

基于SPO语义结构的关键词补充主题识别及演化分析*

于诗睿 李爱花 林紫洛 唐小利
(中国医学科学院医学信息研究所, 北京 100005)

摘要: 弥补基于SPO (Subject-Predication-Object) 语义结构进行文本主题识别及演化分析方法中部分主题信息缺失、无法识别新兴领域主题、语义信息不够具体的缺陷, 提升主题识别及演化分析的效果。首先, 抽取科技文献标题和摘要中的SPO语义结构, 使用关键词作为补充进一步丰富语义; 然后, 结合社会网络分析指标、新颖性和相对增长性指标分阶段进行核心主题、新兴主题识别; 最后, 基于文献引用及各阶段核心主题与新兴主题的变化情况进行主题演化趋势分析。分析发现, 基于SPO语义结构的关键词补充主题识别及演化分析方法在以基因编辑领域为代表的较新的领域效果更佳, 3个阶段的核心主题可从技术和应用2个维度概括为ZFN、TALEN, ZFN、TALEN和CRISPR/Cas9, CRISPR/Cas和碱基编辑在基因编辑系统优化、基础科学、临床疾病治疗和生物技术四大方向的应用, 新兴主题主要包括疾病诊断、高通量功能基因组学、合成生物学代谢工程领域、精准医学精准编辑、基因编辑递送工具和基因编辑中的伦理问题六大方向。提出的方法可有效识别研究领域的核心和新兴主题, 并把握其主题演化趋势, 相较于仅基于SPO语义结构的方法, 效果得到提升。

关键词: SPO结构; 文本语义挖掘; 主题识别; 主题演化; 引文

中图分类号: G353.1 **DOI:** 10.3772/j.issn.1673-2286.2023.06.002

引文格式: 于诗睿, 李爱花, 林紫洛, 等. 基于SPO语义结构的关键词补充主题识别及演化分析[J]. 数字图书馆论坛, 2023 (6) : 13-21.

当今世界正历经新一轮科技革命和产业变革, 新兴技术不断涌现, 学科演变加速, 交叉融合态势加剧。准确识别领域核心研究主题, 前瞻性地发现领域科技发展前沿和新兴热点, 厘清领域主题发展脉络, 把握领域未来发展趋势和演变动向, 对科技战略情报分析、创新资源优化配置等具有至关重要的作用^[1]。此外, 随着各学科快速发展, 基金项目、专利和学术论文等科技文献数据量爆发式增长, 基于科技文献数据的科技创新战略逐渐兴起^[2]。科技文献是记录各学科领域重大成果和最新进展等高质量信息的载体, 已成为分析领

域重要研究方向及可靠数据源^[3]。基于科技文献数据进行有效的核心和新兴主题识别及演化趋势研究, 可以揭示学科领域的研究现状与演化规律, 也可以实现在早期对某一领域某一项研究的潜力进行预测并对其未来研究趋势作出有效预判, 有助于科研人员把握某一具体学科领域发展情况, 跟踪当下的研究热点和前沿动态, 也可以为科研管理机构制定学科发展规划提供有力情报支撑, 以及为企业发展方向提供实际参考, 具有重要现实意义^[4]。本研究考虑基于SPO (Subject-Predication-Object) 三元组结构提供详细具体语义信

收稿日期: 2023-04-26

*本研究得到中国医学科学院医学与健康科技创新工程2021年重大协同创新项目“生物医学文献信息保障与集成服务平台”(编号: 2021-I2M-1-033)资助。

息的特点以及关键词时效性强的特点,整合2种方法,形成互补,提出基于SPO语义结构的关键词补充框架,以对主题进行识别并分析演化趋势。

1 相关研究

主题识别是指从文献中识别蕴含的主题信息,是主题演化分析的前提。而主题演化分析主要围绕主题强度的演化和主题内容的演化两方面,体现主题发展性、动态性和差异性^[5]。当前有关主题识别与演化分析的方法主要分为两类:一类是基于共词分析、引文分析等方法的文献计量方法,另一类是以主题模型为代表的文本挖掘方法。

1.1 主题识别方法

(1) 文献计量方法。基于外部特征指标,较为常用的方法是词频分析法。Chen^[6]使用Kleinberg算法将突发词监测整合到CiteSpace软件中,指出使用突发词可部分展示某领域研究的新兴主题。引文分析旨在通过统计分析文献间的引用与被引用关系,解释文献交流的特点和规律,从而探索学科发展趋势,按照引文模式可分为基于直接引用的主题分析、基于引文耦合的主题分析以及基于共被引的主题分析。Small等^[7]基于直接引用网络聚类结果的可选择性強和共被引网络聚类结果的精简度高的优点,组合构建了一种混合型新兴主题识别方法。在基于共现关系的网络分析方法中,共现关系多以关键词共现建立起来,并基于聚类关键词划分领域主题。You等^[8]通过专利类别的共现关系构建共现网络,从动态知识块转移和创新知识源演化2个角度识别快速增长的知识块,判定技术发展机遇。

(2) 文本挖掘方法。基于语义分析的文本挖掘方法与传统的主题分析方法相比引入了“语义”维度,通过从文本中发掘潜在内容特征和语义关联关系,弥补外部特征语义表达不足的缺陷,从而提升主题识别与演化分析的准确性。主题识别的方法以主题模型为主,还有一些词向量表示及深度学习的方法。其中,以LDA(Latent Dirichlet Allocation)为基础的主题模型是当前较流行的用于解释和分析主题分布的方法,马温如等^[9]使用LDA主题模型并结合熵权法,对领域新兴主题进行有效识别,但该方法存在语义表达不清晰、主题辨识难度大的问题。为解决上述问题,Yoon等^[10]提出了

一种基于SAO(Subject-Action-Object)结构的主题识别模型,该方法能够以概念及其之间的对应关系对学科主题进行描述,提供了广泛的语义信息和更精细的语义结构,实现对语义信息的更深层次揭示,使研究人员能够在此基础上进行详细准确的主题分析、趋势预测和相似度测量。还有一些利用词向量表示、深度学习模型等实现文献主题抽取的方法,如沈思等^[11]引入主题词嵌入表示(Topical Word Embeddings, TWE)模型,联合学习单词和主题的向量表示,从而学习不同主题下词向量的不同表示,提升了医学科技报告主题的挖掘精度。

1.2 主题演化分析方法

(1) 文献计量方法。基于主题词的分析方法主要包括词频分析法和共词分析法。通过对主题词频次、新增词汇、词间关系等方面变化的追踪,在一定程度上反映某一领域主题演变的情况。Havre等^[12]根据词汇出现的频率,用主题河流图(ThemeRiver)的方式可视化研究主题随时间的发展情况,分析主题的发展趋势。基于引文的分析方法利用文献的外部特征,依据文献间的引证关系建立引文网络,通过网络及其动态变化识别学科的热点与演化趋势,包括共被引分析、主路径分析等。Shibata等^[13]利用社团结构探测算法对科学引文网络聚类分析,根据聚类结果中论文簇的平均年指标和不同时间片的论文簇间的父子关系探测领域的新兴主题,并绘制论文簇演变的时间线图。

(2) 文本挖掘方法。主题模型作为重要的文本挖掘工具,已得到广泛应用,如秦晓慧等^[14]使用LDA主题关联过滤规则分析相邻时间窗口间的主题演化关系,实现对免疫领域主题演化轨迹的识别。此外,文本聚类、词嵌入以及深度学习等技术也逐渐应用于演化分析的研究,在文本特征提取和语义表达方面具有巨大优势,能够提高主题演化分析的准确性和深入性。张金柱等^[15]提出了基于引用内容聚类的文献被引主题识别及演化分析框架,并选取了人工智能领域进行实证分析,研究结果表明该方法能够揭示主题时序变化的特征与现实意义,使结果更加科学准确。

近年来,结合自然语言处理技术,提出了基于文本识别SPO三元组知识的方法,该方法能够揭示更深层次的语义信息,能有效识别领域主题。但该方法也存在以下不足:SPO语义结构信息覆盖不全面,导致重要主

题信息缺失; SPO三元组知识存在时滞性, 无法识别新兴主题; SPO语义结构揭示信息不够具体, 无法识别细分领域主题。而论文的关键词具有实时更新、反映具体语义信息、表达主要研究内容的特点, 是识别新兴主题的有效依据, 但包含的语义信息有限, 难以准确解释主题信息^[16]。鉴于此, 本研究提出使用关键词补充SPO三元组语义信息, 并基于引文和主题变化情况实现新兴、核心主题识别与演化趋势分析的方法, 以期提升主题识别与演化分析效果。

2 研究框架及方法

2.1 研究框架

首先, 分别基于从科技文献标题和摘要中抽取的SPO语义结构以及文献的关键词两类数据进行研究; 其次, 对所研究领域的演化发展阶段进行划分; 最后, 识别领域核心主题及新兴主题并分析其演化趋势。具体研究框架如图1所示。

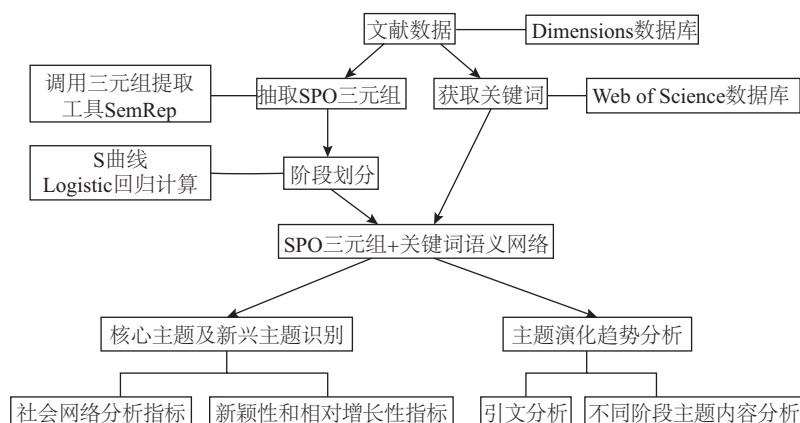


图1 研究框架

2.2 研究方法

2.2.1 数据来源

以文献数据为数据源。鉴于Dimensions数据库收录的数据更为全面, 从Dimensions数据库获取文献数据集, 并基于标题、摘要抽取SPO三元组语义结构。由于Dimensions数据库直接提供的题录数据无关键词信息, 从Web of Science数据库中获取关键词, 舍弃Web of Science数据库未收录的文献。

2.2.2 数据处理

首先, 利用来自Dimensions数据库的数据, 使用基于SPO语义结构的主题识别方法进行分析。使用美国国立医学图书馆(National Library of Medicine, NLM)开发的SemRep工具抽取SPO语义三元组, 并基于Fiszman等^[17]提出的SPO清洗原则筛选出能够表征文献蕴含主要语义信息的SPO。利用Web of Science数据库中DE字段提取的关键词信息, 结合2个数据库共同包含的PMID(PubMed Unique Identifier)和DOI

字段完成文献的对应, 实现关键词对基于SPO语义结构的主题语义信息的补充, 使用Cytoscape软件构建在SPO语义结构基础上补充关键词的语义网络。

2.2.3 发展阶段划分

基于各年份标准化的SPO结构频次累积值, 使用Logistic回归模型拟合, 实现对学科领域发展阶段的划分^[18]。从科技文献中抽取的SPO结构累积频次能够反映学科领域的进展, 因此通过逐年计算SPO结构频次累积值来测度领域进展状况。具体计算公式如式(1)~(2)所示。

$$\sigma_i = \frac{N_i - N_{i-1}}{N_{i-1}} \quad (1)$$

$$\overline{N}_i = \overline{N}_{i-1} \times (1 + \sigma_i) \quad (2)$$

式中: σ_i 为截至第*i*年的SPO结构频次累积值实际增长率; N_i 为截至第*i*年的实际SPO结构频次累积值; $\overline{N}_i$ 为截至第*i*年的标准化SPO结构频次累积值; $i \geq 2$, $\sigma_1 = 0$, $\overline{N}_1 = 1$ 。使用Logistic回归模型对标准化的SPO结构频次累积值进行拟合, 构造S曲线, 从而确定发展阶段。使用Loglet Lab软件完成S曲线的拟合, 根据基因编

辑领域的发展特点划分萌芽期、快速发展期和成熟平稳期3个阶段。

2.2.4 核心主题及新兴主题识别

结合社会网络分析指标来识别各阶段核心主题。通过分阶段构建有向加权的SPO语义网络,使用节点度和边权重2个指标,识别出排名靠前的关键节点及其相邻子网络,再识别出SPO对应文献的高频关键词,将上述结果确定为各阶段的核心主题。在SPO语义网络中,使用SPO结构频次与相对应的关系重要性的乘积作为边权重,其中某一类型关系的重要性为该关系出现的频次与所有关系出现总频次之比。

使用新颖性和相对增长性2个指标识别新兴主题。新颖性的计算公式如式(3)所示。

$$K = \frac{1}{q} \sum_{r=1}^q T_r \quad (3)$$

式中: K 为主题的新颖性值; T_r 为主题内第 r 篇文献的发表/授权时间; q 为主题内文献的数量。主题增长率的计算公式如式(4)所示。

$$Z_i = \frac{A_i - A_{i-1}}{A_i} \quad (4)$$

式中: Z_i 为某一主题在第 i 年的增长率; A_i 为第 i 年的主题文献数量。采用近3年主题的平均增长率来表示主题的相对增长性^[9]。将每个SPO和关键词看作一个主题,分别计算新颖性值和相对增长性值,将由新颖性值和相对增长性值较大的SPO和关键词形成的主题确定为新兴主题。

2.2.5 主题演化趋势分析

基于文献引用以及各阶段核心主题和新兴主题变化情况对基因编辑领域的主题演化趋势进行探究,分析主题的新生、成长、衰减、合并、分裂及消亡情况^[19]。具体方法是将Dimensions与Web of Science数据库中的文献与其参考文献关联起来,根据后一阶段主题包含的SPO和关键词对应文献寻找上一阶段对应的参考文献,分析参考文献包含的SPO和关键词对应的主题,并计算主题下SPO和关键词的占比,依据占比较大的主题确定主要演化路径,从而建立主题演化关联,判断主题的新生、合并、分裂、消亡情况。根据各阶段核心主题和新兴主题的变化情况对主题演化趋势进行补充,

根据上一阶段的新兴主题是否转变为下一阶段核心主题判断主题的成长和衰减情况,从而实现主题演化趋势的分析。

3 实证分析

3.1 实证领域介绍

近年来基因编辑技术掀起了研究的热潮,发展迅速,在农业育种、工业生产、生物医学等领域均取得了新的发展和应用,显示出巨大潜力,推动了各领域的发展完善,是生命科学领域具有颠覆性的先进技术之一。因此,选择基因编辑领域这一较新的领域开展实证分析,通过该领域的核心主题、新兴主题以及主题演化趋势进行探究,梳理领域研究进展,以增进对该技术与未来研究方向的整体认识^[20]。

3.2 数据来源与处理

以“gene editing”“genetic editing”“genome editing”“Clustered regularly interspaced short palindromic repeat”“CRISPR”“transcription activator like endonuclease”“TALEN”“Zinc finger nuclease”“ZFNs”为检索词在Dimensions和Web of Science数据库中检索,分别得到47 083篇和42 557篇文献,识别出SPO结构191 874个、获取关键词121 172个用于后续分析。根据S曲线的拟合情况将基因编辑领域主题演化阶段分为2012年及以前、2013—2017年和2018年至今3个时间段,其可视化的拟合结果如图2所示。

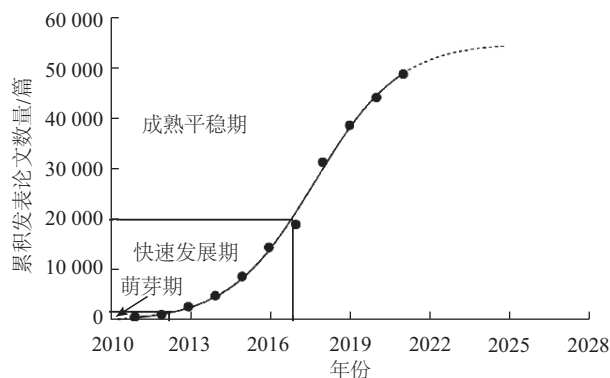


图2 领域主题S曲线图

3.3 主题识别结果

3.3.1 基于SPO语义网络的主题识别结果

基于SPO语义网络进行主题识别, 筛选语义网络中节点度排名前10的节点及边权重排名前50的连接点所构成的子网络, 将其确定为基因编辑领域各阶段的核心主题,

发现该领域的核心主题主要包含基因编辑系统使用酶、临床疾病治疗、干细胞疗法、模式动物、基因编辑工具、基因编辑体系递送和基因编辑技术分子机制。各阶段核心主题代表性SPO结构如表1所示。分析发现, 基于SPO语义网络识别出的主题虽然能够揭示领域主题内容表达的细粒度语义信息, 但存在信息内容不全面、新兴主题缺乏以及表达语义粒度较大、难以揭示领域细分方向的不足。

表1 基于SPO语义网络识别出的核心主题及其代表性SPO结果

核心主题	SPO		
	第一阶段	第二阶段	第三阶段
基因编辑系统使用酶	Targeted Gene Repair USES nuclease	Transcription Activator-Like Effector Nucleases INTERACTS_WITH Conserved Sequence	Uracil-DNA glycosylase AFFECTS Base Excision Repair
临床疾病治疗	Hereditary Diseases PROCESS_OF Homo sapiens	Disease ASSOCIATED_WITH Ubiquitination	Clustered Regularly Interspaced Short Palindromic Repeats DIAGNOSES Malignant Neoplasms
干细胞疗法	Cells LOCATION_OF DNA Repair	Pluripotent Stem Cells INTERACTS_WITH Transcription Factor	Cells LOCATION_OF tumor growth
模式动物	Germ-Line Mutation PROCESS_OF Rattus norvegicus	Mutagenesis, Insertional PROCESS_OF Mus	RNA Interference PROCESS_OF Mammals
基因编辑工具	Zinc Finger Nucleases PART_OF Engineer	Transcription Activator-Like Effector Nucleases compared_with Clustered Regularly Interspaced Short Palindromic Repeats	Clustered Regularly Interspaced Short Palindromic Repeats INTERACTS_WITH Genes
基因编辑体系递送	Transfection USES Electroporation	Probes USES Artificial nanoparticles	In Vivo Model USES Xenograft procedure
基因编辑技术分子机制	Nickase AUGMENTS Non-Homologous DNA End-Joining	DNA INTERACTS_WITH Transposase	DNA Damage PREDISPOSES mitochondrial toxicity

3.3.2 基于关键词的主题识别结果

基于关键词进行主题识别, 筛选各阶段频次排名前20的关键词, 归纳整理后形成基因编辑领域各阶段的核心主题。各阶段核心主题及代表性关键词如表2所示。分析发现, 通过使用关键词进行主题识别, 能更好地识别出领域的新兴主题, 但与SPO三元组相比, 其语义信息的表达有所欠缺, 主题内容辨识度不高、可解释性不强。

主要核心主题及语义网络中代表性SPO和关键词节点如表3所示。通过在SPO三元组基础上补充关键词对领域主题进行识别, 不仅能详细揭示领域重要研究内容的深层次细节, 还能了解该领域的最新研究方向, 故考虑将2种方法结合进行主题识别。

3.3.3 基于SPO语义网络补充关键词的核心主题及新兴主题识别结果

使用社会网络分析指标识别出基因编辑领域各阶段的核心主题。将3个阶段节点度排名前10的节点及其第一连接点所构成的语义子网络, 以及与之对应的排名前20的高频关键词确定为该领域的核心主题。各阶段

根据识别出的SPO和补充的关键词, 进一步归纳基因编辑领域的核心主题, 将基因编辑技术和应用2个维度相结合, 实现对领域主题内容的清晰表示。各阶段的核心主题及其细分领域的具体情况如图3所示。在基因编辑技术维度, 第一阶段核心主题主要为ZFN和TALEN, 第二阶段为ZFN、TALEN和CRISPR/Cas9, 第三阶段为CRISPR/Cas和碱基编辑。在应用维度, 核心主题主要概括为基因编辑系统优化、基础科学、临床疾病治疗和生物技术四大方向。

在临床疾病治疗领域, 第一阶段的基因编辑技术研究内容主要为基因疗法 (gene therapy TREATS Disease、gene correction)。可使用基因敲入、敲除和

表2 基于关键词识别出的核心主题及代表性关键词结果

第一阶段		第二阶段		第三阶段	
主 题	关键词	主 题	关键词	主 题	关键词
基因编辑	gene editing	动物模型	disease modeling	癌症免疫疗法	adoptive T cell therapy
核酸酶结构优化	custom-designed zinc finger nucleases	工程核酸酶	CRISPR/Cas9	功能基因组学	whole-genome sequencing
脱靶效应	off target activity	表观遗传学	epigenome editing	合成生物和谢工程	synthetic biology
基因疗法	gene therapy	基因疗法	targeted mutagenesis	再生发育	development
细胞疗法	induced pluripotent stem cells	递送体系	delivery vehicle	异种器官移植	xenotransplantation
基因调节	gene repair	干细胞移植	cell therapy	疾病诊断	diagnosis
动物模型	animal models	生物制药	drug screening	抗性耐药性	drug resistance
生物制药	drug delivery	动植物育种	transgenic plants	基因疗法	point mutation

表3 基于SPO语义网络补充关键词识别出的主要核心主题及代表性SPO和关键词节点

主 题	第一阶段		第二阶段		第三阶段	
	细分主题	代表性SPO关键词	细分主题	代表性SPO和关键词	细分主题	代表性SPO和关键词
基因编辑技术	ZFN、TALEN	Zinc Finger Nucleases PART_OF Engineer、Homologous recombination	ZFN、TALEN、CRISPR/Cas9	Transcription Activator-Like Effector Nucleases compared_with Clustered Regularly Interspaced Short Palindromic Repeats、CRISPR/Cas9	CRISPR/Cas、Base editing	Cytosine base editor (CBE)、dCas9
基因编辑系统优化	核酸酶结构的优化	Nuclease PART_OF Engineer、Custom-designed zinc finger nucleases	(CRISPR/Cas9) 提高精确修复效率	Homology-directed repair、Precise gene editing	基因编辑引起毒性等副作用	Virulence、Virulence AFFECTS DNA Sequence Rearrangement
	提高基因编辑效率	Zinc Finger Nucleases ASSO-CIATED WITH genotoxicity、Green fluorescent protein	减轻脱靶现象	Off-target、sgRNA	核酸酶传递效率和安全性问题研究	Designer nucleases、Delivery
基础科学	产生基础科学研究新方法	Translocation、DNA Repair	动物模型构建	Zebrafish、Somatic cell nuclear transfer (SCNT)、Microinjection	异种器官移植、再生及发育问题	Xenotransplantation、Regeneration
			表观遗传学研究	Epigenetic Process AFFECTS Gene Expression、epigenetic regulation	微生物基因表达调控	Gene expression、Microorganism
临床疾病治疗	基因疗法	Gene therapy、Gene correction	缓解基因编辑工具引起的免疫反应	Clustered Regularly Interspaced Short Palindromic Repeats AFFECTS Immunity innate immunity、Immune rejection	使用免疫疗法对癌症治疗	Immunotherapy、Chimeric Antigen Receptor T-Cell Therapy TREATS Hematologic Neoplasms
	细胞疗法	Induced pluripotent stem cells、Regeneration	应用基因疗法治疗遗传病	Mutation CAUSES Hereditary Diseases	造血干细胞移植	Stem cell transplant TREATS Patients、Cell therapy
生物技术	生物制药	Pluripotent Stem Cells PART_OF Animals、drug delivery	药物研发	cell line engineering、Pluripotent Stem Cells PART_OF Animals	动植物育种	Animal breeding、Resistance Process PROCESS_OF Plants
	农业育种作物改良	DNA Restriction Enzymes INTERACTS WITH Plant Cells、Transgenic crops	农业	Angiosperm LOCATION_OF Genes、Chloroplast	耐药性问题研究	Multi drug resistant bacteria CAUSES Infection、Drug resistance

替换的治疗策略来解决传统基因疗法的局限性问題, 基因疗法广泛应用于X染色体连锁严重联合免疫缺陷症(SCID-X1)、镰刀状细胞贫血等疾病的治疗, 可以有效降低外源基因随机整合产生的安全隐患, 提高治疗的可靠性。第二阶段的研究内容主要包含体内递送工具研究(delivery method USES Genes、delivery vehicle)、应用基因疗法治疗遗传病(Mutation CAUSES Hereditary Diseases、Chromosomal rearrangement)和解决基因编辑工具引起的免疫反应(Clustered Regularly Interspaced Short Palindromic Repeats AFFECTS Immunity、immune rejection)。体内基因编辑治疗的发展受到递送载体导致的免疫排斥反应的制约, 引起广泛研究。目前, 体内递送工具的研究主要使用病毒和质粒传递基因编辑的DNA序列, 其中AAV(Adeno-Associated Virus)载体是目前常用的病毒载体, 非病毒载体尤其是纳米颗粒也是有效的递送工具, 但其均存在一定局限性, 仍待改进和深入研究。应用基因疗法治疗遗传病的研究通过引入突变点、基因删除、基因插入等方式从根本上治愈疾病, 目前已在治疗神经系统疾病、代谢病等遗传病上取得成功。在缓解基因编辑工具引起的免疫反应研究中, 基因编辑使用的酶大部分来自非人类物种, 人体会对外源蛋白产生免疫, 可利用来自非人类相关微生物的新Cas直系同源物或在一定条件下修饰所用Cas蛋白的抗原表位, 绕过体液和细胞免疫障碍的方式解决此类问题。第三阶段的研究内容主要包含造血干细胞移植

(Haploidentical Hematopoietic Cell Transplantation TREATS Patients、lentivirus)、使用免疫疗法对癌症治疗(Adoptive Immunotherapy TREATS Malignant Neoplasms、T cells)、诱导多能干细胞与基因编辑技术结合(Induced Pluripotent Stem Cells LOCATION_OF Mutation、disease-specific induced pluripotent stem cells)等。在造血干细胞移植的研究中, 体外基因治疗主要采用造血干细胞作为靶点, 使用逆转录病毒或慢病毒颗粒将健康供体基因递送至患者的免疫细胞或造血干细胞上, 然后移植回患者体内, 从而补充免疫系统构建所需的蛋白质, 此方法已成功用于治疗 β -地中海贫血、镰状细胞病等。基于核酸酶的基因编辑工具可实现成功治疗, 碱基编辑器也发挥了重要作用, 且效果更优。在使用免疫疗法对癌症治疗的研究中, 基因编辑技术为了解癌症生物学特征和分子机制、开发新的临床前模型、制定针对性癌细胞消除策略方面的进展提供了潜在机会。目前大多数临床试验使用CRISPR/Cas9系统治疗肺癌、乳腺癌等癌症, 且大多数关注T细胞基因工程、癌症免疫疗法或靶向癌细胞特定基因的方法。在诱导多能干细胞与基因编辑技术结合的研究中, 诱导多能干细胞与基因编辑技术的结合取得了一定研究成果, 已在多个领域有所应用, 包括通过校正致病突变实现疾病细胞治疗, 用于疾病建模药物开发等, 从而指导疾病致病机制研究和疾病治疗。

使用新颖性和相对增长性2个指标识别当下新兴主题, 分析这2个指标排名前10的SPO和排名前50的

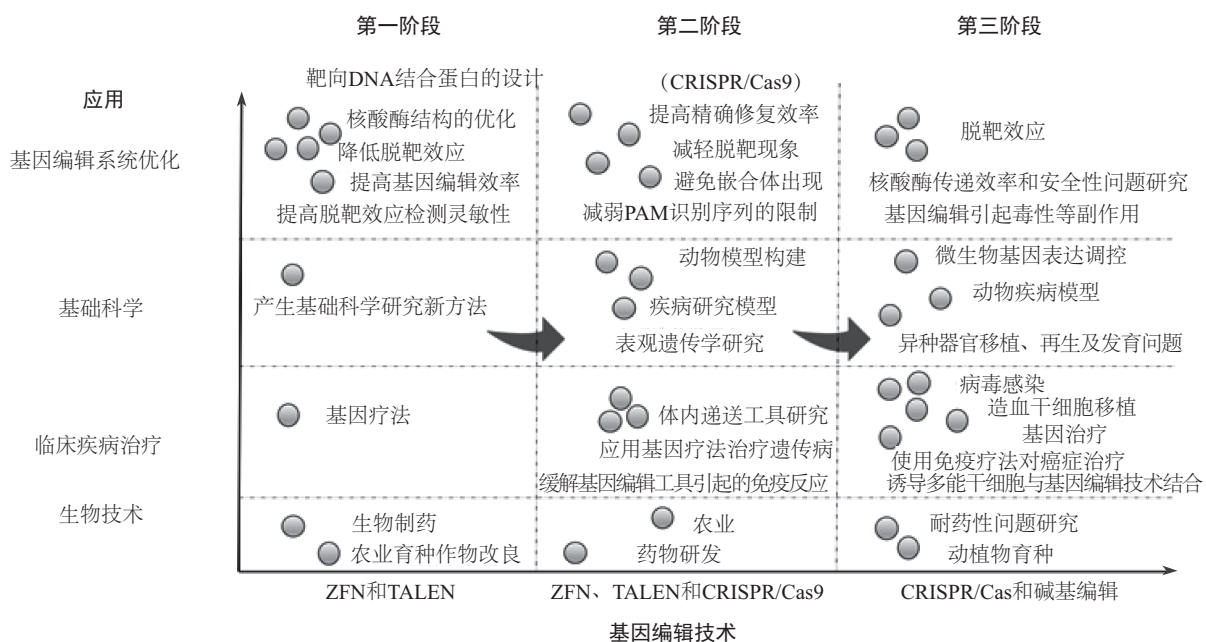


图3 各阶段核心主题及其细分领域

关键词,得出以下新兴主题。①疾病诊断(diagnosis、COVID-19、SARS-CoV-2、Whole Exome Sequencing METHOD_OF Diagnosis),包含核酸检测和CRISPR技术在COVID-19中的应用2个细分领域。②高通量功能基因组学(LncRNA、microRNA、Functional genomics、Mutagenesis Procedure METHOD_OF functional genomics),分为非编码转录物功能研究、元件功能映射、编码基因功能残基研究等细分领域。③合成生物学代谢工程领域(synthetic biology、metabolic engineering、saccharomyces cerevisiae)。④精准医学、精准编辑(Transposon、prime editing、precision medicine、precision editing)。⑤基因编辑递送工具(gene delivery、nanotechnology、nanoparticles、delivery method USES Genes, Viral)。⑥基因编辑中的伦理问题(bioethics)。

3.4 基于SPO语义网络补充关键词的主题演化趋势结果分析

基于引用的方法对基因编辑领域的主题演化趋势进行探究,对主题的新生、合并、分裂、消亡等发展趋势进行剖析,如:第一阶段的动物模型、基因疗法和基因编辑3个领域融合为第二阶段的干细胞移植领域,体现了主题合并;第二阶段的动物模型构建领域在第三阶段发展的基础上又分裂为脱靶效应、再生及发育问题、疾病诊断和异种器官移植4个领域,体现了主题的分裂;第二阶段产生了表观遗传学、功能基因组学等新的研究领域,在第三阶段,第二阶段中的分子生物学、工程核酸酶等消失,体现了主题的新生与消亡;基因疗法、脱靶效应等贯穿了3个阶段,体现主题在不断发展演变。

基因疗法领域由基因编辑工具发展、脱靶效应、干细胞移植、动物模型、基因组测序等领域发展而来。该领域在临床试验中应用动物疾病模型开发疾病治疗策略^[21],关注排除基因治疗引起的免疫排斥反应、致癌风险、脱靶效应、外源DNA整合等危险因素,研发治疗载体以提高递送效率、减少不良反应,研发替代的基因编辑工具以提升编辑精准性,通过对造血干细胞、诱导多能干细胞、T细胞等进行编辑实现干细胞移植治疗疾病,通过基因组学的研究实现致病基因功能探索,通过基因治疗药物研发实现疾病根治^[22]。这些领域的发展将为基因治疗带来全新的局面。

根据各阶段核心主题和新兴主题的变化情况分析,

发现第一阶段的CRISPR/Cas基因编辑体系、通过人工设计核酸酶改进基因编辑工具、治疗性基因编辑中的干细胞移植、递送工具优化和动物疾病模型研究领域得到进一步发展,第二阶段的CRISPR/Cas9基因编辑技术、表观遗传学研究、动物模型构建、高通量功能基因组学研究、治疗性基因编辑中的癌症治疗及干细胞移植、生物技术中的农业生物制药、减轻脱靶效应引起的副作用领域得到深入研究,其余领域的关注度有所减弱。

4 结语

仅基于SPO语义结构识别主题存在识别主题信息缺失、无法识别领域新兴主题、语义信息不够具体的问题,而关键词能够体现领域研究成果的重点内容,且能够及时揭示领域诞生的新兴主题,但在语义揭示方面有所欠缺。提出了基于SPO语义网络的关键词补充主题识别及演化分析方法,并选择基因编辑领域进行实证分析。该方法使用关键词作为补充,可以有效解决SPO语义结构涵盖信息不全、重要主题信息缺失, SPO语义结构存在时滞性、无法实时更新,以及SPO语义结构反映语义信息不够具体、细分领域主题识别存在局限性的问题;实现多维度、细粒度描述领域主题,并发现领域内某个时期的高价值新兴主题,提高主题识别及演化趋势分析结果的可靠性,有助于对学科领域各阶段主题研究情况和发展规律的探索,尤其在较新的领域中能够取得更为显著的效果^[23]。

本研究也存在一定局限性:一是只使用了文献数据,数据来源较为单一,提取出的主题不够全面,未来可考虑使用专利、基金等类型的数据作为补充;二是基于引用关系探究主题演化趋势的方法在揭示复杂动态主题演化方面有所欠缺,后续可考虑使用相似度、关联过滤等主题关联计算的方法建立主题间的演化关系。

参考文献

- [1] 刘盼盼. 基于多源数据和共现网络的新兴技术主题识别研究[D]. 北京:中国科学院文献情报中心, 2022.
- [2] 刘春江, 刘自强, 方曙. 基于SAO的技术主题创新演化路径识别及其可视化研究[J]. 情报学报, 2023, 42(2): 164-175.
- [3] 徐路路, 王效岳, 白如江. 基于PLDA模型与多数据源融合相关性分析的新兴主题探测研究:以石墨烯领域为例[J]. 情报理论与实践, 2018, 41(4): 63-69, 43.
- [4] 杨恒, 王曰芬, 张露. 基于核心专利技术主题识别与演化分析的

- 技术预测[J]. 情报杂志, 2022, 41 (7): 49-56.
- [5] 梁爽, 刘小平. 基于文本挖掘的科技文献主题演化研究进展[J]. 图书情报工作, 2022, 66 (13): 138-149.
- [6] CHEN C M. CiteSpaceII: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57 (3): 359-377.
- [7] SMALL H, BOYACK K W, KLAIVANS R. Identifying emerging topics in science and technology[J]. Research Policy, 2014, 43 (8): 1450-1467.
- [8] YOU H L, LI M J, JIANG J, et al. Evolution monitoring for innovation sources using patent cluster analysis[J]. Scientometrics, 2017, 111 (2): 693-715.
- [9] 马温如, 赵钦, 滕珺. 科学技术新兴技术主题识别研究[J]. 中华医学图书情报杂志, 2021, 30 (12): 1-10.
- [10] YOON J, PARK H, KIM K. Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis[J]. Scientometrics, 2013, 94 (1): 313-331.
- [11] 沈思, 李沁宇, 叶媛, 等. 基于TWE模型的医学科技报告主题挖掘及演化分析研究[J]. 数据分析与知识发现, 2021 (3): 35-44.
- [12] HAVRE S, HETZLER E, WHITNEY P, et al. ThemeRiver: visualizing thematic changes in large document collections[J]. IEEE Transactions on Visualization and Computer Graphics, 2002, 8 (1): 9-20.
- [13] SHIBATA N, KAJIKAWA Y, TAKEDA Y, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications[J]. Technovation, 2008, 28 (11): 758-775.
- [14] 秦晓慧, 乐小虬. 基于LDA主题关联过滤的领域主题演化研究[J]. 现代图书情报技术, 2015 (3): 18-25.
- [15] 张金柱, 仇蒙蒙, 王秋月. 基于引用内容聚类的文献被引主题识别及其演化分析[J]. 情报科学, 2023, 41 (2): 107-117.
- [16] 刘小玲, 谭宗颖. 新兴技术主题识别方法研究进展[J]. 图书情报工作, 2020, 64 (11): 145-152.
- [17] FISZMAN M, RINDFLESCHE T C, KILICOGLU H. Abstraction summarization for managing the biomedical research literature[C]//Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, 2004: 76-83.
- [18] 张琬笛, 胡志刚, 郭佳程, 等. 基于S曲线的研究主题演进分析与可视化技术[J]. 信息资源管理学报, 2020, 10 (3): 70-77, 101.
- [19] 刘倩楠. 基于专利引文网络的技术演进路径识别研究: 以“以太网”技术为例[D]. 大连: 大连理工大学, 2010.
- [20] 马梦遥, 张欣欣, 刘永林, 等. 基因编辑技术的演变及应用[J]. 国际遗传学杂志, 2022, 45 (3): 176-184.
- [21] MEMI F, NTOKOU A, PAPANGELI I. CRISPR/Cas9 gene-editing: research technologies, clinical applications and ethical considerations[J]. Seminars in Perinatology, 2018, 42 (8): 487-500.
- [22] JANIK E, NIEMCEWICZ M, CEREMUGAM, et al. Various aspects of a gene editing system: CRISPR-Cas9[J]. International Journal of Molecular Sciences, 2020, 21 (24): 9604.
- [23] 马铭, 王超, 周勇, 等. 基于语义信息的核心技术主题识别与演化趋势分析方法研究[J]. 情报理论与实践, 2021, 44 (9): 106-113.

作者简介

于诗睿, 女, 硕士研究生, 研究方向: 科学计量学。

李爱花, 女, 博士, 副研究员, 研究方向: 科学计量学。

林紫洛, 女, 硕士研究生, 研究方向: 科学计量学。

唐小利, 女, 硕士, 研究馆员, 通信作者, 研究方向: 医学情报研究与信息咨询, E-mail: tang.xl@imicams.ac.cn。

Keywords Supplementary Topic Recognition and Evolution Analysis Based on SPO Structure

YU ShiRui LI AiHua LIN ZiLuo TANG XiaoLi

(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005, P. R. China)

Abstract: This paper aims to remedy the defects of the original topic recognition and evolution analysis method based on SPO (Subject-Predication-Object) structure, such as lack of important topic, unable to identify topics in emerging fields, and unspecific semantic information, in order to improve the effect of topic recognition and evolution analysis. First, the semantic structure of SPO is extracted from the titles and abstracts of literature, and the keywords are used as supplements to further enrich the semantic. Then, the social network analysis indicators and novelty and relative growth indicators are used to identify the core topics and emerging topics by stages. Finally, the topic evolution trend is analyzed based on the references and the changes of core themes and emerging themes in each stage. It is found that the method of keyword supplementary topic recognition and evolution analysis based on SPO semantic structure is suitable for the emerging new fields represented by gene editing. The core themes of the three stages can be summarized from the two dimensions of technology and application as ZFN and TALEN, ZFN, TALEN, and CRISPR/Cas9, CRISPR/Cas and base editing in gene editing system optimization, basic science, clinical disease treatment, and biotechnology. Emerging topics mainly include disease diagnosis, high-throughput functional genomics, metabolic engineering in synthetic biology, precision editing in precision medicine, gene editing delivery tools, and ethical issues in gene editing. The proposed method can effectively identify the core and emerging topics in the research field, and grasp the evolution trend of their themes. Compared with the method based only on the semantic structure of SPO, the effect is improved.

Keywords: SPO Structure; Text Semantic Mining; Topic Recognition; Theme Evolution; Citation

(责任编辑: 王玮)