基于多元异构数据的前瞻性技术识别和态势研判研究

文◆国网上海市电力公司电力科学研究院　陈予欣　王　娜　张鹏飞
上海久隆企业管理咨询有限公司　李　永

引言

新一轮科技革命和产业变革加速推进，人工智能、新能源等新兴技术不断涌现。如何及时识别并把握其演化趋势，已成为推动新型电力系统建设与能源转型的关键。现有前瞻性技术识别方法虽涵盖专家判断、文献与专利计量分析、机器学习自动化识别等，但仍存在三方面不足。一是依赖单一数据源，难以全面呈现技术演进；二是方法停留在浅层统计，难以捕捉深层语义；三是评价体系不完善，标准主观性强且可复制性差。为此，本文提出融合论文、专利等多元异构数据的识别方法，揭示技术演化趋势，并构建融合语义表示、时间建模与神经网络分析的前瞻性技术识别与研判方法，为战略布局提供数据支持与决策参考。

1 前瞻性技术识别方法研究综述

一是文本分类法。主要用于将大量非结构化文本按照内容划分至对应类别。早期方法依赖专家手工制定分类规则。进入机器学习阶段后，朴素贝叶斯、支持向量机等算法被广泛应用。二是主题识别法。传统统计方法如 TF-IDF 由 Luhn（1957）[1] 提出，后由 SparckJones（1972）[2] 引入逆文档频率（IDF）概念，用于关键词提取。Blei 与 Lafferty（2006）提出动态主题模型（DTM），Wang 等（2008）进一步提出连续时间模型（CTM）[3]，可刻画主题随时间的演化趋势。三是深度学习模型。BERT 模型由 Devlin 等（2019）[4] 提出，具备强大的上下文建模能力，常与聚类算法结合用于语义主题识别。此外，BP 神经网络也被用于前瞻性技术评价指标的建模任务，在本研究中发挥重要作用。

2 研究方法与技术路径

为实现多元异构数据驱动下的前瞻性技术识别与态势研判，本研究构建了涵盖数据获取、主题提取、指标评价与模型预测的系统性研究路径。

2.1 前瞻性技术主题提取

采用 BERTopic 与 CTM 相结合的方法对多元异构文本中前瞻性技术主题识别。首先，通过 BERT 模型对文本进行向量化表示，获取词语的上下文嵌入表示。其次，利用 UMAP 对高维语义向量降维，并通过 HDBSCAN 实现无监督聚类，从而识别出若干潜在的技术主题。最后，将聚类结果输入 CTM 模型，对主题在时间维度上进行连续建模，提取其生命周期曲线与变化趋势。

2.2 构建前瞻性技术主题评价体系

为科学判断潜在技术主题的前瞻性特征，构建包含“技术新颖性、技术持续性、技术影响力、技术扩散性”四大维度的前瞻性

【作者简介】陈予欣（1998—），女，上海人，工程管理学硕士，研究方向：电力市场、智能电网及电力大数据等方面。

技术主题评价指标体系（见表 1）。

（1）篇均非专利引用量。R_{avg} 为篇均非专利引用量，C_i 为第 i 篇文献的非专利类引用次数，N 为技术主题相关文献的总篇数。具体计算公式如式（1）所示。

$$R_{avg}=\frac{\sum_{i=1}^{N}C_i}{N} \quad (1)$$

（2）篇均科学引用量与总引用量的比率。R_{ratio} 为篇均科学引用量与总引用量的比率，$R_{science}$ 为篇均科学引用量，R_{total} 为总引用量，计算公式如式（2）所示。

$$R_{ratio}=\frac{R_{science}}{R_{total}} \quad (2)$$

（3）时间跨度。S_{time} 为时间跨度，单位为月，T_{last} 为技术主题最后一次被文献或专利记录的时间，T_{first} 为技术主题首次被文献或专利记录的时间，具体计算公式如式（3）所示。

$$S_{time}=T_{last}-T_{first} \quad (3)$$

（4）热度持续性。$S_{continuity}$ 为热度持续性得分，T 为技术主题在其生命周期内的总时间窗口数，I_t 为时间窗口 t 的热度值是否达到阈值（阈值为 0 或 1），H_t 为时间窗口 t 的热度值，θ 为热度阈值，表示最小的有效热度值。具体计算公式如式（4）所示。

$$S_{continuity}=\frac{\sum_{t=1}^{T}I_t}{T} \quad (4)$$

（5）学术影响力。$I_{academic}$ 为学术影响力得分，N 为技术主题对应的论文总数，W_p 为第 p 篇论文的权重，若该论文发表在核心期刊上，则 W_p=1，否则 W_p=0，计算公式如式（5）所示。

$$I_{academic}=\sum_{p=1}^{N}W_p \quad (5)$$

（6）专业转化能力。$T_{conversion}$ 为专业转化能力得分，$N_{enterprise}$ 为技术主题下由企业持有的专利数量，N_{total} 为技术主题下的专利总数量，计算公式如式（6）所示。

$$T_{conversion}=\frac{N_{enterprise}}{N_{total}} \quad (6)$$

表 1 前瞻性技术主题识别指标体系

一级指标	二级指标	数据来源
技术新颖性	篇均非专利引用量	论文及专利引用数据
	篇均科学引用量与总引用量的比率	论文及专利引用数据
技术持续性	时间跨度	CTM 曲线
	热度持续性	论文及专利发表数量
技术影响力	学术影响力	核心期刊论文数量
	专业转化能力	专利拥有情况
技术扩散性	学术扩散性	篇均中图分类数量
	专业扩散性	篇均技术分类（IPC）数量

（7）学术扩散性。AE 为学术扩散性，N 为技术主题相关的论文总数，D_i 为第 i 篇论文的中图分类数量，计算公式如式（7）所示。

$$AE=\frac{1}{N}\sum_{i=1}^{N}D_i \quad (7)$$

（8）专业扩散性。PE 为专业扩散性，M 为技术主题相关的专利总数，C_i 为第 i 件专利的 IPC 分类数量，计算公式如式（8）所示。

$$PE=\frac{1}{M}\sum_{i=1}^{M}C_i \quad (8)$$

各指标通过归一化处理计算得分，并结合熵权法确定各指标权重，最终汇总形成前瞻性技术得分，用于不同技术主题间的横向比较与排序。引入 BP 神经网络模型，对各技术主题前瞻性得分进行非线性拟合与预测。

3 实证分析

3.1 数据来源与数据处理

对检索到的论文、专利信息进行数据筛选，共得到 15w+ 论文数据，300w+ 专利数据。最后对专利数据进行 0.5% 抽样得到 15w+ 专利数据，并提取关键信息字段。

3.2 技术主题识别与提取

通过文本嵌入模型，可以计算出文本的特征向量，选择使用 Sentence Transformer 进行文本嵌入，并使用 BERTopic 进行主题建模。经过主题建模后，将每条文本数据放入模型中，最终获得每条数据对应的主题和概率。

3.3 技术主题前瞻性评价

根据设计的技术主题评价体系对提取出的主题进行打分，排名前 5 名主题得分情况如表 2 所示。

3.4 基于大模型的前瞻性技术识别

利用 BP 的神经网络生成能力，为筛选后的主题生成主题名称，针

表 2 排名前 5 主题得分情况

主题	篇均总引用量	篇均科学引用量与总引用量的比率	时间跨度	热度持续性	学术影响力	专业转化能力	学术扩散性	专业扩散性	综合评分
19	0.49	0.001	0.01	0.09	1	0.6	1.667	3.95	0.822
27	0.036	0.002	0.022	0.248	5	0.5	1.495	4	0.522
263	0.07	0.01	0.024	0.043	4	1	1.621	0	0.484
12	0.495	0.001	0.005	0.851	3	0.5	1.182	2.25	0.406
418	0.5	1	0.006	0.03	0	0.849	2	2.161	0.36

表 3 排名前 5 主题及得分情况

主题名称	主题评分
新能源并网消纳	0.822
锂离子电池及储能技术	0.522
有机太阳能电池受体材料研究与应用	0.484
永磁电机设计与控制技术	0.406
摩擦纳米发电机及其应用	0.36

对每个技术主题，将同一个技术主题下的所有数据的篇名、关键词中文摘要合并后输入至 BP 神经网络，然后输出技术主题，排名前 5 主题及得分情况如表 3 所示。

3.5 前瞻性技术分析

通过对高分主题进行分析，为新型电力系统技术布局提供支撑。以新能源并网消纳技术为例，该主题在学术影响力（1.0）和扩散性方面得分较高，学术扩散性 1.667，专业扩散性 3.95，显示其具备跨学科传播与多行业应用能力。但其时间跨度（0.01）与热度持续性（0.09）较低，表明仍处于快速发展初期。作为新型电力系统建设的关键技术，该主题对提升可再生能源接入比例和电网灵活性意义重大。

结语

随着能源电力行业科技的快速发展，早期识别潜在颠覆性技术并评估其影响，成为推动产业升级和能源转型的关键。本研究基于多元异构数据，提出一种前瞻性技术识别方法，结合 BERTopic、CTM 等模型构建系统化识别与评估框架。通过前瞻性技术主题评价体系、时间演化建模与神经网络分析，将技术识别从单一维度拓展至多维综合评价，并利用 BP 神经网络优化结果，提高模型准确性和预测能力。本研究在前瞻性技术识别方法上取得了重要进展，但仍然存在一些不足之处。如何进一步提高模型的实时性和适应性，尤其是在大规模数据处理和复杂环境下，仍然是未来需要解决的重要问题。

引用

[1] Luhn H P.The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Developmen,1958,2(2):159-165.

[2] Jones S K.A Statistical Interpretation of Term Specificity and Its Application in Retrieval[J]. Journal of Documentation,2004,60(5):493-502.

[3] Wang C,Blei D,Heckerman D.Continuous Time Dynamic Topic Models.[C].Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence,Helsinki,Finland,2008:579-586.

[4] Devlin J,Chang M W,Lee K,et al.BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,vol.6.long and short papers:Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies(NAACL HLT 2019),2-7June 2019,Minnea polis,Minnesota,USA.:Association for Computational Linguistics,2019:4171-4186.