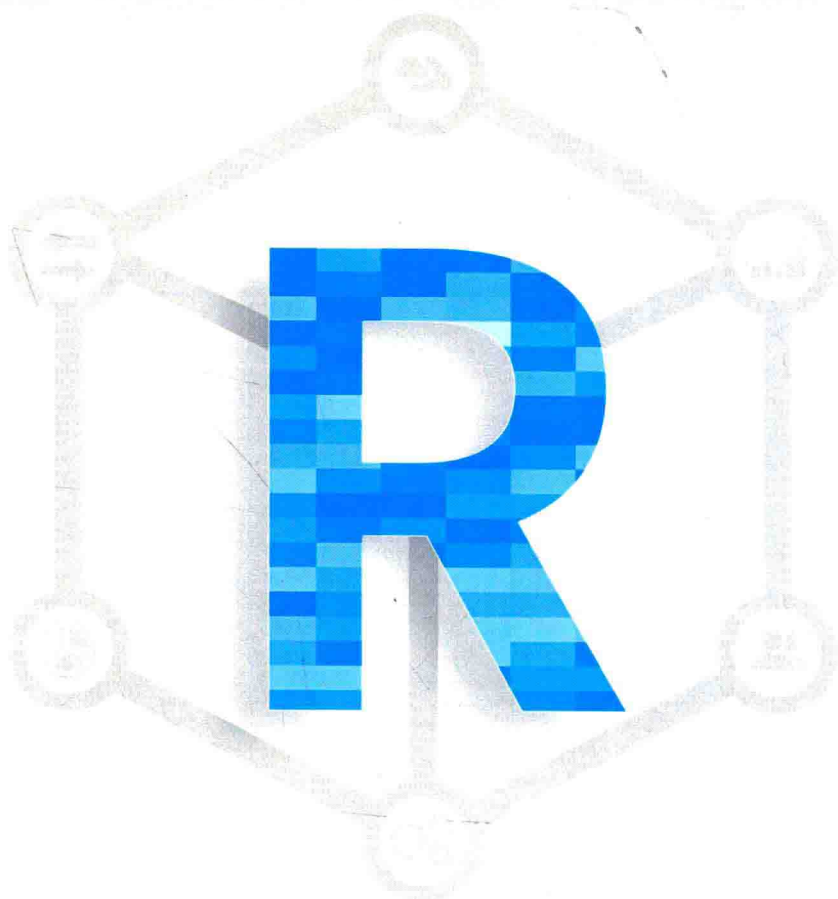




大数据处理技术

——R语言专利分析方法与应用

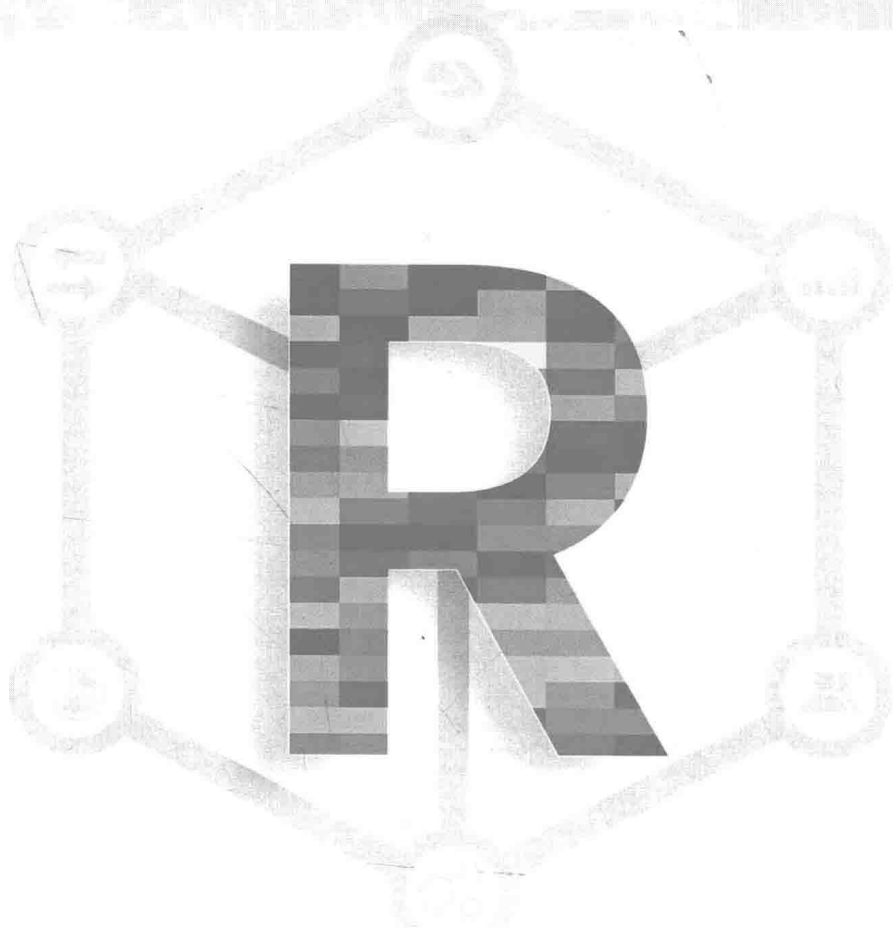
主 编◎屠 忻

副主编◎李立功 左良军 杨 爽 高慧霞 黄 煜 蒋 帆



知识产权出版社

全国百佳图书出版单位



大数据处理技术

——R语言专利分析方法与应用

主 编◎屠 忻

副主编◎李立功 左良军 杨 爽 高慧霞 黄 煜 蒋 帆



知识产权出版社

全国百佳图书出版单位

图书在版编目 (CIP) 数据

大数据处理技术: R 语言专利分析方法与应用/屠忻主编. —北京: 知识产权出版社, 2019. 9

ISBN 978-7-5130-6434-7

I. ①大… II. ①屠… III. ①程序语言—应用—数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2019) 第 189550 号

内容提要

本书是一本关于大数据处理技术的教程, 主要研究 R 语言在专利分析领域的应用方法, 全书从四个方面展开: 首先给出 R 语言快速入门需要掌握的基本知识; 然后从专利分析数据处理角度出发, 总结归纳用 R 语言处理专利数据的几种常用场景; 接着结合专利分析中的数据可视化给出常用专利分析图表的 R 语言制图方法; 最后结合数据挖掘算法介绍了利用 R 语言进行专利数据挖掘与建模的几种常见任务。

责任编辑: 龚卫李叶

责任印制: 刘译文

封面设计: 张冀

大数据处理技术

——R 语言专利分析方法与应用

DASHUJU CHULI JISHU

——R YUYAN ZHUANLIFENXI FANGFA YU YINGYONG

屠忻 主编

出版发行: 知识产权出版社有限责任公司

电 话: 010-82004826

社 址: 北京市海淀区气象路 50 号院

责编电话: 010-82000860 转 8745

发行电话: 010-82000860 转 8101

印 刷: 北京嘉恒彩色印刷有限责任公司

开 本: 720mm×1000mm 1/16

版 次: 2019 年 9 月第 1 版

字 数: 202 千字

ISBN 978-7-5130-6434-7

网 址: <http://www.ipph.cn>

<http://www.laichushu.com>

邮 编: 100081

责编邮箱: laichushu@cnipr.com

发行传真: 010-82000893

经 销: 各大网上书店、新华书店及相关专业书店

印 张: 13

印 次: 2019 年 9 月第 1 次印刷

定 价: 58.00 元

出版版权专有 侵权必究

如有印装质量问题, 本社负责调换。

前言

随着我国供给侧结构性改革地不断推进，具有制度供给和技术供给双重属性的知识产权，尤其是专利对供给侧结构性改革的支撑和保障作用不断增强，专利信息利用也越来越被社会各界所重视，专利在政府决策支撑、行业发展和企业创新等方面越来越发挥着重要作用。

专利分析作为专利信息挖掘的核心手段之一，是充分挖掘专利数据背后隐藏的高价值情报信息的重要支撑。专利分析涉及多个环节，包括数据处理、数据可视化、数据挖掘与建模等。虽然目前专利分析方法很多，但大多是针对不同环节采用不同的处理方式，仍然缺少能够实现多个环节有效衔接、统一处理的模式，这使专利分析的准确性和效率难以保障。

随着全球专利数量地不断提升，专利分析的难度大大增加，这在各个环节中均有所体现。其中，数据处理是专利分析的重中之重，是专利分析工作者获取规范化样本数据的主要途径，数据处理的质量是彰显专利分析成果准确性的核心指标之一。虽然专利格式统一、规范，但不同数据库对专利数据的收录范围、收录方式千差万别，这加大了各数据汇总后处理成规范化数据的难度，而且数据量的增加，也超出了常规数据处理方式的处理能力。因此，提升海量专利数据处理的质量和效率，是做好大数据时代下专利分析工作的重要基石，更是必然要求。

专利数据量的不断提升导致专利数据的可读性不断下降，海量混杂的专利数据信息交织在一起使数据更难于被理解，这不仅需要我们以更细化、更全面的形式来展示数据，还需要我们与数据进行更多的交互，以加强数据的理解，而传统的静态数据可视化形式显然不能满足目前的需求，开发针对海量专利数据的动态可视化方式是解决目前困境的必经之路。

同时，专利数据量地不断提升导致数据类型的不断扩展，数据量及类型的变化也对专利数据的挖掘和建模方式提出了更高的要求。针对特定的应用

场景，建立合适的模型来挖掘海量专利数据背后隐藏的高价值信息，是传统的统计分析工具所不能胜任的。利用深度学习算法进行数据挖掘已经在多个领域展现了巨大的优势，而在专利分析领域尚属空白。因此，探索研究基于深度学习的人工智能专利数据挖掘和建模方法，以满足海量专利数据的要求，势在必行，时不我待。

本书主要从四个方面展开：一是给出 R 语言快速入门需要掌握的基本知识；二是从专利分析数据处理角度出发，总结归纳用 R 语言处理专利数据的几种常用场景；三是结合专利分析中的数据可视化给出常用专利分析图表的 R 语言制图方法；四是结合数据挖掘算法介绍利用 R 语言进行专利数据挖掘的几种常见任务。

本书由国家知识产权局专利局专利审查协作江苏中心光电技术发明审查部组织编写，具体撰写分工如下：

屠忻，参与撰写前言、第 1 章、第 2 章、第 3 章第 3.1~3.2 节；

李立功，参与撰写第 3 章第 3.3~3.7 节、第 4 章第 4.1~4.3.1 节；

高慧霞，参与撰写第 4 章第 4.3.2 节、第 4 章第 4.4~4.5 节；

杨爽，参与撰写第 4 章第 4.3.3 节、第 4.6 节；

蒋帆，参与撰写第 4 章第 4.7 节~4.8 节、第 5 章第 5.1~5.2 节；

黄煜，参与撰写第 5 章第 5.3~5.4 节；

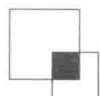
左良军，参与撰写第 5 章第 5.5~5.6 节、本书附录 A~C、参考文献。

全书由李立功、左良军负责统稿、校对。

本书内容为编者在专利分析工作中的经验总结，希望能为提升从业者的专利数据处理、数据可视化及信息挖掘的能力起到一定的积极作用。由于时间仓促，加之编者水平所限，部分成果还需要在实际应用中不断丰富和完善，应用效果也需要经过实践的进一步检验。本书中的观点和内容如若存在偏差和不足之处，敬请广大读者批评指正。

屠忻

2019 年 6 月 10 日



目 录

第1章 简介	001
1.1 关于本书 ▶ 001	
1.1.1 为什么要撰写本书 ▶ 001	
1.1.2 本书的撰写原则 ▶ 002	
1.1.3 本书的读者对象 ▶ 003	
1.1.4 本书的使用方法 ▶ 004	
1.2 专利分析概论 ▶ 004	
1.2.1 专利分析基本流程 ▶ 004	
1.2.2 当前专利分析基本方法 ▶ 010	
1.2.3 专利分析的发展方向 ▶ 011	
1.3 数据科学概论 ▶ 012	
1.3.1 数据取样与探索 ▶ 012	
1.3.2 数据预处理与可视化 ▶ 013	
1.3.3 数据挖掘与建模 ▶ 014	
1.4 小结 ▶ 014	
第2章 R语言入门	016
2.1 本章概述 ▶ 016	
2.2 R语言简介与安装 ▶ 017	
2.2.1 R语言简介 ▶ 017	

2.2.2	R 安装及 RStudio 简介	018
2.3	R 包的使用	021
2.3.1	R 包的介绍	021
2.3.2	R 包的安装和载入	022
2.4	常用 R 包及函数使用介绍	023
2.4.1	数据整理——tidyr	023
2.4.2	表格操纵——dplyr	027
2.4.3	字符处理——stringr	031
2.4.4	时间处理——lubridate	034
2.4.5	数据导入导出——openxlsx	036
2.5	R 语言数据结构	037
2.5.1	向量	037
2.5.2	矩阵	038
2.5.3	数组	039
2.5.4	数据框	040
2.5.5	因子	042
2.5.6	列表	043
2.6	小结	044

第3章 专利数据处理 045

3.1	本章概述	045
3.2	申请年份统计	046
3.2.1	年份申请量统计	047
3.2.2	年份国内外申请量统计	050
3.2.3	年份国别申请量统计	052
3.3	专利申请人统计	055
3.3.1	申请人专利数量统计	056
3.3.2	标准申请人清洗	058
3.3.3	申请人合作关系统计	062
3.4	技术主题统计	069
3.4.1	技术主题分布统计	069

3.4.2	技术主题占比统计 ▶	072
3.5	同族数据统计 ▶	073
3.5.1	同族数据拆分为多列 ▶	073
3.5.2	同族数据拆分为多行 ▶	076
3.6	多维数据联合统计 ▶	080
3.6.1	三维数据的联合统计 ▶	080
3.6.2	四维数据的联合统计 ▶	085
3.7	小结 ▶	086
第4章	专利数据可视化	087
4.1	本章概述 ▶	087
4.2	利用 ggplot2 包制图 ▶	088
4.2.1	柱形（条形）图 ▶	088
4.2.2	折线（路径）图 ▶	098
4.2.3	散点（气泡）图 ▶	103
4.3	利用 Highcharter 包制图 ▶	108
4.3.1	圆环类图 ▶	109
4.3.2	极坐标图 ▶	113
4.3.3	矩形树图及热力图 ▶	118
4.4	利用 Dygraphs 包绘制交互式时序图 ▶	121
4.4.1	折线时序图 ▶	122
4.4.2	折线+条形时序图 ▶	123
4.4.3	堆叠条形+折线时序图 ▶	124
4.5	利用 Circelize 包制图 ▶	126
4.5.1	申请人合作关系弦图 ▶	126
4.5.2	五局技术流向弦图 ▶	128
4.5.3	条形跑道图 ▶	132
4.6	专利地图的绘制 ▶	134
4.6.1	基于 baidumap 和 Remap 包绘制地图 ▶	135
4.6.2	专利地图与其他图表的结合 ▶	140

4.7	利用 NetworkD3 包制图 ▶	145
4.7.1	力导图 ▶	145
4.7.2	网络图 ▶	152
4.8	小结 ▶	155
第 5 章	专利数据挖掘与建模	156
5.1	本章概述 ▶	156
5.2	数据挖掘基础 ▶	157
5.2.1	数据挖掘的基本任务 ▶	157
5.2.2	数据挖掘建模的过程 ▶	158
5.3	变量主成分分析 ▶	159
5.3.1	问题背景 ▶	159
5.3.2	主成分分析方法 ▶	160
5.4	聚类分析 ▶	166
5.4.1	问题背景 ▶	166
5.4.2	K-means 聚类分析 ▶	166
5.5	分类与预测分析 ▶	173
5.5.1	问题背景 ▶	173
5.5.2	人工神经网络算法 ▶	174
5.5.3	支持向量机算法 ▶	181
5.5.4	朴素贝叶斯算法 ▶	185
5.6	小结 ▶	189
附录 A	本书代码索引	190
附录 B	本书用到的扩展包	192
附录 C	R 语言学习资源	195
参考文献		198

第 1 章 简 介

本书是一本关于计算机编程语言在专利分析领域应用方法的教材，主要介绍利用大数据处理技术——R 语言来实现专利分析的方法。在正式开始阅读本书之前，我们有必要对本书涉及的有关专利分析以及数据科学理论做一些基础性、概括性的介绍，以使读者更好地理解并掌握本书知识。

1.1 关于本书

1.1.1 为什么要撰写本书

在撰写本书之前，本书编者已经尝试并使用过多种专利分析的工具，例如，基于传统 EXCEL 软件进行专利数据处理和分析，并绘制相应的分析图表；或者采用 Microsoft Office 最新的商业智能软件，如 Power BI 等进行专利分析。前者在轻量级数据范围（小于 5000 条记录）内基本能满足常规专利分析需求，但是，一旦专利分析工作涉及的数据量级较大，EXCEL 软件运行缓慢，很难高效地完成数据分析任务；而后者基于微软最新推出的商业智能平台，极大地拓展了专利分析人员的分析工具，客观上促进了专利分析工具的发展。

此外，本书编者还曾尝试过直接使用专利检索与分析软件进行专利分析工作，目前市面上被广泛认可的专利分析平台主要有北京索意互动有限公司推出的 Patentics 专利分析客户端、北京合享智慧科技有限公司推出的 IncoPat 分析平台、智慧芽公司推出的 PatSnap 等，它们在很大程度上实现了检索、分析一体化。但上述专利分析平台价格不菲，一般研究人员难以自由使用。此外，利用上述专利分析平台进行分析会受限于平台本身的功能，很难做到分

析方案可定制化,这会影 响专利分析人员的想象力和创造力,容易导致分析报告的模式化。值得赞扬的是,上述分析平台均提供了良好的数据导入、导出接口,分析人员可以利用较低的成本获得分析平台提供的检索数据,这就为我们利用其他分析技术实现高度定制化的自由分析提供了可能。

本书编者在充分调研上述分析工具优缺点的基础上,结合当前大数据技术现状,决定采用计算机编程语言改造传统专利分析工具,即选择一种高效率的数据处理编程语言来实现专利分析工作。一方面,利用程序代码的可移植性,可以极大地减少分析人员在专利分析工作中的低端重复性工作;另一方面,利用程序代码的高度自由性,可以实现专利分析方案的定制化,充分发挥分析人员的想象力和创造力,助力分析人员实现从更深层次挖掘专利数据,进一步拓展专利分析的内涵和外延。

正是基于上述想法,本书编者决定将探索并研究的基于 R 语言的专利分析方法的有关成果整理并集结成书,帮助专利分析工作者进一步拓宽专利分析工具,提高专利分析技能。

1.1.2 本书的撰写原则

事实上,本书属于一本计算机程序设计类教材,更严格的说是一本大数据编程技术教材。编者结合自己的工作内容,将大数据编程技术应用到专利分析领域中,进一步拓宽专利分析人员的分析工具。我们并不苛求每一位专利分析工作者都需要掌握大数据编程技术,但是,希望专利分析从业者,可以用更低的学习成本提升自己的日常工作技能,从而进一步提高工作效率。正是基于上述思路,我们在编写本书时遵循如下两个原则。

1. 实用性

编写本书的首要原则是实用性。读者在看完本书以后,可以很快将本书中的代码应用到实际的专利分析工作中,从而快速地掌握基于 R 语言的专利分析方法。因此,本书编者在专利分析工作经验基础上,总结提炼出一些常规的专利分析应用场景,如关于申请人的统计、技术主题统计、同族数据处理等。在数据可视化章节中,同样是对常见的专利分析图表进行归纳整理,

并选择具有代表性和实用性的图表进行 R 代码的设计。本书后面章节中给出的有关程序设计代码，均可以直接移植并运行在 R 语言编程环境中，读者仅需根据本书说明，安装好 R 语言运行环境，修改个人计算机中的数据文件路径即可实现相应的数据处理或可视化效果。

2. 易读性

考虑到一般专利分析人员并不具备相应的数据编程技能，因此，本书并不过多介绍计算机专业知识，也并不涉及数据编程算法等较为艰深的理论，读者并不需要花费大量时间去学习有关计算机专业知识。本书仅介绍和数据处理相关的必备知识，如 R 语言中的数据结构，常用数据处理包和函数等数据编程基础知识，读者可以零基础学习上述相关知识。在具体代码设计阶段，我们仅对代码的设计逻辑做简单说明，有的代码段给出了封装的函数，读者只需知晓代码中可修改的参数即可，无须过多纠结于代码的设计原理，从而提高本书的易读性。

1.1.3 本书的读者对象

正如上一节中提出的，本书是将大数据编程技术应用到专利分析领域中的教材，主要聚集如何利用 R 语言技术来解决专利分析的任务场景。因此，本书并不对专利分析的基本知识做过多介绍，也不会对专利分析数据处理及可视化的应用背景做过多阐述，需要读者根据实际项目需求进行分析判断，应该对何种数据进行统计分析，应该采用何种分析图表进行信息展示。因此，我们建议阅读本书的读者应该对专利分析中的基本概念和分析内容熟练掌握，并具有一定的专利分析基本技能，至少应该具有利用 EXCEL 进行专利分析的基本经验。

此外，我们并不苛求专利分析人员具有编程基础，但是如果分析人员具有一定的计算机程序设计的通识知识将会有助于其更快、更好地理解并掌握本书的知识内容。如果你想要更进一步地提高专利分析效能，并且是一名对计算机编程技术感兴趣的专利分析工作者，那么，本书将会非常适合你！

1.1.4 本书的使用方法

本书是一本以数据编程技术为基础的教材，作为一本计算机编程技术类教材，读者在学习本书时要加强上机实践训练，在明确分析需求的基础上，读懂代码的功能，并立即动手实践，将代码亲自敲入编程环境，运行代码，观察结果，并根据实际分析需求修改代码中的可变参数，从而完成定制化的分析任务。结合编者的学习经验，在此给出如下推荐的学习路线图（见图 1-1）：

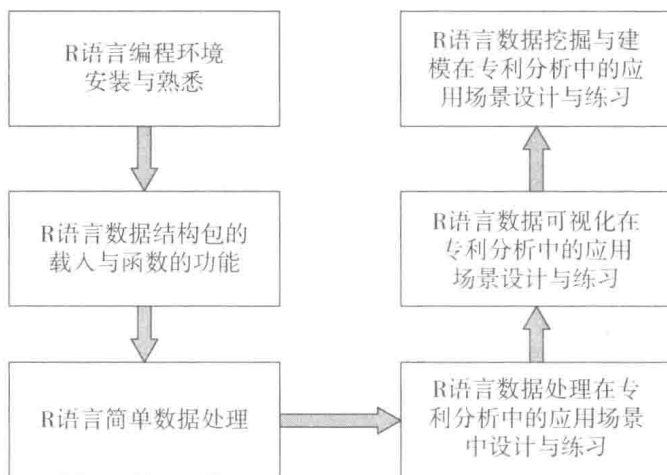


图 1-1 R 语言专利分析方法学习路径图

1.2 专利分析概论

1.2.1 专利分析基本流程

专利分析的基本流程包括前期准备、数据取样和预处理、数据可视化、数据挖掘与建模以及报告撰写五个阶段，每个阶段均包括细化的任务分工和各阶段的衔接。编者结合自身的专利分析工作经验，参考借鉴有关书籍的结论，大致总结归纳出如下专利分析的总体流程，如表 1-1。

表 1-1 专利分析基本流程

专利分析阶段	主要工作
前期准备阶段	开展前期行业和技术调研，收集项目相关技术、经济、法律等信息资料，与技术主体沟通，确定技术分解表
数据取样和预处理阶段	选择检索策略和分析工具，获取项目分析数据所需的数据 对取样数据进行数据探索、清洗、合并、集成、规范处理 对预处理后的数据结果进行标引
数据可视化阶段	针对处理好的数据，结合项目需求，选择合适的可视化形式进行数据呈现，制作可视化图表
数据挖掘与建模阶段	根据项目任务，结合相应的数据挖掘算法，选择相应的挖掘工具开展数据挖掘与建模工作，利用处理好的数据进行信息挖掘
报告撰写阶段	整合数据处理、数据可视化、数据挖掘与建模阶段的成果，解读分析图表，对挖掘的信息进行分析，撰写相应的文字性报告

根据上述专利分析基本流程，我们可以绘制如下专利分析工作流程图，如图 1-2 所示。



图 1-2 专利分析基本流程图

1. 前期准备阶段

前期准备阶段的主要工作包括成立项目组，并根据项目需求开展行业和技术调研，收集与分析项目相关的技术、经济、法律等信息资料，通过与企业、行业技术专家等进行沟通，形成初步的技术分解表。

该阶段要尽量全面掌握反映行业需求的技术、经济、法律等各类信息，为后期正确解读专利图表和数据信息提供必要的背景信息，避免出现分析报告的结论与技术发展的实际情况脱节或背离的现象。

开展必要的行业和技术调研的另一重要目的在于获取准确、清晰的技术主题分解表。要认识到技术主题分解表在前期准备阶段未必能完全确定下来，需要分析人员在充分听取行业和技术专家意见的基础上，结合专利分析的特点，不断调整，制定出符合技术发展实际、技术边界清晰、满足专利检索和分析要求的技术主题分解表。

2. 数据取样和预处理阶段

该阶段是通过检索确定需要分析的样本数据，并利用数据处理的基本手段对采集的样本数据进行处理，包括数据探索、数据清洗、数据合并等数据操纵的基本手段。

由于采集的专利数据中含有很多不规范，数据缺失等情况，因此，只有通过数据预处理，将不规范的专利数据整理、清洗成格式规范、信息完整的数据，才可以供后续数据可视化和数据挖掘阶段使用。这一阶段是专利分析整个流程中最为重要和最为基础的阶段，也是占据整个专利分析流程时间比例最大的阶段，需要分析人员熟练掌握数据处理的基本技能，综合运用数据处理的各种技巧，高效、准确地完成数据预处理工作。

3. 数据可视化阶段

在数据可视化阶段，专利分析人员需要结合项目分析内容，选择恰当的可视化图表展示方案，将纷繁复杂的专利数据用清晰简洁、生动准确的图表反映出来。

从专利数据的属性角度来看，专利分析的内容可以分为如下几类：专利

态势分析、专利技术分析、申请主体分析。基于编者的专利分析项目经验，我们对专利分析中所用到的数据可视化图表进行了简单归纳总结如下。

(1) 专利态势分析

专利态势分析是指针对某一行业或技术领域进行总体专利状况的分析，从而快速了解整个产业、技术、地域的发展趋势。专利态势分析通常是对专利数量进行统计，诸如申请趋势分析、技术构成分析、地域申请量分析以及申请人排名分析，对应这些分析内容的主要图表有折线图、面积图、饼图、柱状图、矩形树图、地图等。示例如下：

图1-3为专利态势分析图示例，其中左图为采用矩形树图进行技术分支申请量展示，右图为采用折线图进行技术分支年份申请量趋势分析。

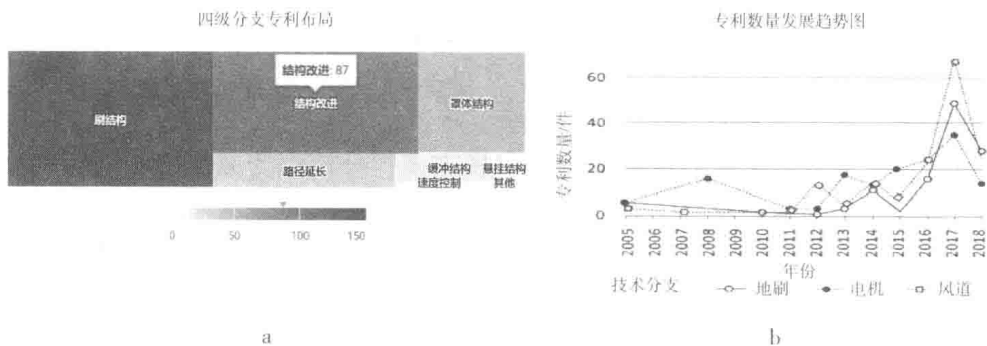


图1-3 专利态势分析示例图

(2) 专利技术分析

专利技术分析是以专利技术内容为分析对象，一般可以分为技术功效矩阵分析、技术路线分析、重点产品专利分析、重点技术引证分析。该分析内容常用的分析图表有气泡图、泳道图、实物分析图、引证网络图等。示例如下：

图1-4为专利技术分析示例图，左图是五局技术合作流向图，用于反映技术国别之间的技术流向；右图为技术分支跑道图，用于对比分析不同技术主题的占比。

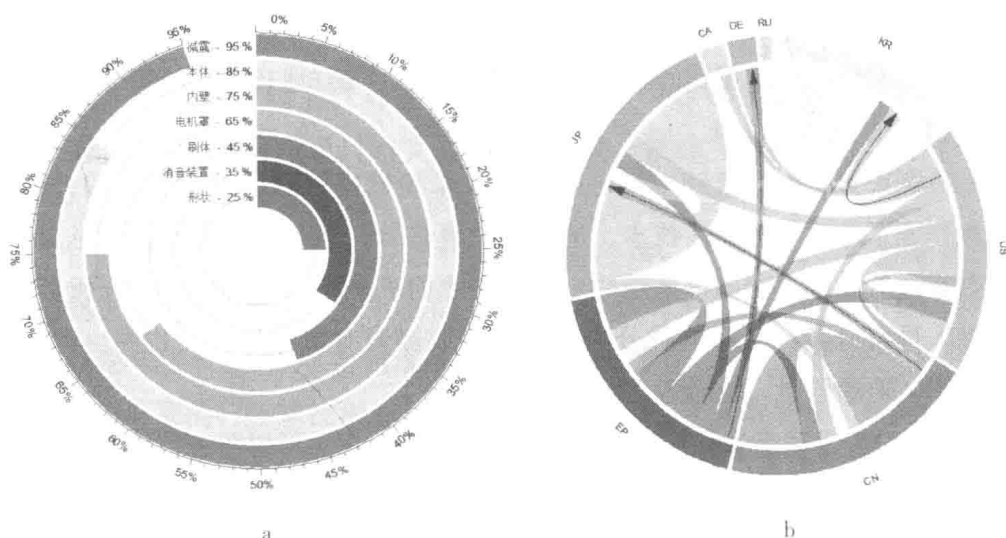


图 1-4 专利技术分析示例图

(3) 申请主体分析

申请主体分析主要包括确定重要市场主体、分析市场主体的专利区域分布、重点技术、重要产品、研发团队、重要发明人、专利合作关系以及企业并购和专利诉讼等。该分析内容所涉及的分析图表主要有弦图、力导图、矩阵图、雷达图、南丁格尔玫瑰图等。

图 1-5 为申请主体分析示例图，左图 of 申请主体合作关系力导图，用于分析重要申请人合作关系；右图 of 申请人合作关系热力图，用于反映不同申请之间合作关系的紧密程度。

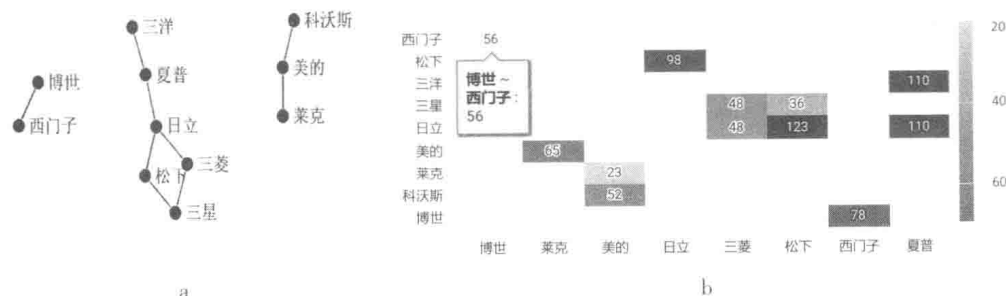


图 1-5 申请主体分析示例图