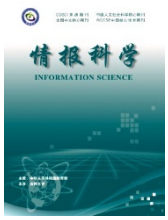

情报科学
*Information Science*
ISSN 1007-7634,CN 22-1264/G2

# 《情报科学》网络首发论文

题目：基于混合方法的“科学论文-专利技术”关联关系模型构建——以生物医药领域为例

作者：冉从敬，田文芳，贾志轩

网络首发日期：2024-05-07

引用格式：冉从敬，田文芳，贾志轩．基于混合方法的“科学论文-专利技术”关联关系模型构建——以生物医药领域为例[J/OL]．情报科学.
https://link.cnki.net/urlid/22.1264.g2.20240506.1927.030

网络首发时间：2024-05-07 15:45:51
网络首发地址：https://link.cnki.net/urlid/22.1264.g2.20240506.1927.030

# 基于混合方法的“科学论文-专利技术”关联关系模型构建*
## ——以生物医药领域为例

冉从敬　田文芳　贾志轩
（武汉大学 信息管理学院，湖北 武汉 430072）

**摘要：[目的/意义]**随着 AI 技术的飞速发展，新兴技术的更新迭代过程日益缩短，与新兴技术相关的论文和专利成为了科研工作者关注的焦点，迫切地需要挖掘科学论文主题与新兴专利技术间的关联关系与演化规律。**[方法/过程]**从多维视角分析出发，以共现网络、文本主题挖掘、文本相似度计算、时间窗口划分为技术支撑，将共现网络分析视角下关联关系分析和主题挖掘视角下的关联关系分析相结合，构建了基于混合方法的科学论文—专利技术关联关系模型。**[结果/结论]**以生物医药技术领域为例，对算法模型进行验证，实验结果表明三个时间区间下“论文-专利”主题之间的余弦相似度值都较高，主题领域形成了科学研究和专利技术的协同创新趋势。**[创新/局限]**本文提出了一套系统的“科学论文-专利技术”关联关系构建方法体系，实现了对科学技术互动规律更加细粒度、动态的剖析，进而挖掘出具有创新潜力的技术主题发展方向，为指导企业开展科技创新活动提供实践路径。
**关键词：**混合方法；关联关系模型；文本共现网络；文本主题挖掘；生物医药

# Modeling of Scientific Paper-Patent Technology Association Relationship Based on Mixed Methods*
## --Taking the Biomedical Field as an Example

RAN Congjing　TIAN Wenfang　JIA Zhixuan
(School of Information Management，Wuhan University，Wuhan 430072，China)

**Abstract:[Purpose/significance]** With the rapid development of AI technology, the updating and iterative process of emerging technology is getting shorter and shorter, and the papers and patents related to emerging technology have become the focus of scientific researchers' attention, and there is an urgent need to excavate the correlations and evolutionary patterns between the topics of scientific papers and emerging patented technology. **[Method/process]** From the perspective of multi-dimensional perspective analysis, using co-occurrence network analysis, theme model, text similarity calculation, and technology life cycle division as the technical support, combining the correlation analysis method under the perspective of social network analysis and the correlation analysis method under the perspective of text mining, we have constructed a scientific paper-patent technology correlation model based on a hybrid method. **[Result/conclusion]** Taking the field of biomedical technology as an example, the algorithm model is validated and the experimental results show that the cosine similarity values between "paper-patent" topics are high in three time intervals, and the topic areas have formed a trend of synergistic innovation between scientific research and patent technology. **[Innovation/limitation]** This paper proposes a set of systematic "scientific paper-patent technology" correlation relationship construction method system, which realizes a more fine-grained and dynamic analysis of the interaction law between science and technology, and then digs out the development direction of technical topics with innovation potential, and provides practical paths for guiding enterprises to carry out scientific and technological innovation activities.
**Keywords:** mixed methods; associative relationship modeling; textual co-occurrence networks; text topic mining; biomedical field

## 0 引言

* 本文为国家社会科学基金重大项目“大数据主权安全保障体系建设研究”的成果，项目编号：21&ZD169。

随着时代的发展，科技创新已成为事关国家命运的重大战略，深入剖析科学论文与专利技术的知识关联和互动机理[1]，对于准确把握科技发展规律及趋势、促进创新成果转化精准施策有着重要的理论和实践意义。在国家层面，党的二十大报告中指出：“强化国家战略科技力量，优化配置创新资源，优化国家科研机构、高水平研究型大学、科技领军企业定位和布局。”在企业层面，新兴科技产业的创新与发展需要同时做到科学研究与技术研发的创新，并且两者之间相互促进，从而加快创新成果的转化，进而促进我国的经济发展。此外，高校是科学研究和技术创新成果产出的重要研究机构，挖掘出技术创新和科学研究二者之间关联关系，对于加快高校科技成果的流通运用有着重要的意义。然而，在我国技术创新与科学研究之间尚未形成良好的互动态势，部分科学研究成果无法及时应用于技术实践，同时诸多技术问题往往因缺少新的科学成果支撑而得不到有效解决，高校、企业等机构科技成果转化面临着很多的困境，阻碍了科技进步和社会发展。对科学研究和专利技术进行关联关系挖掘成为了科学技术互动研究的重要方式[2-6]，在解决某类主题关联关系挖掘关键技术问题时，运用什么样的方法来分析相关主题的科学技术互动行为，挖掘出科学技术创新路径成为了一种迫切的需求。

目前，国内外关于科学论文-专利技术关联关系挖掘已有一定的研究基础，归纳起来主要包括以下三种类型：

一是基于引文网络的分析方法。引文网络分析方法主要是指文献共被引、直接引用和文献耦合引用这三种类型。如：高继平等[7]探析了专利—论文混合共被引网络下的信息交互行为，为研究高科技时代下科学技术相互融合中的知识流动提供了便利。Huang M H 等[8]研究燃料电池领域的论文与专利双向引用情况，发现该领域的科学与技术的关系越来越密切，并且论文更倾向于引用较新的专利。李蓓等[9]在新兴技术主题的核心特征的基础上构建了基于专利引用耦合聚类的新兴技术主题识别模型。

二是基于类目映射的分析方法。类目映射是指结合专利、论文自身的分类属性特征，建立论文和专利之间的类目映射关系。何贤敏等[10]将国际专利分类法、中国图书馆分类法和文本语义相似度计算相结合的方式完成了类目映射。赖院根等[11]通过 IPC 与 CLC 的类目映射、创新主题提取和基于叙词表的语义相似度计算来解决异构科技文献链接中存在的问题。

三是基于主题的分析方法。姜鑫等[12]通过识别不同时间阶段主题领域的核心词、突变词和新生词，并结合共词分析法揭示技术主题领域相关文献研究主题的动态演化趋势。Xie P 等[13]运用 LDA 主题模型对各单元的特定主题和整个文本的全体主题进行主题提取。徐红姣等[14]以 Word2vec 算法为基础，通过对论文和专利关键词聚类、主题相似度计算，探索构建能综合揭示论文和专利主题关系的关联演化图谱。

通过对比分析发现，已有研究主要是从引用网络、类目映射和主题挖掘这三个维度对专利和论文之间的知识流动进行了关联分析。然而，基于引文网络的分析方法存在分析维度太宏观，没有从细粒度的角度来揭示技术主题演化路径的问题；在基于类目映射进行分析时，存在科学和技术的分类系统不同，所以难以形成完全对应关系等问题；而基于主题的分析方法，已有研究主要是基于单一的主题提取模型来进行关联关系挖掘，存在同义词亦或是一词多义等现象，均会对文本挖掘结果造成不可忽视的误差。针对上述问题，本文基于专利和论文数据，将 BERT+Kmeans+LDA 融合模型、社会网络分析、科学知识图谱、关联关系分析这四类研究方法相结合，提出了一套系统的基于混合方法的科学论文-专利技术关联关系挖掘分析的研究框架体系，对今后的科学技术互动行为研究有一定的指导作用，能够帮助研究者及时了解技术创新和科学研究二者之间的技术主题演化规律，进一步发现专利技术与科学研究间的关联关系，对于加快科技成果转化有重要的意义。

## 1 研究方法与关键技术

### 1.1 科学论文-专利技术关联关系分析方法概述

（1）文本共现网络视角下的科学论文-专利技术关联关系分析方法

为了深入挖掘科学研究和专利技术之间的互动规律，本文引入共现网络分析方法，该方法的核心思想是基于科学论文和技术专利的特征属性来展开分析，将科学研究和专利技术这两类关联性不明显的社区联系起来并通过社会网络分析等手段来挖掘两个社区之间知识流动的规律。

（2）文本挖掘视角下的科学论文-专利技术关联关系分析方法

科学论文-专利技术主题关联分析方法是沿着“主题提取”→“关联关系构建”→“关联关系挖掘”这一核心思想来展开的。具体地，首先，利用自然语言处理、主题模型等分析方法对文本内容进行深入挖掘分析；其次，利用余弦相似度算法展开聚类分析；最后，从语义这一细粒度的层面上去探究主题之间的知识流动关系，从而揭示科学论文-专利技术之间的关联规律。

（3）基于混合方法的科学论文-专利技术关联关系分析方法

在使用文本共现网络视角下的科学论文-专利技术关联关系分析方法中存在由于分析角度过于宏观、异构数据源分析对象格式不一致等问题；文本主题挖掘视角下的科学论文-专利技术关联关系分析方法虽然能从细粒度语义的层面进行文本数据挖掘，也存在主题模型计算过程中有主观性的问题，影响关联关系分析结果的准确性。因此，考虑到单一关联关系分析方法存在的局限性[15-17]，在实际应用中，采用基于混合方法的关联关系挖掘，将文本共现网络视角下的科学论文-专利技术关联关系分析方法和文本主题挖掘视角下的科学论文-专利技术关联关系分析方法有机结合，进一步地，将基于社会网络分析的分析结果和基于文本挖掘的分析结果形成对比研究，以更好地揭示科学论文-专利技术关联演化的规律。

1.2 科学论文-专利技术关联关系挖掘关键技术

（1）社会网络分析

众所周知，学术论文是科学研究的成果，而专利文献是技术创新的成果，如何将两类关联性不明显的社区建立起关联关系是在使用社会网络分析方法时需要重点考虑的内容。因此，基于学术论文-专利文献之间关联关系构建这一研究目标，我们对论文和专利文献的内容特征进行分析，具体地，论文关键词是对论文题目、摘要和正文等重要内容的精准提炼，反映着论文的核心主题内容，而专利 IPC 分类号刻画的是专利所属技术领域[18]。基于上述分析，为了挖掘论文和专利之间的关联关系，本文从构建技术主题和论文关键词之间的共现关系这一研究思路出发，在利用自然语言处理技术进行格式转换和时间窗口划分的基础上，借助社会网络分析工具，从技术共现、论文关键词共现、专利技术-论文关键词共现等多维共现分析视角来动态挖掘学术论文和专利文献之间的关联关系。

（2）基于 BERT+K-means+LDA 融合模型提取主题

某一技术主题领域会涵盖多个子技术主题，在进行文本主题挖掘视角下的关联关系构建时，有必要对相应技术领域中的专利文本内容进行主题提取，明确技术领域下包含的子技术主题及各个子主题之间的关联关系，从细粒度语义层面揭示具象领域技术主题的演化规律。已有研究利用单一的 LDA 主题模型来挖掘文本数据的主题信息并建立文档、主题、主题词之间的关联关系，而论文与专利在表述方式等方面存在差异，再加上一词多义或者同义词现象的存在会对文本挖掘的结果产生影响，因此，本文采用 BERT 模型来解决一词多义的问题并构建文本语义特征向量。此外，在基于 BERT 模型获得词向量表示后，需要将文本数据划分到不同的子数据集，之后再进行技术主题提取，而 K-means 算法在实现数据聚类方面有较好的效果。综上，本文将 BERT、K-means 算法和 LDA 主题模型进行融合来实现技术主题的提取。

（3）基于余弦相似度算法计算主题之间的距离

本文需要构建各个时间段技术主题之间的关联关系和演化路径，其关键在于主题之间相似度的度量，因此需要引入相似度计算。常见的相似度计算方法有余弦相似度、欧氏距离、Jaccard 相似系数、皮尔逊相关系数等[19]，其中，余弦相似度算法经常应用于自然语言处理、文本挖掘等领域，适用于文本数据等稀疏数据，因此，选用余弦相似度算法来计算各个主题之间的距离。

2 科学论文-专利技术关联关系挖掘模型构建

文本共现网络视角下的关联关系分析方法是从宏观的层面对论文和专利之间的关联关系展开了分析，难以细粒度地揭示具象领域技术主题的演化轨迹。因此，在社会网络分析的基础上对文本内容进行深入挖掘，从语义层面展开主题关联分析，有利于更加全面和精准地识别出科学研究和专利技术之间的演化路径和动态交互行为。本文将文本共现网络视角下的关联分析和文本主题挖掘视角下的关联分析相结合形成混合方法，构建科学研究-专利技术关联关系挖掘模型，揭示了科学与技术的互动规律和技术发展方向，为指导企业开展科技创新活动提供实践路径，模型图如图 1 所示。

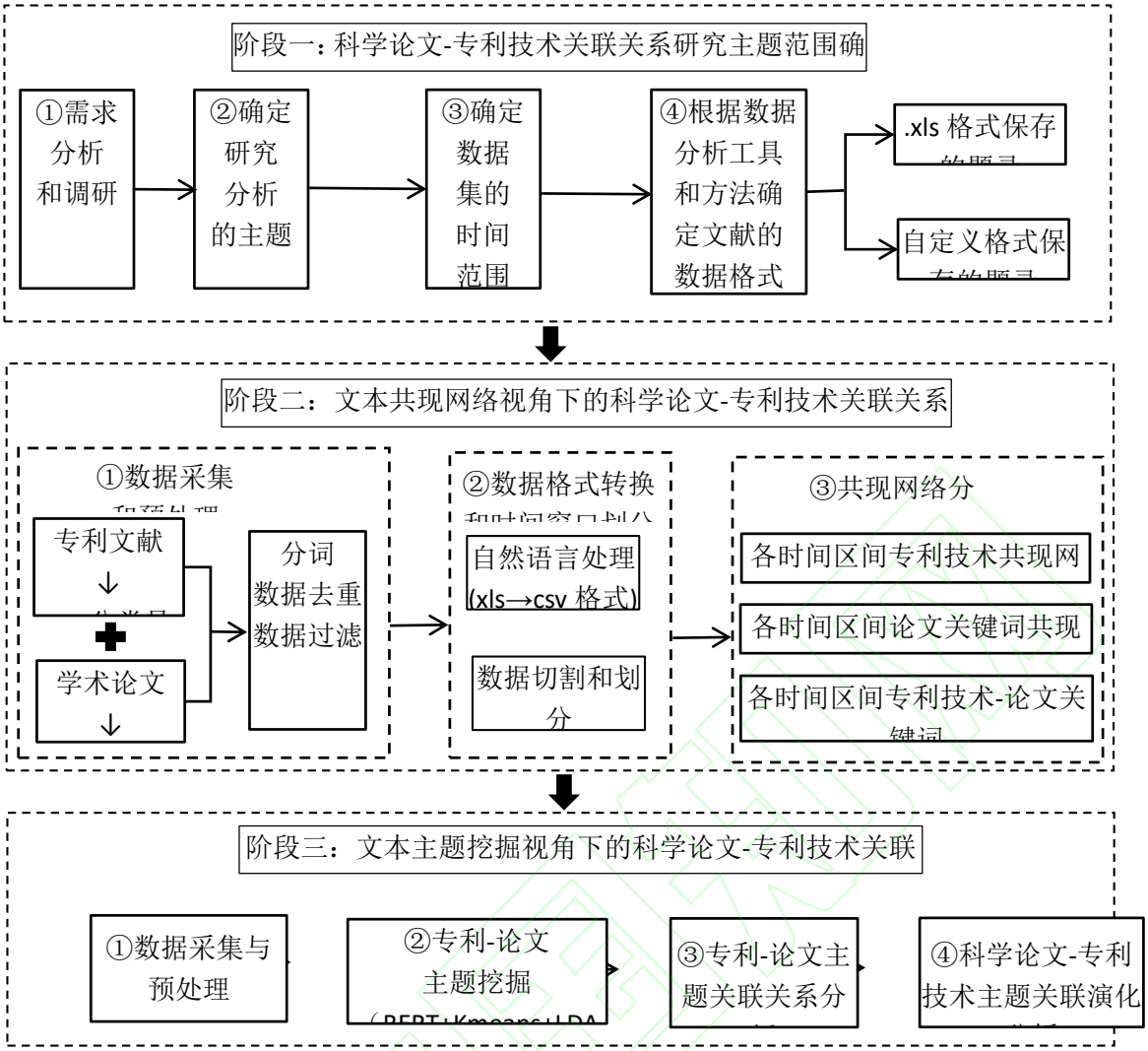


图 1 基于混合方法的科学论文-专利技术关联关系模型思路框架

Figure 1 A framework of ideas for modeling scientific paper-patent technology association relationships based on mixed methods

分析图 1 发现，具体的研究思路主要分为三个阶段，主要内容介绍如下：

第一阶段，根据文献调研和需求分析，确定研究的技术主题、研究的时间范围、数据挖掘工具和文献的数据格式。

第二阶段，进行文本共现网络视角下的科学研究-专利技术关联关系构建。其核心思想是在时间窗口划分的基础上，从动态演化分析的维度出发，通过挖掘 IPC 分类号和论文关键词之间的关系来链接挖掘专利文献与学术论文之间的关联关系。具体地，从 IPC 分类号共现网络、论文关键词共现网络 2 个角度进行多维共现分析，形成了文本共现网络视角下基于动态视角的科学研究-专利技术关联关系挖掘的研究思路框架。

第三阶段，进行文本主题挖掘视角下的科学研究-专利技术关联关系构建。从语义细粒度的维度出发，在分析文本数据的基础上，沿着“主题挖掘”→“关联关系构建”→“主题关联关系分析”这一主线思路，在时间窗口划分的基础上利用 BERT-Kmeans 融合模型和余弦相似度等关键技术进行科学研究-专利技术关联关系构建研究。

2.1 共现网络视角下科学论文-专利技术关联关系构建

以学术论文和专利文献为数据来源，在利用动态时间窗口划分和自然语言处理等技术来进行数据格式转化的基础上，以专利 IPC 分类号和论文关键词之间的关联关系挖掘为目标，从技术共现网络、关键词共

现网络两个维度来动态测度学术论文和专利文献之间的关联关系。

（1）专利技术共现网络分析

专利 IPC 分类号是专利文件的固有属性字段，其是一种对专利技术进行分类和检索的国际标准，因此，本文是将其所属技术类别界定为“技术领域”，进而从专利 IPC 分类号共现的角度来挖掘技术的共现情况。在实际分析过程中发现，一个专利可能有多个 IPC 分类号，表示其对应着多个技术领域，我们对专利的 IPC 分类号进行共现处理，可以反映特定主题下的技术共现情况，进一步地，通过共现网络分析发现技术研究热点领域和技术研究空白领域。

（2）论文关键词共现网络分析

关键词是对论文核心内容的高度凝练和归纳总结，通过对关键词进行分析的方式来挖掘论文研究主题方向。具体地，通过关键词共现分析可以把握主题研究领域研究热点。关键词共现网络是指关键词在论文之间共现的频次所形成的网络[20-21]，共现网络中节点的大小和关键词出现的频次成正相关关系，即节点越大，关键词频次越高；而关键词之间的连线表示其共同出现在同一篇论文当中，连线的粗细和关键词共现强度也成正相关关系。中心度则表示该关键词出现在多篇论文当中，是研究人员关注的重点内容，中心度越大说明该关键词越重要。因此，本文是从关键词共现频次以及中心度两个维度来测度论文的研究热点。

2.2 主题挖掘视角下科学论文-专利技术关联关系构建

在文本共现网络分析的基础上，以时间窗口周期划分和 BERT+Kmeans+LDA 融合模型为技术支撑对文本内容进行深入挖掘，从细粒度语义层面展开技术主题分析，弥补了社会网络分析过于宏观的缺陷。此外，本文将时间窗口划分思想加入研究中，从动态视角来揭示主题演化规律，提高了科学论文-技术专利关联关系挖掘的全面性和准确性。图 2 为基于 BERT+Kmeans+LDA 融合模型的技术主题提取流程。整个数据处理过程是围绕着基于 BERT+Kmeans+LDA 的技术主题提取和基于余弦相似度计算的科学论文-专利技术主题关联关系构建这两个模块的内容来进行的。

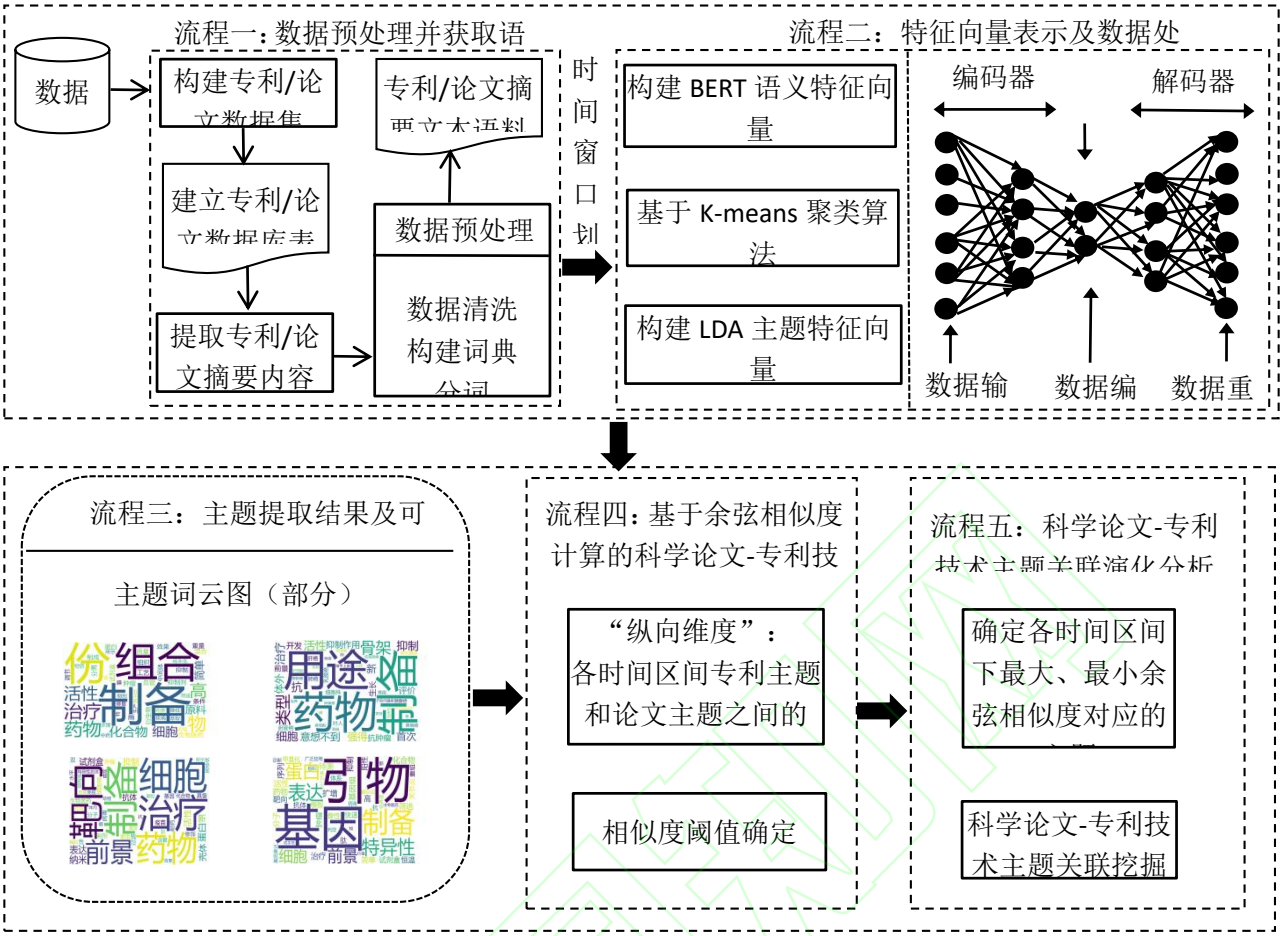


图 2 基于 BERT+Kmeans+LDA 融合模型的技术主题提取流程

Figure 2 Technical topic extraction process based on BERT+Kmeans+LDA fusion model

（1）技术主题提取 BERT+Kmeans+LDA 融合模型

分别对论文和专利数据进行特征提取，基于 BERT+K-means+LDA 主题建模算法，定量地提取技术主题，对应着图 2 中的流程一、流程二和流程三这三个核心模块的内容。LDA 模型采用高效的概率推断算法处理大规模数据，具有较好的文本潜在主题挖掘能力，但是，其在表征上下文语义信息、理解语句结构和单词歧义等方面存在缺陷[22]，对复杂问题的学习效果较差，这为模型的改进提供了机会和空间。已有的研究成果[23-24]证实 BERT 模型在表征上下文语义信息方面表现较好，可解决缺失语义表达能力等问题，将经过预训练的 BERT 模型和 LDA 模型进行融合可以提高主题提取的质量。与此同时，在利用 BERT 进行本文转向量处理之后以及在采用 LDA 模型进行主题提取之前，需要采用聚类算法对技术主题领域数据集进行子领域的划分。下面以专利文本数据为例说明采用 BERT+K-means+LDA 进行数据处理的过程，重点对技术主题提取流程中的关键数据处理步骤进行梳理和解释说明。此外，论文数据的处理步骤也是按照下面的处理流程来进行的。

流程一，数据预处理并获取语料库。

以专利和论文融合数据为研究对象，提取摘要内容并对其进行数据清洗与预处理。具体地，结合文献特点，自定义停用表并使用当下常用中文分词工具对专利摘要文本进行初筛，具体操作包括去除“本发明”“一种”“差一点”等无意义常用词语；去除非中文词汇（包括英语、日语、俄语等）；去除数字及标点；去除非专业词汇，最后根据现有的专业学科词汇对分词结果进行筛选并获取专利/论文摘要文本语料库。

流程二，特征向量表示及数据处理过程。

该部分的数据处理过程包括构建 BERT 语义特征向量、基于 K-means 聚类算法进行专利文本划分和基于

LDA 模型进行技术主题提取三个关键步骤。

首先，构建 BERT 语义特征向量。对筛选后的文本数据使用预训练双向表示模型进行向量化处理，本文是使用哈工大讯飞联合实验室发布的基于全词掩码技术的中文预训练模型 BERT-wwm 作为预训练模型。具体地，在进行文本向量化处理的过程中，将第 i 个摘要文本定义为$T_i$，定义向量化后文本向量$d_i$为：

$$d_i = B(T_i) \quad (1)$$

其次，基于 K-means 聚类算法进行专利文本划分。本文使用 K-means 聚类算法对向量化处理后的专利摘要数据进行聚类分析，将专利文本数据划分到子技术主题当中，聚类公式为：

$$\operatorname{argmin}_d \sum_{i=1}^{k} \frac{1}{|d_i|} \sum_{x,y \in d_i} ||x - y||^2 \quad (2)$$

其中，k 为聚类总数，$d = \{d_1, d_2 \ldots d_k\}$为聚类中数据样本。

最后，基于 LDA 模型进行技术主题提取。将专利文本数据划分到子技术主题后，对专利摘要文本信息进行词频分析，进而通过 LDA 筛选主题技术文本特征以及主题提取，公式组合如下：

$$T_i = \{word_1, word_2 \ldots word_m\} \quad (3)$$

$$d_i = B(T_i) \quad (4)$$

流程三，主题提取结果及可视化呈现。

在利用 BERT+K-means+LDA 融合模型挖掘出主题领域的技术主题后，选择每个技术主题下概率 TOP(5-10)对的主题词进行知识图谱分析，进而确定关键技术主题内容，绘制的部分关键技术主题词云图如图 2 中的流程三所示。

（2）科学论文-专利技术主题关联关系构建

本文是以余弦相似度计算算法为技术支撑，进行同一时间区间论文和专利二者之间的主题关联关系构建和定量测度分析，对应着图 2 中的流程四和流程五两个模块的内容。

第一，基于比对分析视角对同一时间区间不同文献主题关联关系进行构建。在基于融合模型从横向维度挖掘出不同时间区间下的主题内容后，进一步地，从纵向维度，对同一时间区间下论文的研究主题和专利技术主题之间的语义相似度进行定量测度和分析。

第二，基于余弦相似度算法的主题关联关系测度及演化分析。在确定好不同文献主题关联关系构建思路后，采用相似度阈值法来测度主题之间的关联关系，其中余弦相似度是利用两个技术主题向量夹角的余弦值大小来衡量相似程度，两个主题之间的余弦相似度值越大，表明两个主题之间语义相关性越强，余弦相似度的测度公式如（5）所示。

$$CS(T_t, T_{t+1}) = \frac{T_t * T_{t+1}}{\| T_t \| * \| T_{t+1} \|} = \frac{\sum_{i=1}^{n} (p(w_i \mid T_t)\, p(w_i \mid T_{t+1}))}{\sqrt{\sum_{i=1}^{n} p(w_i \mid T_t)^2 \sum_{i=1}^{n} p(w_i \mid T_{t+1})^2}} \quad (5)$$

公式（5）中，$T_t$ 指时间段 t 内的主题的概率值，$T_{t+1}$ 指时间段 t+1 内的主题的概率值。

## 3 实验过程及分析

### 3.1 数据来源及预处理

首先，本文的专利数据来源于 IncoPat 专利数据库，检索范围为“中国发明授权专利”，检索到权利有效的专利 10 581 件。检索时间为 2023 年 11 月 5 日。爬取详细的专利数据信息，包括专利题名、申请号、申请人、发明人、公开号、公开日及 IPC 分类号等字段信息。

其次，为了建立专利文献与学术论文的关联关系，在采集完专利数据的基础上还需要构建论文数据集，本文的论文数据来源于中国知网数据库。以“生物医药”为主题词，检索了中国知网上的期刊数据库，其包括 SCI 来源期刊、EI 来源期刊、核心期刊和 CSSCI 上的数据，共检索到 5806 篇文章，剔除掉检索结果中的新闻、会议通知等非相关文献，得到 5607 篇文章，构成本文案例研究的数据集。

最后，利用 Python 程序设计和 Excel 函数调用工具对数据进行清洗和预处理及格式转换，将输出的结果作为后续社会网络分析软件的数据输入来源。此外，本文是按照各时间段文献数量保持等量的原则将专

利数据集划分为 2010-2016 年、2017-2019 年、2020-2022 年这三个时间窗口。

特别地，针对论文与专利的发表周期存在差异这个问题。首先，本文探讨的是专利-论文之间的主题关联关系及演化规律，而主题的演变是需要经过较长时间来发展和成熟，因此，将数据集划分为 3-6 年的区间作为研究对象，在一定程度上削弱了论文与专利的发表周期的问题；其次，在数据集构建的时候，采集的数据量比较大并且是以时间区间内的数据集为研究对象进行主题挖掘，相比之下，这个较短的时间差异对分析结果的影响有限；此外，在生物医药领域，论文的发表周期通常较长。这是因为生物医药研究往往涉及复杂的实验设计、数据分析和结果验证等步骤，而这些步骤需要经过严格的检查和评审过程，通常需要数个月甚至一年以上的时间。因此，为了方便进行对比分析，将论文数据集的时间窗口划分临界点和专利数据集保持一致。

3.2 共现网络视角下科学论文-专利技术关联关系分析结果

在主题研究领域科学论文-专利技术关联关系挖掘模型构建的基础上，以生物医药技术领域为例，分析该主题下的科学论文-专利技术互动行为。

（1）技术共现网络分析结果

以专利 IPC 分类号作为研究对象，基于 IPC 分类号之间的多维共现关系来探索主题研究领域各时间区间的热点技术领域，各时间区间下专利技术共现网络如图 3、图 4、图 5 所示。在对 2010-2016 年、2017-2019 年、2020-2022 年这三个时间区间的技术共现网络进行分析的基础上，挖掘出各个时间区间的技术共现主题，如表 1 所示。

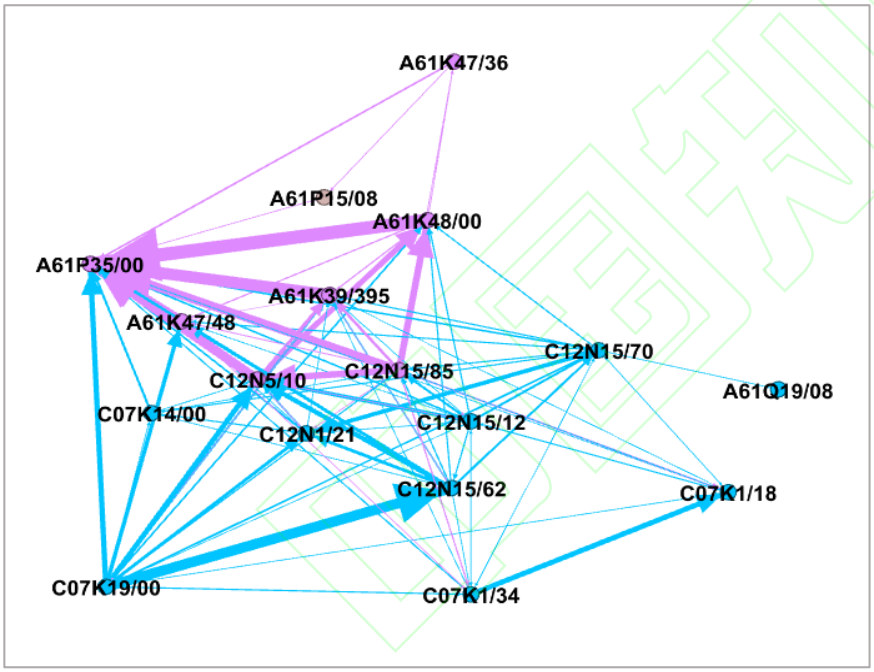


图 3 专利技术共现网络（2010-2016）

Figure 3 Patented technology co-occurrence network (2010-2016)

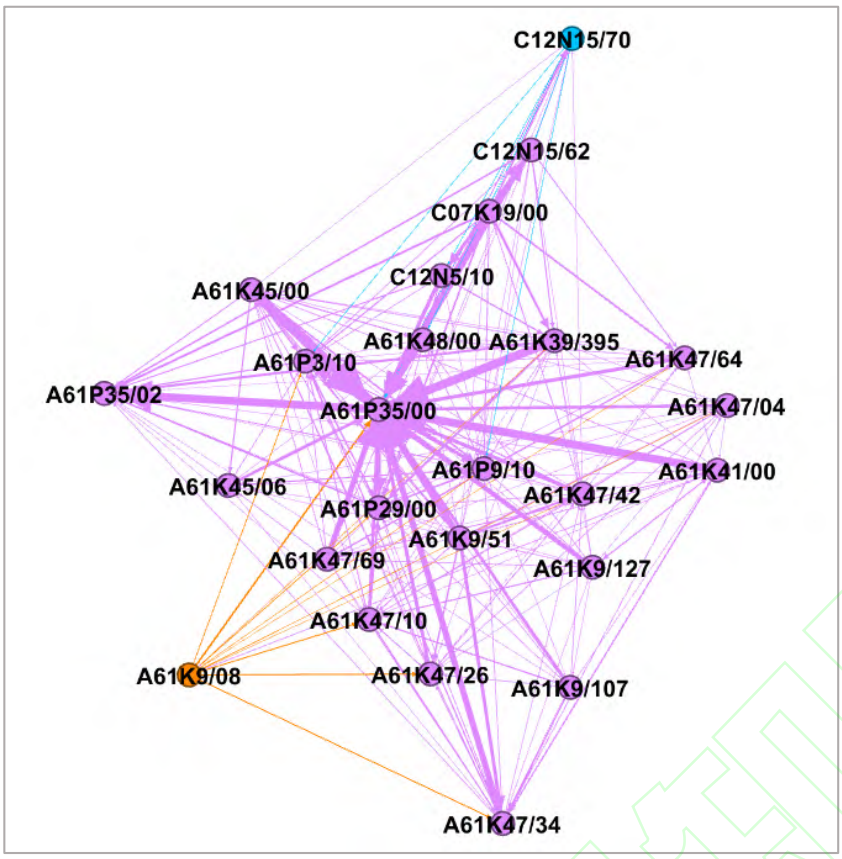


图 4 专利技术共现网络（2017-2019）

Figure 4 Patented technology co-occurrence network (2017–2019)

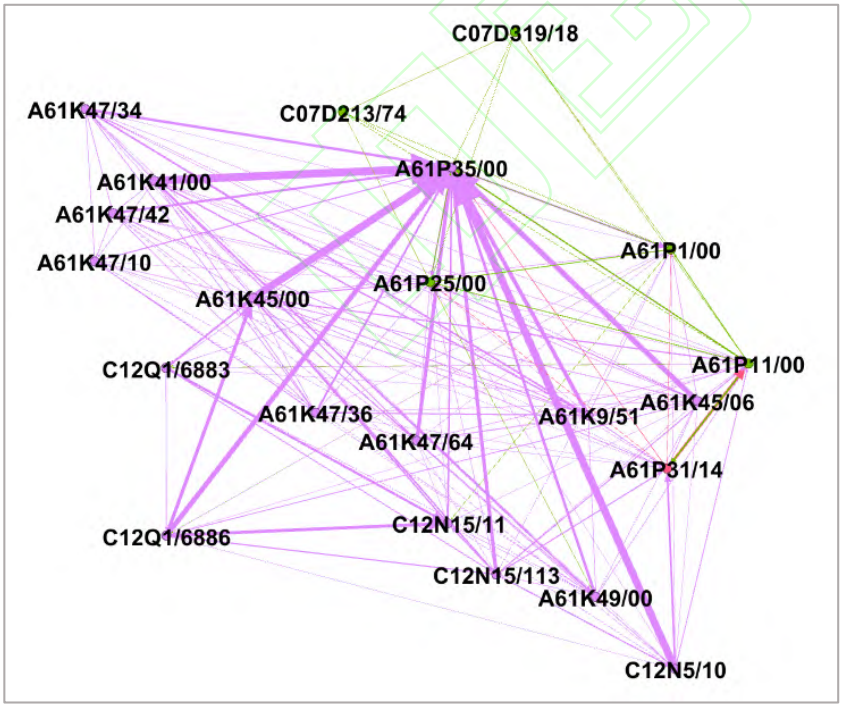


图 5 专利技术共现网络（2020-2022）

Figure 5 Patented technology co-occurrence network (2020–2022)

表 1 不同时间区间下专利技术共现主题分析

Table 1 Analysis of co-occurrence topics of patent technologies under different time intervals

| 时间区间 | 主要专利 IPC | 技术共现主题 |
|---|---|---|
| 2010－2016 年 | A61P35/00、A61K48/00、A61K39/395 、C12N15/85 | 抗肿瘤药物研发、基因治疗、单克隆抗体技术、基因编辑 |
| 2017－2019 年 | A61P35/00、A61P35/02 A61K45/00 、A61K48/00 A61K39/395、 C12N5/10 | 抗肿瘤药物、特异性抗体、小分子抑制技术、基因治疗、经引入外来遗传物质而修饰的细胞 |
| 2020－2022 年 | A61P35/00、A61K41/00 A61K45/00、A61P25/00 A61K47/64、C12N5/10 C12Q1/6886 | 抗肿瘤、纳米粒子、抑制技术、治疗神经系统疾病的药物、药物递送载体、经引入外来遗传物质而修饰的细胞、癌症治疗相关技术 |

2010－2016 年间，专利 IPC 共现网络主要是通过节点 A61P35/00、A61K48/00、A61K39/395、C12N15/85 建立关联关系，形成一个子群。该时间区间主要是聚焦在抗肿瘤药物研发、基因治疗、单克隆抗体技术等技术领域进行技术攻关。该时期的侧重点为创新药物发现、免疫学技术等在癌症治疗领域的实践应用。2017－2019 年间，专利 IPC 共现网络关系变得更加复杂，从图 5 可以很直观地观察出子群的个数以及子群内部的节点数相对 2010－2016 年间大幅度增加。专利 IPC 共现网络主要是通过节点 A61P35/00、A61P35/02、A61K45/00、A61K48/00、A61K39/395、C12N5/10 建立关联关系，该时间区间主要是聚焦在特异性抗体、小分子抑制技术、基因治疗、细胞修饰等技术子领域。该时期的侧重点为免疫检查点抑制剂、治疗性基因编辑技术、重组抗体技术等更广泛地应用于癌症治疗及其他治疗领域。

2020－2022 年间，专利 IPC 共现网络主要是通过节点 A61P35/00、A61K41/00、A61K45/00、A61P25/00、A61K47/64、C12N5/10、C12Q1/6886 建立关联关系。该时间区间主要聚焦在抗肿瘤、纳米粒子、抑制技术、治疗神经系统疾病的药物、药物递送载体、癌症细胞治疗相关技术领域。该时期的侧重点为将细胞治疗、新型药物递送、脑机接口等核心技术在癌症治疗方面进行技术攻关。

（2）论文关键词共现网络分析结果

各时间区间的论文关键词共现网络如图 6、图 7、图 8 所示，主要关键词的分布情况如下：

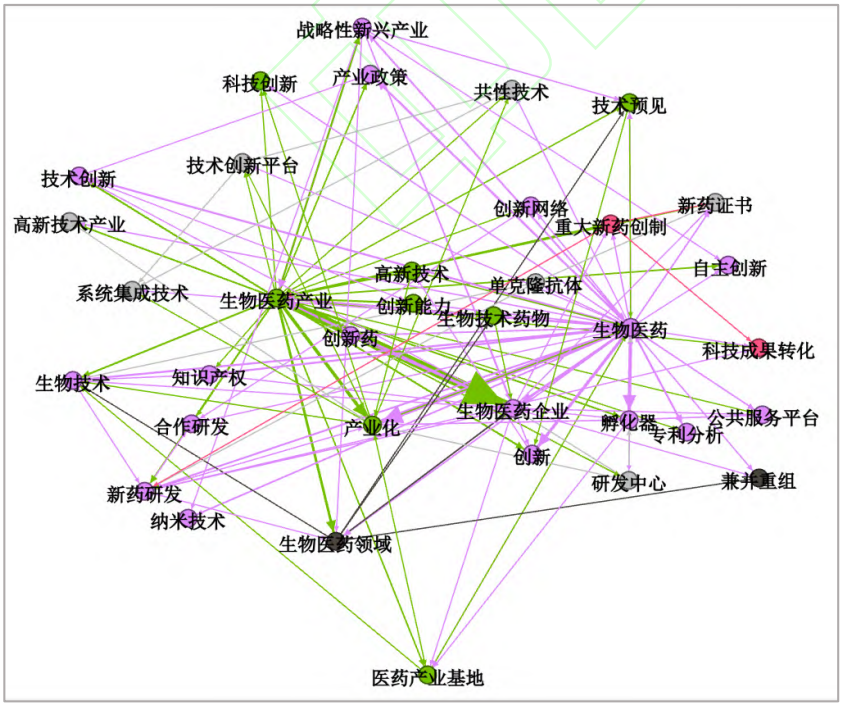

图 6　论文关键词共现网络（2010-2016）

Figure 6 Keyword co-occurrence network of papers (2010-2016)

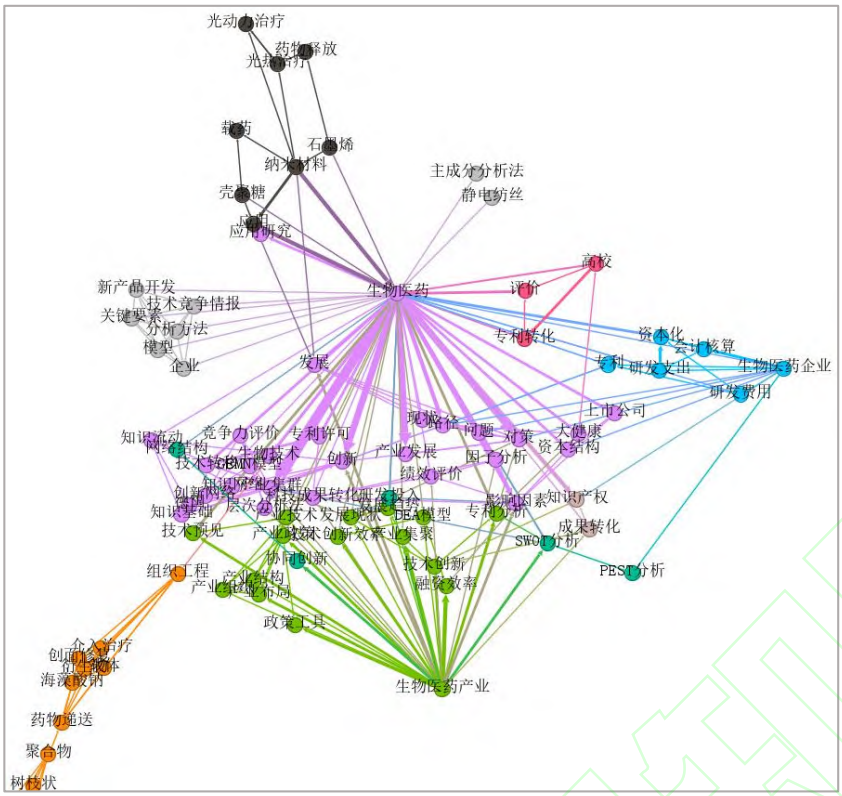


图 7 论文关键词共现网络（2017-2019）

Figure 7 Keyword co-occurrence network of papers (2017−2019)

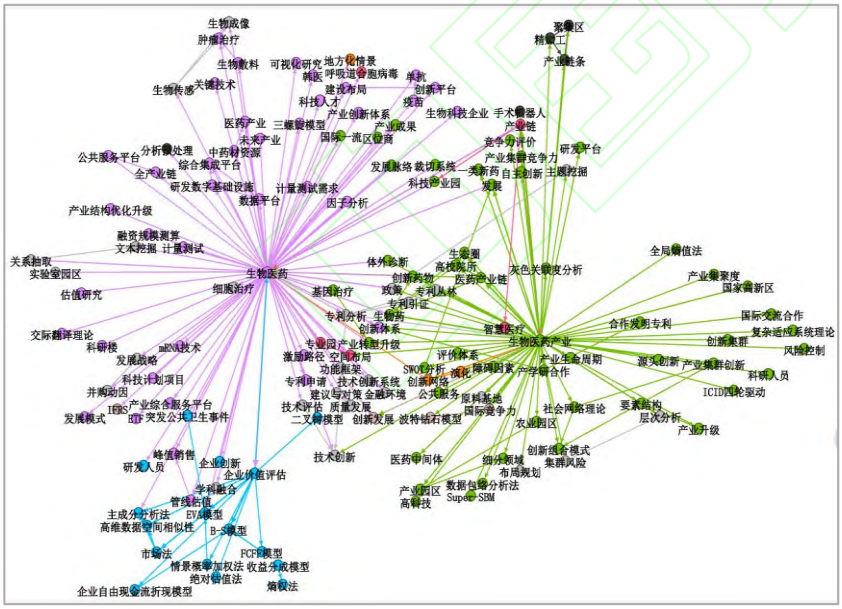


图 8　论文关键词共现网络（2020-2022）

Figure 8 Keyword co-occurrence network of papers (2020−2022)

表 2 不同时间区间下论文关键词共现分析

Table 2 Analysis of keyword co-occurrence in papers under different time intervals

| 时间区间 | 核心关键词 | 关键词共现分析 |
| --- | --- | --- |
| 2010－2016 年 | “重大新药创制”、“生物 | 主要聚焦于新药研发、高新技术和创新能力 |

| | | |
|---|---|---|
| | 技术药物”、“生物技术”、“系统集成技术”“高新技术”“创新能力” | 提升等方面。该阶段人工智能等信息化系统集成技术在生物医药的子领域开始应用，例如，对于新药研发，通过人工智能对大数据的智能分析，辅助研发人员在研究和产业化中的决策制定以及了解新药对不同患者的治疗效应 |
| 2017－2019 年 | “研发创新”、“产业创新平台”、“技术创新”、“创新驱动”、“预测模型”、“分析方法”、“基因”、“细胞” | 主要聚焦于技术创新、模型构建、分析方法等主题。该阶段以“预测模型+生物医药”在基因治疗、细胞治疗等技术领域的深度应用为主线，以智能模型为核心，抽取和挖掘基因治疗等相关关键信息 |
| 2020－2022 年 | “研发平台”“产业创新体系”“医药产业结构优化”“二叉树模型”、“B-S 模型”、“肿瘤治疗” | 主要聚焦于将 B-S 模型、二叉树模型等先进技术嵌入研发平台，全周期助力生命科学领域企业研发、临床、生产的创新需求，可见，该阶段主要目标是致力于生物制药创新能力与效率的提升 |

（3）专利技术-论文关键词关联关系分析

在时间窗口划分的基础上，上文分别对专利技术共现网络和论文关键词共现网络展开了分析，基于此，从主题内容细粒度层面展开“科学论文-专利技术”关联关系挖掘和探讨。

**2010－2016 年间：**专利技术的侧重点为创新药物发现、免疫学技术等在癌症治疗领域的实践应用；与此同时，论文核心关键词主要为“重大新药创制”“生物技术药物”“生物技术”“系统集成技术”“高新技术”。可见，新药研发为该时期科学研究者和技术攻关人员关注的共同主题。

**2017－2019 年间：**专利技术主要是聚焦在免疫检查点抑制剂、治疗性基因编辑技术、重组抗体技术等更广泛地应用于癌症治疗及其他治疗领域。而该时期论文核心关键词主要涉及“技术创新”“基因”“细胞”等相关。经过对比分析可以得出的结论是生命科学的研究重心为基因编辑等核心课题。

**2020－2022 年间：**该时间区间专利技术主要是聚焦在药物递送载体、纳米粒子、治疗神经系统疾病的药物、癌症细胞治疗相关技术领域，重点是将新型药物递送、抗体/蛋白药物偶联物、脑机接口、细胞治疗等核心技术在癌症治疗方面进行技术攻关；核心论文关键词主要是“抗体”“蛋白”“肿瘤治疗”“二叉树模型”“研发平台”等方面相关。这一时期，蛋白质组学等相关为该时期科学研究者和技术攻关人员关注的共同主题。

3.3 主题挖掘视角下科学论文-专利技术关联关系挖掘的分析结果

（1）论文-专利主题提取

各时间区间论文数据主题提取。结合困惑度和肘形的计算结果，三个时间段的最优主题数量分别为 6、6、4，主题提取后，每个 topic 下提取 5-10 个关键词，部分结构如表 3 所示。

表 3 各时间区间下的论文主题识别结果

Table 3 Results of paper topic identification under each time interval

| 三个阶段 | 论文主题识别结果 |
|---|---|
| 2010-2016 年 | Topic#1（生物医药 、医药产业、高新技术、孵化器、产业园）、Topic#2（生物医药、医药产业、研发、新药、高新技术）、Topic#3（生物医药、研发、技术创新、新兴产业、新药）、Topic#4（生物 |

| | |
|---|---|
| | 医药、医药产业、技术产业、知识产权、创新能力）、Topic#5（生物医药、医药产业、研发、新兴产业、传感器）、Topic#6（生物医药、医药产业、产业基地、研发、新药） |
| 2017-2019 年 | Topic#1（生物医药、医药产业、医药企业、研发创新、专利、技术创新、产业政策）Topic#2（生物医药、研发创新、专利、政策、科技成果、新兴产业、细胞）Topic#3（生物医药、医药产业、研发创新、创新能力、纳米、知识产权、细胞）Topic#4（生物医药、医药产业、预测模型、研发、化合物、细胞、富勒烯）、Topic#5（生物医药、医药产业、研发、技术创新、新药、药物、糖胺聚糖）、Topic#6（生物医药、医药产业、创新、政策、新兴产业、纳米、孵化器） |
| 2020-2022 年 | Topic#1（生物医药、生物、协同、研发、创新能力、技术创新、新药、创新平台）、Topic#2（生物医药、医药产业、创新、研发、专利、医药、纳米、新兴产业）、Topic#3（生物医药、医药产业、创新、研发、疫情、抗体、平台、产业链）、Topic#4（生物医药、医药产业、智能模型、创新、研发、新药、DNA 序列） |

（2）各时间区间专利数据主题提取

结合困惑度和肘形的计算结果，三个时间段的最优主题数量为 8、8、5，主题提取后，每个 topic 下提取 10-20 个关键词，部分结构如表 4 所示。

表 4 各时间区间下的技术主题识别结果

Table 4 Results of technical topic identification under each time interval

| 三个阶段 | 技术主题识别结果 |
|---|---|
| 2010-2016 年 | Topic#1（化合物、基因、药物、羟基、抗体、甲基、衍生物、色酮、甲氧基、发酵）；Topic#2（药物、治疗、分子、生物医药、多肽、蛋白、细胞、纳米、抗肿瘤、抑制剂）；Topic#3（基因、多肽、表达、病毒、氨基酸序列、重组、蛋白质、细胞、特异性、单克隆抗体、核苷酸序列、肿瘤）；Topic#4（药物、蛋白、治疗、基因、肿瘤、纳米、抗体、病毒、特异性、壳聚糖、树突状细胞、抗菌）；Topic#5（毒素、蛋白、疾病、氨基酸序列、单胺氧化酶、半胱氨酸、多核苷酸、基因、衍生物、钙离子通道）；Topic#6（药物、抑制作用、体外、抗肿瘤、细胞株、癌细胞、淋巴细胞、胰腺癌、胃癌、膀胱癌）；Topic#7（基因、细胞、治疗、化合物、基因组、重组、抗体、病毒载体、突变型、肿瘤细胞）；Topic#8（药物、抑制、菌株、活性、肝纤维化、病毒、细菌、胰岛素、成纤维细胞、致病菌） |
| 2017-2019 年 | Topic#1（药物、治疗、细胞、纳米、生物医药、干细胞、表达、抑制、活性、质量、蛋白质、提取物、组合、疾病、肽、生物、肿瘤、磷脂、多肽、载体）；Topic#2（化合物、药物、治疗、抗体、组合、生物医药、预防、抗原、蛋白、药学、衍生物、特异性、抗肿瘤、抑制剂、杂环化合物、酸盐、序列、诊断、分子、癌症）；Topic#3（药物、生物医药、基因、单克隆抗体、细胞、活性、疾病、抑制、肿瘤、片段、标志物、靶向、病毒、肽、诊断、培养、中间体、合成、筛选、抗原）；Topic#4（治疗、药物、抑制、基因、抑制剂、细胞、预防、生物医药、分子、组合、化合物、抗体、蛋白、肿瘤细胞、提取物、核苷酸序列、多肽、干扰、体外、抗肿瘤药物）；Topic#5（治疗、纳米、多肽、组合、分子、基因、生物医药、细胞、 |

| | |
|---|---|
| | 表达、抑制剂、肿瘤、活性、药学、疾病、抑制、衍生物、载体、类化合物、体内、特异性）；Topic#6（药物、治疗、组合、作用、细胞、生物医药、抑制、提取物、组织、修复、效果、表达、用途、心肌纤维化、靶向、血管、抗肿瘤、活性、诱导、抗菌肽）；Topic#7（药物、冠状病毒、治疗、纳米、抑制、肿瘤、细胞核、活性、体外、分子、耐药、树枝状、制剂、肿瘤细胞、修饰、释放、聚合物、肽、诱导、胶束）；Topic#8（细胞、治疗、肿瘤、特异性抗体、纳米、编码、疫苗、抑制、活性、结构域、分子、基因、药物、氨基酸序列、免疫、生物医药、蛋白、癌症、受体、病毒） |
| 2020-2022 年 | Topic#1（抗体、药物、化合物、细胞、生物医药、抑制、多肽、治疗、纳米、分子、蛋白、表达、基因、肿瘤、序列、组合、抑制剂、冠状病毒、氨基酸序列、培养基 ）；Topic#2（药物、结构域、抗原受体、嵌合、抗体、诱导、特异性、肿瘤细胞、靶点、分子、核酸、纳米、融合蛋白、跨膜、重组、免疫、干细胞、受体、基因突变、信号肽药物、结构域、抗原受体、嵌合、抗体、诱导、特异性、肿瘤细胞、靶点、分子、核酸、纳米、融合蛋白、跨膜、重组、免疫、干细胞、受体、基因突变、信号肽）；Topic#3（药物、治疗、生物医药、抑制、组合、预防、表达、细胞、提取物、多肽、肿瘤、诱导、基因、冠状病毒、抑制剂、感染、组织、制剂、纤维化、分子）；Topic#4（药物、治疗、生物医药、化合物、衍生物、活性、靶向、纳米、分子、预防、肿瘤、载体、类化合物、特异性、诱导、递送、水凝胶、核酸、释放、联合）；Topic#5（纳米、表达、细胞、特异性、抗原、重链可变区、蛋白、单克隆抗体、编码、病毒、抗原受体、嵌合、轻链可变区、冠状病毒、生物医药、药物、机体、重链、免疫应答、肿瘤） |

（3）论文-专利主题关联关系分析

首先，对各时间区间“论文-专利”主题语义相似度进行计算。基于余弦相似度算法，以论文主题为参照对象，计算论文主题和专利主题之间的余弦相似度值并将余弦相似度数值最大和余弦相似度数值最小的结果输出，如表 5 所示。

表 5 “论文-专利”技术主题之间相似度的计算结果

Table 5 Calculation results of similarity between technical topics of "paper-patent"

| 时间区间 | 论文主题 | 专利主题 | 余弦相似度最大值 | 余弦相似度最小值 |
|---|---|---|---|---|
| 2010-2016 | Topic1 | Topic1<br>Topic2<br>.....<br>Topic8 | 0.95359 | 0.92301 |
| | Topic2 | Topic1<br>Topic2<br>.....<br>Topic8 | 0.98419 | 0.95262 |
| | Topic3 | Topic1<br>Topic2<br>..... | 0.99192 | 0.94353 |

| | | | | |
|---|---|---|---|---|
| | | Topic8 | | |
| | Topic4 | Topic1 | 0.97388 | 0.91673 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic5 | Topic1 | 0.99404 | 0.77374 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic6 | Topic1 | 0.93004 | 0.82693 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic1 | Topic1 | 0.99146 | 0.81061 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic2 | Topic1 | 0.99183 | 0.78527 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic3 | Topic1 | 0.99085 | 0.77374 |
| | | Topic2 | | |
| | | ..... | | |
| 2017-2019 | | Topic8 | | |
| | Topic4 | Topic1 | 0.97349 | 0.74204 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic5 | Topic1 | 0.98429 | 0.82693 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic6 | Topic1 | 0.98781 | 0.82131 |
| | | Topic2 | | |
| | | ..... | | |
| | | Topic8 | | |
| | Topic1 | Topic1 | 0.94984 | 0.90763 |
| | | Topic2 | | |
| | | ..... | | |
| 2020-2022 | | Topic5 | | |
| | Topic2 | Topic1 | 0.98784 | 0.95214 |
| | | Topic2 | | |

| | | | |
|---|---|---|---|
| | ..... | | |
| | Topic5 | | |
| Topic3 | Topic1 | 0.99073 | 0.94861 |
| | Topic2 | | |
| | ..... | | |
| | Topic5 | | |
| Topic4 | Topic1 | 0.90368 | 0.82349 |
| | Topic2 | | |
| | ..... | | |
| | Topic5 | | |

其次，从两个维度对各时间区间“论文-专利”主题关系进行挖掘。

一方面，从数值分布宏观角度来分析：

从表 5 可以直观的看出，在 2010-2016 年间，论文主题和专利主题之间的余弦相似度值最大值为 0.99404，最小值为 0.77374；2017-2019 年间，论文主题和专利主题之间的余弦相似度值最大值为 0.99183，最小值为 0.74204；2020-2022 年间，论文主题和专利主题之间的余弦相似度值最大值为 0.99073，最小值为 0.82349。

可见，这三个时间区间下“论文-专利”主题之间的余弦相似度值都较高，数值分布在 0.74204 和 0.99404 之间，这说明在技术主题网络演化的过程中，大部分科学研究与其技术创新之间是相近的，部分技术主题间呈现高度的统一。

另一方面，从语义内容细粒度维度来看：

2010-2016 年间：在科学研究方面聚焦于对新药研发、技术创新、新兴产业等研究主题的探索；同时，专利发明人围绕着病毒载体类基因治疗药物研究、肿瘤纳米药物的研究、多肽药物在肿瘤免疫治疗中的研究等技术领域进行核心技术攻关。

2017-2019 年间：在科学研究方面聚焦于预测模型、技术创新、研发创新等研究主题；专利发明人围绕着治疗性基因编辑技术、小分子抑制技术、细胞治疗技术等技术领域进行核心技术攻关。相较于上一时期，科学研究重点从新药研发逐步转移到了技术创新、研发创新等方面，开始注重先进技术等在医药领域中的应用，研究如何利用基于特定功能的计算模型来进行基因遗传性疾病的诊断与治疗，与专利发明人聚焦的技术主题领域是强相关的。

2020-2022 年间：在科学研究方面，抗体、DNA 序列、纳米、智能模型、创新平台等成为了热门研究主题；与此同时，专利发明人围绕着高通量测序技术、新型药物递送技术、药物偶联物技术、重组抗体技术等技术领域进行核心技术攻关。生物医药技术领域的科学研究者和专利技术人员的关注重心为蛋白组学相关主题，学术研究主题和技术主题之间存在知识的相互渗透、互相促进。

综上所述，我们从数值分布宏观角度和语义内容细粒度维度这两个层面对各时间区间“科学论文-专利技术”主题关系进行了详细的分析，可见，在 3 个时间段里，“科学论文-专利技术”之间的技术主题大部分是相近的或者高度统一的，科学研究和专利技术之间的知识关联性较强。

## 4 结论

本文将多策略综合分析方法引入到科学论文-专利技术关联关系挖掘中，选取特定技术领域，在社会网络分析、主题模型、相似度计算等核心技术的支撑下，将文本共现网络视角下的关联分析方法和文本主题挖掘视角下的关联分析方法相结合，通过混合方法构建了科学论文-专利技术关联分析模型，揭示了科学论文-专利技术互动规律，对于准确把握科技发展趋势、促进创新成果转化精准施策有重要意义。

主要研究内容总结如下：

第一，在构建科学论文-专利技术关联关系分析模型阶段。首先，基于文献数量分布情况进行时间窗口切割；其次，从文本共现网络视角使用社会网络分析软件对各时间区间的学术论文和专利文献进行多维度共现分析；最后，从文本主题挖掘视角出发，利用 BERT+Kmeans+LDA 融合模型对各时间区间的文本内容进行技术主题提取、利用余弦相似度算法对各时间区间的主题进行距离测度以及后续的基于时间窗口的主题关联关系演化分析。

第二，在实证分析阶段。以“生物医药”技术领域为例进行模型验证分析，从横向维度挖掘出了不同时间区间下“科学论文-专利技术”关注的主题，从纵向维度，进一步地，探讨了主题之间的关联关系和知识流动情况。在不同时间区间下，主题研究领域“科学论文-专利技术”关注的热门主题分别为新药研发、基因编辑、蛋白质组相关；在同一时间区间下，科学知识和技术知识之间相互流动，生物医药领域的科学研究对专利技术发明有着积极的影响，另一方面技术的创新也推动相关科学研究的发展，二者相互关联，互相促进。总之，科学和技术的相互推进和转换促成了知识的创新和科技转移转化。

第三，综上所述，本文较好地将文本共现网络视角的关联关系分析方法和文本主题挖掘视角的关联关系分析方法结合起来并在上述两种关联关系分析方法中融入时间窗口划分的分析方法，采取混合方法进行关联关系分析，实现了多维度、语义细粒度和动态化的统一，提高了科学论文-专利技术关联关系挖掘的准确性和全面性。

需要说明的是，本文是以“生物医药”技术领域为例，在融入动态时间窗口划分思想的基础上从文本共现网络和文本主题挖掘两个维度验证了所混合方法在科学论文-专利技术主题关联演化研究的可行性与合理性，仅实现了单一技术领域的关联关系挖掘和分析。因此，在未来的研究中，可以从多维度分析视角出发对模型结构进行优化和调整并以模型的普适性和稳健性为目标来开发原型系统，为科技成果的转化提供实践路径。

参考文献


[1] PIERRE, AZOULAY, WAVERLY, et al.The impact of academic patenting on the rate, quality and direction of (public) research output[J]. Journal of Industrial Economics, 2009,57(4):637-676.

[2] Han,F.,Magee,C.L.Testing the science/technology relationship by analysis of patent citations of scientific papers after decomposition of both science an-d technology[J].Scientometrics,2018,116(2):767–796.

[3] W Glänzel, Meyer M.Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations[J]. Scientometrics, 2003, 58(2):415-428.

[4] QI Y,ZHU N,ZHAl Y ,et al.The mutually beneficial relationship of patents and scientific literature:topic evolution in nano-science[J].Scientometrics,2018,115(2):893-911.

[5] Meyer M,Debackere K,Glnzel 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.

[6] Wang J J, Ye F Y.Probing into the interactions between papers and patents of new CRISPR/CAS9 technology:A citation comparison[J].Journal of In-formetrics,2021,15(4):101189.

[7] 高继平,丁堃,滕立,等.专利—论文混合共被引网络下的知识流动探析[J].科学学研究,2011,29(8):1184-1189，1146.

[8] HUANG M H,YANG H W,CHEN D Z.Increasing science and technology linkage in fuel cells:a cross citation analysis of papers and patents[J].Journ-al of informetrics,2015,9(2):237-249.

[9] 李蓓,陈向东. 基于专利引用耦合聚类的纳米领域新兴技术识别[J]. 情报杂志,2015,34(5):35-40.

[10] 何贤敏,李茂西,何彦青. 基于孪生 BERT 网络的科技文献类目映射[J]. 计算机研究与发展,2021,58(8):1751-60.

[11] 赖院根.期刊论文与专利文献的链接研究[J].图书情报知识,2011(01):63-69.

[12] 姜鑫,王德庄,马海群.关键词词频变化视角下我国“科学数据”领域研究主题演化分析[J].现代情报,2018,38(1): 141-146，161.
[13] Xie P, Xing E P. Integrating document clustering and topic modeling[EB/OL].[2024-04-08].https://arxiv.org/abs/1309.6874.
[14] 徐红姣,曾文,张运良.基于 Word2vec 的论文和专利主题关联演化分析方法研究[J].情报杂志,2018,37(12):36－42.
[15] Yashuang,Zhu,Na, et al.The mutually beneficial relationship of patents and scientific literature: topic evolution in nanoscience[J].Scientometrics: An In-ternational Journal for All Quantitative Aspects of the Science of Science Policy, 2018,115(2):893-911.
[16] Breschi S, Catalini C. Tracing the links between science and technology: An exploratory analysis of scientists' and inventors' networks[J]. Research Policy, 2010, 39(1):14-26.
[17] Wang, Ming-Yeu, Chang, et al. Exploring technological opportunities by mining the gaps between science and technology: Microalgal biofuels[J]. Technological Forecasting & Social Change, 2015,92(3):182-195.
[18] Tang Y,Lou X,Chen Z,et al. A Study on Dynamic Patterns of Technology Convergence with IPC Co-Occurrence-Based Analysis:The Case of 3D Prin-ting[J].Sustainability, 2020,12(7):2655.
[19] Shibata N, Kajikawa Y, Sakata I. Detecting potential technological fronts by comparing scientific papers and patents[J]. Foresight, 2011, 13(5):51-60.
[20] 宁子晨, 魏来. 专利主体视角下专利文献与学术论文关联关系发现研究——以"数据挖掘"主题为例[J]. 图书情报工作, 2020, 64(12):12.
[21] 叶梦竹.基于专利和论文互引的科学—技术关联研究[D].武汉:华中师范大学,2017.
[22] 席笑文, 郭颖, 宋欣娜, 等. 基于 word2vec 与 LDA 主题模型的技术相似性可视化研究[J].情报学报, 2021, 040(009):974-983.
[23] Taylor R M C,Du Preez J A.SimLDA:A tool for topic model evaluation[EB/OL].[2024-04-08].https://arxiv.org/abs/2208.09299.
[24] Su Y, Zhou X. BERT-LDA for Key Technology Identification: An Experimental Study on Carbon Neutralization[C]//Proceedings of the World Conference on Intelligent and 3-D Technologies (WCI3DT 2022) Methods, Algorithms and Applications. Singapore: Springer Nature Singapore, 2023: 435-445.

**作者简介：**冉从敬（1978—），男，湖北利川人，教授，博士生导师，主要研究领域为大数据治理、知识产权；田文芳（1989—），女，湖北黄冈人，博士研究生，主要研究领域为大数据治理、知识产权，Email：18627035850@163.com；贾志轩（1994—），男，山西运城人，博士后，主要研究领域为人工智能理论、知识产权。

**基金项目：**本文为国家社会科学基金重大项目“大数据主权安全保障体系建设研究”的成果，项目编号：21&ZD169。