

技术领域主题发现研究 ——以基因工程疫苗领域为例*

隗玲^{1,2}, 许海云², 刘春江², 李婧², 方曙²

(1.山西财经大学信息管理学院, 太原 030006; 2.中国科学院成都文献情报中心, 成都 610041)

摘要: 本文梳理基于专利文献进行技术领域主题发现的研究方法发展进程, 提出一套高效获取专利技术主题词、生成战略坐标图的研究流程, 并以基因工程疫苗技术为例, 对我国基因工程疫苗技术主题分布及发展趋势进行实证分析。研究结果显示, 采用本文构建的共词分析流程和方法, 利用专利共词聚类 and 战略坐标图能较好地识别技术主题及其发展现状。

关键词: 技术主题发现; 共词分析; 战略坐标图; 基因工程疫苗

中图分类号: G350

DOI: 10.3772/j.issn.1673-2286.2017.1.006

1 引言

专利技术研发是一项不断创新和更新换代的科学技术活动, 是科学技术发展过程中的重要环节。专利技术主题是专利文献揭露技术内容的主题和核心, 技术领域主题发现指从专利文献中识别该技术领域技术主题的分布或演化, 其中技术主题演化又包括技术主题演变过程、技术发展趋势预测和新兴技术主题发现等内容^[1]。掌握某领域技术主题分布及演化情况, 有助于相关人员了解该技术领域的内容以及研究的热点和发展趋势等, 为科研机构和企业提供前沿预警和参照, 也为国家相关政策制定提供决策支撑。

利用专利文献进行技术领域主题发现的研究方法有多种, 文本挖掘方法因其能深入专利文献的具体内容, 已成为当前技术主题发现研究的重要方法。不同的文本挖掘方法各有其优势和不足, 其中共词分析方法(尤其是战略坐标图)既能反映词与词间的个体关系, 又能反映词群与词群间的整体关系, 原理简单、操作易行, 是一种可用于技术领域主题发现的较成熟科学的方法。国内外学者利用战略坐标图对机器人技术^[2]、燃

料电池^[3]、能源材料^[4]、移动通信技术^[5]、RFID技术^[6]等多个技术领域进行技术主题发现研究, 但仍存在不足。如需要自主开发术语抽取工具, 研究成本高, 方法重用性差^[3]; 需要对技术关键词进行人工清洗, 用于主题发现的关键词数量少^[4,6]; 需要对专利文献的技术术语进行人工预处理, 耗时长、准确性差等^[5]。这些不足之处制约了共词分析方法在情报分析中的有效运用, 鉴于此, 本文结合普适的文本信息处理工具和可视化工具, 提出一套高效获取大量专利技术主题词、生成战略坐标图的研究流程, 并对我国基因工程疫苗领域的技术主题分布及发展趋势进行实证分析。

2 技术领域主题发现研究方法

随着技术发展演化的进程, 根据不同研究方法和特点和作用, 基于专利文献进行技术领域主题发现的研究方法可分为传统的专利分类分析法、基于文本挖掘的技术领域主题发现方法和其他方法(主要指引文分析法和社会网络分析法)三大类。本文重点对前两类方法进行梳理, 对引文分析方法和社交网络分析方法

* 本研究得到成都市软科学项目“基于技术主题识别的产业协同创新模型研究”(编号: 2015-RK00-00058-ZF)资助。

不多赘述。

2.1 传统的专利分类分析法

传统的专利分类分析法主要有统计分析法和共类分析法。统计分析法主要用于统计某技术领域专利的整体产出情况、排名靠前的分类号等以揭示该技术领域在地域、研发机构和核心技术主题的发展变化^[7-8]。共类分析法最早于20世纪80年代提出,是基于文献计量的跨学科的基本研究方法,其原理是首先将论文所在期刊按主题进行归类,然后将主题类别按学科类别或领域进行归类^[9]。在专利分析中,专利共类被定义为“2个或2个以上的分类号在多篇专利文献中共同出现的次数反映了分类号所代表的领域间研究关联程度”^[10]。一般认为,同一论文、期刊和专利可能归属于多个不同的主题类别和技术领域,反映出科技成果的学科交叉性,揭示某一科技领域研究内容的内在相关性和学科领域的微观结构。Spasser^[11]、Kumaresan^[12]、Schummer^[13]、Meyer^[14]和林瑞明等^[15]先后使用共类分析法对制药学领域、日本机器人技术领域、纳米科技领域、LED技术领域的技术主题发展进行研究。共类分析法常和聚类法^[16-17]或社会网络分析法^[18-19]结合使用。

传统的专利分类分析法操作简单易行,可从整体上直观呈现某技术领域主题的构成与变化,但对技术主题的划分过分依赖专利分类号,分析不能深入到专利文本内容,在技术领域主题发现方面的应用具有较大局限性。

2.2 三类基于文本挖掘的技术领域主题发现方法

随着技术领域不断交叉与渗透融合,新技术跨领域性特征越来越突出,技术领域主题发现面临的挑战也越来越大。专利文本除题名、作者等专利元数据字段外,摘要、技术背景、权利要求甚至全文包含丰富的语义内容可供挖掘。

Porter等将专利文本技术挖掘流程总结为确定待挖掘专利数据集、专利文本知识表示、专利技术挖掘场景分析、技术挖掘结果评估与修订四个主要步骤^[20]。在技术挖掘场景设定为技术领域主题发现的前提下,利用专利文本进行技术挖掘的关键工作为专利文本知识表示及基于知识表示单元的主题挖掘。专利文本知

识表示首先需要将非结构化的专利文本信息转化为计算机可识别的结构化信息,然后根据需要进行语义化建模。这些结构化信息称为知识表示单元,简称知识单元。知识单元可以是词语、词组、句子及段落,当前常选用词语作为知识单元。知识单元获取是自动识别文本主题的前提,知识单元的有效提取及关联关系的正确计算共同决定文本主题的识别效果^[21]。本文根据知识单元的形式不同和语义建模的方法不同,将基于文本挖掘的技术领域主题发现方法归纳为三类。

2.2.1 基于知识单元词频统计的发现方法

该方法以知识单元自身为处理对象,基于知识单元获取技术主题,包括术语词频统计。术语词频统计常见的应用是通过提取专利项、摘要或标题等中的技术术语,以技术术语反映技术主题,统计高频术语的分布情况以分析某技术领域的热点分布^[22],也可在分析中加入时间轴统计主题术语词频的时间变化或不同时间段的词频分布以分析技术主题的发展或变化趋势^[23-24]。术语词频统计也常用于共词分析的前期步骤,用来对技术关键词进行筛选和分类^[25-26]。

术语词频统计方法原理简单,操作便利,但是无法体现词与词间的关联,不能揭示技术领域主题间的关联。

2.2.2 基于知识单元共现的发现方法

共现分析术语称为共词分析,是一种较成熟的文献计量方法。共词分析是一种独具特色的内容分析,通过分析在同一文本主体中的单词或名词对共同出现的形式,确认文本所代表的学科领域中主题间的关系,进而探索科学的发展^[27]。共词分析方法具有灵活的优点,它可用图形来表示当前或一段时间内的研究状况和结构,是一种用来发现科学研究各领域间关系大有潜力的方法^[28]。在生成图形过程中,一是可根据研究需求灵活地设定共现频次或词频来确定高频主题词,以其为基础生成共词矩阵;二是可构建灵活的分析指标来测量共现词对间关系的强度^[29-33],最终以这些指标为基础将关键词聚类成组并以网络地图的方式展示。

共词分析主要分为共词网络分析、共词聚类分析和战略图分析三种^[2]。共词网络分析根据词与词间的共现关系,采用不同的指数算法计算共现强度并绘制共现网络图。共词网络分析法通常与社会网络分析法结合

使用以揭示某领域技术主题的分布、关联及演化^[34-35]。共词聚类分析利用词与词间的关联程度进行聚类, 将词聚成多个技术主题。战略坐标图是Law等在共词分析法基础上引入的, 首先通过共词聚类获得代表某领域技术主题的多个子簇, 然后计算各技术主题的中心度和密度来表示不同技术主题内部和相互间的关系^[29]。利用战略图可分析技术主题分布情况及各技术主题的发展趋势^[2-6]。

共词分析法成熟有效, 但是采用此方法依赖一项重要前提, 即主题词的确定。目前, 确定主题词的通用流程是通过自然语言处理和文本分析技术从专利文本中获取高频词, 并对之进行清洗和筛选。其中的问题是清洗和筛选的标准难以确定, 后期检验不易实施, 难以发现和利用相对低频却具有典型技术特征的主题词。

2.2.3 基于知识单元语义相似度的发现方法

随着文本挖掘技术的发展, 利用专利文献进行技术领域主题发现也逐渐从知识单元表面语义分析深入到知识单元间的语义关系识别层面以及知识单元隐含的语义内容识别层面。识别知识单元间的语义关系或隐含的语义内容, 其本质为计算知识单元间的语义相似度, 计算过程一般需要借助领域本体或语义词典, 最后在语义相似度的基础上对知识单元进行聚类获取技术领域主题。本文将此类研究方法统称为专利文本聚类法, 根据语义建模方法不同, 将其分为基于向量空间模型的专利文本聚类和基于SAO三元组的专利文本聚类。

向量空间模型 (Vector Space Model, VSM) 核心思想是将文档、术语、概念等信息对象作为向量空间元素进行建模, 将其表示成向量空间的向量。最简单的VSM是将专利文献直接表示成关键词权重向量, 通过计算特征向量间的距离表示专利文献间的相似度。Kim等基于关键词向量对专利进行聚类和可视化, 对泛在计算技术领域的新兴技术进行预测^[36]。VSM原理简单易懂、算法成熟, 被广泛应用于知识表示。但关键词间缺乏语义关系, 不能从微观层面揭示专利技术主题信息。针对该问题, 有学者基于领域本体概念间语义关系和语义增量对VSM进行改进^[37-38]。

鉴于关键词语义表示能力不足, 研究人员借鉴语义网中利用SAO三元组来表示知识元的思路, 利用SAO结构作为基础知识单元来表示专利。Park等将SAO结

构与TRIZ进化法则相结合, 分析技术演化进程, 识别重要技术主题^[39-40]; Yoon等研究利用SAO结构与离散点探测分析技术演化趋势, 识别新兴技术^[41-42]; Hu等以SAO作为基础语义单元, 通过SAO降维生成问题、方案、功能与效果四类技术主题, 按照技术相关性对专利文档进行聚类分析专利技术演化情况^[43]; 李欣等基于SAO结构语义分析识别新兴技术^[44]; 黄鲁成等使用CiteSpace确定突现文献, 从文献摘要中提取SAO结构, 计算其语义相似度得到文献的相似度矩阵, 然后对文献聚类识别新兴主题^[45]。该方法也可用于技术主题的识别。

基于SAO三元组的专利文本聚类方法可挖掘丰富的技术主题及主题间的相互关系, 但需要使用特定的工具进行SAO抽取与清洗, SAO语义相似度计算也需要借助专业领域的语义词典, 不同技术领域的语义关系不易界定, 因而使用难度较大。

综上所述, 基于文本挖掘的技术领域主题发现方法各有优势和不足, 研究者根据其研究目的需求和技术工具条件采用不同的方法在多个领域开展技术主题识别研究。这些研究一方面发挥作为典型应用学科的情报学对技术创新与发展的支持作用, 另一方面提高了不同领域专家对情报学方法的重视程度与研究兴趣, 情报学方法应用拓展与技术创新发展互相促进、互为补益。

3 基因工程疫苗领域主题发现研究

近年来, 我国明确大力发展生物医药行业^[46-47], 疫苗行业作为生物医药领域的重要子产业, 也是大力发展的对象。经调研发现, 国内尚未有学者对我国基因工程疫苗领域的技术主题进行识别与研究。本文研究的目的是利用情报学中技术领域主题发现研究方法对基因工程疫苗领域技术主题的构成及未来发展趋势进行识别与预测, 以期促进该领域发展。研究中所用到的研究方法、计量指标和技术路线如下。

3.1 研究方法

3.1.1 共词网络和共词聚类分析

本研究采用Gephi自带的Blondel社区探测算法对技术关键词共现网络进行聚类^[48-49]。Blondel社区探测是基于模块性指标的快速压缩社区探测算法, 其本质是将同

属性节点归并到合适社群。该算法分为两个阶段：第一阶段是社群的探测压缩，第二阶段是社群的合并重组。首先，将网络中的每个节点视作一个独立社区，逐步合并邻近节点，如果合并后整个网络的模块度提高，则进行合并，否则撤销；经过循环迭代，直到网络模块度无法提高为止。然后，将每个社区视为一个节点，对每个社区采用此合并算法，直到整个网络的模块度无法提高为止^[50]。该算法通过对模块性指标的判断，不断压缩探测网络的规模大小，以实现群结构的快速探测，设计简单有效，因而被广大研究者重视和使用^[51]。

3.1.2 战略坐标图分析

战略坐标图使用向心度 (Centrality) 和密度 (Density) 指标测量每个技术主题的特性。以向心度指标为X轴，表示技术主题间相互影响的强度；以密度指标为Y轴，表示技术主题内相互连接的程度，以所有技术类别的向心度和密度值作为源点绘制成的二维平面图即为战略坐标图^[27,29]。战略坐标图可清晰地展现各类技术主题在领域发展中的核心或边缘位置以及当前发展程度。

在第一象限，技术主题位于整个研究网络的中心，各主题的向心度、与其他聚类子网连接强度、内部连接密度、研究成熟度均较高。因此，该象限的技术主题是研究领域的核心，具有战略重要性，也表明有研究人员对其进行长期、系统、正规地研究。

在第二象限，技术主题同样位于整个研究网络的中心，向心度较高，但内部连接密度相对较低，有进一步发展的空间。该象限部分主题随着研究推进会日益成熟，成为未来的核心主题，是重点监测、培育的目标。

在第三象限，技术主题处于整个研究网络的外围，向心度较低，内部连接密度较高，研究问题成熟。这些研究问题在早期阶段处于研究网络的中心，现在也是重要投资对象，但逐渐被边缘化。

在第四象限，技术主题处于整个研究网络的外围，外部向心度和内部密度都较低，研究问题不成熟。

3.2 主要指标

(1) 余弦指数。余弦指数是共词网络中测度关键词间共现关系和强度的重要指标，其计算方法见公式(1)^[52]。

$$\cos \theta = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}} \quad (1)$$

$\cos \theta$ 的取值在0—1， $\cos \theta$ 值越趋近1表明两组向量的相似度越高。 X 和 Y 分别为两组空间向量，其中 x_i 和 y_i 分别为 X 和 Y 向量中第 i 个变量的值， n 为空间向量的维度。

本研究中， x_i 表示技术主题词 x 在给定专利文档中出现的次数， y_i 表示技术主题词 y 在给定专利文档中出现的次数， $x_i y_i$ 表示技术主题词 x 、 y 在同一篇专利文档中出现的次数。余弦指数能在一定程度上降低采用共现频次评价技术术语间链接强度的局限性，比较客观地揭示技术术语间的链接强度^[1]。

(2) 特征向量中心度。本文选取特征向量中心度衡量共词网络聚类子网中节点的重要性。节点特征向量中心度值的计算基于一个原则，即一个节点的重要性既取决于其邻居节点的数量（即该节点的度），又取决于其邻居节点的重要性^[53]。

(3) 技术主题外部向心度和内部密度。向心度指标测度一个技术主题与其他技术主题相互作用的强度，该指标强度越大，这个技术主题在整个技术领域越趋于核心地位。对于每个特定技术类别而言，向心度的计算可通过该类别所有关键词与其他类别的关键词间的连接强度进行。外部连接的总和、平均值、平方和的开平方等都可以作为该类别的向心度^[54]。密度指标测度技术主题内部关键词间的连接强度，密度越大，该技术主题内部联系越紧密。某一技术类别密度的计算可通过计算该类内部链接的平均值、中位数或者平方和得出^[54]。

本研究采用郝文宁等研究的方法^[37]，以每个类别与其他类别链接的平均值作为该类别的向心度，取每个类别内部链接的平均值作为该类别的密度，密度与向心度的计算方法见公式(2)与公式(3)。

$$Density = \frac{\sum_{i,j \in \mathcal{Q}_s} E_{ij}}{n-1}, \text{ for } i \neq j \quad (2)$$

$$Centrality = \frac{i \in \mathcal{Q}_s, \sum_{j \in (\mathcal{Q}-\mathcal{Q}_s)} E_{ij}}{N-n} \quad (3)$$

其中， E_{ij} 是共词网络中节点（技术主题词）的相关系数， n 为聚类子网中节点的个数， N 是共词网络

中节点总数, ϕ_s 为聚类子网, ϕ 为共词整体网络。

3.3 技术路线

Thomson Data Anzlyzer (TDA) 是一款强大的文本信息分析工具^[55], 其可从各种丰富的信息来源 (包括汤森路透的Derwent专利数据库和其他科技文献数据库) 获取大量数据进行分析。Derwent的技术领域专家在对专利权利要求和公开事项进行审查后重新撰写结构性摘要, 主要包括新颖性、优势、技术要点和应用领域, 各项含义如下所示^[56]。

Novelty (Abstract-DWPI Novelty/NOV): 描述发明的独特性;

Advantage (Abstract-DWPI Advantage/ADV): 描述发明人所陈述的发明优势;

Focus (Abstract-DWPI Tech Focus/FOC): 描述该发明包含的核心技术要点;

Use (Abstract-DWPI Use/USE): 包括该发明的主要应用领域。

本文主要分析基因工程疫苗的技术主题, 因此选择技术要点字段 (tech focus) 作进一步分析。首先, 使用TDA提取专利文献中的技术要点字段, 对其进行清洗并选择词频在前300的主题词作为分析对象; 其次, 基于主题词间共现关系, 计算任意两个主题词间的相关系数值, 绘制共词网络图; 再次, 综合考虑每个聚类子网中各主题词的含义, 识别目标领域的技术主题; 最后, 结合聚类结果和子网内主题词两两间的相关系数值进行指标计算, 生成战略坐标图, 对识别的技术主题进行可视化表达并分析各技术主题的发展现状和发展趋势。本文采用的技术路线如图1所示。

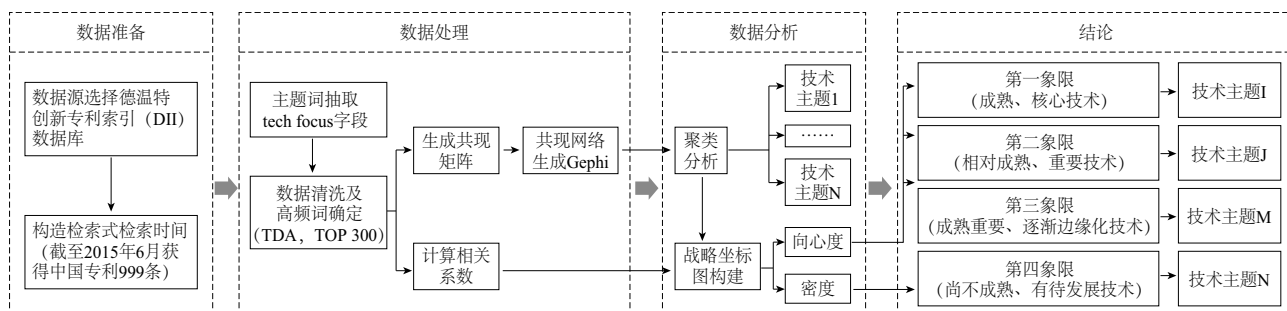


图1 技术路线图

4 基因工程疫苗领域技术主题分析

4.1 数据来源

本文数据来源于ISI Web of Knowledge知识平台中的德温特创新专利索引数据库^[56]。通过背景技术调研对基因工程疫苗领域进行技术主题要素分解, 构建详细的检索式, 检索时间截至2015年6月。通过检索共获取999项中国专利申请作为本研究的数据基础。用TDA提取tech focus字段, 得出按词频排序的关键词列表, 对关键词进行清洗; 借助TDA词表清洗工具和基于词形的相似度合并命令进行初步合并, 通过该步骤可除去停用词以及大量与分析主题无关的词, 此后在领域专家的指导下进行主题词语义合并, 最终得到本文用于共现分析的主题词; 将主题词出现累积频次达到52.2%的TOP 300高频词作为分析对象, 经过进一步分析, 去掉如“is”等无用词, 最终确定关键词数量为282个。

使用Gephi自带的Blondel社区探测算法对技术主题关键词共现网络进行聚类, 模块分割参数 (resolution) 使用默认值1, 聚类结果为5 (antisense RNA、adjuvant component和culturing V 3个孤立点除外), 5个聚类子网的节点数分别为140、28、40、49和12。经专家判读聚类结果发现, 该聚类结果分布不均, 最大子网中的技术主题与其他子网络有交叉和重复。结合专家意见, 经多次实验, 将模块分割参数 (resolution) 设定为0.95, 技术主题数量定为6个, 并请专家辅助解读每个子网代表的技术主题内容。

4.2 技术主题识别

结合关键词间共现情况和聚类结果, 综合考虑聚类子网中各关键词的含义, 对每个类进行命名, 这些类别代表基因工程疫苗领域的技术主题。技术主题、关键词数量和特征向量中心度值TOP 10的关键词见表1。

表 1 基因工程疫苗领域技术主题

技术主题	关键词数量/个	关键词 (TOP 10)
核酸疫苗相关技术	99	nucleic acid、SEQ ID、amino acid、antigen protein、sequence encoding、hepatitis C virus、viral vector、signal peptide、stem cell、T cells
基因工程载体疫苗 技术手段	57	expression vector、fusion protein、PCR、recombinant protein、Escherichia coli、PCR product、DNA sequence、recombinant plasmid、target gene/protein
树突状细胞疫苗 相关技术	56	host cell、tumor antigen、vaccine composition、mammalian cell、protective antigen、influenza virus、insect cell、human cell、plant cell、rabies virus
免疫佐剂相关技术	28	genetic engineering、organic chemistry、polyethylene glycol、therapeutic agent、cholera toxin、disulfide bond、fluorescent labels、solid support、detectable label、preferred agent
病毒类疫苗相关技术	11	antibody fragment、hepatitis B virus、coding reign virus-like particle、antigenic determinant、CpG motif、human papilloma virus、immunostimulatory nucleic acid、Sindbis virus
人源化基因工程 改造技术	28	immune response、chimeric/monoclonal antibody、genomic DNA、pharmaceutical composition、active fragment、amino acid residues、immunostimulatory substance、recombinant expression、detectable signal

主题1: 核酸疫苗相关技术, 关键词包括核酸、抗原结合片段、病毒载体、不完全辅剂、免疫佐剂、表达组件、DNA疫苗 (Deoxyribonucleic Acid Vaccine) 等。该主题中核酸疫苗佐剂、核酸疫苗表达载体及调控元件的构建、核酸疫苗真核表达及免疫效果评价是研究重点。

主题2: 基因工程载体疫苗技术手段, 关键词包括表达载体、融合蛋白、大肠杆菌、PCR (Polymerase Chain Reaction) 产物、重组质粒、引物设计、重组表达载体等。基因工程载体疫苗技术手段一般包括DNA重组技术、蛋白融合和重组表达技术。

主题3: 树突状细胞疫苗相关技术, 关键词包括宿主细胞、肿瘤抗原、保护性抗原、重组病毒、树状细胞以及多类细胞和多种病毒等。该主题包含的内容有树突状细胞疫苗在肿瘤免疫治疗领域的应用和肿瘤抗原及病原微生物保护性抗原的筛选、鉴定、克隆、表达及免疫研究。

主题4: 免疫佐剂相关技术, 关键词包括有机化学、治疗制剂、双硫键、荧光标记、阳离子脂质体等。安全且能有效激活体液与细胞双重免疫的佐剂是目前研发热点, 佐剂疫苗的体外评价 (荧光标记、放射性标记等) 是免疫佐剂相关技术的重点。

主题5: 病毒类疫苗相关技术, 包括关键词病毒样颗粒、抗原决定簇、CpG基元、人乳头状瘤病毒、免疫

刺激酸化、辛德毕斯病毒、热休克蛋白等。该主题的热点有佐剂改良型病毒疫苗研究、基于病毒样颗粒的亚单位疫苗研究、以辛德毕斯病毒作为疫苗载体的研究、无免疫佐剂的重组蛋白疫苗研究。

主题6: 人源化基因工程改造技术, 关键词包括嵌合抗体/单克隆抗体/单链抗体、染色体杂交、氨基酸残基、体液、杂交瘤细胞、抗原结合、体液免疫反映等。该主题包含内容有单克隆抗体的人源化基因工程改造技术及免疫应答反应机理研究。其中, 实现单抗的人源化、解决免疫原性、实现多种抗体重组等问题是该领域的热点和难点。

4.3 技术主题分析

本文利用TDA使用余弦指数计算主题词的相关系数, 结合共词网络聚类结果计算每个类别的向心度和密度值 (见表2), 以向心度为横坐标、密度为纵坐标绘制基因工程疫苗技术战略坐标图 (见图2)。

图2中第一象限暂无技术主题出现, 说明我国基因工程疫苗技术领域研究成熟的核心技术暂时还没有。位于第二象限的技术主题是核算疫苗相关技术和基因工程载体疫苗技术手段, 这两个技术主题具有较高向心度指标, 但是密度指标较低, 与其他技术主题领域的研究联系较多, 内部链接相对较低。位于第三象限的技

表 2 基因工程疫苗技术主题的向心度和密度

技术主题	向心度	密度	所属象限
核酸疫苗相关技术	2.90	0.75	第二象限
基因工程载体疫苗技术手段	2.53	0.61	第二象限
树突状细胞疫苗相关技术	2.35	1.28	第三象限
免疫佐剂相关技术	2.35	0.22	第四象限
病毒类疫苗相关技术	2.21	0.29	第四象限
人源化基因工程改造技术	2.10	2.51	第三象限

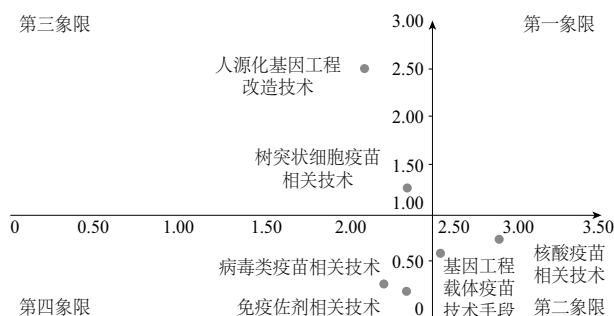


图 2 基因工程疫苗技术主题的战略坐标图

术主题是人源化基因工程改造技术和树突状细胞疫苗相关技术,二者具有较高的密度指标,但是向心度指标较低,内部链接相对较高,与其他技术主题领域的研究联系较少。第四象限的技术主题的向心度指标和密度指标都较低,相关研究尚不成熟,还有很大的发展空间。与国外疫苗研制相比较,我国免疫佐剂的研究相对滞后,能刺激细胞免疫的佐剂缺乏。安全且能有效激活体液与细胞双重免疫、具备免疫性和毒性的最佳平衡的宿主内源性免疫佐剂是我国免疫技术需要关注的方向。病毒类疫苗侧重于以病毒作为疫苗载体,用于治疗或预防危害人类生命与健康的疾病,我国在新型病毒疫苗研究方面有待突破。

5 结语

本文对基于专利文献进行技术领域主题发现的研究方法进行梳理,针对共词分析存在的问题,构造出一套高效获取大量专利技术主题词、生成战略坐标图的研究流程并选用共词分析方法对我国基因工程疫苗领域的技术主题构成和发展现状进行分析。研究发现,我国基因工程疫苗技术可划分为6个技术主题,总体而言,我国在该领域完全成熟的研究还没有,尚不成熟的

研究较突出,我国在该领域还有很多技术空间需进一步探索和突破。研发人员或决策制定者可根据各技术主题的发展情况,预测其发展趋势制定技术布局,使研发活动的开展更有效率。本文改进共词分析中主题词获取流程的规范性,也体现了作为典型应用学科的情报学对技术创新与发展的支持作用。

本研究还限于对关键词的静态分析,虽然可以从整体上反映技术主题的内容,但无法反映技术主题变化或技术进步。未来计划在本研究的基础上,进一步抽取技术新颖性字段(technology novelty)的SAO三元组数据结合语义分析技术,追踪技术的变化或改进,进一步完善技术领域主题发现技术。

参考文献

- [1] 胡阿沛,张静,雷孝平等.基于文本挖掘的专利技术主题分析研究综述[J].情报杂志,2013,32(12):88-92,61.
- [2] LEE B,JEONG Y I.Mapping Korea's national R&D domain of robot technology by using the co-word analysis[J].Scientometrics,2008,77(1):3-19.
- [3] 韩红旗,安小米,朱东华,等.专利技术术语共现的战略图分析方法[J].计算机应用研究,2011,28(2):576-579.
- [4] 郝韦霞,滕立,陈悦,等.基于共词分析的中国能源材料领域主题研究[J].情报杂志,2011,30(6):70-75.
- [5] 沈君,王续琨,陈悦,等.战略坐标视角下的专利技术主题分析——以第三代移动通信技术为例[J].情报杂志,2012,31(11):88-94.
- [6] 张杰,刘美佳,翟东升.基于专利共词分析的RFID领域技术主题研究[J].科技管理研究,2013,33(10):129-132,140.
- [7] 刘立,王博.基于专利情报分析的数控机床产业研究[J].科技管理研究,2010,30(15):149-152.
- [8] 黄晓莉,郑佳,王莹,等.基于专利情报分析的中国数控机床产业研究[J].情报杂志,2012(9):25-29.
- [9] 杨良斌,金碧辉.跨学科研究中学科交叉度的定量分析探讨[J].情报杂志,2009,28(4):39-43,92.
- [10] 高利丹,肖国华,张娴,等.共现分析在专利地图中的应用研究[J].现代情报,2009,29(7):36-39,43.
- [11] SPASSERM A.Mapping the terrain of pharmacy:co-classification analysis of the International Pharmaceutical Abstracts database[J].Scientometrics,1997,39(1):77-97.
- [12] KUMARESAN N,MIYAZAKI K.An integrated network approach to systems of innovation:the case of robotics in Japan[J].Research Policy,1999,28(6):563-585.

- [13] SCHUMMER J. Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology[J]. *Scientometrics*, 2004, 59(3): 425-465.
- [14] MEYER M. What do we know about innovation in nanotechnology? Some propositions about an emerging field between hype and path-dependency[J]. *Scientometrics*, 2007, 70(3): 779-810.
- [15] 林瑞明, 李万辉, 于海洋. 台湾LED技术专利数据挖掘分析[J]. *情报探索*, 2011(3): 59-63.
- [16] 张宪义. 基于专利共类分析的技术领域关联研究[D]. 大连: 大连理工大学, 2013.
- [17] 王贤文, 徐申萌, 彭恋, 等. 基于专利共类分析的技术网络结构研究: 1971~2010[J]. *情报学报*, 2013, 32(2): 198-205.
- [18] YOON J, PARK Y, KIM M, et al. Tracing evolving trends in printed electronics using patent information[J]. *Journal of Nanoparticle Research*, 2014, 16(7): 1-15.
- [19] 祁延莉. 核心专利识别方法研究[J]. *情报理论与实践*, 2016, 39(11): 5-9.
- [20] PORTER A L, CUNNINGHAM S W. Tech mining: exploiting new technologies for competitive advantage[J]. *Technology Analysis & Strategic Management*, 2005, 17(2): 247.
- [21] 许海云, 王振蒙, 胡正银, 等. 利用专利文本分析识别技术主题的关键技术研究综述[J]. *情报理论与实践*, 2016, 39(11): 131-137.
- [22] 梁立明, 谢彩霞. 词频分析法用于我国纳米科技研究动向分析[J]. *科学学研究*, 2003, 21(2): 138-142.
- [23] CHENY H, CHEN C Y, LEE S C. Technology forecasting of new clean energy: the example of hydrogen energy and fuel cell[J]. *African Journal of Business Management*, 2010, 4(7): 1372-1380.
- [24] 贡金涛, 杨帅, 魏晓峰. 基于专利词频和信息可视化的特定竞争对手分析——以通用风力发电技术为例[J]. *现代情报*, 2013, 33(3): 109-114.
- [25] 李阳, 许培扬. 基于专利文献的纳米生物材料领域技术机会识别初探[J]. *生物技术通讯*, 2013(2): 241-247.
- [26] 乔方园, 杨萌萌, 汪雪峰, 等. 纳米技术领域的关键词共现分析研究[J]. *情报杂志*, 2013, 32(5): 150-154, 175.
- [27] HE Q. Knowledge discovery through co-word analysis[J]. *Library Trends*, 1999, 48(1): 133-159.
- [28] 崔雷, 郑华川. 关于从MEDLINE数据库中进行知识抽取和挖掘的研究进展[J]. *情报学报*, 2003, 22(4): 425-433.
- [29] LAW J, BAUIN S, COURTIAL J P, et al. Policy and the mapping of scientific change: a co-word analysis of research into environmental acidification[J]. *Scientometrics*, 1988, 14(3): 251-264.
- [30] KOSTOFF R N, EBERHART H J, TOOTHMAN D R. Database tomography for technical intelligence: a roadmap of the near-earth space science and technology literature[J]. *Information Processing & Management*, 1997, 34(1): 69-85.
- [31] CALLON M, COURTIAL J P, LAVILLE F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry[J]. *Scientometrics*, 1991, 22(1): 155-205.
- [32] CALLON M, LAW J, RIP A. Mapping the dynamics of science and technology: sociology of science in the real world[M]. London: The Macmillan Press LTD, 1998: 103-141.
- [33] COULTER N, MONARCH I, KONDA S. Software engineering as seen through its research literature: a study in co-word analysis[J]. *Journal of the American Society for Information Science*, 1998, 49(13): 1206-1223.
- [34] 秦春娟. 基于专利共现的全球太阳能技术网络及关键技术演进分析[J]. *情报学报*, 2013, 32(1): 68-79.
- [35] 叶春蕾, 冷伏海. 基于社会网络分析的技术主题演化方法研究[J]. *情报理论与实践*, 2014, 37(1): 126-130, 140.
- [36] KIM Y G, SUH J H, SANG C P. Visualization of patent analysis for emerging technology[J]. *Expert Systems with Applications An International Journal*, 2008, 34(3): 1804-1812.
- [37] 郝文宁, 冯波, 陈刚, 等. 基于领域本体的文档向量空间模型构建[J]. *计算机应用研究*, 2013, 30(3): 764-767.
- [38] 胡吉明, 肖璐. 向量空间模型文本建模的语义增量化改进研究[J]. *现代图书情报技术*, 2014(10): 49-55.
- [39] PARK H, REE J J, KIM K. An SAO-based approach to patent evaluation using TRIZ evolution trends[C]//IEEE International Conference on Management of Innovation and Technology. Bali Island: [s.n.], 2012.
- [40] PARK H, REE J J, KIM K. Identification of promising patents for technology transfers using TRIZ evolution trends[J]. *Expert Systems with Applications*, 2013, 40(2): 736-743.
- [41] YOON J, KIM K. An automated method for identifying TRIZ evolution trends from patents[J]. *Expert Systems with Applications*, 2011, 38(12): 15540-15548.
- [42] YOON J, KIM K. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection[J]. *Scientometrics*, 2012, 90(2): 445-461.
- [43] HU Z Y, FANG S, WEI L, et al. An SAO-based approach to technology evolution analysis using patent information: Case study—graphene sensors[J]. *Chinese Journal of Library & Information Science*, 2015, 8(3): 62-75.
- [44] 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.
- [45] 黄鲁成, 张璐, 吴菲菲, 等. 基于突现文献和SAO相似度的新兴主题识别研究[J]. *科学学研究*, 2016, 34(6): 814-821.
- [46] “十二五”国家战略性新兴产业发展规划(全文)[EB/OL]. (2012-07-20) [2016-12-20]. http://www.china.com.cn/policy/txt/2012-07/20/content_25968625.htm.
- [47] 国务院关于印发《中国制造2025》的通知[EB/OL]. (2015-05-19) [2016-12-20].

http://www.gov.cn/zhengce/content/2015-05/19/content_9784.htm.

[48] Gephi.0.8.2 beta[EB/OL].[2016-12-24].<https://gephi.org>.

[49] BLONDEL V D, GUILLAUME J L, LAMBIOTTE J, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 30(2): 155-168.

[50] 朱梦娟, 程齐凯, 陆伟. 基于社会网络的学科主题聚类研究[J]. 情报杂志, 2012, 31(11): 40-45.

[51] FORTUNATO S. Community detection in graphs[J]. Physics Reports, 2009, 486(3/5): 75-174.

[52] 余弦相似度[EB/OL].[2016-07-14]. http://baike.baidu.com/link?url=9dnuj2ao1Ce2ywExtL0u611AbUawEjHVeoBKe_N1Z5yMwal_Mz3eAWEscIF3kC-xEiFIK2NvJI7jA7ozZx-h2f9ItTswnjB2zrm5bO833qbfsKJvBKCRqFeH1lysoS.

[53] 刘军. 社会网络分析导论[M]. 北京: 社会科学文献出版社, 2004.

[54] 张晗, 崔雷. 生物信息学的共词分析研究[J]. 情报学报, 2003, 22(5): 613-617.

[55] Thomson Data Analyzer[EB/OL].[2016-07-14]. <http://www.thomsonscientific.com.cn/productsservices/TDA/>.

[56] Thomson Reuters. Derwent Innovations Index[EB/OL].[2016-12-04]. <http://ip-science.thomsonreuters.com.cn/productsservices/derwentinnovationsindex/>.

作者简介

隗玲, 女, 1981年生, 讲师, 研究方向: 情报计量学理论与实践, E-mail: weiling@mail.las.ac.cn。
许海云, 女, 1982年生, 副研究员, 研究方向: 情报计量学理论与实践, E-mail: xuhy@clas.ac.cn。
刘春江, 男, 1984年生, 馆员, 研究方向: 信息可视化与知识挖掘。
李婧, 女, 1987年生, 助理研究员, 研究方向: 产业技术与专利情报分析。
方曙, 男, 1957年生, 研究员, 研究方向: 科技情报分析与研究。

Study on the Technological Domain Recognition: A Case Study of Gene Engineered Vaccine Technology

WEI Ling^{1,2}, XU HaiYun², LIU ChunJiang², LI Jing², FANG Shu²

(1.School of Information and Management, Shanxi University of Finance and Economics, Taiyuan 030006, China;
2.Chengdu Document and Information Centre, Chinese Academy of Sciences, Chengdu 610041, China)

Abstract: The paper systematically summarizes the development process of research methods on technological domain recognition based on patent documents, proposes a research flow to efficiently obtain patent technical keywords and generate the strategic diagram, and makes an experimental analysis on the technological domain distribution and development trend taking gene engineered vaccine technology for example. The research result proves that, under the co-word analysis flow proposed by the paper, patent co-words clustering and strategic diagram can be used effectively to recognize technological domains and their development currents.

Keywords: Technological Domain; Co-Word Analysis; Strategic Diagram; Gene Engineered Vaccine

(收稿日期: 2016-12-13)

《数字图书馆论坛》青年学者优秀论文资助计划
2016年第10—12期评选结果

奖项	期次	第一作者	标题	起讫页码
一等奖	2016 (10)	胡志刚	从书信沙龙到开放获取——刍议学术论文形态的演化	32—37
二等奖	2016 (12)	浦燕妮	通用型科学元数据标准研究	33—39
优秀论文奖	2016 (12)	林鑫	用户社会化标注中非理性行为的表现及原因分析	48—53