基于关键词语义功能的领域主题演化分析*

俞 琰 郑斯煜 葛 萌

(南京工业大学图书馆 南京 210009)

摘 要:[**研究目的**]领域主题演化分析利用领域科技文献,识别领域主题构成,揭示领域主题变化和演化脉络,具有十分重要的研究意义。[**研究方法**]使用提示学习模型识别关键词的语义功能,通过识别的关键词的不同语义功能,为领域主题演化分析添加不同的语义功能维度。[**研究结果/结论**]相较于传统的机器学习模型,提示学习模型可以使用更少的训练数据更精准地划分关键词的语义功能,而基于不同关键词语义功能进行的主题演化分析能够保留特定的语义信息,为领域主题演化分析提供更细致的分析视角。

关键词:关键词语义功能识别;提示学习;领域主题;主题演化分析;Bootstrapping

中图分类号:G250 **文献标识码**:A **文章编号**:1002-1965(2025)01-0187-11

引用格式:俞 琰,郑斯煜,葛 萌.基于关键词语义功能的领域主题演化分析[J].情报杂志,2025,44(1):187-197.

DOI:10.3969/j.issn.1002-1965.2025.01.024

Analysis of Domain Topic Evolution Based on Semantic Function of Keywords

Yu Yan Zheng Siyu Ge Meng

(Library of Nanjing Tech University, Nanjing 210009)

Abstract: [**Research purpose**] The analysis of domain theme evolution utilizes domain technology literature to identify the composition of domain topics, reveal the changes and evolution of domain themes, and has significant research significance. [**Research method**] This paper proposes using a prompt learning model to recognize the semantic function of keywords, and adding different semantic functional dimensions to domain topic evolution analysis by identifying different semantic functions of the keywords. [**Research result/conclusion**] Compared with the traditional machine learning model, the prompt learning model can more accurately delineate the semantic functions of keywords using less training data, and the topic evolution analysis based on different keyword semantic functions can retain specific semantic information and provide a more detailed analysis perspective for domain topic evolution analysis.

Key words: semantic function identification of keywords; prompt learning; domain topic; analysis of topic evolution; Bootstrapping

领域主题演化分析旨在利用领域文献,识别领域的主题构成,揭示领域主题的变化和演化脉络规律,是科技情报研究中的一项重要内容。目前,国内外领域主题演化分析的相关研究已具有一定规模。特别地,利用文献关键词表征知识具有更细的分析粒度,被广泛应用于领域主题演化分析之中。

然而,该类方法往往忽略关键词语义功能,造成关键词语义功能模糊,使用意图不明,脱离原文后难以解释,从而影响领域主题演化分析结果的准确性[1]。为此,本文利用大语言模型,提出一种优化的提示学习方法,使用少量标注样本识别关键词语义功能,进而从关键词语义功能的各个维度分析领域主题的演化。

1 相关研究

1.1 领域主题演化分析

目前,国内外领域主题演化分析的相关研究已具有一定规模。相关方法主要从频次、引证、内容等视角加以分析。

频次视角以文献发文频次、引证频次以及词语频次等频次统计信息,揭示不同研究主题的发展情况与

收稿日期:2024-05-23 修回日期:2024-06-05

基金项目:国家社会科学基金一般规划项目"数据驱动的高校技术转移供需信息挖掘模式构建研究"(编号:23BTQ098)研究成果。

作者简介:俞 琰,女,1972年生,博士,教授,硕士生导师,研究方向:数据挖掘;郑斯煜,男,1998年生,硕士研究生,研究方向:数据挖掘;葛萌,女,1997年生,硕士研究生,研究方向:数据挖掘。

变化趋势。例如,奉国和等[2]利用学科关键词及其在不同时间段内的出现频次,探讨领域学科内的研究热点以及其变化趋势;王康等[3]通过引入关键词的时间加权以修正词频及其时序趋势,由此识别具有语义特征的学科主题。

引证视角基于直接引证网络、共被引网络以及共引网络等文献引证信息开展领域主题探测与演化分析。如王伟等[4]构建直接引证网络观察不同学科主题之间地知识流动关系;Liu 等[5]利用文献中的作者、机构、国家和关键词等信息,构建共被引网络,分析领域开放创新的发展和演变;Xu 等[6]通过引证网络揭示领域关键演化路径;Dejian 等[7]构建引证网络,通过遍历计权,分别采用局部和全局两种策略提取关键演化路径。

内容视角利用关键词进行浅层共词分析以及深层主题挖掘分析。其中,共词分析通过挖掘两两关键词借助同一中介(如文献、作者、参考文献、引证上下文等)的耦合关联,构建共词网络,在此基础上分析领域主题演化。如吴胜男等[8]构建各生命周期阶段的关键词词共现网络,揭示相关领域不同发展阶段的主题演化过程;Wang 等[9]构建年度共词网络,分析领域主题演化趋势;王康等[10]在关键词时间修正词频测度基础上,构建关键词共现矩阵,绘制领域主题演化路径图;Duan 等[11]在关键词共现网络基础上,通过共邻接指数测度和链接预测揭示领域知识融合的结构性机制。此外,随着文本分析技术的不断发展,基于主题挖掘的学科演化研究引起了学者的广泛关注,其中关于主题的建模和识别主要基于主题模型及其改进方法、词嵌入模型及主题的深度学习表示等方法。如 Wu 等[12]识别不同时间上的主题构成,测度不同主题在不同时间片上的相似度关系,观测主题的演化路径;Li 等[13]通过文档主题分布测度主题强度,分析主题强度的时序演化和学科分布;Jebari 等[14]通过动态主题模型识别学科主题,分析主题的文档比重演化趋势、来源期刊分布和时序演化。

1.2 关键词语义功能识别

目前,关键词语义功能识别方法包括基于规则的方法、基于机器学习的序列标记方法和基于机器学习的分类方法等三大类。

基于规则的方法由专家制定相应的识别规则,以识别文献中特定的关键词语义功能。例如,李贺等[15]提出了基于规则的知识元识别过程,从研究问题、研究理论、研究方法以及结论四个方面构建学术论文的知识元本体。

基于机器学习的序列标记方法将关键词语义功能识别任务转化为序列标记问题,通过机器学习方法,对单词序列的每个单词进行标记,以识别关键词语义功能。例如,程齐凯等[16]提出基于条件随机场的学术文献问题和方法识别模型。

基于机器学习的分类方法将关键词语义功能识别任务转化为分类任务,应用机器学习模型构建分类器来识别关键词语义功能。例如,陆伟等[17]采用深度学习可解的标签判定策略实现关键词的语义功能判别;张国标等[18]提出一种基于多特征融合的词汇功能识别模型,在捕获关键词上下文依赖特征的同时,融合关键词的位置信息以及先验知识信息,继而采用注意力机制和前馈神经网络对关键词进行问题方法的语义功能判别;Garechana 等[19]构建伯努利朴素贝叶斯分类器识别关键词语义功能。

1.3 提示学习

近年来,随着大型预训练语言模型快速发展,提示学习充分挖掘大型预训练语言模型的潜力,通过设计契合下游任务的提示模板,所需要的训练数据显著减少,在小样本或零样本场景下达到理想的效果,归一化预训练语言模型任务,使得所有任务在方法上变得一致。提示学习在各个自然语言处理领域中得到应用,如文本分类[20]、关系抽取[21]、主题分类[22]。

综上所述,目前领域主题演化分析忽略关键词语义功能,而关键词语义功能识别需要大量标注数据,据此,本文提出一种优化的提示学习方法,使用少量样本识别关键词语义功能,并将识别的关键词语义功能用于领域主题分析,从而获得更加准确的领域主题演化分析结果。

2 研究方法

“问题”与“方法”是关键词最主要的两个语义功能。其中,“问题”表明文献研究焦点或研究对象;“方法”表明研究为解决问题采用的技术、工具、材料、手段或方案。本文主要针对关键词的“问题”与“方法”两个语义功能进行识别与分析。

2.1 基于 Bootstrapping 的关键词语义功能标注

在学术文献中,具有不同语义功能的关键词往往具有一些特有的规律,因此,本文采用 Bootstrapping 方法标注关键词语义功能,用于提示学习训练模型,以解决人工标注数据工作。Bootstrapping 以少量人工预先设定的规则获得种子标注数据为基础,然后通过增量迭代,在每一次迭代过程中,产生新的标注数据,如此循环往复,一直到最终收敛结束。具体地,主要包括如下步骤:

步骤一:初始标注规则设定。具有问题与方法语义功能的关键词在学术文献的上下文中具有一定的规律,因此,本文首先人工设置若干初始标注规则,以标

注关键词语义功能。

步骤二:关键词语义功能标注。根据标注规则,对每篇学术文献中的每个关键词,分别考察包含该关键词的该学术文献的标题和摘要中的上下文信息,采用词串匹配方法,从而标注出匹配的关键词的语义功能。

步骤三:标注规则自动生成。利用已经标注语义功能的关键词,通过包含这些关键词的上下文,寻找特定的标注规则。

在每次迭代中,需要控制标注规则的质量。如果规则能够正确识别已经存在的关键词语义功能,并能发现新的关键词语义功能,则认为该标注规则是有价值的。因此,给定一个候选标注规则,其正确标注的关键词语义功能为 n_s,其标注出的关键词语义功能为 n_e 个,若该候选标注模式满足:① $\alpha < \frac{n_s}{n_e} < \beta$;② $n_s > \delta$,其中 α,β,δ 是 预先定义的阈值,则将其设置为标注规则,用于步骤二的往复迭代。

2.2　基于提示学习的关键词语义功能识别

将关键词语义功能任务视为分类任务,给定关键词 k,以及关键词 k 所在的上下文语句 x,形成输入对 (k, x),输出关键词 k 对应的语义功能 $y \in Y = \{$问题,方法$\}$。本部分首先利用基本提示学习识别关键词语义功能,然后针对基本提示学习识别方法存在的问题,提出一种优化的提示学习方法,以提高识别关键词语义功能结果的准确性。

2.2.1　基本提示学习模型

如图 1 所示,基于基本提示学习的关键词语义功能识别包括提示添加、答案搜索和答案映射等三个主要步骤。

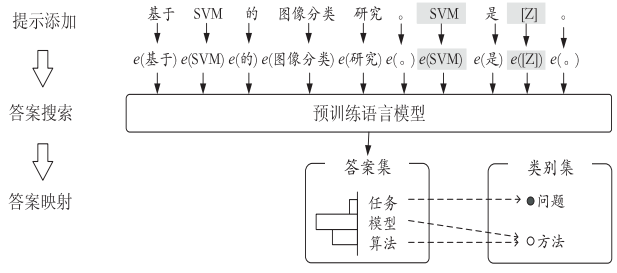


图 1　基本提示学习模型

步骤一:提示添加。通过选择合适的离散提示模板,将关键词语义功能识别问题转化为预训练语言模型的完形填空问题。离散提示模板 T 包含输入关键词 k 的输入槽[k]、输入 x 的输入槽[x]、以及包含生成答案标签 z 的答案槽[z]。使用提示函数 $f_{prompt}(\cdot)$,将关键词 k 与输入文本 x 对(k,x),输出为提示文本 $x' = f_{promt}((k,x))$。

步骤二:答案搜索。将步骤一得到的输出提示文本 x' 输入预训练语言模型,对答案槽[z]进行预测,得到一个表示答案 z 可能词的概率分布,在概率分布中选择最高概率的若干词作为候选答案。

步骤三:答案映射。将候选答案映射到输出类别。答案映射函数 f 将候选答案映射到类别标签中,分类模型基于 $P(y \mid x;\theta)$ 为

$$P(y \mid x) = \prod_j P([z]_j = f(y) \mid x') \tag{1}$$

然而,基本提示学习方法用于识别关键词语义功能存在以下主要问题:①离散提示模板构建问题:离散提示模板细微变动可能引起结果极大的不同,导致实际应用需要尝试许多不同的离散模板;②答案映射问题:由于候选答案标签和最终的分类标签可能不同,需要穷举候选答案与输出类别之间的映射。基于以上两个主要问题,本文提出优化提示学习模型。

2.2.2　优化提示学习模型

优化提示学习的关键词语义功能识别模型如图 2 所示,主要包括连续提示添加、隐藏特征生成和语义功能分类等三个主要步骤。

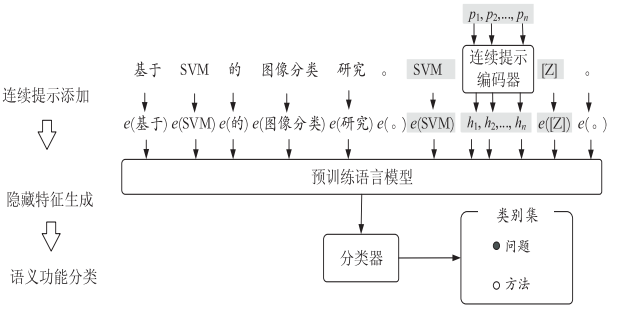


图 2　优化提示学习模型

步骤一:连续提示添加。针对基本提示学习模型存在的第一个问题,即离散提示模板构建困难,采用连续提示模板的方法。本质上,构建提示的目的是使得预训练语言模型有效地执行任务,而不是用于人类使用,因此,采用连续提示模板,使用可训练的伪标签,输入连续向量,训练参数,减少人工构建离散提示模板的成本。

具体地,每个输出提示文本 x' 包括三个部分:①输入关键词 k 和输入语句 x;②连续提示伪标签 p_0, p_1, …, p_n, 以及③一个具有 z 的离散提示。伪标签通过连续提示编码器,以根据训练数据加以调整训练参数。连续提示编码器由两层感知机 MLP 组成,每层感知机由双向长短期记忆模型 BiLSTM 和激活函数 ReLU 构成。连续提示编码器将伪标签转化为连续向量 h_1, h_2, …, h_n。将模板变成可训练参数,增强模型的泛化能力。

步骤二:隐藏特征生成。词嵌入层将 x' 映射成词矩阵,同时伪标签 p_k 通过连续提示编码器生成隐藏向量 h_k。根据伪标签信息将 h_k 插入到 x' 对应位置,通过预训练语言模型已有知识以及输入文本上下文,可以

得到答案槽[z]的隐藏特征。

步骤三:语义功能分类。针对答案映射困难,将基本提示学习的答案映射转化为分类问题。分类学习的目标是依据[z]的隐藏特征,预测关键词语义功能类别,以避免手工答案映射。本文采用一个轻量级的多层感知机 MLP 和 Softmax 层作为分类器,输出将看作模型给每个类别的概率 y'。损失函数使用交叉熵,通过最小化交叉熵误差训练模型。

$$Loss = -\sum_{i} y_i \cdot \log P(y_i' \mid x') \tag{2}$$

2.3 基于关键词语义功能的领域主题演化分析

利用已经识别语义功能的关键词,首先采用层次聚类发现问题主题与方法主题;然后,依据发现的问题主题与方法主题,分别从问题主题维度、方法主题维度以及问题与方法联合维度分析领域主题演化。

2.3.1 基于关键词语义功能的领域主题发现

将所有论文标题和摘要进行分句、分词、去除停用词及标点等预处理操作,利用 Word2Vec 词嵌入模型,计算每个关键词的向量。利用余弦计算关键词向量之间的相似度,将相似度最高的两个簇进行合并,形成新的簇。重复这个过程,直到只有一个簇为止。得到层次聚类后,通过可视化和轮廓系数法确定问题主题,使用每个主题中的高频词表述该主题的主题词,形成多层次问题主题。

同样地,基于凝聚层次聚类,以发现方法主题。

2.3.2 基于单维度关键词语义功能的领域主题演化分析

①问题主题演化分析。问题主题演化从两方面进行分析。

层次演化分析:根据问题主题层次划分,统计每个主题在特定时间阶段的研究文献的数量,研究在特定时间段的研究活跃度与研究兴趣,研究热点等信息。

战略演化分析:基于已经识别出的问题主题,使用中心度和密度两个指标进行问题主题战略演化分析。中心度用来衡量各个主题与其他主题之间的紧密程度,表示其相互影响的程度,反映研究主题的核心度。密度用来度量主题内的紧密程度,反映研究主题的成熟度。进一步地,将横坐标表示中心度,纵轴表示密度,坐标原点是中心度和密度的平均数,形成战略图,四个象限形成战略图的四个种类:a.第一象限主题具有高中心度和密度,表明对领域知识有较大贡献,因此,这类主题被称为发动机主题;b.第二象限主题具有高密度和低中心性,表明该类主题具有高度专业性和领域狭窄性,是成熟且独立的主题;称为孤立主题。c.第三象限主题具有低中心度和低密度,表明该类主题没有被充分关注和充分研究,通常表示消失或可能出现的主题,称为新兴/消失主题;d.第四象限主题具有高中心度和低密度,表明该类主题在领域中的重要性且没有被充分研究,称为基本主题,具有一定的潜能,可能成为研究热点或未来发展趋势。本文重点考察第一象限和第四象限主题。

②方法主题战略演化分析。类似地,方法主题演化分析包括方法主题层次演化分析与方法主题战略演化分析两个部分。

2.3.3 基于多维度关键词语义功能的领域主题演化分析

结合问题主题和方法主题进行联合演化分析,主要进行以下两个研究:

问题方法关联规则演化分析:通过支持度、置信度和提升度挖掘问题主题与方法主题之间的关联关系。其中,支持度(*Support*)表明问题主题 A 与方法主题 B 的频次,支持度高表明该关联规则内 A 和 B 同时出现频次高。支持度公式为:

$$Support(A \Rightarrow B) = P(A \cup B) \tag{3}$$

置信度(*Confidence*):A 发生的前提下 B 发生的条件概率,置信度高说明出现 A 后很可能出现 B,关联规则更可信。置信度公式为:

$$Confidence(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A)} = P(B \mid A) \tag{4}$$

提升度(Lift):描述挖掘出规则的可用性,可以度量此规则的可用程度,即 A 和 B 的置信度与后项 B 的支持度之比,表示 B 的出现对 A 出现的影响程度。提升度公式为:

$$Lift(A \Rightarrow B) = \frac{Confidence(A \Rightarrow B)}{P(B)} = \frac{P(A \cup B)}{P(A)P(B)} \tag{5}$$

满足最小支持度和最小置信度的规则,称为强关联规则。而强关联规则又分为有效强关联规则与无效强关联规则。具体划分为:

若 Lift(A⇒B)>1,则规则 A⇒B 是有效强关联规则,表明 A 与 B 是正相关,表明 A 与 B 实际同时发生的概率大于 A 与 B 独立发生的概率;

若 Lift(A⇒B)≤1,则规则 A⇒B 是无效强关联规则,表明 A 与 B 是负相关;

若 Lift(A⇒B)= 1,则表示 A 与 B 相互独立。

方法应用广度演化分析:使用信息熵计算方法在各问题主题的应用广度。方法主题 X 的信息熵 H 的计算公式为:

$$H(X) = -\sum_{K=1}^{N} P_k \log(P_k) \tag{6}$$

其中,P 表示概率,X 表示信息熵计算集合。N 表示问题类别数目,P_k 表示类别 k 占比。

3　实　验

3.1　实验数据

为了验证本文提出的关键词语义功能识别方法及其在领域主题演化分析中的有效性,本研究选取图书情报领域作为实证分析研究领域,以期获得我国图书情报领域中基于关键词语义功能的主题演化脉络与发展趋势。

具体地,选取中国知网硕博论文数据库中的硕士论文,通过“文献分类目录”中“信息科技”下的“图书情报与数字图书馆”选项对学科标签进行初筛,然后通过“分组浏览”选定“科学专业”,选择“情报学”“图书馆学”“图书、情报与档案管理”等目标学科标签,获得每篇论文的标题、摘要、关键词、年份等信息。删除重复数据,共得到 5470 条图书情报领域的硕士论文数据。同时由于中国知网硕博论文数据库中 2022 年与 2023 年论文数据未完整收录,最终得到中国知网硕博论文数据库收录的 2001—2021 年图书情报领域硕士论文数据 5188 条,其中,数据中共包含 22 775 个关键词(不同硕士论文关键词可重复),平均每篇硕士论文的关键词个数为 4.39 个。

3.2　实验步骤及评估指标

基于下载的硕士论文数据,首先进行 Bootstrapping 的关键词语义功能标注。具体地,首先确定初始模板,如表 1 所示的标注规则,进行精确匹配,其中符号“___”表示关键词所在位置,[*]表示任意长度的字符串,基于确定的标注规则,标注部分关键词的问题与方法语义功能作为种子。接着,基于提示学习的关键词语义功能识别,再识别出标注候选标注规则,然后对其候选标注规则进行过滤,本文设置 $\alpha = 0.6$,$\beta = 0.8$,$\delta = 2$,得到标注规则,进行迭代,一直到没有新的标注规则产生为止。将 Bootstrapping 标注的关键词语义功能作为训练集,训练模型,并使用没有标注的数据作为测试集进行测试。

表 1　初始标注规则

类型	标注规则	实例	关键词
方法	基于___的[*]研究	基于神经网络的专利技术机会分析研究	神经网络
问题	基于[*]的___研究	基于神经网络的专利技术机会分析研究	专利技术机会分析

接着,进行基于提示学习的关键词语义功能识别。具体地,伪标记定义连续提示模板,提示编码器由两层感知机 MLP 组成。每层感知机由双向长短期记忆模型 BiLSTM 和激活函数 ReLU 构成。使用 SciBERT 作为预训练语言模型。接着,将提示模板应用于训练数据之后,可以得到所有标注关键词对应的[z]的隐藏特征,预测关键词语义功能类别。训练模型参数为:Epoch = 10, Dropout = 0.2, Batch_size = 32,激活函数 = ReLu,学习率 = 0.002,全连接神经元个数 100 个。

对提示学习模型进行评估,采用人工标注测试集的方法,邀请 3 位图情领域的研究生进行标注,使用两两交集作为正确结果。并对人工标注结果使用 Kappa 进行评测,Kappa 得分均大于 0.8,当 Kappa 值超过 0.8 时,则认为数据集标注是有效的。使用 P、R 和 F1 评估提示学习方法

最后,根据已经识别语义功能的关键词信息进行领域主题演化分析。将时间分为四个时间段:2001-2005 年(早期)、2006-2010 年(中期)、2011-2015 年(后期)、2016-2021 年(近期)。基于层次聚类方法分别发现问题与方法主题,使用 Word2Vec 模型训练以得到词嵌入,词嵌入的维度设为 100;使用关键词嵌入的余弦计算两个关键词间的相似性,利用 Ward 方法进行层次聚类。

3.3　实验结果

3.3.1　基于 Bootstrapping 的关键词语义功能标注

本文初始人工标注规则以“基于____的[*]研究”标注具有方法语义功能的关键词,以“基于[*]的____研究”标注规则标注具有问题语义功能的关键词。结果,使用该标注规则,共标注 103 各具有问题语义功能的关键词,111 个具有语义功能的语义关键词。接着,基于 Bootstrapping 标注关键词语义功能角色,共标注 2 451 个具有问题语义功能的关键词和 1 724 个具有方法语义功能的关键词,通过人工检查,问题标注的准确率为 98.26%,方法标注的准确率为 96.89%,结果表明基于 Bootstrapping 的关键词语义功能标注具有较高的准确率,但是显然,由于语言表述的多样性,也必定具有低的召回率。表 2 为部分基于 Bootstrapping 方法标注的关键词语义功能实例。

3.3.2　基于提示学习的关键词语义功能识别

将 Bootstrapping 标注结果作为提示学习的训练集,如表 3 所示。由表 3 可见,与传统的监督学习相比,模型学习的训练数据明显少于测试训练数据。

为了评估提示学习识别关键词语义功能的效果,实验将本文提出的方法与其他方法进行比较。欲比较的方法如下:

第一组:离散提示学习

P_Dis1:使用表 4 中的 T1 离散提示,预训练语言模型使用 SciBERT,具体参见 3.2.1 节所述。

P_Dis2:使用表 4 中的 T2 离散提示模板。其他与 P_Dis1 相同。

表 2 基于 Bootstrapping 的关键词语义功能标注实例

类型	标注规则	实例	
		语句	标注关键词
问题	面向____的____研究	面向开放科学的机构知识库内容建设研究	开放科学 机构知识库内容建设
	____研究——以____为例	高校阅读推广活动研究——以高校图书馆阅读推广活动为例	高校阅读推广活动 高校图书馆阅读推广活动
	视角下____研究	用户视角下在线健康社区医生画像研究	在线健康社区医生用户画像
方法	融合____	融合 Altmetrics 的期刊影响力综合评价研究	Almetrics
	引入____	引用网络结构的热点识别方法研究	网络结构
	结合____与____	结合本体与社会化标签的用户动态兴趣建模研究	本体 社会化标签

表 3 实验数据统计

类别	训练数	测试数	总数
问题	2451	8061	10512
方法	1724	5732	7456

P_Dis3:使用表 4 中的 T3 离散提示模板。其他与 P_Dis1 相同。

P_Dis3_BERT:使用表 1 中的 T3 离散提示模板。预训练模型使用 BERT,其他与 P_Dis3 相同。

第二组:连续提示学习

P_Con1_Ver:使用 1 个伪标签进行连续提示学习,最后仍然是答案映射方法。

P_Con2_Ver:使用 2 个伪标签进行连续提示学习,其他与 P_Con1_Ver 相同。

P_Con3_Ver:使用 3 个伪标签进行连续提示学习,其他与 P_Con1_Ver 相同。

P_Con3_Cla:使用 3 个伪标签进行连续提示学习,使用分类优化,具体参见 3.2.2 描述。

第三组:序列标注:将关键词语义功能标识问题转化为序列标注任务,采用典型的模型。

CRF:设计了包含词法、句法、组块等多种特征,结合条件随机场构建针对学术文献研究问题与研究方法的关键词语义功能识别模型。

第四组:分类:将关键词语义功能识别问题转化为分类任务,采用典型的模型。鉴于深度学习方法在多类任务中的优异表现,通过采用深度学习可解的标签判定策略实现关键词的语义功能判别。

BERT_LSTM:利用 BERT 及 LSTM 方法构建分类模型,对关键词所承载的问题和方法语义功能进行了分类。

Bert+TextCNN:利用 BERT 及 TxtCNN 方法构建分类模型,对关键词所承载的问题和方法语义功能进行了分类。

实验结果如表 4 所示。可以发现:

在离散提示学习组中,P_Dis1、P_Dis2、P_Dis3 的 P、R 和 F1 存在较大波动,如,P_Dis1 的 F1 值比 P_Dis2 高 2.04%、比 P_Dis3 的 F1 值高 4.14%,表明离散提示学习中离散提示模板的选择对结果性能会产生较大影响,存在不稳定性。

表 4 离散提示模板

模板名称	离散提示模板
T1	[x]。[k]是[mask]。
T2	[x]。其中,[k]是[MASK]。
T3	[k]是[x]中的[MASK]。

在离散提示学习组中,选择最好的 P_Dis1 更换其预训练语言模型 SciBERT 为 BERT,变为 P_Dis1_BERT。实验结果表明,P_Dis1 的 P、R 和 F1 值比 P_Dis1_BERT 高 5.46%、4.94%和 5.19%。SciBERT 使用科学文献训练预训练语言模型,包含更多科学研究信息,BERT 使用维基百科等普通文本训练预训练语言模型,因此 SciBERT 比 BERT 包含更多科学研究信息,从而获得了更好的识别效果。

对优化提示学习组中的 P_Con1_Ver、P_Con2_Ver、P_Con3_Ver 进行比较,分析不同伪标签的数量对性能的影响。实验结果发现,P_Con1_Ver、P_Con2_Ver、P_Con3_Ver 中,P_Con3 模型性能最好,P_Con3 的 P、R 和 F1 值比 P_Con1 高 1.14%、2.01%和 1.58%,比 P_Con2 的 P、R 和 F1 值高 0.4%、1.87%和 1.15%。三个模型主要区别为连续提示中伪标签的数目,表明伪标签数目为 3 较好。实验结果表明更多的为标签提供更多的可训练参数,增强模型的能力,但是需要更多的数据训练,因此,最终选择 N=3 作为伪标签数量。

在优化提示学习组中,P_Con3_Ver 和 P_Con3_Cla 比较中,P_Con3_Cla 比 P_Con3_Ver 在 P、R 和 F1 值高 7.5%、5.35%和 6.40%。P_Con3_Cla 更好,实验结果表明将其变为分类问题,避免了答案映射的人工过程的繁琐和遗漏,能够显著提高方法的准确性。

与常见的监督学习方法比较,P_Con3_Cla、CRF、BERT+LSTM、BERT+TxtCNN 相比,实验结果表明提示学习结果最好。P_Con3_Clas 比 CRF 的 P 、R 和 F1 值提高 20.44%、14.82%和 17.61%,比 BERT+LSTM 的

P 、R、F1 提高 12.74%、14.02%和 13.42%,比 BERT+TxtCNN 的 P、R 和 F1 提高 10.74%、9.44%和 10.07%。主要的原因在于在标记数据较少的情况下,使用传统的监督学习方法,包括序列标注和分类学习需要较多的标注数据,而提示学习利用预训练模型本身的知识,可在少量标注数据的情况下也获得良好的结果。

表 5　关键词语义功能识别方法比较结果

组	方法	P(%)	R(%)	F1(%)
基本提示学习	P_Dis1	86.43	84.32	85.36
	P_Dis2	84.29	80.44	83.32
	P_Dis3	82.34	80.13	81.22
优化提示学习	P_Dis1_BERT	80.97	79.38	80.17
	P_Con1_Ver	84.23	81.97	83.08
	P_Con2_Ver	84.97	82.11	83.52
	P_Con3_Ver	85.37	83.98	84.67
	P_Con3_Cla	92.87	89.33	91.07
序列标注	CRF	72.43	74.51	73.46
分类	BERT	80.12	75.31	77.65
	BERT+TextCNN	82.13	79.89	80.99

3.3.3　基于关键词语义功能的领域主题演化分析

①基于层次聚类的主题发现。问题主题聚类数目为 2 和 8 时时轮廓系数相对较高。因此,首先将问题主题划分为两大类,称为一级主题;接着,对划分好的一级主题,再进行划分,得到 8 个类别,称之为二级主题。根据聚类结果,每个聚类形成一个问题主题,依据每个问题主题主要包含的关键词,形成各级主题的名称,如表 6 所示。

类似地,方法主题在聚类数目 5 和 10 处存在相对较高轮廓系数,从而得到方法一级主题与方法二级主题。相应地形成如表 7 所示的方法主题。

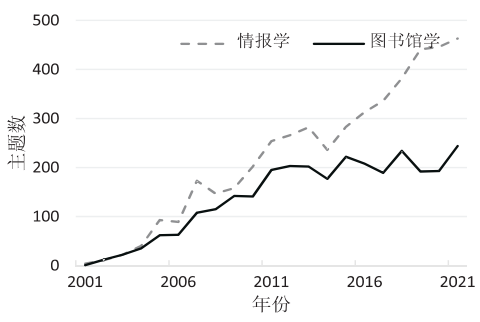
②问题主题演化分析。问题主题层次演化分析:图 3(a)为一级问题主题包含的“图书馆学”“情报学”的演化。“情报学”问题研究略多于“图书馆学”,特别是“情报学”随时间变迁研究的增长更为明显。其主要原因可能是情报学信息源更加多样化,实际应用场景更为丰富。图 3(b)为“图书馆学”问题主题演化,“图书馆学”包括“图书馆建设与管理”与“图书馆服务”两个二级问题主题,其中,“图书馆建设与管理”一直占据主体。图 3(c)“情报学”问题主题演化图。“情报学”中“信息分析”和“信息服务”占据“情报学”研究一直占据研究主体。“信息获取”“信息获取”则处于第二梯度,而“互联网”起步较晚,大约在 2010 年时期开始,增长较快,目前也处于第二梯度。而“信息素养”研究发展较晚,研究总体占比较少,处于第三梯队。

表 6　问题主题

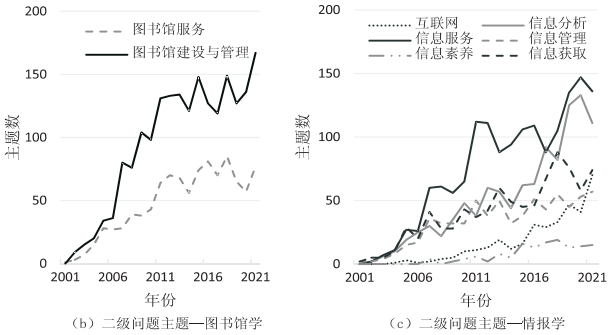
一级主题	二级主题	主题词
图书馆	图书馆建设与管理	公共图书馆服务、公共借阅权、非传统公共图书馆
		高校图书馆、高校智库、高校图书馆网站
		图书馆联盟、图书馆合作、馆藏资源共享
		数字图书馆、智慧图书馆、数字馆藏资源
	图书馆服务	图书馆服务、图书馆+ 、图书馆应用
		学科服务、学科信息门户、学科信息资源整合
		阅读推广、全民阅读 社会化阅读
		公共文化服务、文化信息服务、红色文化资源
		数字参考咨询服务、数字参考咨询系统、
		合作数字参考咨询
情报	信息获取	数据开放存取、开放数据、科研数据共享
		信息检索、搜索引擎、个性化推荐
		信息需求、信息搜寻、信息采纳
	信息管理	数据库建设、元数据、关联数据
		网络信息资源、文献信息资源、信息组织
		共享模式、共建模式、跨界合作服务
	信息分析	网络舆情、社交网络信息、网络学术信息
		用户需求、用户画像、用户满意度
		主题演化、文本分析、文献语义挖掘
	信息服务	个性化信息服务、在线健康信息、健康信息服务
		政府信息公开、政府数据开放平台、政府信息获取
		知识发现、知识共享、知识服务
	互联网	情报服务、情报分析、企业竞争情报
		微信平台、问答社区、科研社交平台
		Web2.0、云技术、新媒体
	信息素养	信息素养、信息素养教育、信息素养评价

表 7　方法主题

一级方法主题	二级方法主题	主题词
逻辑思维	通用思维	归纳法、比较分析、类比分析
	系统思维	逻辑建模、SWOT 法、PEST 法
情报获取	人际交互	问卷调查、专家咨询、访谈
	人机交互	线上检索、网络爬虫
情报分析	数值分析	社会网络分析、数理实证、统计分析
	文献计量	引文网络、文献计量 、专利分析
	大数据技术	云计算、关联数据、5G 技术
	人工智能技术	机器学习、神经网络、自然语言处理
情报展示	情报展示	数据可视化、技术路线图、Citespace
其他	其他	眼动分析、RFID、系统动力学分析



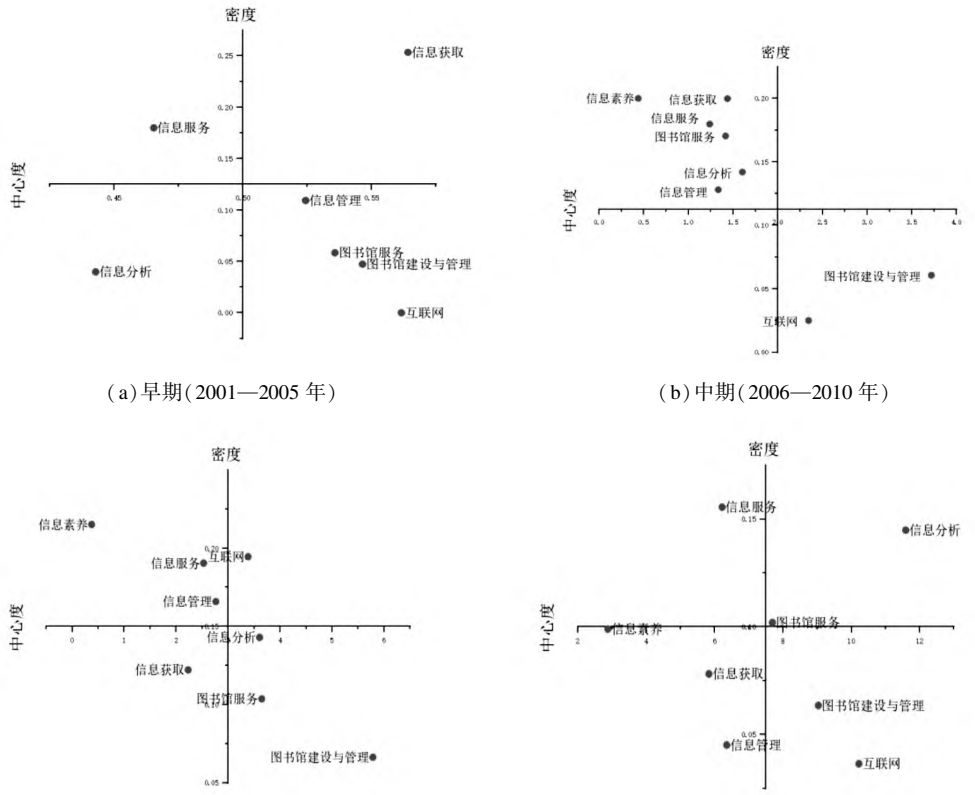
(a) 一级问题主题



（b）二级问题主题—图书馆学　（c）二级问题主题—情报学

图 3　问题主题层次演化分析

问题主题战略演化分析：各时间段问题主题战略图如图 4 所示，本文着重考察第一象限和第四象限主题。图 4(a)早期“信息获取”位于第一象限，表明该主题具有广泛影响力和深入研究；“信息管理”“图书馆服务”“图书馆建设与管理”“互联网”位于第四象限，表明这些主题具有广泛影响力，但缺乏深入研究。图 4(b)中期，“图书馆建设与管理”和“互联网”位于第四象限，表明这些问题主题具有广泛影响力，但缺乏深入研究。图 4(c)后期，“互联网”位于第一象限，表明该主题具有广泛影响力和深入研究；“信息分析”“图书馆服务”“图书馆建设与管理”位于第四象限，表明这些问题主题具有广泛的影响力，却缺乏深入的研究。图 4(d)近期，“信息分析”和“图书馆服务”位于第一象限；“图书馆建设与管理”与“互联网”位于第四象限，表明这些主题具有广泛影响力，但缺乏深入研究。

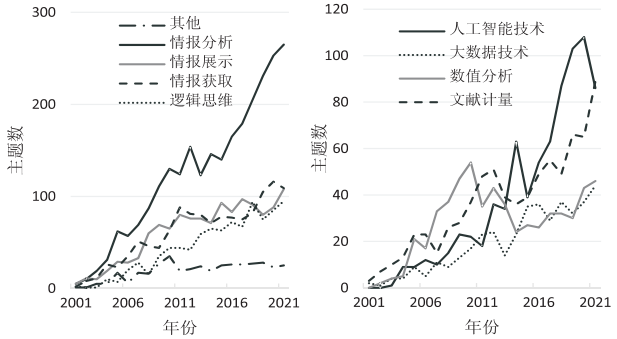


(a)早期(2001—2005 年)　(b)中期(2006—2010 年)

(c)后期(2011—2015 年)　(d)近期(2016—2021 年)

图 4　问题主题战略图

③方法主题演化分析。方法主题层次演化分析：图 5(a)为一级方法主题演化图，“情报分析”一直占据方法的主要部分，“逻辑思维”“情报获取”“情报展示”研究数量一直相当。特别地，对“情报分析”进行进一步分析，图 5(b)为二级方法“情报分析”的主题演化。可以看出，早期“数值分析”和“文献分析”方法占据主体，“大数据技术”和“人工智能技术”很少；到了中期，“数值分析”和“文献分析”仍然占据较大份额，但“大数据技术”和“人工智能技术”已开始增长，特别是“人工智能技术”在 2015 年左右出现较大增长，到了近十年，“数值分析”的份额已明显减少，“人工智能技术”成为图情领域最主要的一种方法。

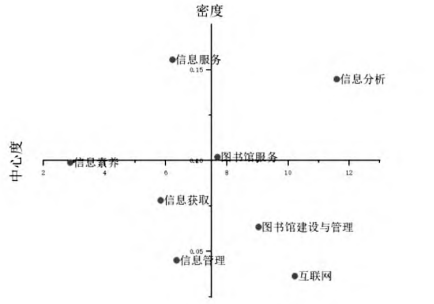


（a）一级方法主题演化　（b）二级方法主题-情报分析

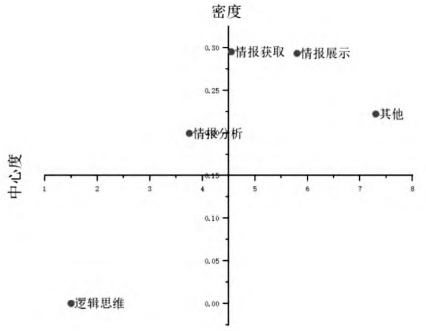
图 5　方法主题层次演化分析

方法主题战略演化分析：图 6 为二级方法主题战略图。早期阶段，其中，“情报获取”“情报展示”“其

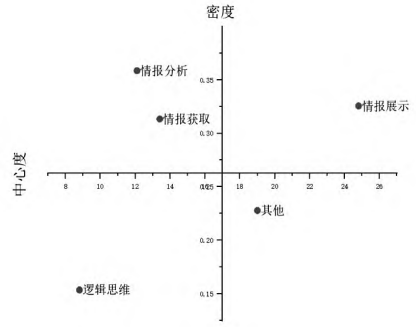
他”位于第一象限,表明这些方法主题具有较大的影响力和较多的研究。中期,“情报展示”位于第一象限,表明该方法主题具有较大影响力和较多研究;后期“情报展示”位于第一象限,表明该类方法具有较大影响力和深入研究;近期,“数据分析”位于第一象限,成为具有较大影响力且深入研究的方法主题,而“情报展示”位于第四象限,表明该类方法主题具有广泛的影响力,却缺乏深入的研究。



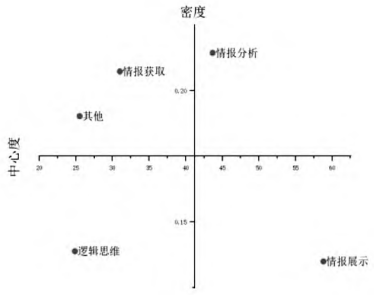
(a)早期(2001—2005 年)



(b)中期(2006—2010 年)



(c)后期(2011—2015 年)



(d)近期(2016—2021 年)

图 6　方法主题战略图

④问题-方法主题演化分析。问题关联方法演化分析:表 8 为问题采用方法挖掘的有效强规则。“高校图书馆管理”在早期使用“数值分析”方法,后期则使用“通用思维”和“文献分析”方法,近期使用“通用思维”。“公共图书馆”在早期使用“人机交互”方法。“数字图书馆建设”早期使用“情报展示”,中后期主要使用“数值分析”,“人机交互”,“情报展示”,“通用思维”,“文献分析”,而近期则集中于“人机交互”和“通用思维”。“图书馆服务”早期主要集中于“数值分析”方法,中期主要集中关于“文

表 7　问题与方法主题有效强关联规则

一级问题主题	二级问题主题	年份	二级方法主题
图书馆	高校图书馆管理	早期(2001—2005 年)	数值分析
		后期(2011—2015 年)	通用思维、文献分析
		近期(2016—2021 年)	通用思维
	数字图书馆建设	早期(2001—2005 年)	情报展示
		中期(2006—2010 年)	人机交互、情报展示
		后期(2011—2015 年)	数值分析、文献计量、人工智能技术
		近期(2016—2021 年)	人机交互、通用思维
	图书馆服务	早期(2001—2005 年)	数值分析
		中期(2006—2010 年)	文献分析
情报	信息获取	早期(2001—2005 年)	人际交互、文献计量、情报展示
		中期(2006—2010 年)	人际交互
		后期(2011—2015 年)	人际交互、情报展示、通用思维
		近期(2016—2021 年)	人际交互、数值分析、人工智能技术
	信息分析	早期(2001—2005 年)	数值分析、文献计量
		中期(2006—2010 年)	数值分析
		后期(2011—2015 年)	数值分析、人工智能技术、文献计量
		近期(2016—2021 年)	数值分析、情报展示、通用思维、文献计量、人工智能技术
	信息服务	早期(2001—2005 年)	数值分析、情报展示
		后期(2011—2015 年)	情报展示
		近期(2016—2021 年)	情报展示、文献计量

献分析”方法。

情报问题主题的有效强关联规则中,“信息获取”主题中,一直使用“人际交互”的方法,“信息分析”中最常用的方法是“数值分析”与“文献计量分析”。近期信息分析方法日趋多元化,且更多使用“人工智能技术”方法。“信息服务”中,“情报展示”是一类重要的分析方法。

方法应用广度演化分析:表 8 为根据信息熵值计算得到的具有较高信息熵的方法,表明该方法可以应用于多种种类主题,具有较大的应用广度。在“情报分析”一级方法主题中,“数值分析”和“文献计量”方法随时间增长,特别是“大数据技术”方法增长很多,表明大数据技术中很多方法在图书情报领域得到广泛的应用。结合前面的分析可知,“人工智能技术”将是下一个方法研究热点,目前来看方法还比较少,值得后续进一步关注和研究。

表 8 方法应用广度

一级方法主题	二级方法主题	年份	具体方法
逻辑思维	通用思维	早期(2001—2005 年)	知识组织、元数据
		中期(2006—2010 年)	知识组织、元数据
		后期(2011—2015 年)	知识组织、元数据、比较分析
		近期(2016—2021 年)	知识管理、元数据、比较分析
	系统思维	早期(2001—2005 年)	层次分析法、用户兴趣模型、评价体系
		中期(2006—2010 年)	层次分析法、评价体系
		后期(2011—2015 年)	评价体系、层次分析法、扎根理论
		近期(2016—2021 年)	评价指标、层次分析法、扎根理论
信息收集	人际交互	早期(2001—2005 年)	
		中期(2006—2010 年)	问卷调查
		后期(2011—2015 年)	问卷调查
		近期(2016—2021 年)	问卷调查
	人机交互	早期(2001—2005 年)	信息检索、搜索引擎
		中期(2006—2010 年)	信息检索
		后期(2011—2015 年)	信息检索、搜索引擎
		近期(2016—2021 年)	信息检索、搜索引擎
情报分析	数值分析	早期(2001—2005 年)	结构方程模型、因子分析
		中期(2006—2010 年)	结构方程模型、因子分析、KANO 模型、社会网络分析、UTAUT 模型
		后期(2011—2015 年)	结构方程模型、因子模型、KANO 模型、社会网络分析
		近期(2016—2021 年)	结构方程模型、因子分析、KANO 模型、UTAUT 模型、社会网络分析
	文献计量	早期(2001—2005 年)	领域本体、文献计量、引文分析、共词分析
		中期(2006—2010 年)	文献计量、共词分析、Altmetrics、引文分析、领域本体
		后期(2011—2015 年)	共词分析、Altmetrics、引文分析、领域本体
		近期(2016—2021 年)	Altmetrics、引文分析、领域本体、共词分析
	大数据技术	早期(2001—2005 年)	知识库、XML
		中期(2006—2010 年)	知识库、XML、云计算、Web2.0、语义网、聚类分析、
		后期(2011—2015 年)	知识库、云计算、Web2.0、聚类分析、LDA 模型、关联数据
		近期(2016—2021 年)	知识库、XML、云计算、语义网、聚类分析、LDA 模型、关联数据、用户画像、情感分析
	人工智能技术	早期(2001—2005 年)	
		中期(2006—2010 年)	文本挖掘
		后期(2011—2015 年)	文本挖掘、机器学习、深度学习、知识图谱
		近期(2016—2021 年)	文本挖掘、机器学习、深度学习、知识图谱
情报展示	情报展示	早期(2001—2005 年)	
		中期(2006—2010 年)	可视化分析
		后期(2011—2015 年)	可视化分析
		近期(2016—2021 年)	可视化分析

4 结 语

本文利用预训练语言模型,提出一种优化提示学习方法,以使用少量标注样本识别关键词语义功能,进而从关键词语义功能的各个维度分析领域主题的演化。通过我国 20 年的图书情报领域硕士论文数据,实证检验本文提出方法的有效性。实验表明,本文提出的优化提示学习方法采用少量标注数据能够准确识别关键词语义功能,比传统方法更具优势。基于关键词语义功能的领域主题演化分析使得分析结果更加清晰准确。

然而,本研究依然不够完善,主要存在以下不足:①仅从问题与方法两种语义功能类型对关键词进行划分,可能忽略了关键词所蕴含的更细粒度的语义信息;②实证分析所用数据来源于知网,缺乏对其他多个论文信息数据库的关注,可能存在数据完整性难以保证的情况。未来的研究可以关注于关键词更多语义信息的挖掘,得到更细粒度的领域主题划分结果;同时可以

进一步考虑纳入论文正文等非结构化的信息,从中拓展关键词,并将其作为扩充以完善论文主题的构成。

参 考 文 献

[1] 程齐凯, 李鹏程, 张国标, 等. 学术文本词汇功能识别——基于标题生成策略和注意力机制的问题方法抽取[J]. 情报学报, 2021,40(1): 43-52.

[2] 奉国和, 孔泳欣. 基于时间加权关键词词频分析的学科热点研究[J]. 情报学报, 2020,39(1): 100-110.

[3] 王　康, 陈　悦, 苏　成, 等. 多维视角下科学主题演化分析框架[J]. 情报学报, 2021,40(3): 297-307.

[4] 王　伟, 杨建林. 基于引文网络重叠社团发现的图书情报领域学科主题结构分析[J]. 情报学报, 2020,39(10): 1021-1033.

[5] Liu T, Tang L. Open innovation from the perspective of network embedding: knowledge evolution and development trend[J]. Scientometrics, 2020, 124(2): 1053-1080.

[6] Xu S, Hao L, An X, et al. Review on emerging research topics with key-route main path analysis[J]. Scientometrics, 2020, 122(1): 607-624.

[7] Dejian Y, Tianxing P. Tracing knowledge diffusion of TOPSIS: A historical perspective from citation network[J]. Expert Systems with Applications, 2021, 168(1): 1-12.

[8] 吴胜男, 卫慧蓉, 于　琦, 等. 结构-内容视角下的学科领域主题演化分析——以肺癌靶向药物领域为例[J]. 信息资源管理学报, 2020,10(5): 112-121.

[9] Wang X, Wang H, Huang H. Evolutionary exploration and comparative analysis of the research topic networks in information disciplines[J]. Scientometrics, 2021,126(6): 1-27.

[10] 王　康, 陈　悦, 苏　成, 等. 多维视角下科学主题演化分析框架[J]. 情报学报, 2021,40(3): 297-307.

[11] Duan Y, Guan Q. Predicting potential knowledge convergence of solar energy: Bibliometric analysis based on link prediction model[J]. Scientometrics, 2021,126(5): 1-25.

[12] Wu H, Yi H, Li C. An integrated approach for detecting and quantifying the topic evolutions of patent technology: A case study on graphene field[J]. Scientometrics, 2021,126(8): 1-21.

[13] Li Y, Chen Y, Wang G Q. Evolution and diffusion of information literacy topics[J]. Scientometrics, 2021,126(5): 1-30.

[14] Jebari C, Herrera-Viedma E, Cobo M J. The use of citation context to detect the evolution of research topics: A large-scale analysis[J]. Scientometrics, 2021,126(4): 1-19.

[15] 李　贺, 杜杏叶. 基于知识元的学术论文内容创新性智能化评价研究[J]. 图书情报工作, 2020,64(1): 93-104.

[16] 程齐凯, 李　信. 面向语义出版的学术文本词汇语义功能自动识别[J]. 数字图书馆论坛, 2017(8): 24-31.

[17] 陆　伟, 李鹏程, 张国标, 等. 学术文本词汇功能识别——基于 BERT 向量化表示的关键词自动分类研究[J]. 情报学报, 2020,39(12): 1320-1329.

[18] 张国标, 李鹏程, 陆　伟, 等. 多特征融合的关键词语义功能识别研究[J]. 图书情报工作, 2021,65(9): 89-96.

[19] Garechana G, RÍO-Belver R, Zarrabeitia E, et al. TeknoAssistant : A domain specific tech mining approach for technical problem-solving support[J]. Scientometrics, 2022, 127(9): 5459-5473.

[20] Shiwen N, Hungyu K. KPT++: Refined knowledgeable prompt tuning for few-shot text classification[J]. Knowledge-Based Systems,2023,74(1): 1-9.

[21] Di Z, Yumeng Y, Peng C, et al. Biomedical document relation extraction with prompt learning and KNN.[J]. Journal of biomedical informatics, 2023,145(1): 1-8.

[22] Zhang Z, Liu S, Cheng J. Exploring prompts in few-shot cross-linguistic topic classification scenarios[J]. Applied Sciences, 2023,13(17): 1-15.

(责编:王平军;校对:王菊)

(上接第 186 页)

[11] Pavon J M , Previll L ,Myung Woo, et al. Machine learning functional impairment classification with electronic health record data[J].Journal of the American Geriatrics Society, 2023, 71(9):2822-2833.

[12] Pradhan M R, Mago B, Ateeq K. A classification-based sensor data processing method for the internet of things assimilated wearable sensor technology[J]. Cluster Computing, 2023, 26(1): 807-822.

[13] Zhou J, Li H, Wang C, et al. Optimizing unbalanced text classification tasks by integrating critical data mining and restricted rewriting techniques[J]. Concurrency and Computation: Practice and Experience, 2020, 32(24): e5952.

[14] 郭　晶,吴应辉,谷　陵,等.国际中文教育数字资源建设现状与展望[J].国际汉语教学研究, 2021(4): 86-96.

[15] 彭玉芳,陈将浩,何志强.基于机器学习和深度学习的南海证据性数据抽取算法比较与应用 [J].现代情报,2022,42(2): 55-69.

[16] 彭玉芳,石　进,徐　浩,等.基于 BERT 和分面分类的多标签的南海证据性数据分类研究[J].图书馆杂志,2022,41(5): 102-108.

[17] Alhuzali H, Ananiadou S. SpanEmo: Casting multi-label emotion classification as span-prediction[J]. ArXiv Preprint ArXiv, 2021:2101.10038.

(责编:王育英;校对:刘影梅)