开源工具支持的专利数据清洗流程研究*

钟 华 李艳梅 安新颖

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 分析专利数据清洗需求，提出专利数据清洗步骤和框架，包括数据导入、规范、字段拆分、机构清洗、数据标引等环节，对可利用的开源工具进行对比分析并以 OpenRefine 为例开展实践研究。

〔关键词〕 专利分析；开源工具；数据清洗

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.05.014

Study on Patent Data Cleaning Process Supported by Open Source Tools *ZHONG Hua, LI Yanmei, AN Xinying, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper analyzes patent data cleaning requirements, proposes patent data cleaning procedures and frameworks including data input, standards, field split, organization cleaning, data indexing and other sections, carries out contrastive analysis on available open source tools and makes a practical study by using OpenRefine as an example.

〔**Keywords**〕 patent analysis; open source tools; data cleaning

1 引言

专利分析是从专利文献中采集零碎、分散、隐藏的信息，通过科学方法对专利信息进行加工整合，利用科学计量和统计的方法将这些信息转化为有价值情报的过程，是支持学科技术发展追踪和产品战略决策参考的重要方法和工具。通过专利分析能够获得可信、客观的技术信息，进行技术追踪和预警，分析技术创新热点和空白点，是了解技术变革和前沿发展的领先指标。目前常用的专利分析工具如德温特数据分析软件（Derwent Data Analyzer，TDA）、德温特创新平台（Derwent Innovation，DI）、Delphion 等商业软件和平台虽然具有较强的分析处理功能，但是其较高的收费限制应用实践。开源工具虽然在稳定性和安全性方面存在不足，但其免费、开放、便捷等特点也使其成为一类普遍应用的数据处理工具。本文对开源工具支持的专利数据清洗流程进行分析，能够为短期实践、教学培训、研究实习等场景下流畅高效地进行专利数据处理操作提供参考。

〔修回日期〕 2018－12－07

〔作者简介〕 钟华，助理研究员，发表论文 10 余篇。

〔基金项目〕 中国医学科学院中央级公益性科研院所基本科研业务费“科技创新环境下医学科研机构科技成果转化能力评价研究”（项目编号：2017PT63004）；中国医学科学院医学与健康科技创新工程“医学科技创新评价与卫生服务体系构建研究”（项目编号：2016－I2M－3－018）。

2 专利数据清洗需求

2.1 数据清洗的必要性

专利分析流程包括确定分析目标、检索下载数据、加工处理数据、统计分析总结、撰写分析报告

等步骤，数据清洗是专利分析工作的重要步骤[1]。专利数据清洗指对从专利数据库中检索并下载的数据进行规范，包括格式规范、去重合并、字段拆分、提取删除等操作，从而保证清洗后的数据准确、完整、规范，以便于后续数据批量处理和分析。从专利数据库下载的专利记录字段包括专利号、专利名称、发明人、专利权人、申请时间、技术分类、引证信息、摘要等多项信息。从数据库中以纯文本格式或制表符分隔格式下载的专利记录普遍在各字段中存在如大小写、冗余、空格等各种程度的问题[2]，不能直接进行分析处理，需要利用数据处理工具并人工介入下进行数据清洗。一般而言专利数据清洗需求主要有数据规范和拆分抽取需求两种。

2.2 具体需求

一是数据规范需求。即对专利数据采集导入过程中由于人为或系统因素而造成的无用或不规范数据进行去除、修正和规范，如对记录的机构名称规范合并、时间日期格式统一、名字全称简写核对、中英文拼写错误修正以及空值、异常值等数据规范问题进行统一处理，形成规范的数据集合。二是拆分抽取需求。拆分抽取处理有一次和二次拆分两种情况。一次拆分是指一条字段内容中包含多项同类信息，如发明人字段中有多个发明人姓名，需将这条记录项拆分成与其对应的多条记录。二次拆分指专利记录的一个字段中包含多项不同类型的信息，需要进行二次细化拆分，如优先权字段中包括优先权国和优先权日，数据清洗时需要进一步拆分成两个字段。

3 专利数据清洗框架

3.1 概述

数据清洗是保证专利分析科学准确的前提。在针对专利数据清洗的相关研究中路霞[3]等针对专利地址信息相关的中文专利数据建立清洗框架，提出算法，利用对照法对该框架进行验证优化。在专利数据清洗角度，翟东升[4]等以文本形式的专利信息为数据源，在对各字段内容进行分别抽取的基础上综合运用表达式清洗策略、循环清洗策略和基于正则表达式的脚本清洗策略对各字段进行清洗转换。王永红[5]提出的专利数据清洗步骤包括：选择数据来源、限定数据范围、生成样本空间、数据规范、字段拆分以及数据标引。根据研究经验和实践总结，本研究总结出专利数据清洗框架，见图 1。

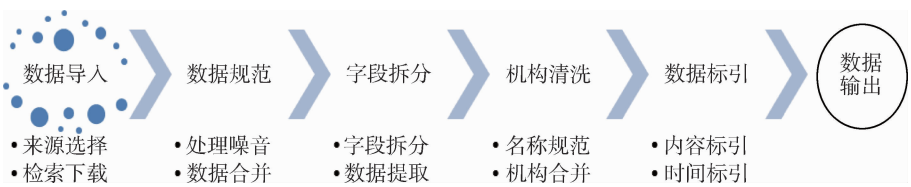

图 1 专利数据清洗框架

3.2 数据导入

是指选择专利分析指标数据，下载并导入数据处理工具的过程，正确、完整、可靠的数据导入是专利分析工作的前提。在数据导入阶段，结合唯一性、完整性、一致性原则，应注意保持不同数据表的专利申请号作为唯一标识的准确性，还需要统计不同数据表的字段格式，特别是日期和分隔符号格式、姓名写法等，以便于数据调用。此外需要人工核对和补充缺漏的数据项，保证数据完整准确。

3.3 数据规范

由于来源于不同时间、专利申请主体、申请国、代理机构及专利数据库数据存在外部特征及内容方面的不一致性或错误，导入后一般会存在分隔符不统一、数据格式不一致、一词多形等不同程度的数据问题，如未进行规范而直接进行统计分析会产生一定的误差，影响统计结果。因此需要根据存在问题的类型和规律制定处理规则，对发明人、专利权人、专利申请号等字段进行规范化。数据规范

的内容包括统一大小写或全角半角、删除前置空格、修订内容性乱码、错行、文字性错误等。

3.4 字段拆分

在进行专利数据统计前需将包含多个统计项的字段进行拆分处理，如在统一各项日期格式的基础上可对优先权日字段中的国别代码和日期属性进行拆分提取，得到优先权国和年份信息。

3.5 机构清洗

需将机构官方名称、缩写名、别名、变更名及其直属、附属机构名称纳入集合中，通过数据清洗、抽取、切分、合并、去重和人工处理对每个机构名称及其直属、附属机构名称进行规范，提取机构别名关系、名称变更关系等，便于后续对机构的完整统计。

3.6 数据标引

标引是数据清洗的重要环节，通过标引赋予专利以检索和分类标识，标明其外部特征和内容特征的类属，数据标引质量直接关系到后续各类统计分析的准确性，需结合分析需求明确需要标引数据的属性值。一般而言，专利数据的主要标引项包括申请日、优先权日、发明人、专利权人、国家地区、IPC 国际专利分类号、其他分类号、同族专利信息、被引频次等[6]。由于数据标引工作量大，以技术内容分类标引为例，首先需要明确学科领域下的技术分类划分规则，明确各分支下专利文献所包括技术主题的内涵和外延，对数据集合内的专利文献赋予最合适和准确的技术分类号，保证后续技术内容统计的准确性。此外在必要时可通过人工复审和交叉标引进一步提高正确率。

4 开源数据清洗工具比较

近年来开发了较多适用于专利数据清洗的开源或免费工具，常用工具有 Trifacta Wrangler、Talend、OpenRefine、DataCleaner 等，各款软件功能及优缺点分析，见表 1，利用这些工具能够更快、更简单、更准确地进行专利数据清洗。其中 OpenRefine 是在数据清洗、探索、转化方面非常有效的工具。它是一个开源的网络应用，具备数据清洗和批量标引功能，可在专利数据清洗这一过程中实现对不同来源数据的归一化处理。

表 1 开源/免费的专利数据清洗工具比较

名称	导入数据格式	数据清洗功能	缺点
Trifacta Wrangler	支持 Excel、JSON 和原始的 CSV 文件的导入导出	数据操作基于列进行半自动化数据整理并确定结构，可分析数据存在值缺失、不匹配或不一致的情况，按类型对数据进行直观分类	需要手动识别数据中的错误和问题
OpenRefine	支持 CSV、XML、XHTML、JSON 等多种数据格式的导入导出	数据操作基于行、列或者单元格。通过删除重复项、空白字段和其他错误来清理数据。记录数据处理的操作，可随时进行浏览和撤销	需具备一定数据处理基础
DataCleaner	只支持 Comma - separated CSV 文件导入	数据操作基于列，可根据智能算法自动决定数据清洗的方法，将半结构化数据集转换为数据可视化工具可读取数据集	不是针对所有数据类型的工具，主要针对科学数据
Microsoft Excel	支持 XML、CSV、TXT 等多来源的数据导入	提供一系列函数，如删除重复、查找替换以及拼音检查，此外提供消除重复、查找、替换、拼写检查以及用于转换数据的许多公式	不适用于大数据集

5 嵌入专利数据清洗流程的应用方案

5.1 概述

本文以肿瘤干细胞（Tumor Stem Cells，TSCs）领域专利数据清洗实践为例，说明利用开源工具 OpenRefine 进行专利数据清洗的实现过程。使用 OpenRefine 的原因是它是一个简单、有效的免费工具，与 Excel 或 OpenOffice 相比在执行相同的清理数据任务时性能更加优越，使用更有效率。OpenRefine 可以制定和开发有效的数据工作流来清理和重新整理数据，创建和利用特定任务所需的自定义代码。因篇

幅所限，本文以关键步骤的数据清洗为例进行说明。通过德温特（Derwent）专利数据库检索得到 2007 – 2016 年肿瘤干细胞领域相关专利 866 个，记录内容选择全记录，数据以制表符分隔格式导出。

5.2 数据导入

利用 OpenRefine 创建项目操作简单，通过点击创建项目标签页、选择数据集、点击下一步来创建新项目，完成文件导入。

5.3 数据规范

是处理专利数据的第 1 步。关键步骤包括：规范化字符（如小写、大写）；删除前导和尾随空格；地址编码和相关问题；转换日期；添加备注信息并创建新的列和/或行。导入数据集中的第 1 列是专利申请号（PN），以对此列数据进行规范为例，首先选择列菜单并从 Text Facet 下拉列表中进行选择，生成包含统计数据的侧面菜单面板，然后检查该列问题，当向下滚动侧面板时可以看到一些申请编号具有小写的国家代码。为解决问题，选择列菜单 Edit Cells > Common Transforms > To Uppercase。此操作可以使所有该专利号被转换为大写，这一步骤的数据规范将使后续提取国家代码的步骤更方便，见图 2。

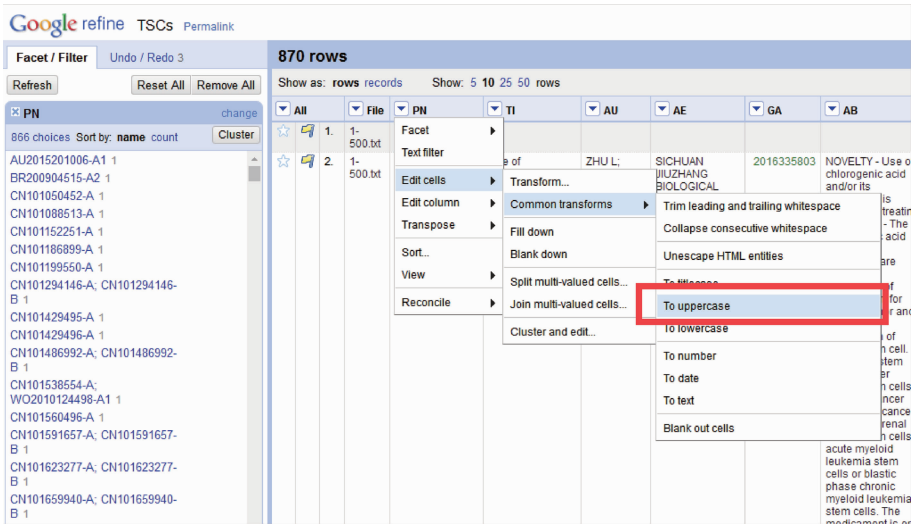

图 2 数据规范

5.4 字段拆分

利用 OpenRefine 可以方便地进行行列处理，如对包含多值字段分行分列、多值字段去重处理、单值字段分列处理、行列倒置合并等，可根据分析需要灵活转换数据。在处理专利数据时 OpenRefine 的优点之一是可以简便地将专利权人分成单独行，选择 Edit Cells，然后拆分多值单元格。在弹出菜单中选择“;”作为分隔符，此时数据集有 2 578 行并且所有专利权人都在一列，但是其余数据尚未复制到新行，之后选择 PN > Edit Cells > Fill Down。数据集中以每条专利的申请号作为关键标识。需要注意的是使用 Fill down 功能是数据填充的作用，谨慎使用以防数据变得混乱，这也是在开始数据处理工作时首先进行数据规范的重要原因。

5.5 机构清洗

未经清洗的专利数据普遍存在机构命名不规范现象，机构合并、改名、上下属机构、分支机构、缩写简写等各种问题造成机构名称的多样性，如不进行机构名称规范会导致针对专利申请人的统计分析不够准确。OpenRefine 在机构规范方面提供关键词碰撞和邻近取样两种方法，通过对机构名称自动聚类处理，可为分析人员提供机构名称聚类结果作为机构清洗的参考。再经过机构名称自动聚类处理后，如还存在不规范字段，可以通过人工识别和调

整，以及对机构进行重命名的个别特殊处理得到最终数据清洗结果。在肿瘤干细胞专利数据处理中，从专利权人“（AE）”列中可以拆分出1 276个项目，然后将相似的值进行聚类分析，“聚类（Cluster）”选项便于系统对相似名称的机构进行聚类，定义新的单元格值（New Cell Value），然后点击 Merge Selected & Re－cluster，逐个机构进行归并，完成机构清洗步骤。

5.6 数据标引

是拟定量分析的基础。数据标引通常可用人工判读标引和机器辅助标引。在处理专利数据时可利用 OpenRefine 的文本过滤功能进行机器辅助标引，减少人工标引的工作量[7]。以标引专利的国家信息为例，可选择优先权申请信息和日期（PI 列），在列中的数据中隐藏一系列信息，可以使用简单的代码对 PI 列的信息进行处理，根据返回的值创建一个新列。如可用 substring 函数提取国家代码，输入 substring（value，0，2），即代码从 0 开始计数（如 0，1 = U，2 = S）。代码的第 1 部分在值字段中查找。0 表示代码从 0 开始计数，2 表示从 0 读取两个字符，提取国家代码，实现对专利所属国家信息的标引，见图 3。

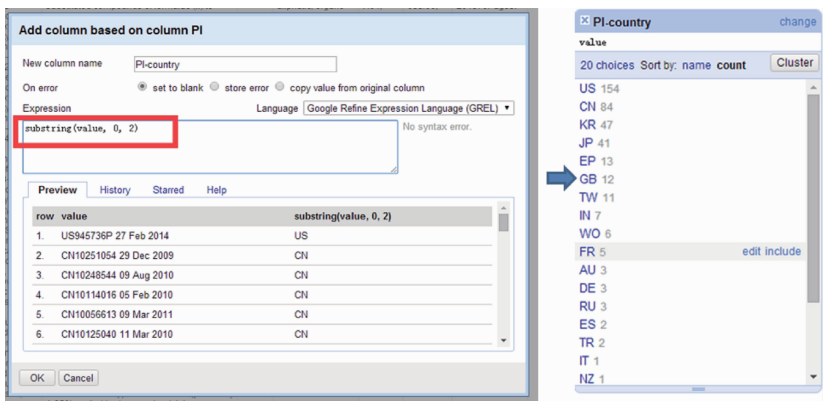

图 3 国家标识的数据标引

6 结语

数据清洗是专利分析的重要环节，其任务繁琐、耗时长，需要大量细致认真的工作。本文对开源工具支持的专利数据清洗流程开展研究，对比常用的开源数据清洗工具，提出有针对性的清洗策略和步骤，为高质量的专利分析数据集合的形成提供可参考的应用示范。实践证明利用 OpenRefine 等开源工具可以完成对专利数据的清洗、标注与规范化存储功能，为专利分析前期的清洗工作提供便捷的处理手段，生成更规范、准确、清晰的分析数据集合。

参考文献

1 左良军. 专利分析中样本选取与数据清洗环节的探究［J］. 中国发明与专利，2016（9）：69－71.

2 刘喜文，郑昌兴，王文龙，等. 构建数据仓库过程中的数据清洗研究［J］. 图书与情报，2013（5）：22－28.

3 路霞，吴鹏，王曰芬，等. 中文专利数据地址信息清洗框架及实现［J］. 情报理论与实践，2016，39（4）：128－132.

4 翟东升，李倩，张杰，等. 德温特专利信息清洗与标注模型研究［J］. 情报杂志，2013，32（8）：150－154，203.

5 王永红. 定量专利分析的样本选取与数据清洗［J］. 情报理论与实践，2007，30（1）：93－96.

6 王丽，张冬荣，张晓辉，等. 利用主题自动标引生成技术功效矩阵［J］. 现代图书情报技术，2013，29（5）：80－86.

7 王丽. 开源/免费工具比较及专利分析全流程解决方案研究［J］. 情报理论与实践，2016，39（1）：118－122.