基于知识元的科学-技术知识关联指标与测度方法研究*

唐晓波[1,2]　陈俭静[2]　周禾深[2]　杜鑫[2]
（1. 武汉大学信息系统研究中心，武汉 430072；2. 武汉大学信息管理学院，武汉 430072）

摘要：通过科学-技术知识关联指标与测度方法研究，能够细粒度分析科学与技术的互动关系，为科技评价奠定基础。提出一种基于知识元的科学-技术知识关联指标与测度方法。以论文和专利为数据来源，在知识元抽取基础上，基于科技术语在不同类型知识元中的共现情况实现科学-技术知识关联指标测度。以糖尿病领域为例开展实证研究，结果验证提出的科学-技术知识关联指标与测度方法能够在高质量论文识别中有效发挥作用，所提指标和方法对于完善科技评价体系、促进创新驱动发展战略实施有参考意义。

关键词：知识元；知识关联；科学-技术关联；指标；测度方法
中图分类号：G301　DOI：10.3772/j.issn.1673-2286.2024.02.006
引文格式：唐晓波，陈俭静，周禾深，等. 基于知识元的科学-技术知识关联指标与测度方法研究[J]. 数字图书馆论坛，2024，20（2）：58-69.

科学技术的发展是人类社会文明进步的重要标志。习近平总书记在党的二十大报告中强调：加快实施创新驱动发展战略，加快实现高水平科技自立自强，以国家战略需求为导向，集聚力量进行原创性引领性科技攻关，坚决打赢关键核心技术攻坚战[1]。在“大科学”的时代背景下，知识呈现出网络化交织的特点，科学知识与技术知识在交互共进中孕育科技创新，为生产力的发展与文明的进步奠定坚实的基础。随着科学与技术关系的日益密切，科学技术互动关系在科技评价中逐渐引起学界的重视。本文探索基于知识元的科学-技术知识关联指标与测度方法：以论文和专利为数据来源，在知识元层面抽取科技术语，基于共词关系测度科学-技术知识关联。通过聚焦科学-技术知识关联在高质量论文识别中的作用，检验所提知识关联测度方法的有效性。

1　相关研究

1.1　知识元抽取

知识元是组成知识的基本单位和结构要素。知识元作为知识的基本组分，是知识在微观领域的存在形态[2]。在对知识基本组分的研究中，Brookes[3]的认知地图思想、Popper[4]的知识进化思想有着重要的理论影响。温有奎等[5]指出，在知识元的概念内涵界定中，最小颗粒度、独立性、完备性等元素是必备条件。基于知识元的特性，文庭孝等[6]将知识元构建的现实意义概括为知识的自由切分与存取、知识的自由组织与检索、知识的自由组合与创造、知识的准确计量与评价。在情报学研究中，知识元模型被引入知识组织理论。对文献知识元的内容、功能等的挖掘有助于建立文献资源间的语

收稿日期：2023-12-28
*本研究得到国家社会科学基金重大项目“基于大数据的科教评价信息云平台构建和智能服务研究”（编号：19ZDA349）资助。

义关联，为知识管理与创新提供有效的途径[7]。此外，在知识工程领域，知识元在知识表示[8]、知识标引[9]和语义关联抽取[10]等任务中的运用得到学者们的重视。

知识元的抽取方法主要分为3类：基于规则的方法、基于统计的机器学习方法和基于深度学习的方法。基于规则的方法耗时耗力，有限的规则难以应对自然语言表达的灵活性。叶光辉等[11]提出基于“人工标注—规则归纳—机器识别—规则补充”流程的知识元抽取方法，较为准确地从学术文献中提取知识元文本片段。基于统计的机器学习方法往往需要领域专家构建稳定的特征模板[12]，且在小规模语料下该方法的鲁棒性不足。王忠义等[13]根据科技文献资源中方法知识元的特征制定初始描述规则，并基于PreFixSpan算法实现方法知识元及其描述规则的动态更新，抽取结果的准确率、召回率和F1值分别达到0.71、0.80和0.73。深度学习作为当下的热门技术，在知识元抽取任务中表现优秀。余丽等[14]基于多类型知识元的标注语料库训练LSTM-CRF模型，自动抽取的平均正确率达到91%。

1.2　知识关联

知识关联是指知识与知识之间的联系。揭示和利用知识关联是知识组织、知识管理、知识发现和知识创造的起点[15]。高继平等[16]认为，知识元之间存在着隶属、交叉、共现、引用、共被引、耦合等多种类型的关联，将知识关联应用于知识计量学、知识网络等方面的分析有助于实现知识的增殖。洪亮等[17]指出，知识关联的发现、认知和描述是一个从隐性到显性的演化过程，知识关联具有可描述、可计算、可演化的特征。知识关联广度和深度的变化构成了知识系统的自组织自演化过程。

国内外研究表明，知识关联在揭示知识联系、挖掘知识价值等方面有着广阔的应用空间。唐旭丽等[18]基于知识关联视角开展金融知识表示及风险识别研究，构建金融知识表示框架。Liang等[19]将知识关联推理应用于财务报表的信息融合，提出了一种挖掘关联交易的图网络模型和算法，为财务管理提供决策支持。Xiong等[20]通过整合药物实体和疾病实体在知识图中的语义关联与结构关联，提高了药物重定位的预测准确率。刘鑫[21]将知识关联应用于专利估值研究，构建了由专利数据与市场数据组成的异构知识关联网络，实现对专利的价值评估。鲁云蒙等[22]探讨了知识关联和交互策略对隐性知识扩散的影响机理，研究结果表明：知识关联强度在科研合作网络知识扩散初期有重要影响，强知识关联下有着更高的知识扩散速率。

1.3　科学技术互动关系

目前，已经有不少学者对科学技术互动关系测度开展研究。在定性研究中，Guan等[23]提出科学-技术互动双螺旋模式，将科技创新演化趋势解释为科学与技术创新主题在相互作用下螺旋式上升发展的过程。在定量研究中，Ke[24]基于生命科学相关专利与生物医学研究之间的引文联系，揭示了基础研究与应用研究之间的关联演变；Ranaei等[25]将机器学习与LDA主题模型结合，基于主题相似性实现科技关联识别并探究专利与科学出版物之间的语义联系；刘自强等[26]从共词、作者与引用关系出发，探究关联主题在科学与技术路径上的时间分布，归纳出4个科技互动模式。整体上看，现有研究主要集中在文献层次的引用关系和主题层次的相似关系两个方面，对于文献内知识元间关联的揭示不够深入。

在科学技术互动关系对科学技术影响力、创新能力的影响方面，也有学者进行了探讨。有研究证实了吸收科学知识能够提高专利的影响力，如：Wang等[27]以纳米领域为例，探究专利文献的非专利引用与专利质量之间的关系，研究结果表明高质量的学术研究为高质量专利的发展作出了重大贡献；Fernández等[28]基于包含54 578个能源专利家族的数据集开展研究，结果表明包含更多科学文献引用的专利倾向于有更高的知识扩散强度；Ahmadpoor等[29]对美国专利商标局（United States Patent and Trademark Office，USPTO）登记的480万件专利以及1945—2013年出版的3 200万篇期刊论文的引用情况进行分析，从引用距离的角度揭示了科学进步对技术进步的促进作用；Ding等[30]通过分析引用频次的增长轨迹，探究科学论文引用对专利影响力的正向影响，并探讨了论文引用频次、论文发表期刊的排名、论文类型等因素的作用。也有研究证实了技术与科学影响力之间的正相关关系，如：Meyer等[31]探讨了纳米科学专利引文与引文影响之间的关系，结果表明被专利引用的论文倾向于获得更高的引用频次；Patelli等[32]通过探究全球范围内多个国家科学与技术的关系，揭示了科学与技术的协同发展关系。

从现有研究可以看出，科学-技术知识关联能够为高质量论文和高质量专利的识别提供新的视角，有助

于完善科技评价体系，对于科技创新战略布局具有一定的意义。然而，现有研究缺少对科学与技术内在关联的考虑，难以全面把握创新特征[26]，同时也尚未提出统一的科学技术互动关系测度指标与测度方法。针对目前研究的不足之处，本文从知识元视角探索科学-技术知识关联指标与测度方法，并通过实证研究论证该指标与方法的有效性。

2 基于知识元的科学-技术知识关联指标与测度模型

为了细粒度地识别科学与技术之间的关联，基于深度学习方法实现知识元抽取，应用知识关联理论探究科学与技术互动关系，构建基于知识元的科学-技术知识关联指标与测度模型（见图1）。

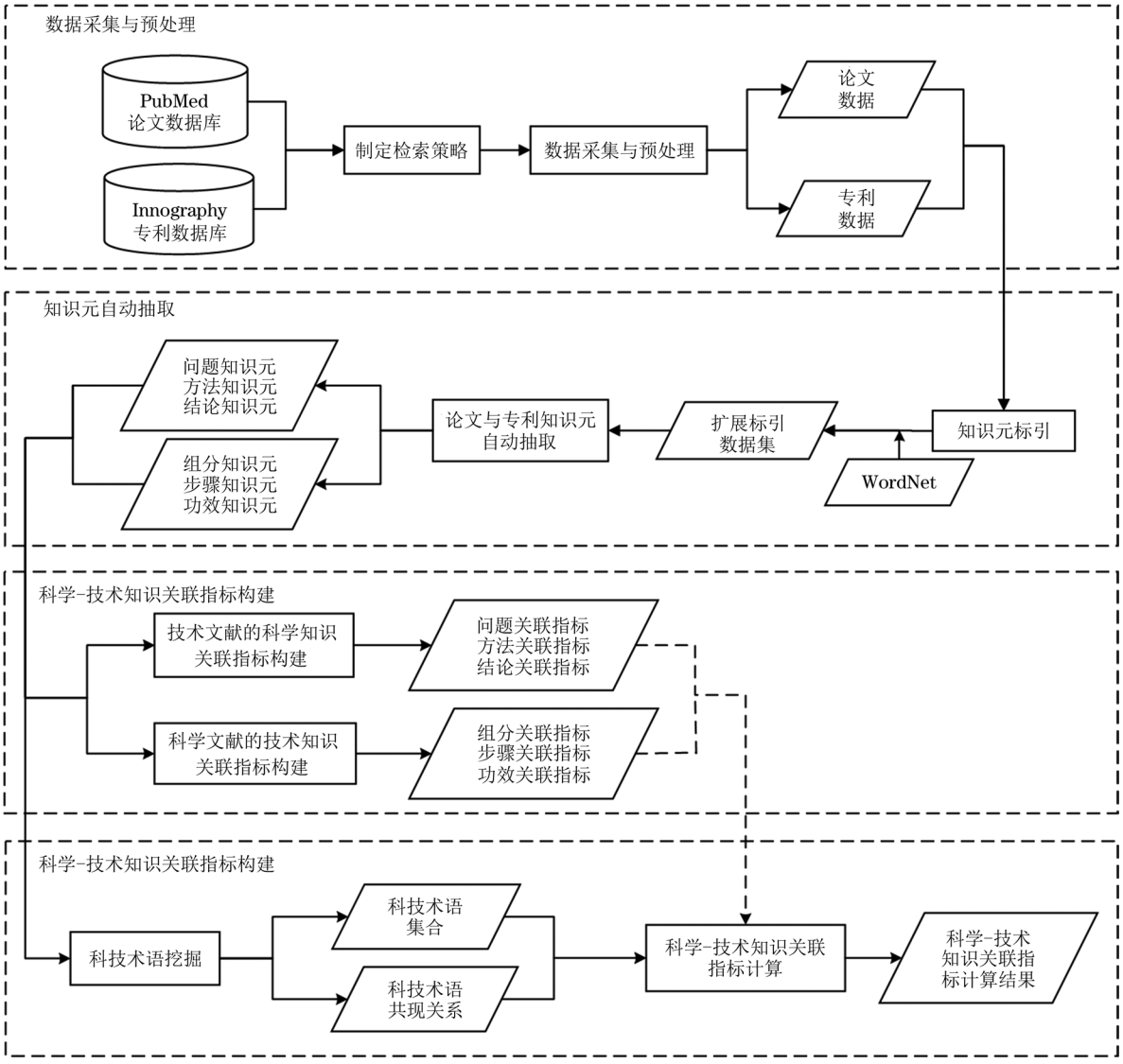

图1 基于知识元的科学-技术知识关联指标与测度模型

2.1 数据采集与预处理

论文是科学知识的载体，专利是技术知识的载体[33]。由于论文及专利数据能够反映科学研究成果与技术创新成果，且具有易获取、时间序列长等特点，选取论文、专利分别作为科学与技术的表征。数据采集与预处理阶段具体包括如下步骤。

（1）制定检索策略。根据选择的领域制定相应的检索策略，收集用于分析的论文和专利数据。论文数据的字段包括标题、作者、摘要、发表时间等，专利数据的字段包括专利号、标题、发明人、公示时间、摘要等。

（2）数据清洗。对数据去重后，筛除标题、摘要、

发表时间（公示时间）字段缺失以及摘要文本长度较短的数据，得到论文和专利数据集。

2.2　科学与技术知识元自动抽取

秦春秀等[34]基于科技文献的知识结构和特征提出了一种基于知识元的科技文本内容描述框架，使用主题/类别知识元、研究领域知识元、背景知识元、问题知识元、理论/原理知识元等13个大类的知识元对科技文本知识对象的语义信息进行揭示。参考这一框架进行知识元抽取对象的选取。为了让选取的知识元类型能够反映科技文献的创新特征，对13个大类的知识元进行筛选，构建科技文献内容描述框架（见图2），其中：科学文献包含的科学知识元类型有问题知识元、方法知识元和结论知识元；技术文献包含的技术知识元类型有组分知识元、步骤知识元和功效知识元。

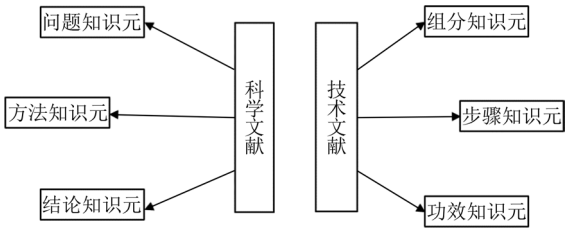

图2　科技文献内容描述框架

（1）科学知识元类型选择。杜杏叶[35]在学术论文知识元库构建基础上，提出了基于研究问题创新性、理论创新性、方法创新性、结论创新性进行学术论文智能化评价的基本方法。其中：理论创新性侧重描述科学文献在科学理论方面的继承与创新[36]，反映科学文献与科学文献之间的关联，与技术文献关联程度相对较低；问题、方法、结论创新性对应杨中楷等[37]提出的“基础科学研究作为科技创新源头”观点，分别为技术创新提供了科学研究需求、科学研究方法体系以及科学发现3个方面的知识供给，与技术文献存在密切关联。

（2）技术知识元类型选择。Heinze等[38]将技术研究的创新性概括为发明新仪器、提出和使用新方法、整合现有理论的应用等，上述3种创新性分别对应技术文献中的组分、步骤与功效3个知识元类型。其中：组分、步骤知识元描述了技术文献为解决实际问题而提出的新设计、新方案；功效知识元描述了技术文献在所属领域的实际应用成效，同时也反映了技术文献所解决的实际问题。上述3个知识元类型分别为科学创新提供了技术产品、技术工艺以及技术应用方向3个方面的知识供给，与科学文献存在密切关联。

在自动抽取模型方面，Rei等[39]提出“词语级向量与字符级向量组合的LSTM-CRF模型”，采用注意力机制将字符级别的信息融入序列标注体系结构，提升了模型在序列标注任务中的性能。余丽等[14]的研究表明，该模型将CRF模型嵌入LSTM模型，在研究范畴、研究方法、实验数据、评价指标及取值4类知识元的抽取实践中相当有效。沈思等[40]的研究表明，LSTM-CRF模型在摘要结构功能识别任务中效果显著，与RNN、CRF、LSTM和SVM模型相比整体性能分别提升了33.63%、39.13%、32.81%和38.33%。

本研究在LSTM-CRF模型的基础上实现科学文献与技术文献知识元的自动抽取。在知识元方面，相比词粒度知识元，句粒度知识元在反映科技文献中与目的、方法、结论、贡献等结构功能内容相关的关键信息方面更具优势[41]，因此基于论文和专利的摘要文本，对科技文献中的句粒度知识元进行抽取。首先，通过人工标注部分摘要文本中出现的各类型知识元得到原始标引数据集，结合WordNet词典实现标引数据集的扩展。其次，将扩展标引数据集输入LSTM-CRF模型，实现多类型的知识元抽取，并基于词性标注结果制定特征抽取规则，用于知识元中的科技术语挖掘。

2.3　科学-技术知识关联指标构建

在上文提出的科技文献内容描述框架下，结合已有研究成果构建科学-技术知识关联指标，包括技术文献的科学知识关联指标和科学文献的技术知识关联指标。已从科技文献内容描述的角度筛选得到3类科学知识元与3类技术知识元，下面从科学-技术关联的角度说明将这6类知识元用于科学-技术知识关联指标构建的合理性。

在科学文献内容分析方面，现有研究指出，科学研究本质上是一种解决问题的行为活动，问题知识元是对科学文献中研究问题的描述[42]；方法知识元是指调查研究问题所采取的行动，以及应用特定程序或技术来识别、选择、处理和分析研究问题相关信息的基本原理[43]；结论知识元用于反映研究结论的关键信息[44]，是学术文献摘要结构功能的重要组成部分[45]。

在科学与技术之间的关联方面，Schmoch等[46]采

用文献计量方法对以科学为基础的新兴技术项目的影响力开展研究，指出一部分技术项目的新颖性产生于科学界关于如何解决某个问题的持续争论，基础学科和应用学科之间的一致性对于技术项目中新概念的产生有重要影响。Wang等[47]指出，研究方法、实验、结果分析等知识元是科学文献创新性测度的有效资源。因此，综合考虑科学文献的内容描述框架与科学-技术关联关系，在构建技术文献的科学知识关联指标时，分别计算特定文献包含的技术术语与领域内科学术语在问题、方法、结论3类科学知识元中的共现频次，形成问题关联、方法关联和结论关联3个科学知识关联指标。

在技术文献内容分析方面，Yoon等[48]提出TrendPerceptor系统，实现对专利关键信息的自动提取，其定义的专利关键信息包括属性和功能两个方面：属性指的是为发明目的而应用的方法，功能则与方法的应用效果相关。Wang等[49]在对专利文本进行主体-行动-客体（SAO）分析的基础上，构建了包括材料、技术、影响因素、成分、产品、目标和未来方向在内的技术路线图。本研究将技术文献中包含的知识元归纳为组分、步骤与功效3类。其中：组分知识元是专利文献知识描述框架的构成要素之一，蕴含专利各元件间的连接关系和作用关系[50]；步骤知识元来源于专利说明书中的具体实施方式模块，是发明标的物实现过程或流程步骤的细粒度知识描述[51]；功效知识元是专利通过技术特征所实现的预期效果和对专利产生作用的评价，反映了专利的创新程度[51]。

在技术与科学之间的关联方面，Han等[52]通过关注技术域的分解来理解和分析科学和技术的复杂相互作用，将领域中技术成果的目的与基本效用作为区分科学研究的技术相关性的因素。Basnet等[53]指出，与技术领域相关的交互作用是绩效改进率的重要影响因素，其探讨的交互作用主体具体包括技术研究的组件、系统以及功能需求等。因此，综合考虑技术文献的内容描述框架与技术-科学关联关系，在构建科学文献的技术知识关联指标时，分别计算特定文献包含的科学术语与领域内技术术语在组分、步骤、功效3类技术知识元中的共现频次，形成组分关联、步骤关联、功效关联3个技术知识关联指标。

2.4 科学-技术知识关联指标测度

相比于引文网络分析，共词分析粒度更细，侧重于概念内容层面的知识关联[54]，大多数学者在知识关联研究中将共词网络作为研究对象[31-32]。在现有研究的基础上，基于不同类型知识元下的科技术语集合以及科技术语共现关系来计算科学-技术知识关联指标。

在特定领域数据集中，科学术语集合为$S=\{s_1, s_2, \cdots, s_M\}$，技术术语集合为$T=\{t_1, t_2, \cdots, t_N\}$。科学知识元类型集合为$K_S=\{$问题，方法，结论$\}$，技术知识元类型集合为$K_T=\{$组分，步骤，功效$\}$。依据每个术语在不同类型知识元中出现的最高频次确定其归属，将科学术语集合S划分为3个互不相交的集合（$S_{问题}$、$S_{方法}$以及$S_{结论}$），技术术语集合T划分为3个互不相交的集合（$T_{组分}$、$T_{步骤}$以及$T_{功效}$）。

对于某篇特定的科学文献，其科学术语集合为S_{sub}，依据知识元类型划分为$S_{sub_问题}$、$S_{sub_方法}$以及$S_{sub_结论}$。考虑不同类型科学知识元与技术知识元的两两组合，该科学文献的技术知识关联指标计算公式如式（1）所示。

$$R_{Tec}=\frac{N_{Avg,S}}{N_{patents}N_S}\sum_{p\in K_S}\sum_{q\in K_T}F_{co\_occur,p-q}\left(s\in S_{sub\_p},t\in T_q\right) \quad (1)$$

式中：$N_{patents}$表示技术文献的数量；N_S表示该科学文献所包含的科学术语总数；$N_{Avg,S}$表示平均每篇科学文献所包含的科学术语数量；p和q分别表示科学知识元类型与技术知识元类型；$F_{co\_occur,p-q}$（$s\in S_{sub\_p}$，$t\in T_q$）表示S_{sub_p}集合中的术语s与T_q集合中的术语t在该科学文献中的共现频次。式（1）描述了特定科学文献在每个科学术语上与领域内每篇技术文献产生的术语共现频次。

考虑知识元类型，可将技术知识关联指标分为组分关联、步骤关联、功效关联指标，其中组分关联描述特定科学文献在每个科学术语上与领域内每篇技术文献的组分知识元产生的术语共现频次，计算公式如式（2）所示。

$$R_{Tec,组分}=\frac{N_{Avg,S}}{N_{patents}N_S}\sum_{p\in K_S}F_{co\_occur,p-组分}\left(s\in S_{sub\_p},t\in T_{组分}\right) \quad (2)$$

同样的，对于某篇特定的技术文献，其技术术语集合为T_{sub}，依据知识元类型划分为$T_{sub_组分}$、$T_{sub_步骤}$以及$T_{sub_功效}$。该技术文献的科学知识关联指标计算公式如式（3）所示。

$$R_{Sci}=\frac{N_{Avg,T}}{N_{papers}N_T}\sum_{p\in K_S}\sum_{q\in K_T}F_{co\_occur,p-q}\left(s\in S_p,t\in T_{sub\_q}\right) \quad (3)$$

式中：N_{papers}表示科学文献的数量；N_T表示该技术文献所包含的技术术语总数；$N_{Avg,T}$表示平均每篇技

术文献所包含的技术术语数量；$F_{\text{co\_occur},\,p-q}$（$s\in S_p$，$t\in T_{\text{sub}\_q}$）表示$S_p$集合中的术语$s$与$T_{\text{sub}_q}$集合中的术语$t$在该技术文献中的共现频次。式（3）描述了特定技术文献在每个技术术语上与领域内每篇科学文献产生的术语共现频次。

考虑知识元类型，可将科学知识关联指标分为问题关联、方法关联、结论关联指标，其中问题关联描述特定技术文献在每个技术术语上与领域内每篇科学文献的问题知识元产生的术语共现频次，计算公式如式（4）所示。

$$R_{\text{Sci,问题}}=\frac{N_{\text{Avg},T}}{N_{\text{papers}}N_T}\sum_{q\in K_T}F_{\text{co\_occur,问题}-q}\left(s\in S_{\text{问题}},t\in T_{\text{sub}\_q}\right)\quad(4)$$

在科学-技术知识关联测度指标的基础上，基于PubMed医学领域论文数据库和Innography专利数据库的数据进行实证分析，通过聚焦科学-技术知识关联指标在高质量论文识别中的作用，检验所提知识关联测度方法的有效性。

3　实验与结果分析

3.1　数据采集

在医学领域，论文与专利数据丰富，基础研究与临床应用之间存在着紧密联系。糖尿病相关研究在反映科学与技术知识关联方面具有代表性，因此本研究选取糖尿病领域的论文和专利作为研究数据。论文数据来源于医学领域论文数据库PubMed，构建检索式MeSH Major Topic=“diabetes”，文献类型设定为Clinical Trial和Randomized Controlled Trial，共收集了46 485条数据。专利数据来源于Innography数据库，构建检索式TI=“diabetes”，共收集了167 579条数据。检索时间为2022年10月25日。经数据清洗后，删除了重复数据、标题缺失以及摘要文本长度小于50个词的数据，得到27 738条论文数据和157 563条专利数据。论文与专利数据的年度分布情况如图3所示。

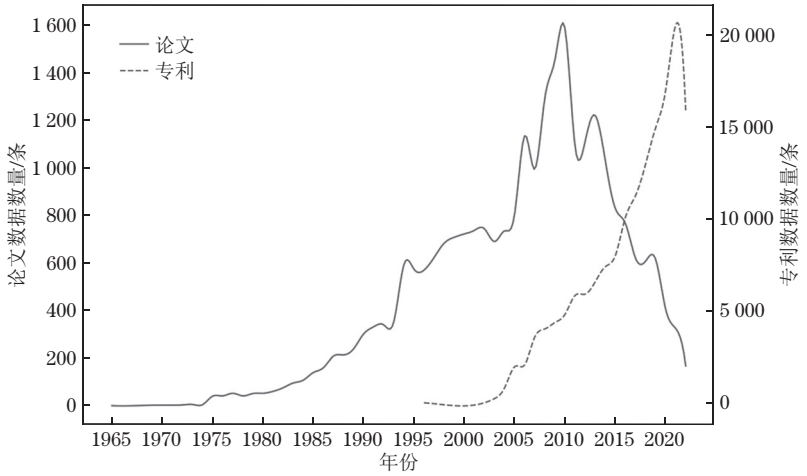

图3　糖尿病领域论文与专利数量的年度分布

从上述结果可以看出，糖尿病领域的论文数量在2009年达到峰值，随后开始下降，而专利数量在1996—2021年持续上升。在总数量上，专利数量远多于论文数量。总体上看，专利数量的增长晚于论文，在增长趋势上滞后于论文，但增长速度较论文更快。

3.2　知识元抽取

在知识元抽取方面，考虑到标注成本较高，借鉴了赵冠壹等[55]提出的知识元识别与抽取流程：首先基于描述规则得到少量带标注的数据集，接着在小数据集的基础上通过Bootstrapping方法筛选得到增量样本，对知识元词库进行扩充，从而解决标注样本过少的问题。对100篇论文和100篇专利摘要文本中的知识元进行人工标注，扩充后的标注样本数量为3 000条，人工标注结果示例如图4所示。需要注意的是，一篇科技文献的摘要文本并不一定完整包含3类知识元，可能缺失某些知识元类型。此外，一篇科技文献的摘要文本中可能包含多个同一类型知识元。

在人工标注结果的基础上，基于WordNet同义词

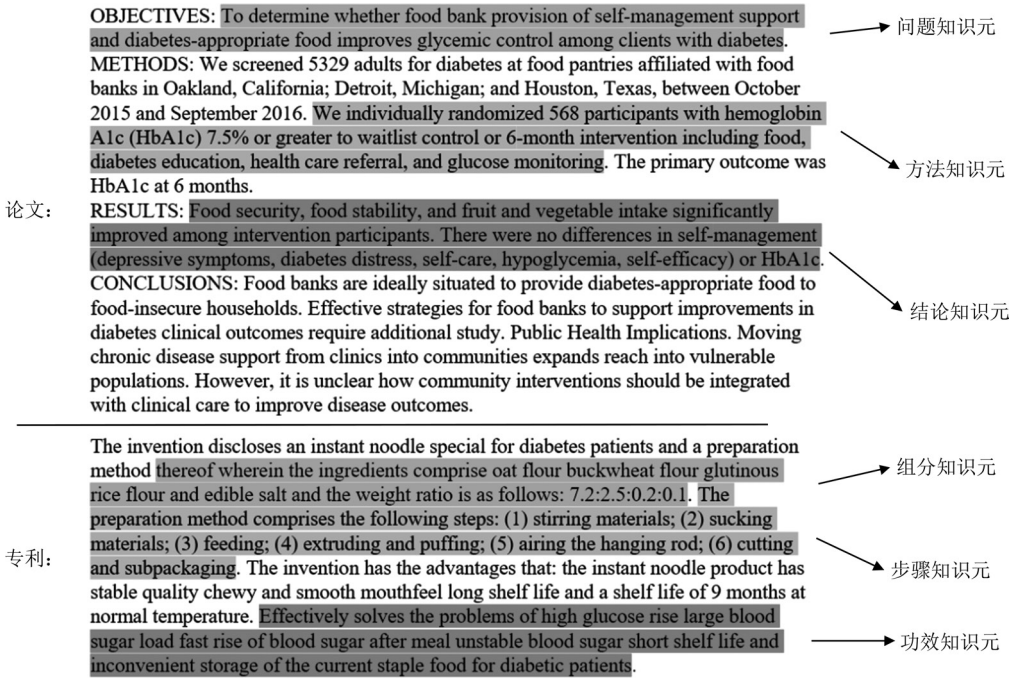

图4　科技文献知识元人工标注结果示例

替换5~15个单词，对每篇摘要文本进行14次替换操作，得到扩展标引数据集作为LSTM-CRF模型的输入。采用glove向量作为词嵌入层的输入，构建LSTM-CRF模型进行知识元抽取。在抽取结果方面，以问题知识元为例，抽取结果的评价指标分别为准确率0.80、召回率0.74、F1值0.77，这与余丽等[14]实验得到的准确率0.83、沈思等[40]实验得到的准确率0.84基本持平，效果达到预期。同时，在相同数据集上与各种大语言模型的抽取效果进行比较，发现大语言模型在知识元抽取任务中表现不佳，ChatGPT、Claude以及ChatGLM-6B的F1值分别为0.64、0.57、0.48，这可能与特定领域专有名词较多、实体组分复杂、大模型缺少相关训练语料有关[56-57]。上述结果表明该模型在科技文献多类型知识元抽取任务中具有有效性。

在知识元自动抽取结果的基础上，结合词性标注筛选出科学术语和技术术语，用于后续的科学-技术知识关联指标计算。经实践，归纳出以下特征抽取规则（见表1），用于科学知识元与技术知识元的术语挖掘。

对基于上述规则挖掘出的科学术语与技术术语进行清洗，包括去除部分无关词（如diagram、method、formula、image、figure、day、week、month等）、去除特殊符号（如/、*、<等）、规范单复数、合并同义词等，得到的科学术语与技术术语部分结果示例如表2所示。

从结果可以看出，科学文献与技术文献中存在大量共用术语，但其术语使用在知识元层次存在异质性。科学文献的常用术语往往聚焦于该领域的某一研究问题，如代谢紊乱、炎症性疾病、糖尿病多发性神经病等，表明研究者倾向于从问题出发描述其科学发现。与之相比，技术文献的常用术语多涉及具体对象，如胰岛素、血液样本、肾功能等，表明研究者倾向于从实验对象出发描述其技术发现。

表1　知识元术语挖掘的特征抽取规则

编 号	抽取规则	编 号	抽取规则
1	Adjective+Noun（singular or plural）	5	Noun+Noun
2	Adjective+Proper noun（singular or plural）	6	Noun+Noun+Noun
3	Adjective+Cardinal number+Noun	7	Noun+（+Proper noun+）
4	Noun	8	Proper noun+Noun+Noun

表2　科学术语与技术术语部分结果示例

术语类型	术语来源	结果示例
科学术语	问题知识元	diabetic retinopathy、diabetic polyneuropathy
	方法知识元	blood pressure、acid derivative
	结论知识元	phosphorylase activity、metabolic disorders
技术术语	组分知识元	blood pressure、blood samples
	步骤知识元	insulin secretion、renal function
	功效知识元	insulin、metabolic control

3.3　科学-技术知识关联指标测度

筛选出现频次高于3次的科技术语，统计得到不同类型知识元中科技术语的共现频次。采用前文所述方法对科学文献的技术知识关联指标以及技术文献的科学知识关联指标进行测度，部分结果如表3所示。

3个科学知识关联指标、3个技术知识关联指标的总体统计结果如表4所示。从统计分布结果来看，功效关联指标的平均值最大（0.083），组分关联指标的平均值最小（0.002）；6个指标的取值区间为[0，1.323]；偏度均为正值，表明数据分布相对于正态分布右偏；峰度较大，表明分布不均匀且陡峭，结合百分位数可以看出，大部分值集中在[0，0.017]区间。

已有研究指出，科学-技术知识关联与论文的高影响力之间存在正相关关系[31]；同时，科学-技术知识关联也是高颠覆性专利预测任务的重要特征[58]。为了说明所提科学-技术知识关联指标与测度方法的有效性，可以从技术知识关联指标在高质量论文识别中的作用以及科学知识关联指标在高质量专利识别中的作用两个方面进行验证。下面只从技术知识关联指标在高质量论文识别中的作用方面来说明所提科学-技术知识关联指标与测度方法的有效性，科学知识关联指标有效性分析与之类似。

参考杨杰等[59]的研究，对技术知识关联指标排名前6和后6的论文进行人工检视。通过分析筛选出论文的被引频次，以说明所提的科学-技术知识关联测度方法能够有效反映科技文献对科学或技术知识的利用情况，有助于高质量论文的识别，进而说明所提科学-技术知识关联指标与测度方法的有效性。

将3个技术知识关联指标的平均值作为排名依据，得到排名前6、后6的论文如表5所示。其中有多篇论文的技术知识关联指标均为最小值0，因此采取随机抽取方式得到排在后6名的论文。由表5可知，技术知识关联指标排名较高的6篇论文具有较高的被引频次，相比之下排名较低的6篇论文的被引频次显著更低。从技术知识关联指标与论文质量之间较高的一致性可以看出，将所提的技术知识关联指标作为特征，能够在一定程

表3　部分科技文献的科学知识关联指标、技术知识关联指标

文献类型	标 题	科学知识关联指标				技术知识关联指标			
		整 体	问 题	方 法	结 论	整 体	组 分	步 骤	功 效
论文	Gastroesophageal reflux in diabetes mellitus					0.054	0.000	0.023	0.031
论文	Long-term effect of ACE inhibition on development of nephropathy in diabetes mellitus type II					0.145	0.022	0.036	0.087
专利	Systems and methods for managing diabetes care data	0.053	0.027	0.010	0.016				
专利	Food for curing diabetes	0.882	0.302	0.272	0.308				

表4　科学知识关联指标、技术知识关联指标的统计分布

指 标		均 值	标准差	偏 度	峰 度	最小值	25%	50%	75%	最大值
科学知识关联	问题	0.028	0.070	4.682	29.914	0.000	0.000	0.001	0.017	0.825
	方法	0.017	0.051	5.933	47.248	0.000	0.000	0.001	0.008	0.677
	结论	0.022	0.061	5.592	44.845	0.000	0.000	0.001	0.008	0.824
技术知识关联	组分	0.002	0.006	6.811	71.492	0.000	0.000	0.000	0.000	0.115
	步骤	0.012	0.020	2.688	9.612	0.000	0.000	0.000	0.016	0.184
	功效	0.083	0.124	2.465	8.423	0.000	0.000	0.029	0.120	1.323

表5　技术知识关联指标排名前6与后6的论文

编 号	标 题	论文作者（发表年）	组分关联	步骤关联	功效关联	均 值	NPRs	Topic Interaction	被引频次/次
1	Canagliflozin and cardiovascular and renal events in type 2 diabetes	Neal等（2017）	0.015	0.076	1.323	0.471	3	1 079	5 064
2	Empagliflozin，cardiovascular outcomes，and mortality in type 2 diabetes	Zinman等（2015）	0.027	0.097	1.168	0.431	1	988	4 796
3	Cardiovascular and renal outcomes with empagliflozin in heart failure	Packer等（2020）	0.030	0.182	0.960	0.391	1	1 231	2 351
4	Canagliflozin and renal outcomes in type 2 diabetes and nephropathy	Perkovic等（2019）	0.023	0.118	1.027	0.389	0	162	2 736
5	The oregon experiment—effects of medicaid on clinical outcomes	Baicker等（2013）	0.024	0.090	1.017	0.377	0	267	861
6	Primary care-led weight management for remission of type 2 diabetes（DiRECT）：an open-label，cluster-randomised trial	Lean等（2018）	0.026	0.052	1.035	0.371	0	563	1 017
7	Study of lacrimal fluid trace elements in several eye diseases	Vinetskaya等（1994）	0.000	0.000	0.000	0.000	0	441	12
8	Role of the indigenous drug saptamrita lauha in hemorrhagic retinopathies	Sharma等（1992）	0.000	0.000	0.000	0.000	0	1 204	5
9	Augmentation laser for proliferative diabetic retinopathy that fails to respond to initial panretinal photocoagulation	Doft等（1992）	0.000	0.000	0.000	0.000	0	381	11
10	Pharmacodynamic considerations with recombinant human insulin-like growth factor-I in children	Ferry等（2005）	0.000	0.000	0.000	0.000	0	882	3
11	Early ST-segment recovery，infarct artery blood flow，and long-term outcome after acute myocardial infarction	French等（2002）	0.000	0.000	0.000	0.000	0	1 577	24
12	Mitral valve surgery in patients with extensively calcified mitral annulus long-term echocardiographic and clinical follow-up	Steuer等（2012）	0.000	0.000	0.000	0.000	0	720	0

度上判断高质量论文。

为进一步验证研究结果，将所提的技术知识关联指标与以往研究中反映科学-技术互动关系的引文关联指标（NPRs）[24]、主题相似性关联指标（Topic Interaction）[25]对比。其中NPRs为论文被专利引用的次数，Topic Interaction为LDA模型下主题分布概率从高到低排列前3项差值之和小于5%的专利数量（LDA主题数为20个）。对比实验发现，由于专利对论文的引用次数普遍较低，大部分论文的NPRs值为0，NPRs对高质量论文的区分能力不及所提的技术知识关联指标；被引频次的分布差异在Topic Interaction值高于1 000和低于300的两个论文集合中不明显，例如表5中论文8、11的Topic Interaction值高于1 000、被引频次低于50次，论文4、5的Topic Interaction值低于300、被引频次高于500次，而类似的不一致现象在技术知识关联指标中并未出现，表明Topic Interaction对高质量论文的区分能力不及所提的技术知识关联指标。

为进一步分析技术知识关联指标的测度有效性，按照被引频次为0~30次、31~150次、151~500次、501次及以上划分4个论文数据集，对每个集合中技术知识关联指标（取3个指标的平均值）的分布情况进行分析，其核密度分布情况如图5所示。由图5可以看出，不同被引频次的论文在技术知识关联指标上的分布呈现明显的异质性，被引频次低的论文倾向于有更低的技术知识关联指标，这补充证明了上文的结论。以上分析说明了所提技术知识关联指标与测度方法的有效性。

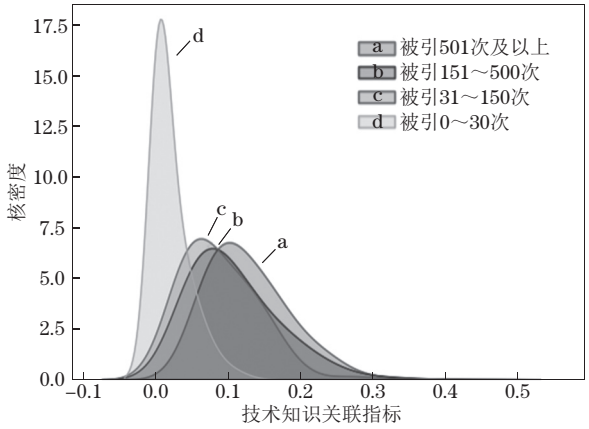

图5　不同被引频次论文的技术知识关联指标核密度分布

4　结语

基于知识元理论，本研究提出了科学-技术知识关联指标与测度方法。在抽取出问题、方法、结论3类科学知识元和组分、步骤、功效3类技术知识元的基础上，形成3个科学知识关联指标与3个技术知识关联指标。基于科技术语在不同类型知识元中的共现情况，实现科学-技术知识关联指标的测度。通过聚焦科学-技术知识关联指标在高质量论文识别中的作用，检验所提知识关联指标与测度方法的有效性。

从理论上讲，在知识元层次对科学-技术知识关联进行测度，有助于在微观内容层面量化科学与技术之间的知识联系和结构联系。与以往通过论文和专利的被引频次来反映科技交互的研究不同，本研究深入科技文献的语义内容层面捕捉科技关联，丰富了当前科技关联测度的方法。从现实意义上讲，研究构建的科学-技术知识关联指标为高质量论文的识别提供了参考，对于完善科技评价体系、促进创新驱动发展战略实施有参考意义。

提出的科学-技术知识关联测度方法还存在一定的不足，例如研究仅在糖尿病领域展开验证，研究结论的普适性还需在多组具有学科或主题差异性的数据集上进一步验证。此外，通过扩大标注语料库、改进自动抽取模型，还可以实现科技文献知识元自动抽取效果的进一步提升。对引用关系与语义关系的测度依赖完备的论文-专利引用关系数据集和领域本体，现阶段工作难以满足该条件，因此在计算科学-技术知识关联指标时，只从共词关系层面考虑了科技术语之间的关系，并未从引用层面和语义层面出发进行探索。在未来的研究中，将进一步考虑引用关系和语义关系，在具有学科或主题差异性的多组数据集上探索科学与技术之间的广泛联系。

参考文献

[1]　加快实施创新驱动发展战略[N]. 经济日报，2022-10-22（10）.

[2]　索传军，戎军涛. 知识元理论研究述评[J]. 图书情报工作，2021，65（11）：133-142.

[3]　BROOKES B C. The foundations of information science[J]. Journal of Information Science，19813（1）：3-12.

[4]　POPPER K R. Objective knowledge：an evolutionary approach[M]. Oxford：Clarendon Press，1972.

[5]　温有奎，焦玉英. 基于知识元的知识发现[M]. 西安：西安电子科技大学出版社，2011.

[6]　文庭孝，侯经川，龚蛟腾，等. 中文文本知识元的构建及其现实意义[J]. 中国图书馆学报，2007，33（6）：91-95.

[7]　石湘，刘萍. 基于知识元语义描述模型的领域知识抽取与表示研究：以信息检索领域为例[J]. 数据分析与知识发现，2021，5（4）：123-133.

[8]　BENSE H. The unique predication of knowledge elements and their visualization and factorization in ontology engineering[J]. Frontiers in Artificial Intelligence and Applications，2014，267：241-250.

[9]　JIANG L，YANG Z K，WANG J X. Knowledge indexing of Chinese text based knowledge element[C]//2008 International Symposium on Knowledge Acquisition and Modeling，2008：35-38.

[10]　XIE N F，WEI X，HAO X N. Research on knowledge element relation and knowledge service for agricultural literature resource[C]//Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence，2019：172-176.

[11]　叶光辉，彭泽，陈国梁，等. 学术文献中的知识单元抽取及其分布特征识别研究[J]. 情报理论与实践，2023，46（4）：90-98.

[12]　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.

[13]　王忠义，沈雪莹，黄京. 科技文献资源中方法知识元的抽取研究[J]. 情报科学，2021，39（1）：13-20.

[14]　余丽，钱力，付常雷，等. 基于深度学习的文本中细粒度知识元抽取方法研究[J]. 数据分析与知识发现，2019，3（1）：38-45.

[15]　文庭孝，刘晓英，刘进军. 知识关联的理论基础研究[J]. 图书

馆，2010（4）：9-11.

[16] 高继平，丁堃，潘云涛，等. 知识关联研究述评[J]. 情报理论与实践，2015，38（8）：135-140.

[17] 洪亮，马费成．面向大数据管理决策的知识关联分析与知识大图构建[J]. 管理世界，2022，38（1）：207-219.

[18] 唐旭丽，马费成，傅维刚，等. 知识关联视角下的金融知识表示及风险识别[J]. 情报学报，2019，38（3）：286-298.

[19] LIANG Z Q，PAN D，DENG Y. Research on the knowledge association reasoning of financial reports based on a graph network[J]. Sustainability，2020，12（7）：2795.

[20] XIONG Z K，HUANG F，WANG Z Y，et al. A multimodal framework for improving in silico drug repositioning with the prior knowledge from knowledge graphs[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics，2022，19（5）：2623-2631.

[21] 刘鑫. 基于知识关联的专利估值模型与算法研究[D]. 呼和浩特：内蒙古大学，2021.

[22] 鲁云蒙，刘铁忠. 基于知识关联性的科研合作网络隐性知识扩散模型研究：以重大科技工程为例[J]. 数据分析与知识发现，2021，5（9）：10-20.

[23] GUAN J C，HE Y. Patent-bibliometric analysis on the Chinese science：technology linkages[J]. Scientometrics，2007，72（3）：403-425.

[24] KE Q. An analysis of the evolution of science-technology linkage in biomedicine[J]. Journal of Informetrics，2020，14（4）：101074.

[25] RANAEI S，SUOMINEN A，DEDEHAYIR O. A topic model analysis of science and technology linkages：a case study in pharmaceutical industry[C]//2017 IEEE Technology & Engineering Management Conference（TEMSCON），2017：49-54.

[26] 刘自强，许海云，罗瑞，等. 基于主题关联分析的科技互动模式识别方法研究[J]. 情报学报，2019，38（10）：997-1011.

[27] WANG L L，LI Z X. Knowledge flows from public science to industrial technologies[J]. The Journal of Technology Transfer，2021，46（4）：1232-1255.

[28] FERNÁNDEZ A M，FERRÁNDIZ E，MEDINA J. The diffusion of energy technologies. evidence from renewable，fossil，and nuclear energy patents[J]. Technological Forecasting and Social Change，2022，178：121566.

[29] AHMADPOOR M，JONES B F. The dual frontier：patented inventions and prior scientific advance[J]. Science，2017，357（6351）：583-587.

[30] DING C G，HUNG W C，LEE M C，et al. Exploring paper characteristics that facilitate the knowledge flow from science to technology[J]. Journal of Informetrics，2017，11（1）：244-256.

[31] MEYER M，DEBACKERE K，GLÄNZEL W. Can applied science be ‘good science’？exploring the relationship between patent citations and citation impact in nanoscience[J]. Scientometrics，2010，85（2）：527-539.

[32] PATELLI A，CIMINI G，PUGLIESE E，et al. The scientific influence of nations on global scientific and technological development[J]. Journal of Informetrics，2017，11（4）：1229-1237.

[33] 王诗炜，陈春. 基于科学论文和技术专利关联关系识别潜在知识发现方法研究综述[J]. 数据分析与知识发现，2023，7（7）：18-31.

[34] 秦春秀，刘杰，刘怀亮，等. 基于知识元的科技文本内容描述框架研究[J]. 图书情报工作，2017，61（10）：116-124.

[35] 杜杏叶. 学术论文关键指标智能化评价研究[D]. 长春：吉林大学，2020.

[36] COSTANZO B P，SÁNCHEZ L E. Innovation in impact assessment theory and practice：how is it captured in the literature？[J]. Environmental Impact Assessment Review，2019，79：106289.

[37] 杨中楷，梁永霞，刘则渊. 重视技术科学在科技创新供给侧改革中的作用[J]. 中国科学院院刊，2020，35（5）：629-636.

[38] HEINZE T，SHAPIRA P，ROGERS J，et al. Creative capabilities and promotion of highly innovative research in Europe and the United States[EB/OL]. [2023-12-01]. https://www.crea.server.de/finalreport/CREA_ Final_Report.pdf.

[39] REI M，CRICHTON G K O，PYYSALO S. Attending to characters in neural sequence labeling models[C]//Proceedings of the 26th International Conference on Computational Linguistics（COLING-2016），2016.

[40] 沈思，胡昊天，叶文豪，等. 基于全字语义的摘要结构功能自动识别研究[J]. 情报学报，2019，38（1）：79-88.

[41] 李广建，袁钺. 基于深度学习的科技文献知识单元抽取研究综述[J]. 数据分析与知识发现，2023，7（7）：1-17.

[42] HEFFERNAN K，TEUFEL S. Identifying problems and solutions in scientific text[J]. Scientometrics，2018，116（2）：1367-1382.

[43] WANG Z Y，SHEN X Y，HUANG R，et al. Extracting method knowledge elements from scientific literature：a rule-based

approach[J]. Proceedings of the Association for Information Science and Technology，2019，56（1）：805-807.

[44] 柴庆凤，史霖炎，梅珊，等. 基于人工特征和机器特征融合的科技文献知识元抽取[J]. 数据分析与知识发现，2021，5（8）：132-143.

[45] 孟旭阳，白海燕. 文献摘要结构功能识别在关键词抽取中的应用[J]. 情报工程，2022，8（1）：79-89.

[46] SCHMOCH U，BECKERT B，SCHAPER-RINKEL P. Impact assessment of a support programme of science-based emerging technologies[J]. Scientometrics，2019，118（3）：1141-1161.

[47] WANG Z Y，WANG K Y，LIU J Y，et al. Measuring the innovation of method knowledge elements in scientific literature[J]. Scientometrics，2022，127（5）：2803-2827.

[48] YOON J，KIM K. TrendPerceptor：a property-function based technology intelligence system for identifying technology trends from patents[J]. Expert Systems with Applications，2012，39（3）：2927-2938.

[49] WANG X F，QIU P J，ZHU D H，et al. Identification of technology development trends based on subject-action-object analysis：the case of dye-sensitized solar cells[J]. Technological Forecasting and Social Change，2015，98：24-46.

[50] 韩爽，江屏，牛志伟，等. 基于公理设计的专利规避创新知识挖掘[J]. 计算机集成制造系统，2016，22（6）：1387-1395.

[51] 傅柱，丁玮珂，关鹏，等. 基于知识元的外文专利文献知识描述框架[J]. 数据分析与知识发现，2022，6（Z1）：263-273.

[52] HAN F，MAGEE C L. Testing the science/technology relationship by analysis of patent citations of scientific papers after decomposition of both science and technology[J]. Scientometrics，2018，116（2）：767-796.

[53] BASNET S，MAGEE C L. Artifact interactions retard technological improvement：an empirical study[J]. PLoS ONE，2017，12（8）：e0179596.

[54] 王倩，钱力，刘细文. 知识演化分析的技术方法研究综述[J]. 图书情报工作，2023，67（7）：121-134.

[55] 赵冠壹，韩松花. 科技文献的多粒度知识组织研究[J]. 情报科学，2023，41（8）：134-138，161.

[56] 赵浜，曹树金. 国内外生成式AI大模型执行情报领域典型任务的测试分析[J]. 情报资料工作，2023，44（5）：6-17.

[57] 鲍彤，章成志. ChatGPT中文信息抽取能力测评：以三种典型的抽取任务为例[J]. 数据分析与知识发现，2023，7（9）：1-11.

[58] 梁镇涛，毛进，李纲. 融合“科学-技术”知识关联的高颠覆性专利预测方法[J]. 情报学报，2023，42（6）：649-662.

[59] 杨杰，邓三鸿，王昊. 科学研究的颠覆性创新测度：相对颠覆性指数[J]. 情报学报，2023，42（9）：1052-1064.

作者简介

唐晓波，男，博士，教授，博士生导师，研究方向：知识组织与情报研究，E-mail：xbtang2010@126.com。
陈俭静，男，硕士研究生，研究方向：知识组织与情报研究。
周禾深，男，博士研究生，研究方向：知识组织与情报研究。
杜鑫，女，博士研究生，研究方向：知识组织与情报研究。

Indicators and Measurement Method of Science-Technology Knowledge Linkage Based on Knowledge Elements

TANG XiaoBo[1,2]　CHEN JianJing[2]　ZHOU HeShen[2]　DU Xin[2]
(1. Center for Information System Research, Wuhan University, Wuhan 430072, P. R. China; 2. School of Information Management, Wuhan University, Wuhan 430072, P. R. China)

Abstract: Through the research of science-technology knowledge linkage indicators and measurement method, the interaction between science and technology can be analyzed from a finer granularity level, which lays a foundation for science and technology evaluation. Based on knowledge element theory, this paper proposes science-technology knowledge linkage indicators and measurement method. Taking papers and patents as data sources, based on the extraction of knowledge elements and the co-occurrence of scientific and technological terms in different types of knowledge elements, the indicators of science-technology knowledge linkage are measured. An empirical study is carried out in the field of diabetes. The empirical research results in the field of diabetes show that the science-technology knowledge linkage indicators and measurement method proposed in this paper can play an effective role in the identification of high-quality papers. The research results have significance for improving the evaluation system of science and technology and promoting the implementation of innovation-driven development strategy.

Keywords: Knowledge Element; Knowledge Correlation; Science-Technology Linkage; Indicator; Measurement Method

（责任编辑：王玮）