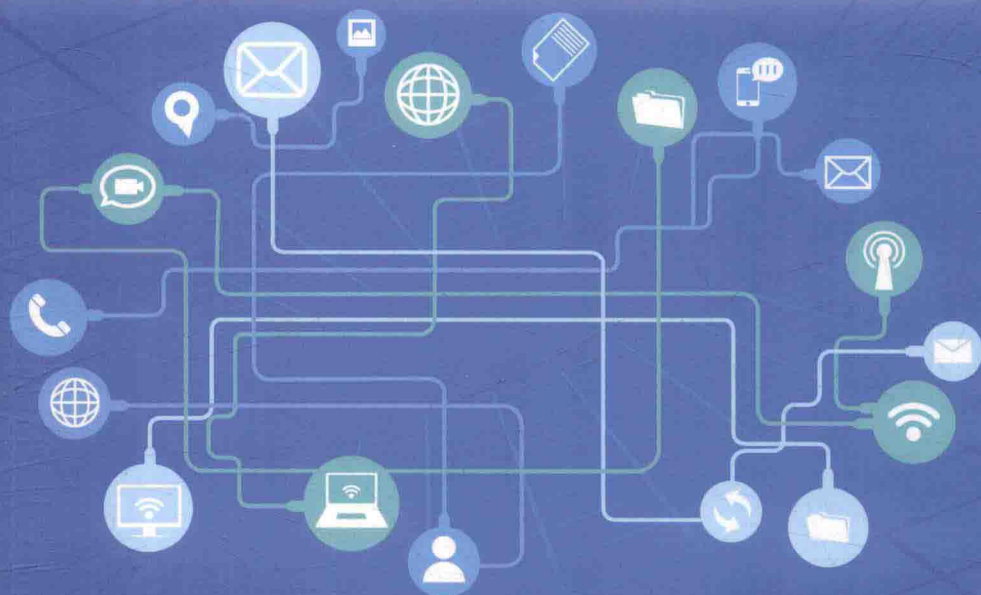


专利信息资源挖掘与发现 关键技术研究

刘 耀 朱礼军 靳 玮 © 著



国家科技支撑计划课题“专利信息资源挖掘与发现关键技术研究”
(2013BAH21B02) 特别资助

专利信息资源挖掘与发现 关键技术研究

刘 耀 朱礼军 靳 玮 著



科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

· 北 京 ·

图书在版编目 (CIP) 数据

专利信息资源挖掘与发现关键技术研究 / 刘耀, 朱礼军, 靳玮著. —北京: 科学技术文献出版社, 2018.2

ISBN 978-7-5189-3616-8

I. ①专… II. ①刘… ②朱… ③靳… III. ①专利—信息资源管理—研究
IV. ①G306.3

中国版本图书馆 CIP 数据核字 (2017) 第 284114 号

专利信息资源挖掘与发现关键技术研究

策划编辑: 周国臻 责任编辑: 周国臻 白建刚 责任校对: 文 浩 责任出版: 张志平

出 版 者 科学技术文献出版社
地 址 北京市复兴路15号 邮编 100038
编 务 部 (010) 58882938, 58882087 (传真)
发 行 部 (010) 58882868, 58882870 (传真)
邮 购 部 (010) 58882873
官 方 网 址 www.stdp.com.cn
发 行 者 科学技术文献出版社发行 全国各地新华书店经销
印 刷 者 北京教图印刷有限公司
版 次 2018年2月第1版 2018年2月第1次印刷
开 本 710×1000 1/16
字 数 189千
印 张 12.25 彩插4面
书 号 ISBN 978-7-5189-3616-8
定 价 58.00元



版权所有 违法必究

购买本社图书, 凡字迹不清、缺页、倒页、脱页者, 本社发行部负责调换

前 言

专利文献是一种承载海量关键技术的信息资源。根据世界知识产权组织（WIPO）发布的报告显示，自2012年起中国已成为全世界发明专利申请数量第一大国。在如此庞大的数据量下，如何将专利文献的关键信息呈献给用户，使用户能快速掌握信息要点，是优化专利服务的关键科学问题。

本书是国家“十二五”科技支撑计划课题“专利信息资源挖掘与发现关键技术研究”（课题编号：2013BAH21B02）研究成果的组成部分。课题研究的目标是研究数字化、网络化环境下专利信息资源的知识化组织、高效率存储、智能化分析和可视化展示等关键技术并开发系列软件工具，形成面向技术创新的专利信息资源挖掘与发现的关键技术支撑。

围绕以上目标，课题组通过以国家和企业斥巨资建成的海量专利信息资源和服务体系为基础，在建设专利深度标注与知识抽取的资源基础上，研究了专利文本内容自动标注与集成、专利关联知识网络构建和领域专利知识组织工具等专利信息深度分析与处理技术，开发相关软件工具集，并搭建应用服务平台，以检索、分析、可视化的方式面向行业领域提供专利信息资源挖掘与发现服务。

在课题开展的过程中，课题组注入了夜以继日、砥砺深耕的心血。这也是一个不断思考、实践、提升的过程。我们将研究过程中所知所学所获整理为一本书，希望能为相关读者提供一些可供参考、讨论的方法与实例。

本书内容主要介绍专利信息挖掘与发现涉及的框架、方法与技术研究等内容，包括：

①面向智能信息服务的专利信息组织与自动处理研究。研究专利文本内容自动标注与集成技术、专利关联知识网络构建和领域专利知识组织工具构建等关键技术。

②专利信息分析评价模型及应用模型研究。采用定性分析、定量分析、拟定量分析结合的方式,构建宏观、中观、微观3个层面的专利分析模型,并从多个表现特征入手构建了包含引证分析、技术功效矩阵分析、技术成熟度分析等的预测方法和体系。

③专利信息融合分析研究。整合专利文献中的申请人、法律状态、同族专利、引证文献等关键信息,研究专利信息动态关联分析关键技术。

④专利信息智能检索与语义导航研究。研究基于概念检索的海量专利数据智能搜索引擎、多维索引体系的构建、交互性立体式检索结果揭示、个性化推荐等关键技术。

⑤专利信息可视化研究。研究支持专利语义内容关联分析的专利信息可视化关键技术。

本书在编写过程中,参阅了大量的图书和文献,汲取了很多精髓,特别是引用了部分图表、数据等,在此向有关作者表示诚挚的感谢。在研究过程中也得到诸多业界专家、领导、同事的帮助和支持,在此一并致谢。同时,也向课题组成员表示感谢。

由于书中涉及内容广泛,加之笔者水平有限,书中若有不足、错谬之处,恳请同行专家和读者批评指正,以便再版时进一步修订完善。

刘耀

2018年1月

Contents 目 录

1 绪 论	1
1.1 研究背景与意义	1
1.2 研究思路	1
1.3 研究内容	2
1.4 关键问题	3
1.5 创新之处	3
2 面向智能信息服务的专利信息组织与自动处理关键技术及工具研发	5
2.1 专利文本内容深度标注研究	5
2.1.1 专利获取与清洗	6
2.1.2 专利句法分析	10
2.1.3 专利语料错误检测	20
2.1.4 专利标引	34
2.2 专利动词语义框架库建设	39
2.2.1 物性结构	39
2.2.2 语义角色	40
2.2.3 专利动词语义框架	41
2.3 专利知识抽取模板规则知识库建设	43
2.3.1 基于条件随机场的专利术语抽取研究	43
2.3.2 专利术语语义关系自动标注	48
2.3.3 专利部件图库建设	52
2.4 管线技术研究	63
2.4.1 定制化管线技术	63

2.4.2	专利文献管线处理	64
2.4.3	应用研究	70
3	专利信息分析评价模型及应用模型研究与工具 研制和开发	73
3.1	专利信息分析理论与模型研究	73
3.1.1	宏观分析	74
3.1.2	中观分析	76
3.1.3	微观分析	77
3.2	专利引证分析	78
3.3	技术功效矩阵分析	80
3.3.1	技术词	81
3.3.2	功效词	81
3.4	技术成熟度分析	82
3.4.1	技术生命周期判断	83
3.4.2	技术特征分析	83
3.4.3	技术发展趋势	84
3.4.4	新兴技术发现模型	85
4	专利信息融合分析核心技术研发	86
4.1	基本框架	86
4.2	分类法映射	89
4.2.1	解决思路	89
4.2.2	实现流程	90
4.2.3	功能展示	92
4.3	同族专利分析	93
4.4	专利相似度分析	96
4.4.1	总体思路	97
4.4.2	实现流程	98
4.4.3	功能展示	99
4.5	多模态信息融合分析	100
4.5.1	表格检索	101
4.5.2	图像检索	115

5 专利信息智能检索与语义导航关键技术研发	117
5.1 智能检索与语义导航框架	117
5.2 语义资源自动构建技术	119
5.2.1 语义资源结构获取	119
5.2.2 语义资源关系获取	122
5.2.3 语义资源进化	124
5.2.4 语义资源管理	128
5.3 语义标注与索引技术	130
5.3.1 思路与框架	131
5.3.2 语义标注流程	132
5.3.3 结果与分析	134
5.4 技术应用与展示	137
6 专利信息可视化技术研发	141
6.1 可视化流程与设计	141
6.1.1 数据采集	142
6.1.2 数据处理和变换	142
6.1.3 可视化映射	143
6.1.4 用户感知	144
6.2 基于统计图表的可视化	144
6.2.1 柱状图	145
6.2.2 折线图	151
6.2.3 饼图	157
6.2.4 气泡图	160
6.2.5 雷达图	162
6.3 基于拓扑结构的可视化	165
6.3.1 基本原理	165
6.3.2 生成步骤	166
6.3.3 应用	167
6.4 基于地图的可视化	168
6.4.1 基本原理	168
6.4.2 生成步骤	168

6.4.3 应用	168
6.5 基于聚类结构的可视化	169
6.5.1 基本原理	169
6.5.2 生成步骤	169
6.5.3 应用	170
6.6 基于标签云的可视化	170
6.6.1 平面标签云	171
6.6.2 3D 标签云	172
7 总 结	177
参考文献	178

图表目录

图 1-1	技术路线	2
图 2-1	一体化语义爬虫流程	7
图 2-2	一体化爬虫系统界面	8
图 2-3	前 100 条准确率	9
图 2-4	前 1000 条准确率	10
图 2-5	不同词性的依存准确率	12
图 2-6	MSTParser 特征空间	14
图 2-7	通用词性和依存句法标注示例	14
图 2-8	通用不同规模训练语料准确率	15
图 2-9	通用不同规模训练语料准确率变化趋势	15
图 2-10	专利中文依存关系的 dot 图示	16
图 2-11	专利词性和依存标注示例	17
图 2-12	专利不同规模训练语料准确率	17
图 2-13	专利不同规模训练语料准确率变化趋势	18
图 2-14	专利标注示例的短语表示	19
图 2-15	专利语料用标注格式转换工具	21
图 2-16	专利依存错误检测的网页输出结果	26
图 2-17	专利标引流程	35
图 2-18	专利摘要标注	36
图 2-19	分词标注选择及处理结果	36
图 2-20	句法分析选择及处理结果	37
图 2-21	XML 结构化处理及结果	38
图 2-22	自动标引处理结果及校对	38

图 2-23	专利动词语义框架库截图 1	43
图 2-24	专利动词语义框架库截图 2	43
图 2-25	线链条件随机场模型的图形结构	44
图 2-26	模板文件片断	46
图 2-27	专利文献中的装置结构	52
图 2-28	专利部件图示例 1	53
图 2-29	专利部件图示例 2	53
图 2-30	消解错误示例 1	60
图 2-31	消解错误示例 2	61
图 2-32	消解错误示例 3	61
图 2-33	消解错误示例 4	62
图 2-34	切分词流程	64
图 2-35	切分词部件切分词结果	64
图 2-36	词性标注流程	65
图 2-37	词性标注结果	65
图 2-38	句法分析流程	66
图 2-39	处理前文本	67
图 2-40	句法分析部件处理结果	67
图 2-41	篇章关系标注流程	68
图 2-42	篇章内容标注流程	69
图 2-43	分段分局处理结果	69
图 2-44	平台基础流程	70
图 2-45	专利知识抽取步骤	71
图 2-46	专利知识抽取线索	71
图 2-47	专利知识抽取结果展示	72
图 3-1	专利详情页面	80
图 3-2	专利引证分析页面	80
图 3-3	以“石墨烯”为主题的功效矩阵	82
图 3-4	以“燃料电池”为主题的功效矩阵	82
图 4-1	专利信息融合分析技术框架	87
图 4-2	信息融合分析系统业务流程	88

图 4-3	分类法映射流程	90
图 4-4	分类法映射交互界面	93
图 4-5	同族专利分析技术方式和流程	94
图 4-6	同族专利页面	95
图 4-7	同族专利链接结果页面	96
图 4-8	专利相似度分析技术方式和流程	97
图 4-9	专利相似度分析系统框架结构	98
图 4-10	专利详情页面	100
图 4-11	专利相似度对比微观分析页面	100
图 4-12	表格识别与检索框架	102
图 4-13	导入的 Word 格式文档示例	103
图 4-14	Word 和 Excel 文档处理流程	104
图 4-15	HTML 处理流程	104
图 4-16	导入的网页示例	105
图 4-17	网页数据	105
图 4-18	转换结构化数据示例	106
图 4-19	图像类表格处理流程	107
图 4-20	输入的图像类表格	108
图 4-21	输出的 Excel 结果示例	108
图 4-22	输入的 PDF 文档类型	109
图 4-23	输出结果示例	109
图 4-24	索引后的结构文件示例	110
图 4-25	索引后的内容数据示例	111
图 4-26	内容数据分在不同的文档中存储	111
图 4-27	检索入口展示	112
图 4-28	检索结果页面展示	112
图 4-29	点开标题超链接结果展示	112
图 4-30	查看详情示例	113
图 4-31	点击油耗显示结果示例	113
图 4-32	输入二氯甲烷的显示结果	113
图 4-33	点击二氯甲烷标题之后的显示结果	114

图 4-34 图像检索的基本框架	115
图 4-35 建立图像索引	116
图 5-1 基于领域本体的专利信息智能检索与语义导航框架	118
图 5-2 词表转换整体流程	119
图 5-3 基于代码的结构化词表导入流程	120
图 5-4 层级关系转换流程	121
图 5-5 语义资源结构获取结果	122
图 5-6 文件学习整体流程	122
图 5-7 导入属性 1 流程	123
图 5-8 提取属性 2 流程	123
图 5-9 语义资源进化整体流程	125
图 5-10 关键词获取流程	126
图 5-11 新知识添加流程	127
图 5-12 语义资源构建系统与知识管理系统的交互流程	128
图 5-13 写入语义资源时同步流程	129
图 5-14 读取语义资源时同步流程	129
图 5-15 选择要发布的类	130
图 5-16 选择命名空间	130
图 5-17 语义索引目录结构	134
图 5-18 领域语料来源示例	135
图 5-19 语义标注算法结果	135
图 5-20 功能模块结构	137
图 5-21 专利信息智能检索与语义导航页面	138
图 5-22 专利基本信息检索页面	138
图 5-23 专利高级检索页面	139
图 5-24 高级检索结果页面	139
图 5-25 专利智能检索与导航页面	140
图 6-1 可视化流程	142
图 6-2 可视化流水线	142
图 6-3 专利可视化呈现形式	144
图 6-4 基于统计图表的可视化呈现形式	145

图 6-5	2012—2016 年专利申请数量	147
图 6-6	石墨烯领域发明人专利数量	148
图 6-7	石墨烯领域申请人专利申请数量	148
图 6-8	石墨烯技术分类专利数量	149
图 6-9	某技术领域重点申请人发明人专利产出率	149
图 6-10	石墨烯领域各个专利申请人专利类型专利申请量	150
图 6-11	石墨烯领域各个专利申请人技术分类专利数量	150
图 6-12	某公司专利申请地域分布柱状图	151
图 6-13	专利量趋势分析	153
图 6-14	石墨烯领域专利申请人数量	154
图 6-15	石墨烯领域专利发明人数量	154
图 6-16	石墨烯领域专利类型	155
图 6-17	石墨烯领域各个专利申请人专利数量	155
图 6-18	石墨烯领域各个技术分类专利数量	156
图 6-19	生命周期曲线	156
图 6-20	石墨烯领域专利法律状态	159
图 6-21	石墨烯领域各个技术分类占比	159
图 6-22	各国石墨烯制备技术	160
图 6-23	化工领域技术主题专利申请量	164
图 6-24	专利引证关系 1	167
图 6-25	专利引证关系 2	167
图 6-26	石墨烯领域的专利主题聚类	170
图 6-27	绕 X 轴旋转示意	173
图 6-28	绕 Y 轴旋转示意	174
图 6-29	绕 Z 轴旋转示意	174
图 6-30	搜索热词 3D 标签云	176
表 2-1	专利语料的词频统计	12
表 2-2	专利语料的依存距离	13
表 2-3	MSTParser 通用语料实验结果	15
表 2-4	MSTParser 专利依存实验结果	17

表 2-5	通用语料与专利语料的词比较	19
表 2-6	MSTParser 专利对比实验依存结果	20
表 2-7	专利 POS 错误检测评价	23
表 2-8	专利 POS 检测后处理 N 变元统计	24
表 2-9	专利 POS 错误检测后处理评价	24
表 2-10	专利依存错误检测评价	27
表 2-11	专利依存错误类型	28
表 2-12	引入语义信息化的分析器优化	31
表 2-13	专利 MSTParser 词性错误率	32
表 2-14	引入词性语义信息后的分析器优化	32
表 2-15	引入词类语义信息后的依存关系优化	33
表 2-16	专利动词及其论元结构库	42
表 2-17	专利动词的句法特征	42
表 2-18	术语抽取结果	47
表 2-19	与基于词标注方法的比较	48
表 2-20	语义关系标记集	49
表 2-21	语义关系抽取正则表达式示例	50
表 2-22	测试数据集统计分布	50
表 2-23	经典 RST 关系集	57
表 2-24	修辞结构与零形回指指向对照	59
表 2-25	规则 1 效果	60
表 2-26	规则 2 效果	60
表 2-27	规则 3 效果	61
表 2-28	规则 4 效果	62
表 2-29	规则 5 效果	62
表 2-30	规则 6 效果	62
表 3-1	结点信息	79
表 3-2	存储节点之间的引用信息	79
表 4-1	领域文本的各种层次关系对照	91
表 4-2	IPC 类目层级	91
表 4-3	专利文档相似度计算结果样例	99

表 4-4	关系表格	110
表 4-5	内容表格	110
表 4-6	测试结果	114
表 5-1	概念编码示例	121
表 5-2	知识关系获取结果 1	124
表 5-3	知识关系获取结果 2	124
表 5-4	互联网进化结果	127
表 5-5	互联网进化结果 F 值	128
表 5-6	搜狗搜索引擎与语义标注算法逆序数对比	136
表 5-7	语义标注后化工本体属性 2 统计	136

1 绪 论

1.1 研究背景与意义

伴随科学技术的迅猛发展,新兴学科、边缘学科不断涌现,使得包括专利文献在内的各种科技信息资源数量急剧增加。缺失专利文献的分析意味着无法回答科技发展过程中的某些重大问题,意味着难以使面向科技创新与科技管理的相关政策达到理想的效果,也难以对其进行测度与评价。目前,科研项目管理及创新主体科研实践过程中,专利信息服务与决策支持能力不够深入和系统。因此,通过开展专利信息资源挖掘与发现关键技术研究,在整合多种资源、工具、技术的基础上,进行专利语义知识内容的自动抽取技术及工具的研发,突破海量知识数据的存储、组织、索引和检索的效率与性能,构建专利知识关联网络,并以检索、可视化的方式进行信息展示,旨在为专利信息挖掘与发现相关技术的研发及关键技术在热点监测评价、专利预警、科研项目管理和重点领域的应用提供有益参考,以促进科技信息、技术资源的传播与交流,提升专利信息服务的社会及行业影响力,吸引企业、科研机构等创新主体对专利信息的关注与实践应用,实现创新资源的优化配置、合理共享与高效利用,提升创新能力与质量。

1.2 研究思路

本书的主要研究目标是研究数字化、网络化环境下专利信息资源的知识化组织、高效率存储、智能化分析和可视化展示等关键技术并开发系列软件工具,形成面向技术创新的专利信息资源挖掘与发现的关键技术支撑。

在建设专利深度标注与知识抽取的知识资源基础上,研究专利文本内容自动标注与集成技术、专利关联知识网络构建技术和领域专利知识组织工具等专利信息深度分析与处理的技术并开发相关软件工具集。结合专利评估与监测的理论研究与模型构建,以及上述专利自动标注与知识抽取的基础资源与工具集的建设等研究成果,进行专利信息融合分析核心技术、专利信息智能检索与语义导航关键技术研发、专利信息可视化技术研发的开发工作。