现代情报
Journal of Modern Information
ISSN 1008-0821,CN 22-1182/G3

《现代情报》网络首发论文

题目：基于多源数据弱信号分析的早期新兴研究主题识别
作者：唐超，许海云，杨俊浩，谭晓，刘春江
网络首发日期：2025-08-12
引用格式：唐超，许海云，杨俊浩，谭晓，刘春江．基于多源数据弱信号分析的早期新兴研究主题识别[J/OL]．现代情报．
https://link.cnki.net/urlid/22.1182.g3.20250812.1133.002

网络首发时间：2025-08-12 16:48:12　网络首发地址：https://link.cnki.net/urlid/22.1182.g3.20250812.1133.002

基于多源数据弱信号分析的早期新兴研究主题识别

Early Identification of Emerging Research Topics through Weak Signal Analysis of Multi-Source Data

唐超[1]　许海云[1]　杨俊浩[1*]　谭晓[2]　刘春江[3]

Tang Chao[1]　Xu Haiyun[1]　Yang Junhao[1*]　Tan Xiao[2]　Liu Chunjiang[3]

（1. 山东理工大学管理学院，山东 淄博，255000；2. 北京市科学技术研究院科技情报研究所，北京，100089；3. 中国科学院成都文献情报中心，四川 成都，610213）

(1. Business School, Shandong University of Technology, Zibo 255000, China; 2. Science and Technology Information Institute, Beijing Academy of Science and Technology, Beijing 100089, China; 3. National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610213, China)

摘　要：[目的/意义] 从新兴研究主题早期特征和弱信号的特性出发，通过多源数据的弱信号分析方法实现新兴研究主题的早期识别。[方法/过程] 首先，采用专利、临床、新闻和论文数据4种类型数据，利用BERTopic模型获取主题，构建新兴度综合指标识别新兴研究主题；其次，基于主题可见度和主题扩散度，构建主题涌现图、主题分配图，识别弱信号新兴研究主题，并在多源数据交叉验证下，测度其潜在影响力；最后，使用干细胞领域开展领域实证分析。[结果/结论] 实证发现，本文识别的弱信号新兴研究主题较其他类型主题更符合权威报告、权威期刊论文、专业学术指南的技术方向，具有较强的跨领域影响力。本文构建的基于多源数据弱信号分析的早期新兴研究主题识别方法，可以实现新兴研究主题的早期识别，且通过多源数据整合提升了识别的准确性与覆盖度。

关键词：新兴研究主题；弱信号；BERTopic模型；早期识别；多源数据

分类号：G350；G353

Abstract: [Purpose/Significance] Starting from the early features of emerging research topics and the characteristics of weak signals, this study aims to achieve early identification of emerging research topics through weak signal analysis based on multi-source data.[Methods/Process] First, four types of data sources—patents, clinical data, news articles, and academic papers—are utilized to extract topics using the BERTopic model and to construct a composite indicator of “emergence” for identifying emerging research topics. Next, based on topic visibility and topic diffusion, topic emergence maps and topic allocation maps are constructed to identify emerging research topics characterized by weak signals. Under the framework of multi-source data cross-validation, the potential influence of these topics is assessed. Finally, an empirical analysis is conducted in the field of stem cells.[Results/Conclusions] The empirical findings indicate that the weak signal-based emerging research topics identified in this study align more closely with the technological directions outlined in authoritative reports, high-impact journal publications, and specialized academic guidelines compared to other types of topics, and they exhibit strong cross-domain influence. The proposed early identification method for emerging research topics,

基金项目：国家自然科学基金项目“基于弱信号时效网络演化分析的变革性科技创新主题早期识别方法研究”（项目编号：72274113）；山东省自然科学基金“基于弱信号分析的变革性创新主题早期识别方法研究”（项目编号：ZR202111130115）；山东省泰山学者工程“变革性科技创新：动因解析、早期识别与预测”（项目编号：tsqn202103069）。

作者简介：唐超（2000-），男，硕士研究生，研究方向：科技情报分析。**许海云**（1982-），女，教授，博士，博士生导师，研究方向：科技与产业情报分析。**谭晓**（1983-），女，博士，副研究员，研究方向：文本挖掘，科学计量，安全情报研究。**刘春江**（1984-），男，博士，高级工程师，研究方向：科技文献挖掘。

通信作者：杨俊浩（2001-），男，硕士研究生，研究方向：科技情报分析。

based on weak signal analysis of multi-source data, not only enables early detection but also enhances the accuracy and coverage of topic identification through integrated multi-source data.

Key words: emerging research topics; weak signals; BERTopic model; early detection; multi-source data

从欧盟“2020 地平线”计划到中国“十四五”国家科技创新规划，再到美国“国家科学基金会战略计划”，科技创新日益成为国家关注的重点。在各国（地区）政策的推动下，新一轮科技革命加速演进，技术涌现频率加快，但社会资源有限，亟需对投入的技术方向进行准确筛选。新兴研究代表了一项技术的最新动态或一项科学的前沿，其创新程度较高，是科学发展过程中最具前瞻性的研究课题，可能带来巨大的科学进步[1]。从海量科技文献中识别新兴研究主题，对科技政策的制定以及企业的战略布局至关重要，现已成为科技情报工作者亟待解决的研究课题。科技情报领域学者常使用德尔菲法、文献计量学以及机器学习等方法识别潜在的新兴研究主题，以期推动新兴研究主题的识别研究[2-3]。目前，新兴研究主题的识别大多依赖专家的定性解读，但随着各学科的发展，科学知识体量呈现指数级增长趋势，依靠专家力量无法高效发现所有的新兴研究主题。此外，新兴研究主题的识别大多从其可能出现的来源视角开展，如知识基础的新颖性、多种知识的交叉融合、弱关系、突破性创新对知识结构的影响、可能的词频突变等[4]。这些研究视角较为单一，并且大多使用论文和专利数据进行识别，而论文和专利的发表公开存在滞后性，难以实现新兴研究主题的前瞻性识别[5]。因此，需要考虑扩充数据来源并引入新的理论方法开展新兴研究主题早期识别。

弱信号具有新颖性、不确定性等特征，与新兴研究主题的特性高度契合[6]，因此，可将弱信号分析融入到新兴研究主题识别中。本文采用多源数据，首先，通过构建新兴度综合指标识别新兴研究主题。其次，结合弱信号分析筛选出弱信号新兴研究主题，实现新兴研究主题的早期识别。最后，尝试通过不同数据主题间的交叉测度，测度弱信号新兴研究主题的潜在影响力，以期推动前沿新兴领域预见方法的发展。

1 相关研究

1.1 新兴研究主题识别研究

新兴研究主题或新兴研究话题（emerging research topic/ emerging topic）也被称为新兴主题、新兴研究趋势、新兴研究领域等[7]。Wang Q[8]在 Rotolo[9]关于新兴技术研究的基础上，认为新兴研究主题是一个具有一定连贯性并具有显著科学影响、根本上新颖且相对快速增长的研究主题，其与“新兴技术”属于不同的概念，但存在语义重叠。新兴研究主题与前沿主题也存在交叉，二者都是具备一定创新性和影响力的主题，但前者强调时间维度上的“新”以及发展维度上的“兴”，后者强调研究水平的“高”、研究难度的“大”以及研究质量的“优”[3]。近期，许海云等[4]对新兴研究主题及其多个相似概念进行了辨析，发现不同概念间存在

广泛交叉但各有侧重，颠覆性技术和前沿技术关注技术创新，突破性创新和研究前沿关注科学研究，并基于科学—技术交互的背景下，认为新兴研究主题包含科学研究的新兴前沿和技术创新的新兴技术。

已有研究主要通过新兴研究主题的特征构建识别指标进行识别。有研究从某一特征出发构建单一识别指标进行识别，如通过技术主题的专利和论文引用率指标[10]表征其影响力特征，借助权利要求总数或技术创新性[11]等指标确定其新颖性特征，利用主题词数量及专利数增长率[12]衡量其是否具有高增长性，使用 Kullback 散度或 Jensen-Shannon 散度[13]等判断技术发展是否具有连贯性，基于论文与专利的前向引用及后向引用[14]等估计未来的不确定性等。但是，利用单一指标识别多具有片面性，可能会造成识别结果不准确。为提高新兴研究主题识别的准确性与全面性，有研究基于多种特征构建综合指标进行识别，如基于影响力、高增长率和新颖性等特征，利用 CRITIC 法构建综合指标识别新兴研究主题[15]、从新兴研究主题的“新”“兴”“热”3 个维度构建新兴潜力综合指标，识别人工智能领域的新兴研究主题[16]等。

无论是单一指标还是多指标融合，都是以新兴研究主题特征作为评判标准。卢超等[7]通过对新兴研究主题相关研究进行梳理，将识别指标划分为成长性指标、影响力指标、新颖性指标和其他类型指标。柴文越等[3]总结了新兴研究主题识别的 8 个特征及识别指标，包括新颖性、学科交叉性、不确实性、成长性、高主题强度、高主题影响力、突变性和持续性。基于新兴研究主题五大公认特征，本文对相关研究中的部分指标进行概述，如表 1 所示。

表 1 新兴研究主题特征及识别指标

Table 1 Characteristics and Identification Indicators of Emerging Research Topics

特征	描述	识别指标	相关研究
新颖性	一个原创且前所未有的主题，即与过去主题的相似性较低	技术创新性 先验知识量 权利要求总数	银路等[17] Upham S P 等[18] Small H 等[19]
高增长率	有关新兴研究主题术语的词汇数量在短时间内快速增长	专利数量增长率 主题词数量	沃顿商学院[20] Cozzens S 等[21] 许海云等[22]
影响力	能够在未来对科学、社会、产业产生巨大影响	论文和专利引用率 论文或专利平均关注人数	Porter A L 等[12] Meyer M 等[23]

连贯性	与新兴研究主题有关的词汇在时间切片上呈现连续性	主题关注度	Kwon H 等[24]
		词频变化率	Wang Q[8]
		Kullback 散度	丁敬达等[25]
		Jensen-Shannon 散度	Stahl B C[26]
不确定性	新兴研究主题在未来发展上具有一定的风险，即不一定发展成为新兴研究主题	前向引用	Altuntas S 等[14]
			李仕明等[27]
		后向引用	Day G S 等[28]

新兴研究主题识别方法包括定性分析和定量分析，定性分析依赖专家意见，具有一定的科学性和权威性，代表方法有德尔菲法[29]、层次分析法[30]、情景分析法[31]等。定量分析主要基于文献计量学、统计学等方法，通过挖掘数据背后的关系及规律实现识别，代表方法有共现分析[32-33]、引文分析[34]、文本分析[35]等。共现分析通过揭示关键词、作者等文献特征信息之间的关联，实现对文献间关系的表征，张光宇等[36]利用共现分析梳理了 25 年来颠覆性创新国际研究领域的知识脉络，探索领域未来发展趋势。引文分析通过文献间的被引、共被引关系，研究文献中主题的演化关系，如 Xu H Y 等[37]利用引文扩散特征评估新兴研究课题的突破潜力。文本分析基于文本表示及其特征项的选取进行分析，如使用文本中提取的特征词及逆向量化表示文本信息。李欣等[38]提取文本 SAO 结构，依据文本相似度进行聚类，进行新兴技术的识别。

1.2 弱信号分析

弱信号是表示未来潜在变化的候选指标，但不能提供对未来的完整预测，其可能发展为强信号，也可能在未实现显著增长或维持现有状态的情况下消失[39]。Ansoff H I[40]将弱信号定义为“未来可能发生变化的征兆”，也有研究将其视作来自外部或内部的警告信号，是未来趋势、变化和新兴现象背后的面向未来的信息[41]。Veen B L 等[42]将不同学科的定义统一为“在环境中检测到的或在解释过程中产生的战略现象的感知，其与感知者的参考框架相距甚远”。韩盟等[43]通过梳理弱信号相关研究，认为其本质是一种预示未来变化的某种现象、事件、机会或威胁，对其识别与分析有助于提前了解和增加未来发展的确定性。

Hiltunen E[44]基于符号三元模型提出了未来征兆的三元模型，包括信号（signal）、问题（issue）和解释（interpretation）3 个维度。其中，信号指信号的真实可见性，问题指各种事件，解释是接收者对信号含义的理解。信号和问题为客观维度，与来源无关，取决于事件的可见性，而解释是唯一的主观维度，与弱信号的接受者和解释者有关，其参与者基于问题

本身的外部信号形成自己的内部信号，并进一步转化为次级外部信号。Yoon J[45]基于该理论框架，提出可见度和扩散度指标，构建关键词组合图，通过新闻数据识别弱信号，该方法已广泛应用于弱信号研究[46]。Kim J 等[47]则从新颖性视角出发，使用专利和未来数据构建结构化的关键词—文档矩阵，将关键词和文档视为信号，使用局部异常因子进行稀有性和范式无关性评估，从而构建信号组合图识别弱信号，实现关键词和文档级别的分析。

作为研究未来变化现象的方法，弱信号分析在新兴技术以及其他突变性识别研究中起到了举足轻重的作用[48]，通过对弱信号的挖掘，可以发现那些尚未被广泛关注但具有巨大创新潜力的技术点，从而识别出潜在技术机会。在相关研究中，文本挖掘[49]、指标体系构建[50–52]与网络分析[53-54]是常用方法，如唐虎林等[55]依据颠覆性技术弱信号模糊性、前瞻性的主要特征，构建颠覆性技术弱信号测度指标体系。刘俊婉等[56]将基于专利的弱信号探测模型和技术颠覆性潜力测度体系相结合，实现颠覆性技术的早期识别。多源数据视角下，张慧玲等[6]立足早期情报感知方法流程，结合定量与定性、主观与客观、因果与相关分析 3 个方法维度，探究了可用于各个阶段的方法及其优缺点。从弱信号数据源的特征出发，可以通过文档离群性和文本相异性指标筛选异类数据，借助组合图确定弱信号关键词，并基于意义建构理论抽取其上下文语境，实现新兴技术弱信号的语义建构与解析[57]。此外，弱信号也被应用于关键核心技术的识别，如通过组合图法与 SAO 语义分析，识别隐性“卡脖子”关键核心技术[46]。

弱信号作为“未来信号”，在内涵特征上，其不确定性与新兴研究主题早期特征高度契合，且后者多以弱信号形式存在，对弱信号的分析有助于新兴研究主题的早期识别。现有研究多依赖单一数据来源，导致弱信号识别不全。同时，弱信号的解释在很大程度上取决于数据的来源，为了提高弱信号的可解释性，应引入多元化的数据来源[58-59]。

综上，现有研究已对新兴研究主题及弱信号的定义、特征进行阐释，二者在内涵特征上具有天然相通性[4]，使得利用弱信号实现新兴研究主题早期识别具备可行性。但新兴研究主题及弱信号的融合研究较少，且多使用单一数据源，识别结果不够全面。因此，本文基于多源数据，融合弱信号与新兴研究主题，实现对新兴研究主题的早期识别。

2 研究方法

针对新兴研究主题在早期阶段呈现的零散性与模糊性等特征，本文将弱信号分析方法引入新兴研究主题识别中，以期前瞻识别早期的新兴研究主题。同时，本文将通过多源数据分析，来避免单一数据源的倾向性，并提高早期识别结果的准确性、覆盖度。此外，本文计算弱信号新兴研究主题在多源数据中的分布广度，量化其在不同领域和场景中的传播程度与影响力。本文的研究内容与技术框架如图 1 所示。

首先，获取论文、专利、新闻和临床四类数据并预处理为按年份划分的文本数据，采用BERTopic 主题模型进行主题建模，计算不同年份主题的相似度，构建时间序列下的主题集合。其次，基于新兴研究主题的主要特征构建测度指标，完成初步筛选。但仅利用多维指标难以捕捉其在技术生命周期早期阶段的特征，本文进一步构建弱信号识别指标，识别主题集合中的弱信号主题。将同时具备新兴性与弱信号特征的主题界定为弱信号新兴研究主题，并尝试进一步测度其潜在影响力。

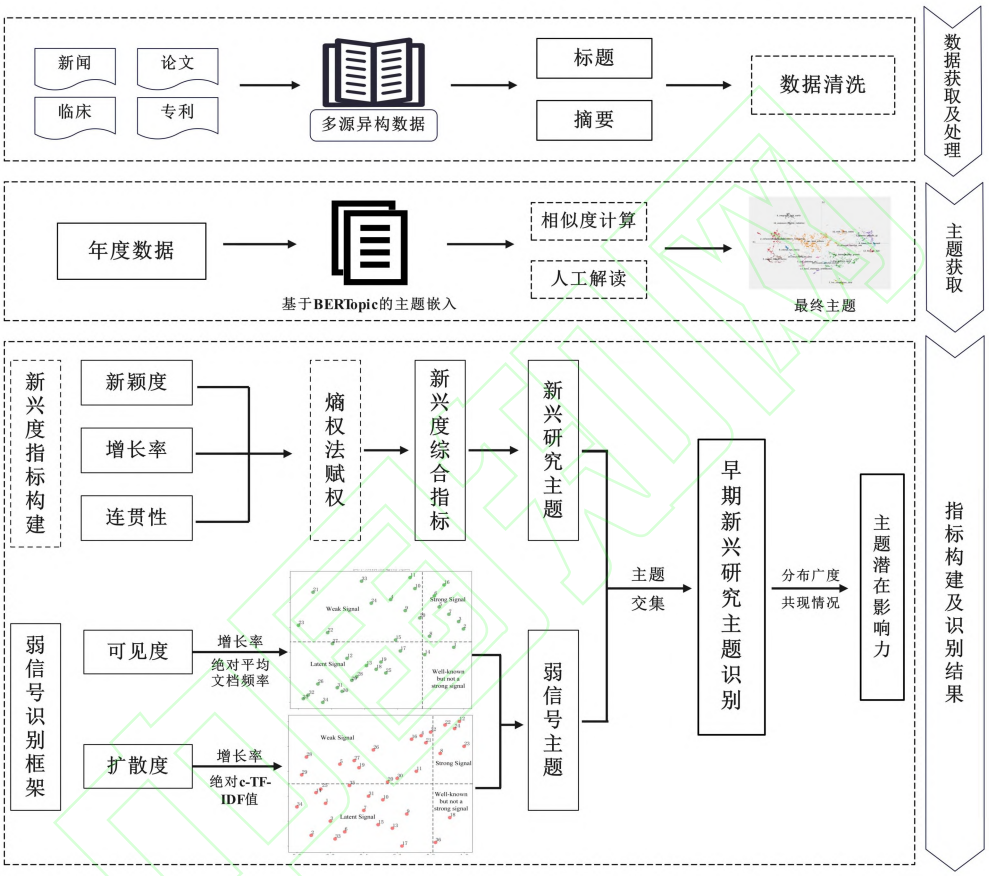

图 1 早期新兴研究主题识别路线

Fig.1 Early Identification Pathway for Emerging Research Topics

2.1 数据获取与处理

本文选取干细胞领域进行实证分析，获取该领域的论文、专利、临床数据、新闻。其中，论文数据作为科学研究的基础成果，体现了学界对特定技术问题的关注与探索；专利作为技术创新的产出，是重要的技术信息载体[60]；临床数据可以体现医疗措施的实际效果，为临床实践提供科学依据及指导[61]；新闻作为社交媒体数据，具有较强的时效性，可反映社会对技术发展的态度。

本文选择标题和摘要作为研究所需的文本数据，首先删除重复、摘要缺失的记录，并进

行分词、去除停用词等处理，将其转化为单词组成的文本，之后按照年份划分文本数据，形成年度数据。

2.2 基于 BERTopic 模型的主题识别

BERTopic 通过结合嵌入模型与聚类算法实现文档中的主题提取，相较传统主题模型具有更加强健的稳定性和离散性[62]，并有效弥合了密度聚类和中心采样方法间的不兼容性[63]。Egger R 等[64]基于推特数据比对多种模型，验证了 BERTopic 模型的优越性能，同时国内学者[65]也利用 BERTopic 模型取得了良好的识别效果。本文采用 BERTopic 对四类数据预处理后的年度文本进行主题建模，得到年份主题，并基于主题向量计算主题间的余弦相似度，结合人工研读确定最终主题集合。

2.3 新兴研究主题识别

2.3.1 新兴度指标构建

新兴研究主题具有新颖度、高增长性、影响力、连贯性、不确定性 5 个特征，单一指标有助于更清晰地理解识别结果，但结果本身可能存在较大的偏差，而复杂的综合指标通常更精确，但可能难以解释[66]。考虑到指标数据的获取以及本研究侧重的新兴研究主题特征等因素，本文从“新”“兴”“持续”3 个维度，使用新颖度、增长率和连贯性指标构建新兴度综合评估指标，衡量研究主题的发展潜力。其中，新颖度反映主题的时效性，值越高表明出现时间越晚；增长率衡量主题的扩张速度，值越高表示成长潜力越强；连贯性评估主题在时间维度上的延续性与一致性，值越高说明研究持续性与主题一致性越强。

（1）新颖度指标

科技文献具有老化规律，随着时间的推移，其内容会变得陈旧过时。相较于陈旧文献，新生文献具有更强的新颖性。通常来说，主题出现的时间越晚，其新颖性越高，为了综合考虑主题中相关文档的数量和发表时间，本文采用主题内所有文档的平均发表时间计算新颖度[67]，如公式（1）所示：

$$N_i = \frac{\sum_{j=1}^{N} T_j}{N} \#(1)$$

其中，N_i为第i个主题的新颖度，T_j为该主题中第j篇文档的时间，N为该主题在所有时间范围内包含的所有文档的总数。

（2）增长率指标

主题中的文献增长率是评估其成长潜力的一个关键标准，一个主题内的文献增长率越快，该主题就越有可能发展为新兴研究主题，增长率计算如公式（2）所示：

$$F_{i,t}=\frac{\left(Fre_{i,t}-Fre_{i,t-1}\right)}{Fre_{i,t-1}}\#(2)$$

其中，$F_{i,t}$是主题i在时间t到$t+1$期间的增长率，$Fre_{i,t}$表示主题i在时间t包含的文档数量。当某一年度主题包含的文档数量为 0 时，增长率无法计算，因此使用主题i中文档的平均增长率作为该主题的增长率指标，如公式（3）所示，其中n为主题i所出现的年份数。

$$G_i=\frac{1}{n}\sum_{i=1}^{n}F_{i,t}\#(3)$$

（3）连贯性指标

新兴研究主题通常在特定领域不断拓展和持续深化，连贯性越高，说明研究持续性与聚焦性越强。本文参照前人研究[1]，采用 Jaccard 系数作为连贯性评价指标，衡量相邻时间段主题一致性；考虑到部分主题在某一时间段高度相关，而其他时间段关联较弱的情况，本研究对各相邻时间段的 Jaccard 系数平均化处理，以获得整体连贯性。计算方法如公式（4）和公式（5）所示：

$$J_{i(A,B)}=\frac{|A\cap B|}{|A\cup B|}\#(4)$$

$$J_i=\frac{\sum_{t=1}^{T_f}J_{i(A,B)}}{T_f}\#(5)$$

其中，A和B分别表示主题i在相邻时间段t和$t+1$内的高频主题词集合，$|A\cap B|$为两个集合的交集大小，$|A\cup B|$为两个集合的并集大小，J_i表示主题i的连贯性，T_f表示主题i出现的年度数。

（4）新兴度指标构建

在完成所有主题的新颖度、增长率和连贯性指标计算后，本文对指标值进行最大最小值归一化处理，并采用熵权法计算四类数据中各指标的熵值与差异系数，以此分配合适的权重W_n、W_g、W_s，最终构建新兴度综合评价指标。

2.3.2 弱信号主题识别

为了实现弱信号主题早期识别，本文将弱信号引入新兴研究主题识别体系中，参照 Yoon J[45]的弱信号识别框架，对主题的可见度和扩散度进行计算。本文在对扩散度计算时引入了主题词共现网络中主题词的特征向量中心度，然后从可见度和扩散度两个维度识别弱信号，具体过程如下：

（1）可见度指标

在 Hiltunen E[44]提出的三元模型中，信号（signal）维度与研究对象对外界的可见性有关，主题在文档中的出现频率表明了该主题的曝光程度，利用主题的出现频率衡量主题的可

见程度，主题的出现频率越高，表示可见度越高，主题i的可见程度（Degree of visibility，DOV）计算公式如公式（6）所示：

$$DOV_i = \frac{TF_{i,t}}{N_t} \times \left(1 - w \times (n - t)\right) \#(6)$$

其中，$TF_{i,t}$是主题i在第t年包含的文档总数，N_t表示第t年内的所有文档总数，n是所选时间范围内所有的年份数；w是时间权重，参照 Yoon J[45]的研究将w设置为 0.05。

（2）扩散度指标

问题（issue）轴显示了弱信号主题向外传播的程度[44]，即扩散性，包含主题词的文档越多，则该主题词在文本集合中越具有一般性且越重要，说明该主题词对应的主题的扩散程度越高，本文使用主题词的 c-TF-IDF 值代替主题词的频率。同时，为了实现对主题词重要程度的度量，利用每一年中所有主题的主题词构建主题词共现网络，通过主题词在网络中的特征向量中心度衡量其重要程度，以此对主题词进行加权。主题i的扩散程度（degree of diffusion，DOD）计算如公式（7）和公式（8）所示：

$$\mathrm{c-TF-IDF} = DF_{t,c} * \log\left(1 + \frac{A}{DF_t}\right) \#(7)$$

$$DoD_{ij} = \frac{\sum (c - TF - IDF)}{N_{i,t}} * \left(1 - w \times (n - t)\right) * \mathrm{a} \#(8)$$

公式（7）计算每个主题词的 c-TF-IDF 值，其中，$DF_{t,c}$是指主题词t在所有类中出现的总频率，A是每个类的平均单词数。公式（8）中的$N_{i,t}$是指主题i在第t年的主题词总数，w是时间权重，其数值与公式（6）一致，n是所选时间范围内的年份数，a是每一年中所有主题的主题词构建的共现网络中的主题词的特征向量中心度。

（3）弱信号主题识别

通过上述公式计算每一年主题的可见度和扩散度指标，并以几何平均法计算每个主题在时间范围内的平均时间加权增长率。根据 Hiltunen E 的观点，潜在弱信号主题常表现为具有异常模式但鲜受关注或很少扩散的主题，具体表现为出现频率较低但增长幅度较高，且 c-TF-IDF 值较低但扩散度增长率较高。因此，本研究构建主题涌现图（基于平均时间加权可见度增长率与绝对平均文档频率）与主题分配图（基于平均时间加权扩散度增长率和绝对平均 c-TF-IDF 值），据此识别弱信号主题。

对于主题涌现图和主题分配图，横坐标分别为绝对文档频率的平均值和绝对平均 c-TF-IDF 值的平均值，纵坐标分别为可见度平均增长率和扩散度平均增长率。本文以主题在横坐标的平均值和纵坐标的 50%位数为阈值，当主题在主题涌现图和主题分配图中同时位于纵坐标的前 50%，且在横坐标上低于平均值时，被视为弱信号主题。

2.4 结果识别及潜在影响力测度

本文基于新兴研究主题与弱信号主题的识别结果，选取二者交集，以实现对新兴研究主题中弱信号主题的筛选，进而识别出具有早期弱信号特征且已发展为新兴研究主题的弱信号新兴主题。

单一数据来源通常具有侧重性，如学术论文更侧重理论创新，新闻更强调关注度，导致其识别结果存在局限，难以反映主题在其他领域的影响力。在多源融合背景下，若某一主题在不同类型数据中表现出较高的相似性，往往意味着其具备跨场景适应性与广泛关注度。基于此，本文尝试通过多源数据间识别结果的相似度，评估其潜在影响力。首先，获得弱信号新兴研究主题后，计算其在不同年份向量的平均值，作为整体表征向量；其次，对不同数据中的主题进行余弦相似度计算；最后，通过其在四类数据中的共现情况计算其分布广度，衡量其潜在影响力。

3 实证研究

3.1 数据获取与处理

干细胞是一类具有自我更新和多向分化能力的生物细胞，被视为医学“工具箱”的重要补充，是全球生命科学和医学研究的核心方向。该领域交叉融合性强、发展潜力大、发展路径复杂，具有高度的不确定性与前沿性。其早期研究热点往往呈现分散、微弱、前瞻性强的弱信号特征，且该领域经过长期发展积累了丰富得多源数据，契合本文的研究需求。因此，本文选取干细胞领域相关数据开展实证研究，以验证所提出方法的可行性。

本文选取德温特专利数据库作为专利文本的数据来源，专利检索式为TAB=(“stem cells” OR “stem cell” OR ESC near cells OR iPS near cells OR PGC near cells OR MSC near cells OR CSC near cells OR TSC near cells OR ADSC near cells OR HSC near cells OR ESC near cell OR iPS near cell OR PGC near cell OR MSC near cell OR CSC near cell OR TSC near cell OR ADSC near cell OR HSC near cell OR “totipotent cell” OR “pluripotent cell” OR “multipotent cell” OR “unipotent cell” OR “progenitor cell” OR “precursor cell” OR “totipotent cells” OR “pluripotent cells” OR “multipotent cells” OR “unipotent cells” OR “progenitor cells” OR “precursor cells”) NOT TAB=(“stem cellulose” OR “stem Cellular” OR “cello” OR “cellar” OR “cellphone” OR “nonpluripotent” OR “fuel cell” OR “in-plane switching” OR “Intrusion Prevention System”)。

Proquest 数据库涵盖了新闻报道、学术期刊等，内容广泛且更新及时，可以从中获取最新的新闻报道，选择该数据库获取新闻数据。Dimensions 数据库涵盖了专利、学术论文、临

床试验和政策文档等多种类型的数据，该平台数据量庞大且更新及时，本文选择该数据库获取临床数据。OpenAlex 数据库整合了全球期刊论文与预印本等开放学术资源，具有免费获取、多维关联和实时更新等优势，本文选择该数据库获取论文数据。

通过上述数据库检索 2013—2022 年的数据，共检索到专利数据 49 765 条、新闻数据 41 327 条、临床数据 12 143 条、论文数据 30 106 条。导出年份、标题、摘要等内容，删除重复、内容缺失等记录后，共获得 48 665 条专利数据、41 220 条新闻数据、7 198 条临床数据、19 825 条论文数据，然后对文本进行预处理，并构建年份数据。

3.2 主题获取

本研究使用 BERTopic 进行主题建模，在确定专利、新闻、论文数据的主题数量时，首先由模型自动确定主题数，当主题间的相似度超过 0.9 时，BERTopic 模型自动迭代减少主题数量[63]，重复运行主题建模，并对建模结果进行可视化，通过主题距离地图发现存在主题高度重合的情况，如图 2 中 a 图所示。因此，通过基于主题嵌入向量计算的主题间余弦距离矩阵并结合人工判断，合并相似高的主题，最终确定专利、新闻、论文数据各年份最终的主题数量，如图 2 中 b 图所示。临床数据较少，由模型自动确定主题数。

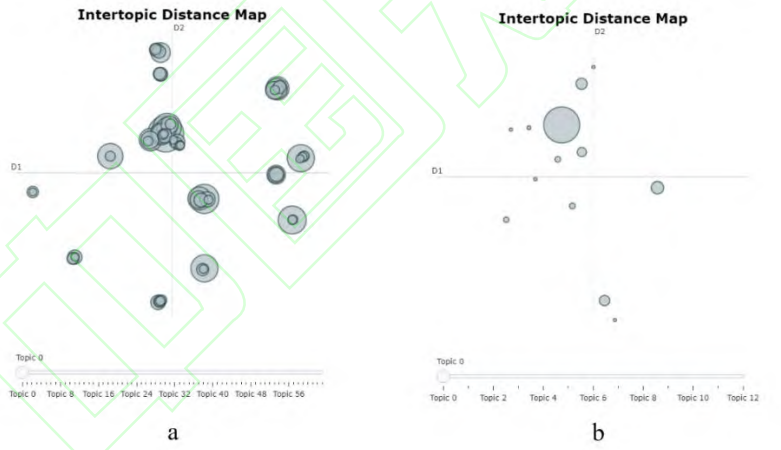

图 2 2013 年主题距离地图——专利

Fig.2 2013 Topic Distance Map – Patents

通过对四类数据中每一年的文本数据进行主题建模，输出相关数据，四类数据中每年的主题数量如表 2 所示。本研究兼顾主题数量以及信息完整度进行多次试验后，当专利、新闻、临床数据两个不同年份主题间的相似度超过 0.9 时，认为是同一主题，而论文涉猎范围较为广泛，将阈值设定为 0.6，在结合人工研读后确定不同年份中相同的主题，并设置为同一主题编号，最终得到专利主题 36 个、新闻主题 34 个、临床主题 35 个、论文主题 14 个。

表 2 四类数据各年的最终主题数

Table 2 Final Number of Topics for Each Year in the Four Types of Data

年份	论文主题数	专利主题数	新闻主题数	临床主题数
2013	20	14	14	15
2014	25	13	18	17
2015	22	11	19	16
2016	19	12	16	24
2017	21	11	11	16
2018	17	14	13	18
2019	20	13	14	18
2020	22	16	18	23
2021	18	16	15	21
2022	21	13	14	15

3.3 新兴研究主题初筛

本文在计算各主题的新颖度、增长率与连贯性指标后，采用标准化处理，并通过熵权法为四类数据分别赋权：论文主题（0.155、0.569、0.276）、专利主题（0.417、0.084、0.498）、临床主题（0.719、0.141、0.138）、新闻（0.196、0.277、0.525），据此计算主题的新兴度综合指标，部分主题的指标计算结果如表 3 所示。

表 3 部分主题新兴度指标测度结果

Table 3 Some Topic Emerging Index Measurement Results

主题编号	数据类型	主题新颖度	主题增长率	主题连贯性	新兴度
#1	论文	0.102 6	0.145 7	0.275 8	0.524 3
	专利	0.219 7	0.000 0	0.498 1	0.717 9
	新闻	0.102 5	0.029 4	0.204 2	0.336 1
	临床	0.000 2	0.004 3	0.073 0	0.077 6
#2	论文	0.107 1	0.168 1	0.257 4	0.532 7
	专利	0.180 5	0.000 0	0.069 7	0.250 2
	新闻	0.136 5	0.023 7	0.159 1	0.319 5
	临床	0.000 1	0.000 0	0.109 9	0.1100 8

#3	论文	0.137 1	0.161 4	0.117 8	0.416 4
	专利	0.326 7	0.000 7	0.209 6	0.536 9
	新闻	0.070 1	0.025 1	0.525 4	0.620 8
	临床	0.000 1	0.008 5	0.055 3	0.064 0

注：表中四类数据的主题编号仅作为相同数据主题间的区分标志，相同编号不代表四类数据的共同主题（下文同）。

基于上述指标结果，对新兴研究主题进行筛选，依据二八法则筛选出新兴度综合指标得分前 20%的主题，视为新兴研究主题，识别结果如表 4 所示。

表 4 新兴研究主题识别结果

Table 4 Results of Identification of Emerging Research Topics

主题类别	论文数据	专利数据	新闻数据	临床数据
新兴研究主题	#2、#12	#1、#3、#25、#26、#29、#35、#36	#3、#5、#17、#19、#24、#30	#13、#16、#17、#20、#27、#33、#34

3.4 弱信号主题识别

3.4.1 主题可见度和扩散度指标计算

根据每个主题中的文档频率及年份数据，对四类数据中主题的可见度及增长率进行计算，为下文主题涌现图的构建提供数据基础。表 5 是部分主题的可见度指标及增长率计算结果。

表 5 部分主题可见度指标及增长率结果

Table 5 Visibility Indicators and Growth Rates of Some Topics

主题		2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	增长率
#1	论文	0.281	0.255	0.286	0.281	0.441	0.398	0.413	0.379	0.483	0.439	0.357
	专利	0.238	0.254	0.307	0.342	0.369	0.289	0.358	0.285	0.304	0.283	0.300
	新闻	0.124	0.029	0.213	0.077	0.083	0.129	0.037	0.084	0.079	0.039	0.075
	临床	0.138	0.278	0.274	0.287	0.377	0.002	0.073	0.014	0.002	0.013	0.049
#2	论文	0.013	0.013	0.030	0.018	0.017	0.020	0.021	0.036	0.029	0.038	0.022
	专利	0.028	0.022	0.015	0.011	0.014	0.010	0.017	0.011	0.010	0.018	0.015
	新闻	0.147	0.028	0.012	0.013	0.087	0.124	0.038	0.000	0.084	0.082	0.050
	临床	0.287	0.827	0.013	0.024	0.028	0.083	0.002	0.025	0.083	0.394	0.057
#3	论文	0.008	0.015	0.000	0.000	0.020	0.030	0.018	0.032	0.029	0.047	0.022
	专利	0.023	0.000	0.036	0.019	0.000	0.089	0.103	0.169	0.156	0.150	0.070

	新闻	0.014	0.037	0.007	0.087	0.082	0.000	0.000	0.000	0.000	0.000	0.031
	临床	0.295	0.385	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.064

通过主题建模的数据结果，获得每个主题中主题词的 c-TF-IDF 值，对四类数据中主题的扩散度及增长率进行计算，为下文主题分配图的构建提供数据基础。表 6 是部分主题的扩散度指标及增长率计算结果。

表 6 部分主题扩散度指标及增长率结果

Table 6 Results of Some Topic Diffusion Indicators and Growth Rates

主题		2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	增长率
#1	论文	0.115	0.191	0.178	0.153	0.203	0.223	0.299	0.236	0.232	0.286	0.204
	专利	0.069	0.083	0.081	0.094	0.104	0.125	0.119	0.118	0.121	0.161	0.104
	新闻	0.025	0.040	0.286	0.005	0.135	0.326	0.101	0.046	0.100	0.101	0.071
	临床	0.028	0.028	0.024	0.021	0.038	0.002	0.073	0.014	0.007	0.013	0.018
#2	论文	0.011	0.000	0.010	0.000	0.020	0.000	0.021	0.041	0.020	0.024	0.019
	专利	0.038	0.055	0.045	0.064	0.091	0.077	0.103	0.095	0.060	0.114	0.070
	新闻	0.164	0.079	0.087	0.046	0.101	0.275	0.105	0.000	0.066	0.106	0.100
	临床	0.017	0.173	0.013	0.013	0.091	0.013	0.001	0.072	0.022	0.012	0.021
#3	论文	0.190	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.022	0.022	0.009
	专利	0.062	0.000	0.082	0.079	0.000	0.100	0.108	0.099	0.102	0.136	0.094
	新闻	0.104	0.086	0.036	0.035	0.007	0.000	0.000	0.000	0.000	0.000	0.037
	临床	0.183	0.138	0.000	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.062

3.4.2 主题涌现图和主题分配图构建

根据小节 2.3.2 中的方法构建主题涌现图和主题分配图。为了更好地呈现，本文对数据进行均匀化处理，使计算结果映射到 0~1 之间的一个均匀分布的值上。主题涌现图和主题分配图分别如图 3、图 4 所示，不同的颜色圆点表示四类数据中的主题，数字表示主题序号。

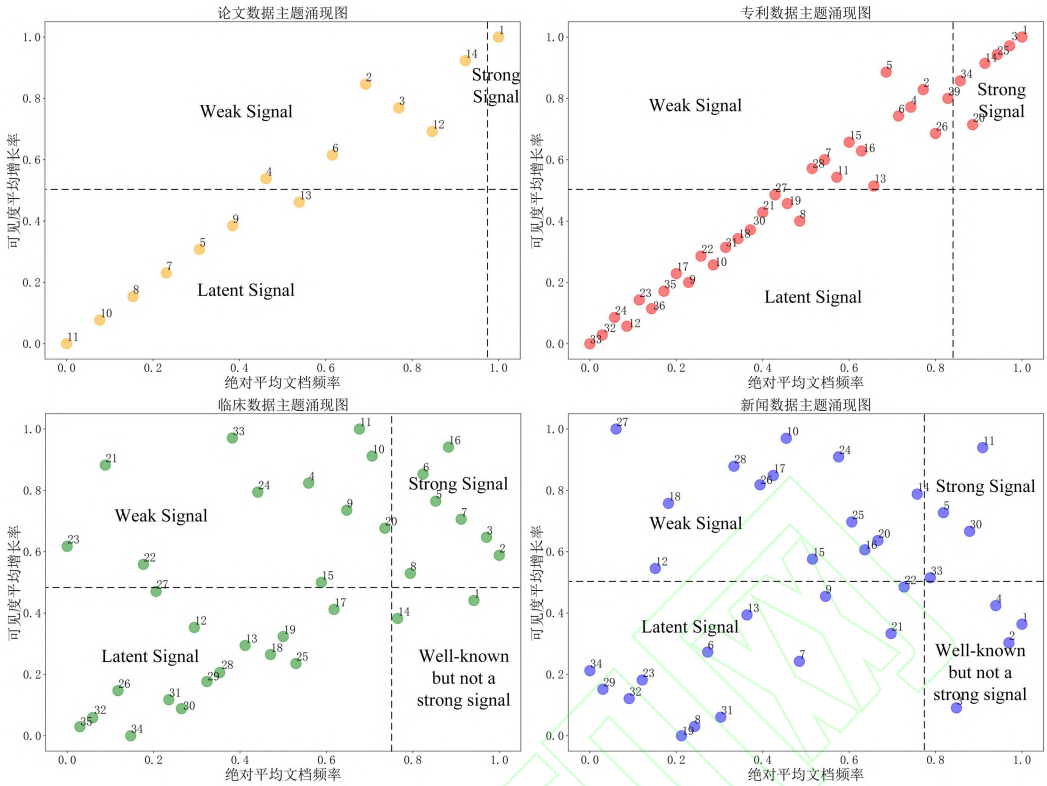

图 3 四类数据的主题涌现图

Fig.3 Topic Emergence Diagram of Four Types of Data

主题涌现图和主题分配图被横纵坐标上的阈值形成的两条垂线分为 4 个象限，可以帮助识别 4 种信号[68]，即弱信号（Weak Signal）、强信号（Strong Signal）、潜在信号（Latent Signal）和众所周知但不强的信号（Well-known but not a strong signal）。从图 3 中可以看到，论文主题与专利主题多分布于弱信号区域、潜在信号区域，强信号较少且无众所周知但不强的信号；临床及新闻数据在 4 个区域均有分布，但主要分布于弱信号区域、潜在信号区域。

从主题分配图来看，论文主题在各区域分布数量基本不变；专利主题与临床主题在强信号区域分布数量减少，在弱信号区域分布数量增加，在众所周知但不强的信号数量明显增加；新闻主题在潜在信号区域分布数量减少，在众所周知但不强的信号区域分布数量增加。

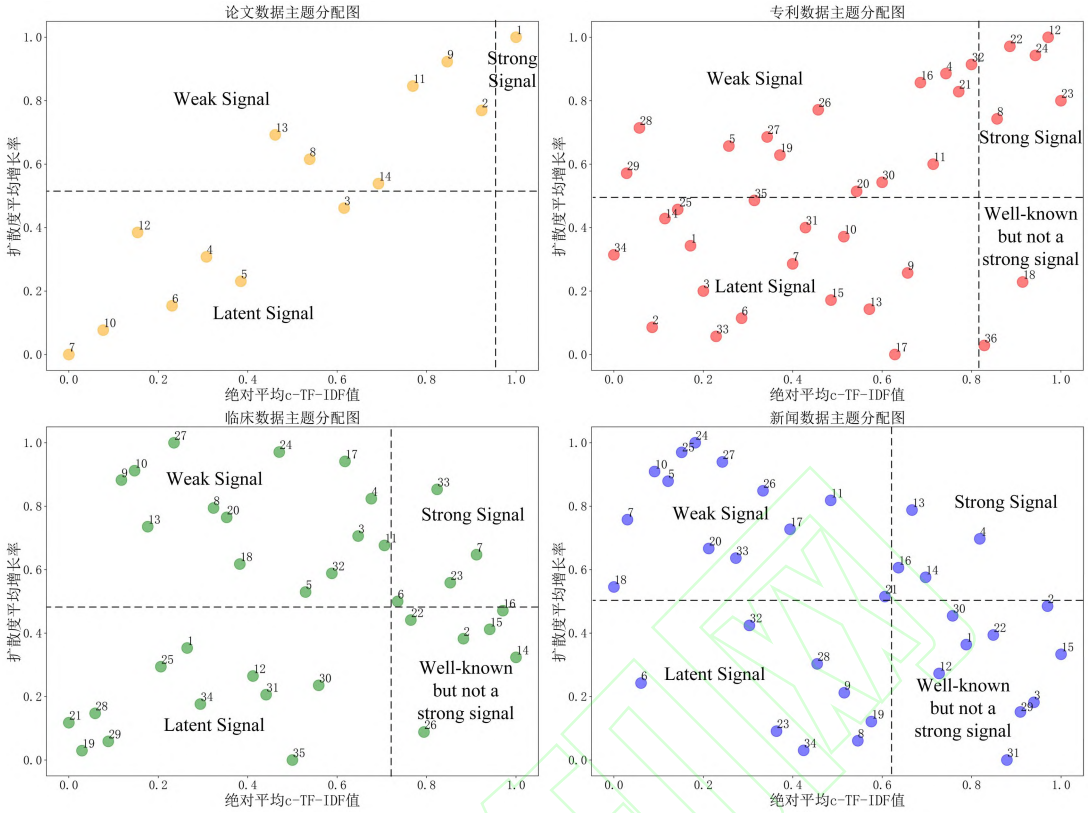

图 4 四类数据的主题分配图

Fig.4 Topic Distribution Diagram of Four Types of Data

两类图中左上角区域为弱信号主题，根据划分标准，当主题同时归属于主题涌现图和主题分配图的弱信号区域时，为本文识别的四类数据中的弱信号主题，结果如表 7 所示。从识别结果来看，新闻数据的弱信号主题最多，有 8 个主题；专利数据弱信号主题数量次之，有 7 个主题；临床数据中的弱信号主题共有 6 个主题；论文数据中的弱信号主题仅有 2 个主题。

表 7 弱信号主题识别结果

Table 7 Weak Signal Topic Recognition Results

主题类别	论文数据	专利数据	新闻数据	临床数据
弱信号主题	#2、#14	#4、#5、#11、#16、#26、#28、#29	#10、#17、#18、#20、#24、#25、#26、#27	#4、#9、#10、#11、#20、#24

3.5 识别结果及潜在影响力测度

在获得四类数据中新兴研究主题与弱信号主题的识别结果后，取其交集作为弱信号新兴研究主题，实现对新兴研究主题中弱信号主题的过滤。其中，专利数据识别出主题 26 与主题 29，新闻数据识别出主题 17 与主题 24，临床数据识别出主题 20，论文数据识别出主题 2。

根据弱信号新兴研究主题的主题向量，计算不同数据间的余弦相似度，并设置相应阈值：专利与临床为 0.5，临床与新闻、专利与新闻为 0.45，论文与其他数据为 0.15。当不同来源主题间的相似度超过对应阈值时，视为跨数据的共现主题，根据主题共现情况计算其分布广度以衡量潜在影响力。相关结果如表 8 与图 5 所示。

表 8 主题识别结果交集及分布广度

Table 8 Intersection and Coverage of topic Recognition Results

主题	论文数据	专利数据	新闻数据	临床数据	分布广度
论文主题 2	●	●	●	⊗	0.75
专利主题 26	●	●	●	●	1.00
专利主题 29	⊗	●	⊗	⊗	0.25
新闻主题 17	●	⊗	●	●	0.75
新闻主题 24	⊗	●	●	⊗	0.5
临床主题 20	⊗	●	●	●	0.75

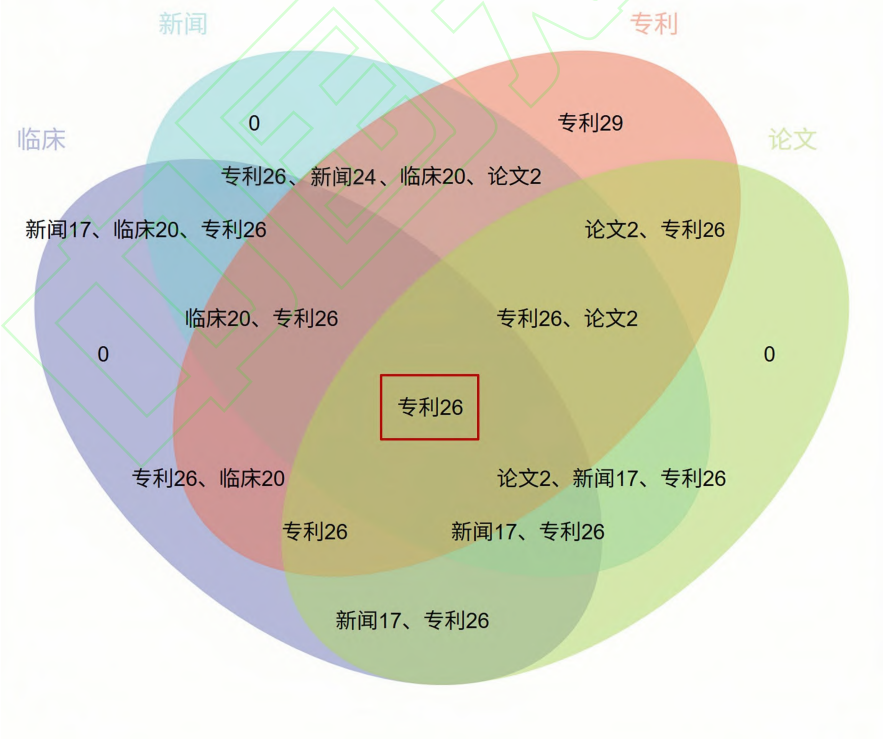

图 5 四类数据中主题识别结果的交集

Fig.5 The Intersection of Topic Identification Results in Four Types of Data

由表 8 和图 5 可知，论文主题 2、专利主题 26、新闻主题 17、新闻主题 24、临床主题

20 同时在多类数据中共现，分布较为广泛。其中，专利主题 26 在四类数据中均有出现，表明其在学术研究、临床应用及公共传播中均受到关注，潜在影响力较大；而专利主题 29 仅出现在专利数据中，辐射范围有限，潜在影响力较弱。

因篇幅所限，本文仅选取专利主题 26 进行解读。该主题聚焦于细胞受体相关技术在疾病治疗中的应用，涵盖 T 细胞受体（TCR）结构（如α链和β链可变域的互补决定区（CDR）氨基酸序列）、多价 TCR 复合物、核酸分子、载体、宿主细胞及其相关药物组合物在肿瘤和自身免疫疾病治疗中的应用。此外，该主题还包含抗γδ T 细胞受体嵌合抗原受体（CAR）的结构（如胞外域、跨膜域和胞内域），以及其载体、基因改造细胞在治疗γδ T 细胞相关疾病中的应用。目前，CAR T 细胞疗法已成为复发或难治性 B 细胞恶性肿瘤患者的标准治疗方法[69]，而 TCR-T 细胞疗法也在实体瘤治疗中显示良好前景，成为多种肿瘤抗原的有潜力的工具[70]。

3.6 识别结果验证

研究报告因编写与发布周期较短，内容需经专家判断筛选，且权威机构具备广泛的信息获取渠道与前沿敏感性。因此，其发布内容在一定程度上反映了领域内的弱信号新兴研究主题，可为学术研究提供重要参考。鉴于专利主题 26 系四类数据共现主题，分布广度最广，潜在影响力显著，本文选其作为主要验证对象。

本文考虑到权威报告通常涉及领域内普遍使用的词汇（如“cell”“stem cell”等），此类词汇虽广泛出现，但并不能表明主题的独特性，容易导致相似度值虚高。为更客观地反映研究主题与权威报告之间在研究方向上的契合度，本文引入三类对照主题的关键词，计算其与报告的平均相似度，作为领域通用词汇相似度基准对照。通过与该基准进行对比，剔除通用词汇造成的噪声影响，确保专利主题 26 与权威报告之间的高相似度来源于其特定的研究焦点与技术路径，而非由常见词汇的叠加效应引起。

本文筛选权威机构于 2020—2025 年发布的干细胞报告，共计获取 10 篇，部分报告信息如表 9 所示。

表 9 部分报告信息

Table 9 Reports the Information in Part

报告名称	发布机构
2024 ANNUAL REPORT	California Institute for Regenerative Medicine
Annual Report 2023	University of Wisconsin Stem Cell & Regenerative

	Medicine Center
Nebraska Stem Cell Research Annual Report 2022	Nebraska Department of Health and Human Services

关键词在一定程度上可反映文章的研究主题。本文首先构建领域停用词表，借助 NLTK 进行分词，并采用 TF-IDF 方法提取报告关键词。为避免领域通用词汇造成相似度虚高，本研究随机选取三类对照主题进行比较：一是仅具新兴特征的临床主题 27，二是仅具弱信号特征的新闻主题 10，三是具备弱信号特征的新兴研究主题——新闻主题 24，并与专利主题 26 进行对比分析。部分提取结果及对应主题部分关键词如表 10 所示。

表 10 部分关键词提取结果

Table 10 Extraction Results of Some Key Words

文件类型	文件名称	关键词
权威报告	2024 ANNUAL REPORT	cirm, regenerative, california, medicine, regenerative medicine, clinical, cell, stem, stem cell, therapies
	Annual Report 2023	cell, stem, cells, stem cell, lab, scrmc, medicine, regenerative, brain, regenerative medicine
	Nebraska Stem Cell Research Annual Report 2022	stem, cell, stem cell, nebraska, cancer, cells, project, university, committee, stem cells
目标验证主题	专利主题 26	specific cell receptor, cell, plant, lettuce, hair, follicle, alkyl, 6c, sequence, gene, substitute, compound
对照主题	临床主题 27	covid, 19, covid 19, cell, patient, respiratory, study, severe, treatment, mesenchymal, covid
	新闻主题 10	oms, hematopoietic stem cell, hematopoietic stem, hematopoietic, thrombotic, treatment hematopoietic stem, treatment hematopoietic, omeros, designation, orphan
	新闻主题 24	apoptosis, stress, cell, oxidative stress, oxidative, er, induced, er stress, stem, university

提取关键词后，本文采用 TF-IDF+余弦相似度综合评估各主题与相关报告之间的相似性，

部分结果如表 11 所示。

表 11 部分报告与主题相似度

Table 11 Similarity between Some Reports and the Topic

主题名	平均相似度	最高相似度
专利主题 26	0.337	0.391
临床主题 27	0.067	0.076
新闻主题 10	0.146	0.207
新闻主题 24	0.121	0.139

由表 11 可知，专利主题 26 的平均相似度最高为 0.337，仅具备新兴特征的临床主题 27 平均相似度最低仅为 0.067，其余两个对照主题的平均相似度也呈现较低水准，均与专利主题 26 平均相似度相差较大。进一步选取临床主题 27、新闻主题 10 和新闻主题 24 的平均相似度（0.111）作为通用词汇相似度基准。相比之下，专利主题 26 的平均相似度为 0.337，显著高于基准值，且其与“Annual Report on the Nebraska Stem Cell Research Act”的相似度接近 0.4，表明该主题在技术表达和研究方向上与权威报告高度契合。

本文通过定量验证证实识别结果的有效性及专利主题 26 具有的代表性，但单一定量验证难以全面证明专利主题 26 与实际研究前沿具有高度一致性。本文在定量验证的基础上，进一步开展定性分析，以增强验证效果。

本文进一步研读 2023 年以后发布的干细胞领域权威文章、专业学术指南等资料进行验证。

《Nature Communications》为高质量《Nature》子刊。Saotome K 等[71]在 2023 年 4 月于该刊发表题为“Structural Analysis of Cancer-Relevant TCR-CD3 and Peptide-MHC Complexes by cryoEM”的文章。该文章指出 T 细胞受体（TCR）对抗原肽-MHC（pMHC）分子的识别可启动 T 细胞介导的免疫反应，报道了两种不同的全长α/β TCR-CD3 复合物与其 pMHC 配体（癌症-睾丸抗原 HLA-A2/MAGEA4（230–239））结合的低温电子显微镜结构。该研究内容与专利主题 26 聚焦的 T 细胞受体（TCR）结构及α、β链可变域的互补决定区（CDR）高度契合。

此外，美国食品药品监督管理局（FDA）下设机构 Oncology Center of Excellence 于 2024 年 3 月发布了名为“Oncology Cell and Gene Therapy”[72]的专业学术指南，明确指出将重点

推进具有治愈潜力的转化性癌症治疗技术的临床评估与开发，包括但不限于嵌合抗原受体 T 细胞疗法（CAR-T）、重定向特异性 T 细胞受体 T 细胞疗法（TCR-T）、转录激活因子样效应核酸酶（TALEN）开发的 T 细胞、造血干细胞移植（HSCT）等内容。

综上，专利主题 26 与权威报告的平均相似度为 0.337，显著高于领域通用词汇相似度基准（0.111），且与权威期刊论文、专业学术指南相似程度较高。专利主题 26 可判定为具有早期弱信号特征的新兴研究主题，验证了本文所提出的融合弱信号分析的新兴研究主题识别方法的有效性。

4 结语

本文提出一种基于多源数据融合的弱信号新兴研究主题识别方法，利用新兴研究主题早期特征与弱信号特征的高度契合，实现其早期识别。首先，利用 BERTopic 对年份数据进行主题建模，并结合相似度计算与人工解读确定主题集合；其次，从新颖性、成长性与连贯性特征出发，构建新兴度指标体系，识别新兴研究主题，并基于可见度与扩散度识别弱信号主题；最后，通过交集分析识别出具有弱信号特征的新兴研究主题，并基于跨数据主题相似度测度其潜在影响力。

本研究选取干细胞领域进行实证，涵盖论文、专利、新闻、临床四类数据，共识别出 6 个弱信号新兴研究主题，并以专利主题 26 为例进行验证。结果表明：所提出的方法在弱信号新兴研究主题识别中，具有较强的适用性与有效性，且在多源数据交叉测度支持下，所识别的代表性主题不仅具备前沿性与新颖性，还展现出较强的跨领域影响潜力。

本文存在局限性。第一，研究未针对不同数据构建更符合其数据特征的指标体系，可能导致具有独特数据特征的主题被忽略。第二，受限于弱信号本身的不确定性，增加数据来源有助于提升弱信号的识别与解释能力，但本文仅选取论文、专利、新闻、临床四类数据，未能涵盖更广泛的数据类型。未来研究可进一步拓展数据来源，如引入政策文件、网站评论等，并结合不同数据特性构建更为细化的指标体系，以提升新兴研究主题早期识别的精准度与覆盖度。

参 考 文 献

[1] Xu H Y, Winnink J, Yue Z H, et al. Multidimensional Scientometric Indicators for the Detection of Emerging Research Topics[J]. Technological Forecasting and Social Change, 2021, 163: 120490.

[2] 卢小宾, 张杨燚, 杨冠灿, 等. 新兴技术识别中的不均衡分类研究——基于代价敏感的随机森林算法[J]. 情报学报, 2022, 41(10): 1059-1070.

[3] 柴文越, 刘小平, 梁爽. 新兴主题识别方法研究综述[J]. 现代情报, 2023, 43(12): 164-177.
[4] 许海云, 龚兵营, 杨俊浩, 等. 新兴研究主题识别方法研究进展与前瞻[J]. 图书情报工作, 2025, 69(3): 135-150.
[5] 李欣, 谢前前, 洪志生, 等. 基于社会感知分析的新兴技术发展趋势研究——以钙钛矿太阳能电池技术为例[J]. 科技进步与对策, 2018, 35(10): 15-24.
[6] 张慧玲, 许海云, 刘春江, 等. 科技创新弱信号早期感知方法探究与前瞻[J]. 情报学报, 2024, 43(10): 1129-1141.
[7] 卢超, 侯海燕, Ying D, 等. 国外新兴研究话题发现研究综述[J]. 情报学报, 2019, 38(1): 97-110.
[8] Wang Q. A Bibliometric Model for Identifying Emerging Research Topics[J]. Journal of the Association for Information Science and Technology, 2018, 69(2): 290-304.
[9] Rotolo D, Hicks D, Martin B R. What Is an Emerging Technology?[J]. Research Policy, 2015, 44(10): 1827-1843.
[10] 任佳妮, 张薇, 杨阳, 等. “人工智能+医疗” 新兴技术识别研究——以医疗机器人为例[J]. 情报杂志, 2021, 40(12): 45-50.
[11] Jarić I, Knežević-Jarić J, Lenhardt M. Relative Age of References as a Tool to Identify Emerging Research Fields With an Application to the Field of Ecology and Environmental Sciences[J]. Scientometrics, 2014, 100(2): 519-529.
[12] Porter A L, Garner J, Carley S F, et al. Emergence Scoring to Identify Frontier R&D Topics and Key Players[J]. Technological Forecasting and Social Change, 2019, 146: 628-643.
[13] Boyack K W, Newman D, Duhon R J, et al. Clustering More Than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches[J]. PLoS One, 2011, 6(3): e18029.
[14] Altuntas S, Erdogan Z, Dereli T. A Clustering-Based Approach for the Evaluation of Candidate Emerging Technologies[J]. Scientometrics, 2020, 124(2): 1157-1177.
[15] 唐恒, 邱悦文. 多源信息视角下的多指标新兴技术主题识别研究——以智能网联汽车领域为例[J]. 情报杂志, 2021, 40(3): 81-88.
[16] 朱相丽, 张敬, 李伟伟, 等. 国际顶级会议视角下人工智能领域的新兴主题发现研究[J]. 情报理论与实践, 2024, 47(9): 147-155.
[17] 银路, 王敏, 萧延高, 等. 新兴技术管理的若干新思维[J]. 管理学报, 2005(3): 277-280, 300.
[18] Upham S P, Small H. Emerging Research Fronts in Science and Technology: Patterns of New Knowledge Development[J]. Scientometrics, 2010, 83(1): 15-38.
[19] Small H, Boyack K W, Klavans R. Identifying Emerging Topics in Science and Technology[J]. Research Policy, 2014, 43(8): 1450-1467.
[20] Day G S, Schoemaker P J H. Avoiding the Pitfalls of Emerging Technologies[J]. California Management Review, 2000, 42(2): 8-33.
[21] Cozzens S, Gatchair S, Kang J, et al. Emerging Technologies: Quantitative Identification and Measurement[J]. Technology Analysis & Strategic Management, 2010, 22(3): 361-376.
[22] 许海云, 张慧玲, 武华维, 等. 新兴研究主题在演化路径上的关键时间点研究[J]. 图书情报工作, 2021, 65(8): 51-64.
[23] 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.
[24] Kwon H, Kim J, Park Y. Applying LSA Text Mining Technique in Envisioning Social Impacts of Emerging Technologies: The Case of Drone Technology[J]. Technovation, 2017, 60: 15-28.
[25] 丁敬达, 钟建兰. 新兴主题属性量化研究综述[J]. 图书情报工作, 2023, 67(9): 12-22.
[26] Stahl B C. What Does the Future Hold? A Critical View of Emerging Information and Communication

Technologies and Their Social Consequences[C]//Chiasson M, Henfridsson O, Karsten H, et al. Researching the Future in Information Systems. Berlin, Heidelberg: Springer, 2011: 59-76.
[27] 李仕明, 李平, 肖磊. 新兴技术变革及其战略资源观[J]. 管理学报, 2005, 2(3): 304-306, 361.
[28] Day G S, Schoemaker P J. Peripheral Vision: Detecting the Weak Signals That Will Make or Break Your Company[M]. Harvard Business Press, 2006.
[29] Cho Y Y, Jeong G H, Kim S H. A Delphi Technology Forecasting Approach Using a Semi-Markov Concept[J]. Technological Forecasting and Social Change, 1991, 40(3): 273-287.
[30] Lee S, Kim W, Kim Y M, et al. The Prioritization and Verification of IT Emerging Technologies Using an Analytic Hierarchy Process and Cluster Analysis[J]. Technological Forecasting and Social Change, 2014, 87: 292-304.
[31] 张金柱, 王秋月, 仇蒙蒙. 颠覆性技术识别研究进展综述[J]. 数据分析与知识发现, 2022, 6(7): 12-31.
[32] 叶春蕾, 冷伏海. 基于共词分析的学科主题演化方法改进研究[J]. 情报理论与实践, 2012, 35(3): 79-82.
[33] Dotsika F, Watkins A. Identifying Potentially Disruptive Trends by Means of Keyword Network Analysis[J]. Technological Forecasting and Social Change, 2017, 119: 114-127.
[34] Momeni A, Rost K. Identification and Monitoring of Possible Disruptive Technologies by Patent-Development Paths and Topic Modeling[J]. Technological Forecasting and Social Change, 2016, 104: 16-29.
[35] 黄鲁成, 成雨, 吴菲菲, 等. 关于颠覆性技术识别框架的探索[J]. 科学学研究, 2015, 33(5): 654-664.
[36] 张光宇, 曹阳春, 戴海闻, 等. 颠覆性创新国际研究 25 年回顾: 基于文献计量分析[J]. 科技管理研究, 2021, 41(15): 1-10.
[37] Xu H Y, Winnink J, Pang H S, et al. Breakthrough Potential of Emerging Research Topics Based on Citation Diffusion Features[J]. Journal of Information Science, 2023, 49(5): 1390-1416.
[38] 李欣, 王静静, 杨梓, 等. 基于 SAO 结构语义分析的新兴技术识别研究[J]. 情报杂志, 2016, 35(3): 80-84.
[39] Ha T, Yang H, Hong S. Automated Weak Signal Detection and Prediction Using Keyword Network Clustering and Graph Convolutional Network[J]. Futures, 2023, 152: 103202.
[40] Ansoff H I. Managing Strategic Surprise by Response to Weak Signals[J]. California Management Review, 1975, 18(2): 21-33.
[41] 李金泽, 夏一雪, 张鹏, 等. 突发舆情事件的情报感知模型研究[J]. 情报理论与实践, 2021, 44(10): 119-128.
[42] Veen B L, Ortt J R. Unifying Weak Signals Definitions to Improve Construct Understanding[J]. Futures, 2021, 134: 102837.
[43] 韩盟, 陈悦, 王玉奇, 等. 弱信号识别研究综述: 寻找微弱的未来信号[J]. 情报学报, 2023, 42(8): 996-1008.
[44] Hiltunen E. The Future Sign and Its Three Dimensions[J]. Futures, 2008, 40(3): 247-260.
[45] Yoon J. Detecting Weak Signals for Long-Term Business Opportunities Using Text Mining of Web News[J]. Expert Systems with Applications, 2012, 39(16): 12543-12550.
[46] 刘鹏, 王炳森, 张珂, 等. 基于弱信号感知的“卡脖子”关键核心技术甄别与竞争态势分析——以深海潜水器为例[J]. 情报理论与实践, 2025, 48(6): 102-112.
[47] Kim J, Lee C Y. Novelty-Focused Weak Signal Detection in Futuristic Data: Assessing the Rarity and Paradigm Unrelatedness of Signals[J]. Technological Forecasting and Social Change, 2017, 120: 59-76.
[48] Mendonça S, Cardoso G, Caraça J. The Strategic Strength of Weak Signal Analysis[J]. Futures, 2012, 44(3): 218-228.
[49] Bzhalava L, Kaivo-oja J, Hassan S S. Digital Business Foresight: Keyword-Based Analysis and CorEx Topic

Modeling[J]. Futures, 2024, 155: 103303.
[50] 韩盟, 陈悦, 王玉奇, 等. 新兴技术弱信号识别: 理论模型与测度方法[J]. 科学学研究, 2024, 42(11): 2262-2274.
[51] 王莉晓, 陈伟, 邱含琪. 基于机器学习的颠覆性技术弱信号识别模型研究[J]. 数据分析与知识发现, 2024, 8(S1): 63-75.
[52] 陈伟, 王莉晓. 颠覆性技术弱信号识别漏斗模型研究[J]. 图书情报工作, 2024, 68(8): 97-111.
[53] 吴柯烨, 孙建军, 张力, 等. 弱链接突变视角下的技术机会识别研究[J]. 图书情报工作, 2024, 68(10): 81-96.
[54] Ma M, Mao J, Li G. Discovering Weak Signals of Emerging Topics With a Triple-Dimensional Framework[J]. Information Processing & Management, 2024, 61(5): 103793.
[55] 唐虎林, 苏成, 李曼迪, 等. 基于弱信号的颠覆性技术早期识别方法研究[J]. 图书情报工作, 2025, 69(10): 42-61.
[56] 刘俊婉, 庞博, 徐硕. 基于弱信号的颠覆性技术早期识别研究[J]. 情报学报, 2023, 42(12): 1395-1411.
[57] 韩盟, 陈悦, 王玉奇, 等. 基于异类数据和语义建构的新兴技术弱信号识别研究[J]. 情报学报, 2024, 43(3): 302-312.
[58] Griol-Barres I, Milla S, Millet J. Improving Strategic Decision Making by the Detection of Weak Signals in Heterogeneous Documents by Text Mining Techniques[J]. AI Communications, 2020, 32(5/6): 347-360.
[59] Griol-Barres I, Milla S, Cebrián A, et al. Detecting Weak Signals of the Future: A System Implementation Based on Text Mining and Natural Language Processing[J]. Sustainability, 2020, 12(19): 7848.
[60] 翟东升, 郭程, 张杰, 等. 基于专利的企业潜在研发伙伴推荐方法研究[J]. 数据分析与知识发现, 2017, 1(3): 10-20.
[61] 陈友琼, 师庆科, 王觅也, 等. 临床科研大数据共享平台建设与应用[J]. 医疗卫生装备, 2024, 45(4): 27-31.
[62] 逯万辉. 科学文献主题建模方法及其效果评估研究[J]. 现代情报, 2024, 44(4): 22-31.
[63] 曹树金, 曹茹烨. 基于研究主题和引文分析的信息资源管理学科发展探究[J]. 信息资源管理学报, 2023, 13(2): 12-29.
[64] Egger R, Yu J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts[J]. Frontiers in Sociology, 2022, 7: 886498.
[65] 杨思洛, 于永浩. 基于 BERTopic 模型的国内信息资源管理研究主题挖掘与演化分析[J]. 情报科学, 2024, 42(8): 12-21.
[66] Shanmugam R. Measuring Research: What Everyone Needs to Know[J]. Journal of Statistical Computation and Simulation, 2019, 89(3): 557-558.
[67] 陈稳, 陈伟. 基于计量指标多变量LSTM模型的新兴主题热度预测研究[J]. 数据分析与知识发现, 2022, 6(10): 35-45.
[68] Krigsholm P, Riekkinen K. Applying Text Mining for Identifying Future Signals of Land Administration[J]. Land, 2019, 8(12): 181.
[69] Schett G, Müller F, Taubmann J, et al. Advancements and Challenges in CAR T Cell Therapy in Autoimmune Diseases[J]. Nature Reviews Rheumatology, 2024, 20(9): 531-544.
[70] Baulu E, Gardet C, Chuvin N, et al. TCR-Engineered T Cell Therapy in Solid Tumors: State of the Art and Perspectives[J]. Science Advances, 2023, 9(7): eadf3700.
[71] Saotome K, Dudgeon D, Colotti K, et al. Structural Analysis of Cancer-Relevant TCR-CD3 and Peptide-MHC Complexes by cryoEM[J]. Nature Communications, 2023, 14: 2401.
[72] Oncology Center of Excellence. Oncology Cell and Gene Therapy[EB/OL]. (2024)[2025-05-24]. https://www.fda.gov/about-fda/oncology-center-excellence/oncology-cell-and-gene-therapy.