基于本体语义增强和多源数据融合的石墨烯医学应用前沿探测*

靳 杨 徐路路

（首都医科大学附属北京安贞医院 北京 100029） （南开大学信息资源管理系 天津 300071）

〔摘要〕 提出基于 WordNet 本体语义增强和多源数据主题贡献度分析的科学研究前沿探测方法，介绍相关研究情况、方法框架并进行实证分析，采用论文、基金项目及专利等多源数据综合识别出石墨烯医学应用领域未来发展方向。实验表明该方法能够弥补利用单一数据源进行医学前沿探测的不足，为石墨烯医学发展架构调整提供参考。

〔关键词〕 本体语义；多源数据；石墨烯医学探测；科学研究前沿

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.02.014

Frontier Detection of the Medical Application of Graphene Based on the Integration of Ontological Semantic Enhancement and Multi－sources Data

JIN Yang, *Beijing Anzhen Hospital Affiliated to Capital University of Medical Sciences*, *Beijing 100029*, *China*; *XU Lulu*, *Department of Information Resources Management*, *Nankai University*, *Tianjin 300071*, *China*

〔**Abstract**〕 The paper puts forward the scientific study frontier detection method based on WordNet ontological semantic enhancement and analysis of the contribution degree of multi－sources data topic, introduces related study situation and method framework, performs empirical analysis, as well as identifies the future development direction of graphene in the medical application field by using multi－sources data such as papers, fund projects and patents comprehensively. Experiments show that the method is able to make up for the defects in the implementation of medical frontier detection with single－sources data, and provides references for the adjustment of the medical development architecture of graphene.

〔**Keywords**〕 ontological semantics; Multi－sources data; graphene medical exploration; frontier of scientific study

1 引言

分析医学领域科技文献研究前沿主题信息可有效揭示出该领域新材料、新技术和新方法，从而优化布局发展[1]。如何从海量医学科技文献中识别研究前沿并对未来发展方向进行预测分析成为亟需解决的问题。但目前科学研究前沿存在诸多问题，如数据源单一（论文为主）、语义理解不足、多数据源无法交叉融合等，制约文本内容主题探测的有效性和准确度[2]，论文数据主题丰富但其研究前沿探测的前瞻性受到广泛质疑，规划文本等蕴含更多前瞻价值信息但主题粒度较大[3]。本文分析医学科学研究前沿中存在的主要问题和不足，提出 WordNet 本体语

〔修回日期〕 2018－09－14

〔作者简介〕 靳杨，硕士研究生；通讯作者：徐路路，博士。

〔基金项目〕 国家自然基金面上项目“LPR6 基因调控区变异调控血管内皮细胞 Wnt 通路的机制研究”（项目编号：81770353）。

义增强和多源数据主题贡献度分析，识别论文、基金项目数据以及专利文献中的石墨烯材料在医学领域的前沿主题。利用本体库 WordNet 丰富和拓展主题词语义信息，基于不同文本特征要素分析进行主题贡献度融合，构造多源数据融合的科学研究前沿探测公式，从而揭示石墨烯新材料领域竞争发展态势，为优化战略部署和重点领域大势研判提供情报支撑[4]。

2 相关研究

2.1 WordNet 研究与应用

WordNet 是普林斯顿大学 Miller 于 1985 年组织语言及心理学相关领域专家开发的大型英文词汇数据库[5]，采用语义网络作为其词汇概念本体的基本组成形式，将不同词汇以不同分类组织形式关联融合，形成语义本体。多年来众多学者基于 WordNet 丰富的语义描述能力及词汇覆盖度展开相关研究。1998 年 Fellbaum C 等基于基准语义消歧方法与融合 WordNet 相关词语进行语义相似度计算，实验证明该方法使排歧准确度有所提高[6]。2011 年王瑞琴等将 WordNet 本体和 WordNet Domains 扩展库作为消歧数据源，利用查询扩展技术建立查询关键词和本体概念的映射，提高信息检索准确度，满足多样化检索需求[7]。2013 年张泽宇等针对语义标注效率低下的问题提出基于 WordNet 语义知识的文档标注方法，实现对科技文献的有效标注与识别[8]。2015 年 X Zhu 利用 WordNet 在线语义词典提出基于语义和边权重的相似度计算方法，MC30 和 RG65 测试集实验分析表明该方法在计算性能和效率的优越性[9]。针对主题粒度较大的文本，如规划文本、基金项目数据等，利用 WordNet 拓展其语义信息作为主题内容的补充，进而利用主题概率识别模型识别其蕴含的前沿主题是未来前沿探测的有效方法之一。

2.2 科学研究前沿

1965 年 Price 从引用次数维度首次定义科学研究前沿[10]。1973 年 H. Small 将同被引文献的聚类分布结果定义为科学研究前沿[11]，围绕研究前沿内涵展开研究的还有 O. Persson 提出的高同被引文献关联的施引文献群以及 E. Garfield 提出的被引聚类的核心文献和引用该论文的最新文献研究前沿的概念[12-13]。2011 年张士靖等利用共被引分析和共词聚类分析方法对医学健康领域研究热点和前沿主题进行追踪并利用 Ucinet 进行可视化分析[14]。2012 年冷伏海等提出基于案例分析的科学前沿探测新方法，利用因子分析、战略坐标等多种方法综合分析学科领域研究前沿[15]，相关研究者还有白如江[16]、牟冬梅[17]等。研究前沿的有效探测对于学科未来发展规划具有重要指导意义。

3 方法框架

3.1 概述

为更加准确前瞻地识别出多种科技文本中蕴含的科学研究主题，本文提出基于 WordNet 语义增强和多源信息主题贡献度分析的科学研究前沿探测方法，对基金项目、论文、专利等数据进行主题贡献度分析并利用本体语义研究技术对探测得到的主题信息进行语义增强以提高主题探测的科学性和准确度。

3.2 WordNet 语义增强

主题概率识别模型可模拟科技文献生成过程，通过参数估计和先验概率抽取其主题信息，实现文本内容深度挖掘，是目前前沿识别中重要方法。然而该方法也存在不足，侧重于量化统计和概率分布研究，忽略科技文本语义理解和词汇语义关联，如对 energies 和 energy、application 和 using 等词形不同但词义相同的主题词无法有效识别并权重叠加，另外也产生较多的噪音数据，降低前沿准确性和科学性。语义角色标注可对科技文献内容信息进行分析及解读，增强语义信息理解，目前主要方法有语义角色标注（句子粒度浅层语义分析）和基于本体语义增强研究（词语粒度概念映射）两种。基于本体语义分析方法可将表征研究前沿信息的主题词语义映射，进而识别其上位词（hypernym）、近义词（homoionym）等语义信息，归类同语义信息关键词并调整权重分配，深入挖掘概念语义类型，其中 WordNet 是较为成熟的英文语义本体库。本文提出

基于主题概率识别模型的语义增强方法，将主题词袋概念映射为概念词袋以增强其语义信息，调整主题词分布及权重，以提高科学研究探测的准确度。基于主题概率模型语义增强处理，见图1。

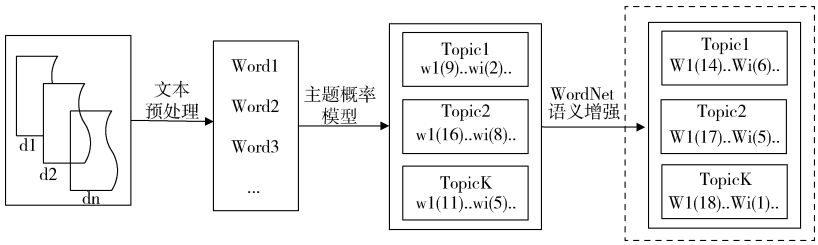

图1 基于主题概率模型语义增强处理

3.3 多源科学研究前沿分析

科学研究前沿蕴含在不同的科技文本数据源中，如基金项目、专利文本、论文数据等，且不同数据源具有不同文本特征和要素，因此不应以单一论文数据为分析数据源，扩展数据源类型进行多源数据主题交叉融合识别是重要发展趋势。基金项目是由国家组织部署相关研究机构实施的科技创新任务，其经过专家论坛探讨往往代表相关学科优先领域，研究主题具有“将来时”属性，利用基金项目进行科学研究前沿分析在主题新颖度和前瞻性方面贡献权重大，但同时基金项目数据文本量较小，主题较为宏观，粒度较大而主题丰富度不足。专利文献数据庞大且蕴含着丰富的技术信息，是进行情报分析和前沿识别的重要数据源，反映某学科领域的关键技术和方法流程，专利可视为基金项目当前阶段的具体部署和解决方案，具有“现在时”属性，其前瞻信息价值较高，主题较为丰富。而论文数据从产生到发表再到引用需要一定时间，存在一定的滞后性，其“过去时”特征使其在前沿识别中前瞻性较低，但论文数量较多，主题丰富度较高。因此有效融合多源数据提高研究前沿水平十分必要。

3.4 基于多源数据的科学研究前沿公式构建

本文分析研究前沿相关研究，利用不同阶段的主题发展特性可以对科学研究前沿多主题类型进行区分，即分为新兴研究、热点研究和未来研究前沿主题3 种多源数据分析主题类型。主题强度是指学科主题的主题研究热度及发展程度，可通过主题内部关键词频次及权重表征学科主题强度。融合多源数据特征首先识别其主题强度并根据上述分析进行贡献度分析融合。主题强度指标如下：

$$TII_t^z = \sum_{i=1}^{n} weight(k_i) \quad (1)$$

其中 $weight(k_i)$ 表示在 t 时间段主题 z 的主题词 k 所占权重比值；$\sum_{i=1}^{n} weight(k_i)$ 反映 t 时间段内该主题累计主题词权重值，即为该主题权重值，表示为 TII_t^z。该指标的平均值即为平均主题强度指标（Average Topic Intensity Index，$ATTI_t$），可反映基金项目中主题强度的平均水平。

考虑多源数据主题丰富度和新颖度两个参量，在主题粒度可利用主题概率模型予以表征，粒度大的主题其识别出的数值较大；主题前瞻价值则是根据上述分析设定相应的主题前瞻价值系数。因此本文提出针对多源数据类型的科学研究前沿探测公式：

$$RFDF_z = \alpha \times \frac{TII_{t\_f}^z}{ATTI_{t\_f}} + \beta \times \frac{TII_{t\_p}^z}{ATTI_{t\_p}} + \lambda \times \frac{TII_{t\_t}^z}{ATTI_{t\_t}} \quad (2)$$

公式中 α，β，λ 为不同数据源的主题贡献度系数以表征基金数据及专利论文在前沿探测中的主题贡献度大小，3 个子项分别为基金项目、专利及论文主题强度指标，利用贡献度系数调谐统一，最终得到科学研究前沿探测公式（Research Front Detection Formula，$RFDF_z$）。

4 实证研究

4.1 数据集获取

石墨烯具有独特的蜂窝纳米结构，目前在分子

化学、航空航天等领域取得广泛应用，分散性、生物相容性、亲水性等特质使其在生物医学领域具有广阔的应用前景和价值。因此本文利用科学研究前沿探测方法识别石墨烯在生物医学领域研究动向。石墨烯生物医学领域数据检索，见表1。

表1 石墨烯生物医学领域数据检索

数据类型	项目文本	论文文本	专利文本
数据库	NSF 数据库	Web of Science 核心合集	Derwent Innovation Index
检索方式	Keyword = " graphene * " and" biomedicine"	Keyword = " graphene * " and" biomedicine"	Keyword = " graphene * " and" biomedicine"
时间跨度	2008 – 2017 年	2008 – 2017 年	2008 – 2017 年
检索结果	439 项	10 954 篇	1 054 件

4.2 数据预处理

新兴主题探测在于第一时间发现具有较大潜力而未引起广泛关注的主题，因此将子时期单位设置为1 年可较早识别短时间内突发主题词。为保证足够数据进行主题分析，本文以2008 年为时间起始，以每年为时间单位进行细粒度时间切片处理，得到10 个子时期。实验发现权重系数 α 取 0.4，β 取 0.35，λ 取 0.25 效果最好。

4.3 基于 WordNet 语义增强的前沿识别

4.3.1 参数设置与主题表征 选用 Kmine 实验平台的 LDA 模型进行主题识别。相关参数设置：No of topic 主题数 40；No of words per topic 每个主题的词数 10；Alpha 0.5；Beta 0.1；No of iteration 迭代次数 2 000；No of thread 线程数 8；复杂度为 100。对 10 个子时期（2008 – 2017 年）的基金项目数据集进行主题建模，得到主题 – 主题词 – 项目序号的多维映射关系。对利用 LDA 模型得到的文档 – 主题及主题 – 主题词映射进行语义处理，将主题词袋概念映射为概念词袋以增强其语义信息，合并同语义信息主题词并调整主题词分布及权重，使主题识别实验更为准确和科学。WerdNet 语义增强处理，见表2。选取 2011 年度主题识别对基于传统主题概率模型方法和语义增强处理主题识别方法进行对比，由表2 中 Topic0 相关主题词可知该主题主要描述石墨烯生物化学相关特性与纳米级衍生物材料研究，其中材料（material）和材质（materials）以及电子（electronic）和电流（electro）存在语义相关，将同语义主题词权重叠加使主题表达更为准确，同时一定程度上增加低权重主题词的识别效果，语义处理可细化主题识别效果。

表2 WordNet 语义增强处理

Topic 0				Topic 2			
Word	Pro.	Word	Pro.	Word	Pro.	Word	Pro.
graphene	134	graphene	134	project	80	project	80
materials	78	materials	97	optical	42	remedy	6
electronic	48	electronic	66	device	35	optical	42
properties	43	properties	43	remedy	31	electrical	31
application	42	application	42	electrical	31	graphene	29
mechanical	30	material	30	semiconductor	27	semiconductor	27
antibiosis	30	antibiois	30	graphene	23	growth	22

续表 2

project	28	project	28	growth	22	transport	20
material	19	hydroxyl	14	transport	20	property	6
electro	18	filter	9	graphite	6	ion	5

4.3.2 石墨烯生物医学前沿分析 其识别出 3 个热门研究前沿主题，即两个新兴科学研究前沿主题及 1 个未来科学研究前沿主题。本部分结合探测主题词及强度值进行生物医学领域应用分析。（1）热门研究前沿主题 topic 0、topic 4 和 topic 8。该主题目前阶段的重要研发热点和科技竞争区域主要围绕氧化石墨烯生物探测器设备研发用于多肽蛋白质等生物分子检测；分析羧基、羟基等诸多功能基团对于荧光淬灭效率以及信号自动放大等石墨烯生物应用方面的探索。该领域目前研究成果较多、主题强度较多，是目前及未来一段时间内的科技竞争领域。（2）新兴科学研究主题 topic 2、topic 9。该主题属于新兴、具有较大未来发展潜力的前瞻科学研究前沿主题，主要围绕氧化石墨烯光学特性、生物光热治疗以及光储存和数据保存等方面展开：光敏剂的载体对于肿瘤等细胞的周期作用机制探索以及石墨烯与亚甲蓝等多种复合物光数据的保存等相关研究。该研究主题未来发展潜力巨大，研究逐步开展在未来有望成为热门主题。（3）未来科学研究主题 topic 5。目前该主题的主题探测值低于平均水平，相关研究有待于进一步开展，但在未来有较大的研究潜在价值和应用场景。主要围绕石墨烯氧化抗菌性能、细胞膜结构破坏以及石墨烯生物安全性和毒性作用机理研究，探究石墨烯颗粒大小、状态以及其氧含量在生物毒性响应研究；石墨烯材料对于红细胞的脂质双分子层破坏作用研究。

5 结语

本文针对目前研究中主要利用论文数据进行科学研究前沿识别中存在的时滞性问题以及在主题识别中欠缺语义理解而导致探测准确度不足的问题，提出基于本体 WordNet 语义增强和多源数据主题贡献度分析的科学研究前沿探测方法，利用石墨烯生物医学领域的实证研究，采用文献调研方法，验证本文提出多源数据分析的科学研究前沿识别方法的可行性和有效性。未来将围绕石墨烯生物医学应用研究展开进一步研究，拓展分析数据源并构建针对多源数据的综合研究前沿识别框架，为我国科学研究提供决策支撑和部署建议。

参考文献

1 崔雷，陈东滨. 国外医学信息学科研热点的文献计量学分析［J］. 医学信息学杂志，2007，28（2）：97－102.

2 徐路路，王效岳，白如江. 一种基于 TDT 模型的基金项目科学研究前沿识别方法研究［J］. 情报理论与实践，2018，41（8）：72－78.

3 刘小平，冷伏海，李泽霞. 国际科技前沿分析的方法和途径［J］. 图书情报工作，2012，56（12）：60－65.

4 徐路路，王效岳，白如江，等. 基于 DTM 模型和文本特征分析的基金项目新兴趋势探测研究——以 NSF 石墨烯领域为例［J］. 数据分析与知识发现，2018，2（3）：87－97.

5 Miller G A，Beckwith R，Fellbaum C，et al. Introduction to WordNet：an on－line lexical database［J］. International Journal of Lexicography，1990，3（4）：235－244.

6 Fujii A，Inui K，Tokunaga T，et al. Selective Sampling of Effective Example Sentence Sets for Word Sense Disambiguation［J］. Computer Science，1997，24（4）：573－597.

7 王瑞琴，孔繁胜. 基于无导词义消歧的语义查询扩展［J］. 情报学报，2011，30（2）：131－137.

8 张泽宇，李莉，谭凤，等. 基于语义的文档标注方法研究［J］. 计算机工程与科学，2013，35（9）：151－156.

9 郭小华，彭琦，邓涵，等. 基于边权重的 WordNet 词语相似度计算［J］. 计算机工程与应用，2018，54（1）：172－178.

10 Price D J D S. Networks of Scientific Papers［J］. Science，1965，149（3683）：510.

（下转第 85 页）

http：//www. nhfpc. gov. cn/xcs/s3582/201612/1c81b23264df4880829263b18a2f0947. shtml.

7 中华人民共和国国家卫生健康委员会宣传司. 2016 年我国居民健康素养监测结果发布［EB/OL］.［2018 - 09 - 10］. http：//www. nhfpc. gov. cn/xcs/s3582/201711/308468ad910a42e4bbe9583b48dd733a. shtml.

8 董子畅. 2017 年中国居民健康素养水平为 14. 18% 呈持续上升态势［EB/OL］.［2018 - 10 - 10］. http：//www. chinanews. com/jk/2018/09 - 19/8631500. shtml.

9 云南省卫生计生委宣传处，云南省健康教育所. 2014 年云南省居民健康素养水平监测结果通报［EB/OL］.［2018 - 09 - 10］. http：//www. ynjkjy. com/cms/document/detail/id/3141. html.

10 云南省卫生计生委宣传处，云南省健康教育所. 2015 年云南省居民健康素养水平监测结果通报［EB /OL］.［2018 - 09 - 10］. http：//www. ynjkjy. com/cms/document/detail/id/3400. html.

11 陈鑫龙. 全省提高人均预期寿命工作稳步推进［EB /OL］.［2018 - 10 - 10］. http：//www. yn. gov. cn/yn_ynyw/201807/t20180709_ 33322. html.

12 Sakai Y. Health Literacy Research and the Contribution of Library and Information Science：to aspects of consumer health information services［J］. Library & Information Science, 2008 (59)：117 - 146.

13 中华人民共和国教育部. 教育部关于印发《普通高等学校图书馆规程》的通知［EB/OL］.［2018 - 09 - 10］. http：//www. moe. edu. cn/srcsite/A08/moe_ 736/s3886/201601/t20160120_ 22848 7. html.

14 仇晓春. 医学院校图书馆的职能与展望［J］. 中华医学图书情报杂志，2018，27 (1)：54 - 57.

15 尹明章. 医学院校图书馆与公众健康信息素养［J］. 中华医学图书情报杂志，2011，20 (1)：34 - 35，59.

16 昆明医科大学教务处. 学校慕课《健康生活 预防癌症》入选 2017 年国家精品在线开放课程［EB /OL］.［2018 - 09 - 10］. http：//www. kmmc. cn/Pages_ 2_ 29751. aspx.

17 徐海燕. 高校图书馆社会化信息服务实践探索——以医学院校图书馆为例［J］. 图书馆理论与实践，2017 (4)：75 - 77，112.

18 李朝阳. 省情概述人口及民族［EB /OL］.［2018 - 09 - 10］. http：//www. yn. gov. cn/yn_ yngk/gsgk/201509/t20150923_ 22230. html.

19 陈习琼. 云南省人口发展现状及特征分析——基于第六次人口普查数据分析［J］. 经济研究导刊，2016 (33)：113 - 116.

20 焦玲霞，谭世芬. 医学院校图书馆开展公众健康信息服务举措探讨［J］. 管理观察，2017 (8)：160 - 161，164.

21 张士靖，周彦霞，陶亚萍. 医学图书馆服务的典范——美国 NN/LM 的服务及其启示［J］. 图书馆建设，2008 (3)：105 - 108.

(上接第 74 页)

11 Small H, Griffith B C. The Structure of Scientific Literatures I：identifying and graphing specialties［J］. Science Studies, 1974, 4 (1)：17 - 40.

12 Morris S A, Yen G, Wu Z, et al. Time Line Visualization of Research Fronts［J］. Journal of the Association for Information Science & Technology, 2003, 54 (5)：413 - 422.

13 Chen C. CiteSpace II：detecting and visualizing emerging trends and transient patterns in scientific literature［J］. Journal of the Association for Information Science & Technology, 2006, 57 (3)：359 - 377.

14 张士靖，郭海红，刘小利，等. 国际健康素养领域研究现状、热点与前沿的可视化分析［J］. 医学信息学杂志，2011，32 (4)：36 - 41.

15 张英杰，冷伏海. 基于案例的科学前沿探测方法比较研究［J］. 图书情报工作，2012，56 (20)：42 - 46.

16 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.

17 潘玮，牟冬梅，李茵，等. 关键词共现方法识别领域研究热点过程中的数据清洗方法［J］. 图书情报工作，2017，61 (7)：111 - 117.