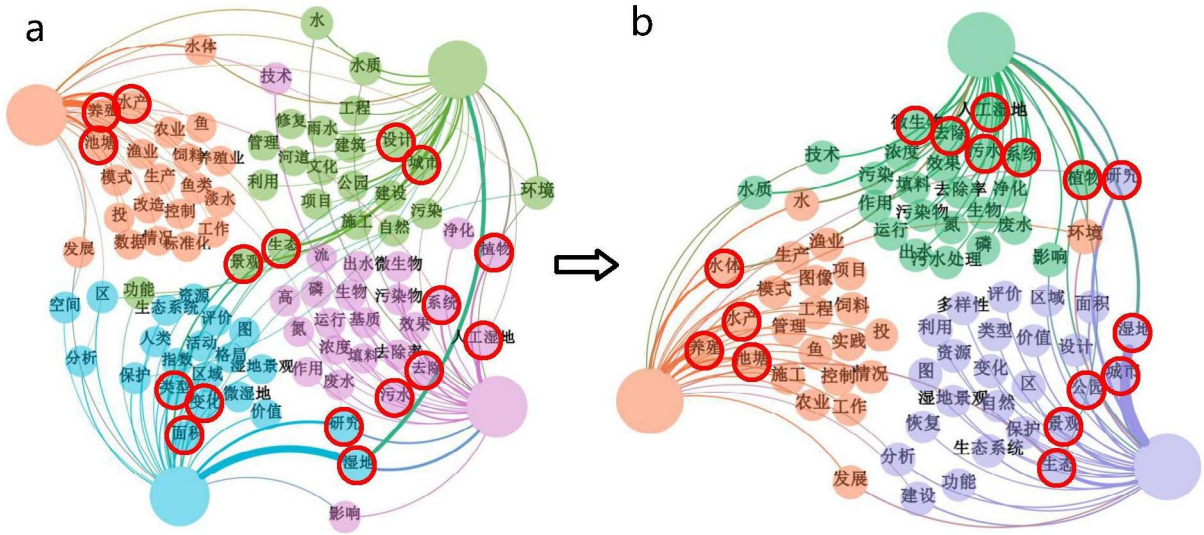

a
b
水
水体
技术
水质
工程
水产
养殖
池塘
农业
鱼
渔业
饲料
养殖业
管理
雨水
建筑
设计
河道
文化
城市
模式
生产
鱼类
利用
公园
建设
环境
淡水
项目
投
改造
控制
施工
自然
污染
净化
工作
发展
数据
情况
标准化
生态
流
出水
微生物
植物
景观
空间
区
功能
资源
生态系统
评价
高
磷
生物
污染物
系统
人类
活动
图
氮
运行
基质
效果
人工湿地
分析
指数
格局
浓度
填料
去除率
去除
保护
区域
湿地
景观
作用
废水
类型
变化
微湿地
污水
价值
研究
面积
湿地
影响
水
水质
技术
水体
养殖
池塘
水产
模式
生产
渔业
图像
项目
工程
饲料
管理
投
实践
鱼
施工
控制
情况
农业
工作
发展
分析
建设
功能
恢复
生态系统
保护
景观
生态
湿地景观
自然
公园
城市
湿地
评价
区域
面积
多样性
利用
类型
价值
设计
资源
图
变化
区
人工湿地
微生物
去除
污水
系统
浓度
效果
去除率
净化
植物
研究
水质
水
作用
污染物
生物
废水
运行
氮
磷
出水
污水处理
环境
影响

图5

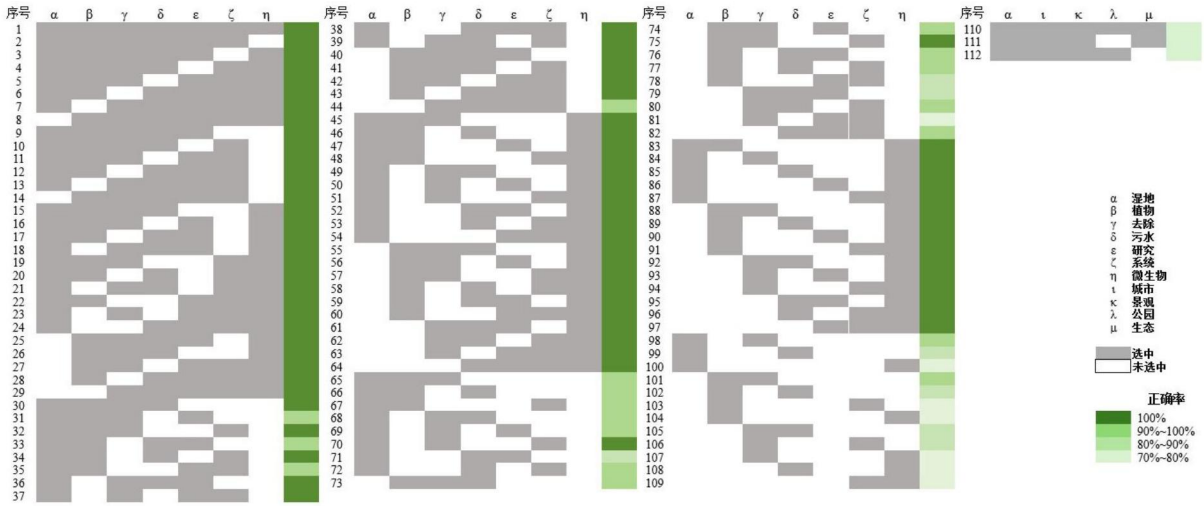

序号
α
β
γ
δ
ε
ζ
η
ι
κ
λ
μ
湿地
植物
去除
污水
研究
系统
微生物
城市
景观
公园
生态
选中
未选中
正确率
100%
90%~100%
80%~90%
70%~80%

图6