高校战略性新兴产业专利识别与分析系统的构建*

——从IPC自动识别到语义相似度分析

于曦[1] 方胜宇[2,3]
（1. 天津师范大学图书馆，天津 300387；2. 天津仁爱学院信息与智能工程学院，天津 301636；3. 天津大学电气自动化与信息工程学院，天津 300072）

摘要：高校战略性新兴产业专利的识别与分析对于促进高校学科发展与布局，增强自主创新竞争力，实现知识产权强国的战略目标具有重要的现实意义。首先通过文献调研法分析战略性新兴产业专利的研究现状和识别难度，其次解构国家知识产权局制定的《战略性新兴产业分类与国际专利分类参照关系表（2021）（试行）》，以德温特专利数据为分析对象，采用SAO与NLTK语义相似度算法，利用信息熵词典加权，构建集自动识别、内容分析、质量价值评估与可视化研究于一体的战略性新兴产业专利半监督分析系统。高校可利用该系统开展战略性新兴产业专利的识别与分析。

关键词：战略性新兴产业；专利；知识产权；语义分析；高校
中图分类号：G258.6；G255.53　DOI：10.3772/j.issn.1673-2286.2023.11.009
引文格式：于曦，方胜宇. 高校战略性新兴产业专利识别与分析系统的构建：从IPC自动识别到语义相似度分析[J]. 数字图书馆论坛，2023（11）：74-82.

“战略性新兴产业”的概念最早是在2009年国务院首次召开的战略性新兴产业发展座谈会上正式提出的：“战略性”强调的是其在国民经济和产业结构调整中的重要性，对经济社会的发展具有引领和推动作用；“新兴产业”是指新建立的或是重新塑形的产业。2010年10月国务院正式发布纲领性文件《国务院关于加快培育和发展战略性新兴产业的决定》，确定了新能源、新材料、信息通信、新医药、生物育种、节能环保、电动汽车等七大产业为未来重点培育的战略性新兴产业[1]。自2016年国务院颁布《“十三五”国家战略性新兴产业发展规划》，明确了九大战略性新兴产业之后，我国对战略性新兴产业的支持力度不断加大，战略性新兴产业的发展也在不断加速。

高校产生的创新资源是国家科技创新的主要组成部分，尤其是在高新技术和新兴产业方面的创新能力和研究成果，能够极大地提高国家和地区的经济发展水平。高校图书馆通过开展对高校产生的战略性新兴产业专利的深度分析服务，能展现出高校战略性新兴产业研究的重点和特色，发现存在的问题和不足，制定改进措施，从而提高高校开展战略性新兴产业研究的水平。同时高校战略性新兴产业研究水平的提高，对我国的经济增长和产业转型也有一定的促进作用。

1 战略性新兴产业专利研究现状

为贯彻落实党的十九届五中全会关于发展战略性

收稿日期：2023-08-25
*本研究得到天津市哲学社会科学规划（重点）项目“‘双一流’学科建设背景下全域学科分类映射理论与实践”（编号：TJTQ21-001）资助。

新兴产业部署要求，加强战略性新兴产业专利分析及动向监测，满足战略性新兴产业专利活动的统计需要，2021年国家知识产权局制定了《战略性新兴产业分类与国际专利分类参照关系表（2021）（试行）》（以下简称《关系表》）。针对九大战略性新兴产业领域以及脑科学、量子信息和区块链等关键核心技术领域，建立战略性新兴产业与国际专利分类（IPC）的参照关系，为实现战略性新兴产业专利与经济活动的关联分析提供统计依据。

自《关系表》颁布以来，先后有学者开展了基于《关系表》的战略性新兴产业专利分析研究，可以归纳为以下3类。①全国或不同地域范围的战略性新兴产业专利的分析：张晶等[2]研究了我国战略性新兴产业创新发展情况；李楠楠等[3]、任树刚等[4]和赵华等[5]分别分析了承德市、宁波市和苏北地区战略性新兴产业专利发展现状及对策建议。②面向高校的战略性新兴产业专利的分析：吴晶晶等[6]和冯劭华等[7]对青岛市高校战略性新兴产业发展现状、创新能力与产业化前景进行分析并提出质量提升的对策。③战略性新兴产业专利的识别及深加工分析：王淼等[8]提出了战略性新兴产业专利数据的筛选模式和具体的深加工方法，开展了加工后数据的应用研究。

上述文献主要是对战略性新兴产业专利的描述统计分析与比对研究，战略性新兴产业专利识别和内容分析的相关研究较少。就高校战略性新兴产业专利识别和分析而言，虽然现有专利数据库incoPat可以提供对战略性新兴产业专利的一次性检索服务，但作为商业数据库，其费用并非每所高校都能负担的，尤其是在后疫情时代各高校的资金都相对匮乏的情形下。用于专利分析的数据库很多，每个数据库都有其特点，每个高校购买的专利分析数据库也各不相同，不一定都是incoPat。商业数据库为方便操作，其分析模式一般相对固定，但每所学校的专利特色不同，为满足一些学校的特定需求，专利分析者更倾向于使用操作相对自由的分析工具。因此本文构建基于IPC自动分析与SAO（Subject-Action-Object）语义相似度的半监督专利分析系统，开展对高校战略性新兴产业专利的识别和分析研究，方便高校尤其是中小型专利体量高校的专利分析工作。

2 战略性新兴产业专利识别与分析框架

《关系表》是战略性新兴产业专利的识别与分析的基石。依据《关系表》提取出符合IPC范畴的专利，根据《关系表》提供的对应关键词，筛选出战略性新兴产业专利。对专利进行语义相似度分析、SAO提取和专利语境提取。在此基础上构建信息关系矩阵（Information Relation Matrix，IRM），开展专利可视化，最终实现对专利的全面分析和评估。其系统逻辑如图1所示。

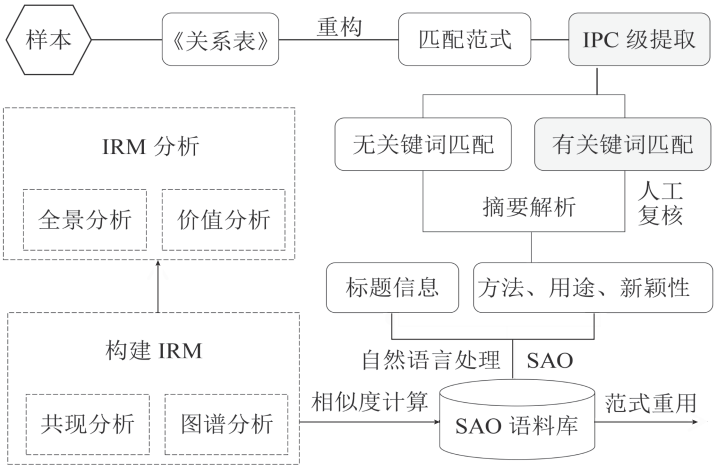

图1　战略性新兴产业专利的识别与分析系统

2.1 战略性新兴产业专利识别难度

根据国家知识产权局关于战略性新兴产业的表述与《关系表》，相关专利的精确检索分析难度较高。用战略性新兴产业专利IPC分类号制定检索式较复杂，不方便专利检索。更为重要的是，从表1可见国家知识产权局的中国专利标题无特征化描述，不利于对核心技术的识别，而德温特专利数据库对原始专利重新进行

表1 德温特专利标题与国家知识产权局专利标题对比

专利号	标题分类	专利标题
CN108987218-A、CN108987218-B	德温特专利标题	Improving field emission performance of graphene sheet-silicon nanowire array composite material involves preparing silicon nanowire array by using metal catalytic corrosion method, and bombarding silicon nanometer by using silver ion
		提高石墨烯片-硅纳米线阵列复合材料的场发射性能，包括采用金属催化腐蚀法制备硅纳米线阵列，并采用银离子轰击硅纳米
	国家知识产权局标题	一种提升石墨烯片-硅纳米线阵列复合材料场发射性能的方法
CN108359452-A、CN108359452-B	德温特专利标题	Water-soluble graphene quantum dots for detecting e.g. epinephrine, comprise macromolecular polymer, and have UV absorption at preset wavelength, fluorescence in visible blue light region, and surface with large number of amino groups
		用于检测例如肾上腺素的水溶性石墨烯量子点由大分子聚合物组成，在预设的波长下有紫外线吸收，在可见蓝光区域有荧光，表面有大量的氨基
	国家知识产权局标题	一种水溶性类石墨烯量子点及其制备方法与应用

著录，提供了特征明显的专利标题和摘要。

为解决上述问题，实现战略性新兴产业专利的方便、快速和精准检索，也为进一步的专利分析，利用德温特专利开展战略性新兴产业专利的识别与检索，从而提升准确率，便于后期的语义相似度分析。样本数据是某高校的专利产出，该校属于地方性综合类高校，专利主要来自化学、生物和环境学科。专利检索自德温特专利数据库，共有1 767条专利数据。

2.2 《关系表》数据结构

《关系表》共建立了1 872个关系，有8个部、89个大类、317个小类、2 893个大组、35 473个小组涉及IPC。《关系表》中包括战略性新兴产业分类代码、战略性新兴产业名称、IPC分类号和关键词概述共4项内容（见表2）。战略性新兴产业分类代码及战略性新兴产业名称相关表述均与《战略性新兴产业分类（2018）》相一致。战略性新兴产业均对应一个或多个IPC分类号，表示该IPC分类号下的专利与所述战略性新兴产业相关。IPC分类号后加“*”表示包括该层级IPC分类号及以下所有分类号，对于需排除的分类号会加括号予以说明。关键词概述是指IPC对应的关键词，可对该分类进行进一步的限定和说明，以实现更为准确的统计。《关系表》具体内容如表2所示。

表2 国家知识产权局《关系表》（部分）

分类代码	产业名称	IPC分类号	关键词概述
1.3	新兴软件和新型信息技术服务	G06F8*（不含G06F8/60、G06F8/61、G06F8/65、G06F8/654、G06F8/656、G06F8/658、G06F8/70、G06F8/71、G06F8/72、G06F8/73、G06F8/74、G06F8/75、G06F8/76、G06F8/77）、G06F30*	
		G06F11/36	基础类网络与信息安全软件、网络与边界安全类软件等其他软件开发；互联网安全服务

2.3 战略性新兴产业专利提取与解析

从《关系表》结构可知，无关键词概述的专利只要符合对应分类的IPC范围即是战略性新兴产业专利，对有关键词概述的专利还需要判断专利是否包含指定的关键词。为此，对《关系表》进行重构，根据表2生成唯一性匹配范式（见表3），共1 839个战略性新兴产业专利匹配范式，应用于系统的IPC级自动识别。

系统根据表3，从样本数据中自动提取出符合匹配范式的专利信息，得到自动抽取结果（见表4）。其中“建议”列为系统分析后给出的人工干预建议：专利“无匹配项”则可以直接删除忽略；如果某范式存在《关系表》指定的关键词，则对专利标题与关键词进行语义相似度判断，给出匹配的关键词，后期进行人工干预。

样本专利CN101073311-B覆盖IPC的A、B、C 3个大部，经自动处理后，符合国家知识产权局“生物农业及相关产业”类型，细分为3个小类（A01G、B09B、C05F），可以看出只符合范式198，对应国家知识产权局《关系表》“4.3 生物农业及相关产业 C05F*”，无指

表3 战略性新兴产业专利匹配范式（部分）

ID	分类代码	产业名称	IPC_IN（包含）	IPC_OUT（不包含）	关键词
15	1.3	新兴软件和新型信息技术服务	G06F8*	G06F8/60、G06F8/61、G06F8/65、G06F8/654、G06F8/656、G06F8/658、G06F8/70、G06F8/71、G06F8/72、G06F8/73、G06F8/74、G06F8/75、G06F8/76、G06F8/77	
15	1.3	新兴软件和新型信息技术服务	G06F30*		
16	1.3	新兴软件和新型信息技术服务	G06F11/36		基础类网络与信息安全软件、网络与边界安全类软件等其他软件开发；互联网安全服务

注：ID是系统根据《关系表》的每行数据赋予的序号。

定关键词项。C05F为“C05B、C05C子类未涵盖的有机肥料，例如废物或垃圾肥料”。该专利的原始标题为“一种采用农作物秸秆协同草坪植物修复生活垃圾重金属的应用方法”，德温特专利数据库改写后的标题为“用于修复生活垃圾重金属（如‘铬’）的草坪植物栽培基质，是由包括城市生活垃圾堆肥和农业秸秆的材料形成的”。后者更契合C05F分类的定义，其专利标题如前文所述更具特征和准确性。因此，利用德温特专利数据库的标题与《关系表》指定的关键词进行语义匹配与分析是一个效率上的良好选择，进一步可以利用德温特专利的结构化摘要作更为细化的分析。

样本CN103436263-B的IPC分类号C09K-011/85符合国家知识产权局“C09K-011”条件，专利标题“水溶性红绿可调法转换的掺稀土的纳米材料”同时符合“稀土相关”关键词要素，“建议”列给出了匹配项“Rare earth”，供人工审核。

根据系统自动处理和人工干预，样本数据中“符合要求”的有186条，“无匹配项”的有1 202条，需要进行人工干预的有379条，经人工审核后共发现512条战略性新兴产业专利记录。

表4 系统自动抽取结果（部分）

专利号	IPC分类号	类 型	德温特改写专利标题	ID	关键词	建 议
CN101073311-B	A01G-001/00、A01G-031/00、B09B-003/00、C05F-017/00	生物农业及相关产业	Grass lawn plant cultivating substrate for repairing domestic garbage heavy metal, e.g. chromium, is formed by materials comprising city domestic garbage compost, and agricultural straws	199	林木育种和育苗；种子种苗培育	无匹配项
				277	清淤机械；环保技术、资源循环利用技术推广服务	无匹配项
				198		符合要求
CN103436263-B	C09K-011/85、G01N-021/64	先进有色金属材料	Preparation of a water-soluble red green tunable method conversion of rare earth-doped nanomaterials, by oxidizing yttria, ytterbium oxide and erbium oxide powder, adding concentrated nitric acid and pure water, heating and evaporating	92	稀土相关关键词	Rare earth

2.4 战略性新兴产业专利内容分析

通过IPC级匹配后又经人工甄选，得到的只是形式上的战略性新兴产业专利。鉴于专利文本的复杂性，仍需要利用自然语言处理、人工智能等技术进行内容层面的分析。从专利的标题、摘要部分利用SAO提取核心表述、特征以及专利文本的技术语境，进行相似度分析。

2.4.1 语义相似度

有三大类常见的语义相似度计算算法：①基于最短路径的算法，包括Wu-Palmer、Leacock & Chodoraw等；②基于信息量（Information Content，IC）的算法，包括nuno、DN、概念拓扑以及熵模型；③路径与IC混合模型。NLTK（Natural Language Toolkit）自然语言处理包中包含WordNet组件，这是由普林斯顿大学的心

理学家、语言学家和计算机工程师联合设计的一个基于认知语言学的英语词典，其根据单词的意义组成一个“单词的网络”。NLTK提供6个语义相似度计算模块：计算词典层次结构中最短路径的path_similarity[9]；基于Wu-Palmer的最短路径算法wup_similarity[10]；类别信息加权的最短路径算法lch_similarity[11]；计算层次结构中公共包容节点位置，即最特定祖先节点深度的算法res_similarity[12]；使用公共包容节点和两个同义词集的IC计算两词的相似度的算法jcn_similarity [13]；在IC的基础上，把共性IC与完整描述所需的IC比值作为相似度分值的算法lin_similarity[14]。

IC的概念被广泛应用于各种各样的信息度量中，用来衡量能从一个信息来源中学到多少东西。一个概念的IC是该概念在大型语料库中的频率，通过最大似然估计来计算[15]。当事件发生的概率增大时，IC的值就减小。Resnik[16]提出基于IC的res_similarity算法，在测量语义相似度时引入了概念的概率：建议计算语料库中某一类型词的出现次数，将该计数除以与该词相关的不同概念/语义的数量，将这个值分配给每个概念，即概念的概率，根据IC在最相关的语义结构上测量语义相似度，如概念S_1和S_2之间的语义相似度是两个概念共享的IC。算法如式（1）所示。

$$RES(S_1, S_2)=IC[LCS(S_1, S_2)] \quad (1)$$

式中：$RES(S_1, S_2)$指概念S_1和S_2之间的语义相似度，$IC[LCS(S_1, S_2)]$指概念S_1和S_2之间最小共同子项（LCS）的IC。最小共同子项在WordNet中指两个概念之间的最短距离。

信息熵将消息中的信息量化为消息接收者的新信息，通过将概率与IC相乘，放大了一个具有较大的IC的罕见事件的小概率，但抑制了具有小的IC的普通事件的大概率。人们更希望从稀有事件，而不是普通事件中学到更多东西。信息熵相比IC更能体现概念的价值。

本系统基于NLTK和WordNet，将基于IC的语义相似度算法（res_similarity、jcn_similarity和lin_similarity）公式中的IC替换为信息熵，对相似度的计算进行加权。系统提供了bnc（http://www.hcu.ox.ac.uk/BNC/）、brown（http://www.hit.uib.no/icame/cd）、treebank（http://ldc.upenn.edu）、semcor（WordNet 3.0）、semcorraw（http://www.cs.unt.edu/~rada/downloads.html#semcor）等预置IC语料库用于提高关键词与专利语义匹配的精度。语义相似度分析可帮助找到表述不同但内容相近的专利，便于后续分析。

2.4.2 SAO三元组提取

SAO源于生物研究领域，国内的专利特征分析研究广泛应用SAO。李晓曼等[17]对常见的研究进行了综述，评析了常见的专利相似度算法与SAO抽取工具。任雪菁[18]总结了SAO技术的主题演化路径。翟东升等[19]在词向量的基础上结合SAO，构建专利技术功效图。基于SAO的专利分析通过分析专利中的句子来提取技术信息，常见的SAO抽取范式见表5。

表5 常见SAO抽取范式

范 式	句子类型	样 例	派生子句
S1	SA	S died.	（S，died）
S2	SAad	S remained in Princeton.	（S，remained，in Princeton）
S3	SAC	S is smart.	（S，is，smart）
S4	SAO	S has won the Nobel Prize.	（S，has won，the Nobel Prize）
S5	SAOO	RSAS gave S the Nobel Prize.	（RSAS，gave，S，the Nobel Prize）
S6	SAOad	The doorman showed S to his office.	（The doorman，showed，S，to his office）
S7	SAOC	S declared the meeting open.	（S，declared，the meeting，open）

注：S表示主语，A表示谓语，O表示宾语，ad表示状语，C表示补语。

大多数句子都有主语（S）、谓语（A）和宾语（O），基于此可以识别重要的信息，如技术的功能、效果、特点、解决方案、组件和语境。事实上，功能即一个系统或技术可以执行的任务或行动，可以用A-O的形式来表达。在包含技术描述的句子中，专利中规定的发明的对象、工具、方法和系统可以用S的形式表达。专利分析中A尤其重要，主要表述专利的核心方案、技术手段；O是技术方案或制备产生的结果，以及核心素材，它也是专利特征的重要元组。目前专利SAO相关学术研究多聚焦于理论[20-22]，鲜有研究运用于知识分析实践。借鉴Reverb、Clauses、Ollie等三元组知识抽取工具框架，结合专利应用语境，使用预先训练的能够预测语

言特征的统计模型en_core_web_sm（https://spacy.io/models/en），设计专利核心特征提取方案。表6列出了部分专利的SAO提取结果。

表6 专利SAO提取内容（部分）

德温特专利号	确信度	S	A	O	附加项
DIIDW：1992260761	0.799	lithium	does not exist in	the battery	[enabler=so the battery does not ignite even when broken]
DIIDW：2002519092	0.756	electrodes	allow redox reactions by	exchange of Li ions	

2.4.3 专利语境与IRM

一项专利包含各种类型的信息，但其对技术功能、语境和组件信息的描述是最基本和最重要的，可以用矩阵的形式来表达。语境在语言学的环境下，大多是S与O的定语或者句子的状语（ad），表明专利的应用环境、应用领域、应用场景、技术范围等。这些内容在相似专利的特征提取以及相似度分析时，比SAO本身更具研究与实践价值。

结合专利语境和结构化文摘（方法、用途、新颖性）开展对战略性新兴产业专利的IRM分析[23]，是战略性新兴产业专利深度分析必不可少的环节。系统根据预定义的语法分析器，从512条专利样本数据的摘要中抽取出语境信息（见表7）。

表7 专利语境（部分）

语 境	计 数
as catalyst	39
as fluorescent material	82
for adsorption	15
for detection	23
for fluorescent material	13
for preparation	12
for use	31
in airport	12
in detection	27
in fluorescent material	26
in preparation	15
in water	11
of adsorption amount	25
of dye	54
of Festuca arundinacea	15
of heavy metal	19
of lawn plant	23
of ryegrass	14

注：从摘要的用途部分提取。

可以看到，出现频次最高的是“as fluorescent material”，可知“荧光材料”是样本专利的主要应用场景。构建“语境-功能矩阵”（见图2），从专利数据中提取符合需求的技术信息，并简单直观地显示各类技术信息之间的相互关系。

Context-Function Matrix	detect adsorption	detect dye	detect cadmium ion	detect kanamycin	detect sulfur isotope	reduce number of insect
in airport	0	0	1	2	0	10
in coupling reaction	2	0	1	3	2	0
in detecting dye	4	9	13	12	12	1
in fluorescent material	9	9	8	5	6	12
in preparation	13	11	2	1	0	6
in the field	10	5	11	9	4	7
in water	5	7	7	5	13	12

图2 语境-功能矩阵（部分）

2.5 专利质量、价值与可视化分析

尽管抽取出符合国家知识产权局《关系表》定义的战略性新兴产业专利，但还需进一步确定专利的质量，而不是只作形式上的界定。专利质量的分析是一项繁琐的工作，因为有许多因素影响其质量。为了进行适当的专利估值，知识产权分析师必须了解哪些因素会影响专利的质量。

2.5.1 专利质量与价值

专利质量的影响因素有很多，包括专利强度、广度和兴趣。从现有文献[24-26]中整理出专利质量与价值的主要影响因素，见表8。

结合专利质量价值因子，分析并判断战略性新兴产业专利的价值。很多文献详细表述了专利质量与价值的评价方法，限于篇幅，本文不再赘述。高校知识产权分析师在确定专利质量时必须考虑许多因素。然而，如果市场不相信该发明提供的效用优于市场上的其他发

表8 专利质量价值因子

专利质量价值因子	描 述
律师或代理人的专利申请	相比发明人，律师或代理人可能更了解如何提出能够承受诉讼的权利要求
从属和独立权利要求数量	拥有大量这两种权利要求的专利描述了更多潜在的发明和变化
专利申请过程	专利越细化，就越不可能面临授权后的有效性挑战。专利申请过程越详尽，专利的质量就越高
专利转让	没有既得利益的实体之间发生专利转让，这可能表明专利质量高
维护费支付	为现有专利支付经常性维护费的公司发现了这些专利的价值
现有技术披露	大量的现有技术披露意味着申请人对专利空间有深刻的理解。它意味着申请人在起草专利申请时考虑了现有技术，可能更难对已授权专利的有效性质疑
授权后的挑战	能经受住授权后挑战的专利质量较高，不容易受到后续挑战的影响。这种成功也进一步保证了专利对现有技术的有效性
维权的胜利	打赢侵权官司的专利权人证明其专利具有良好的质量

明，那么专利可能没有什么价值。因此，虽然专利质量是一个重要的指标，也是知识产权分析师必须考虑的指标，但它只是专利评估过程中的一小部分。

2.5.2 专利可视化

除了文字语言的表述，专利质量、价值或许可以可视化的方式进行展示。可视化分析不是统计数字的简单列表，而是基于专利内部与外在特征进行关系揭示，除了传达表达质量价值的特征，还应对某些模糊而难以精确表达的潜在意义予以图示。专利可视化常用共现矩阵、聚类和分类、空间概念图谱、分层或堆叠信息、地理表示法、网络分析与语义分析等方法来实现。限于篇幅仅列出层次聚类和专利的全域科学叠加图谱。

从1 767条样本数据中抽取出无关键词匹配的战略性新兴产业专利186项，使用jcn_similarity算法计算，并进行层次聚类分析，得到186项战略性新兴产业专利层次聚类图（见图3）。图3中每个点代表1项专利，这些专利在不同的层级构成树状关系，表明其以相似度计算为基础的层次聚类。其中，较大且相邻的节点即为相似度较高的专利，以专利CN105859751-A和CN105859786-A为例。这两项专利的标题分别为“New 1，4-dimethyl-2，5-di-1H-1，2，4-bistriazole copper terephthalate complex crystals used as catalyst for preparing 4，4’-difluorobiphenyl，have structure analyzed by graphite monochromated ray，and preset unit cell parameters”和“New manganese 1，4-dimethyl-2，5-di-1H-1，2，4-bistriazole 5-methylisophthalate single crystal complex used as catalyst for preparing 4，4’-difluorobiphenyl，has structure analyzed by graphite monochromated ray，and preset cell parameters”。在此基础上进一步探索，发现这两项专利的著作权人是来自同一所高校的WANG Y和WANG Z。通过相似度聚类可以在大量专利文献中发现专利之间的亲疏关系，找到内容相似的专利。通过对专利原始内容的查看找到具有相同研究方向的研究者，发现潜在的合作伙伴，如上述两位发明人。此外还可发现专利申请中可能存在的学术不端行为。

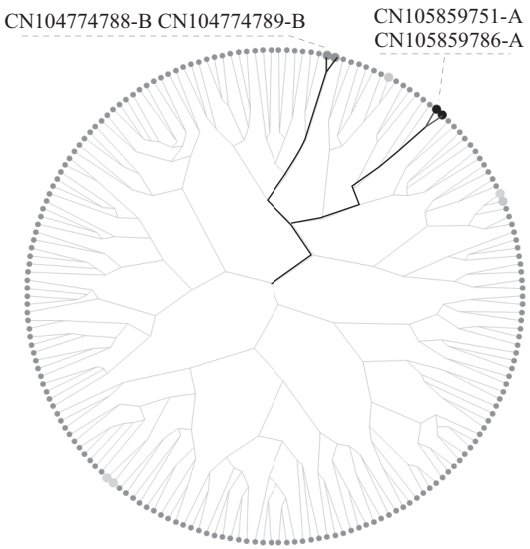

图3 186项战略性新兴产业专利层次聚类图

通过可视化分析中的叠加图谱可以在全域科学的场景中观察到目标样本的分布与关系。研究所用的专利全域基础图是Leydesdorff等[27]基于美国专利商标局的

专利引文关系构建的，可以为分析人员提供一个稳定的思想框架，以便在不同年份的检索中跟踪所关注领域的发展情况。从图4可见，这186项专利集中在IPC的C部“化学，冶金”（Chemistry，Metallurgy），以及A部的“人类生活需要”（Human Necessities）。更深入的，专利集中在“C07C无环或碳环化合物”“C12N微生物或酶、繁殖、保藏或维持微生物；变异或遗传工程；培养基”等，反映了这些专利的布局重心在于化学、生物领域，说明化学和生物学科是产生战略性新兴产业专利的主阵地，为校企合作战略布局提供决策支持。

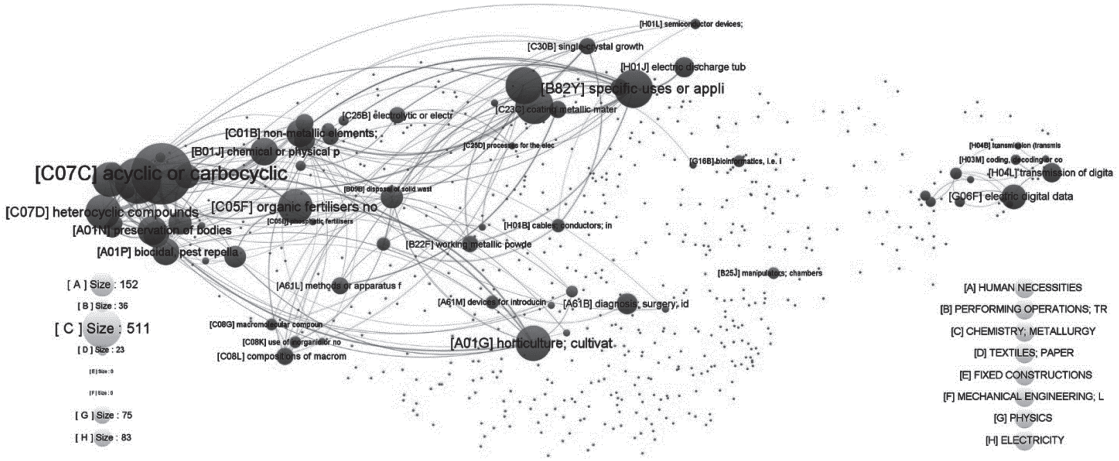

图4　186项战略性新兴产业专利IPC全域叠加图

3　结语

在数字经济时代，知识产权的创造和运用是自主创新的主要指标，申请战略性新兴产业专利必将成为新兴产业参与全球竞争的重要手段。高校是产生战略性新兴产业专利的主要机构，也是推动我国新兴产业快速、持续发展的重要力量。因此，如何构建新兴产业的专利战略已成为高校面临的突出问题。国家知识产权局认定的战略性新兴产业专利在实践中具有战略指导意义，然而鲜有研究关注其具体应用与实践，这是由于专利文本自身复杂性以及《关系表》的复杂结构导致了实际操作困难。本文构建的战略性新兴产业专利半监督分析系统对战略性新兴产业专利的识别与分析具有重要的理论价值和现实意义，对分析高校战略性新兴产业专利布局特点、存在的问题和未来发展方向具有指导意义。

参考文献

[1]　国务院关于加快培育和发展战略性新兴产业的决定[EB/OL]. [2023-11-21]. https://www.gov.cn/gongbao/content/2010/content_1730695.htm.

[2]　张晶，苏源哲. 我国战略性新兴产业创新发展探究：专利信息分析视角[J]. 科技创业月刊，2021，34（6）：74-79.

[3]　李楠楠，寇成，马闯，等. 承德市战略性新兴产业专利分析[J]. 质量与市场，2022（18）：4-6.

[4]　任树刚，金燕. 宁波市战略性新兴产业发明专利质量提升研究[J]. 中国市场监管研究，2022（2）：74-77，73.

[5]　赵华，赵雪雅，顾瑞婷. 专利视角下苏北地区战略性新兴产业发展现状及对策[J]. 科技中国，2021（5）：57-61.

[6]　吴晶晶，张根. 青岛市高校战略性新兴产业发展现状与质量提升对策：基于2011—2020年专利数据的分析[J]. 中国高校科技，2022（6）：68-72.

[7]　冯劭华，咎栋，廖巍，等. 战略性新兴产业专利的创新能力与产业化前景：以青岛市12所高校院所为例[J]. 科技管理研究，2022，42（10）：160-167.

[8]　王淼，秦璐，彭茂祥. 新兴产业分类与国际专利分类对照方法及专利数据深加工方法模式研究[J]. 科学管理研究，2020，38（5）：87-92.

[9]　RADA R，MILI H，BICKNELL E，et al. Development and application of a metric on semantic nets[J]. IEEE Transactions on Systems，Man，and Cybernetics，1989，19（1）：17-30.

[10]　WU Z B，PALMER M. Verbs semantics and lexical selection[C]//Proceedings of the 32nd annual meeting on Association for Computational Linguistics. 1994：133-138.

[11]　LEACOCK C，CHODOROW M. Combining local context and

WordNet similarity for word sense identification[M]//FELLBAUM C. WordNet. Cambridge，MA：The MIT Press，1998.

[12] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1. 1995：448-453.

[13] JIANG J J，CONRATH D W. Semantic similarity based on corpus statistics and lexical taxonomy[C]//Proceedings of International Conference on Research in Computational Linguistics（ROCLING X）. 1997：19-33.

[14] LIN D. An Information-theoretic definition of similarity[C]//Proceedings of the Fifteenth International Conference on Machine Learning. 1998：296-304.

[15] ALEXANDER G. Computational Linguistics and Intelligent Text Processing[M]. Berlin：Springer，2003.

[16] RESNIK P. WordNet and Class-Based Probabilities[M]. Cambridge，MA：The MIT Press，1998.

[17] 李晓曼，宋红燕. 面向专利情报研究的SAO语义结构分析方法述评[J]. 情报科学，2020，38（10）：168-176.

[18] 任雪菁. 基于SAO结构的技术主题演化路径研究[D]. 北京：北京协和医学院，2021.

[19] 翟东升，张京先，胡等金. 基于SAO结构和词向量的专利技术功效图自动构建研究[J]. 情报理论与实践，2020，43（3）：116-123.

[20] 张永真，吕学强，申闫春，等. 基于SAO结构的中文专利实体关系抽取[J]. 计算机工程与设计，2019，40（3）：706-712.

[21] 张玉洁，白如江，刘明月，等. 融合语义联想和BERT的图情领域SAO短文本分类研究[J]. 图书情报工作，2021，65（16）：118-129.

[22] 徐惟康. 基于SAO结构的专利创造性检索系统的设计与实现[D]. 南京：南京邮电大学，2020.

[23] KI W，KIM K. Generating information relation matrix using semantic patent mining for technology planning：a case of nano-sensor[J]. IEEE Access，2017，5：26783-26797.

[24] 罗家豪，孙巍. 基于专利价值的技术成熟度测度与分析方法研究[J]. 数字图书馆论坛，2022（1）：17-25.

[25] 徐明，陈亮. 基于文献综述视角的专利质量理论研究[J]. 情报杂志，2018，37（12）：28-35.

[26] 许鑫，赵文华，姚占雷. 多维视角的高质量专利识别及其应用研究[J]. 现代情报，2019，39（11）：13-22，45.

[27] LEYDESDORFF L，KUSHNIR D，RAFOLS I. Interactive overlay maps for US patent（USPTO） data based on International Patent Classification（IPC）[J]. Scientometrics，2014，98（3）：1583-1599.

作者简介

于曦，女，硕士，副研究馆员，研究方向：学科服务、数据分析，E-mail：yuxisd_2004@126.com。
方胜宇，男，硕士，讲师，研究方向：人工智能、深度学习。

Construction of Patent Identification and Analysis System for Strategic Emerging Industries in Universities: From IPC Automatic Identification to Semantic Similarity Study

YU Xi[1] FANG ShengYu[2,3]
(1. The Library of Tianjin Normal University, Tianjin 300387, P. R. China; 2. School of Information and Intelligence Engineering, Tianjin Renai College, Tianjin 301636, P. R. China; 3. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, P. R. China)

Abstract: The identification and analysis of strategic emerging industry patents in universities have important practical significance in promoting the development and layout of disciplines, enhancing independent innovation competitiveness, and achieving the strategic goal of building a strong intellectual property country. This paper first analyzes the research status and identification difficulty of strategic emerging industry patents through literature research, and then deconstructs the Reference Relationship Table of Strategic Emerging Industry Classification and International Patent Classification (2021) (for trial implementation) formulated by the China National Intellectual Property Administration. Taking the data of Dwent patents as the analysis object, using the semantic similarity algorithm of SAO and NLTK, and using information entropy dictionary weighting, this paper constructs a semi supervised analysis system for strategic emerging industry patents that integrates automatic recognition, content analysis, quality value evaluation, and visualization research. This system can be used for universities to identify and analyze patents in strategic emerging industries.

Keywords: Strategic Emerging Industry; Patent; Intellectual Property Right; Semantic Analysis; University

（责任编辑：王玮）