

多源专利信息资源集成研究 ——以美国专利数据为例*

刘会景^{1,2}, 凌锋², 朱新超², 齐萍², 杨冠灿¹

(1. 中国科学技术信息研究所, 北京 100038; 2. 北京万方数据股份有限公司, 北京 100038)

摘要: 从专利情报分析的一致性、直接性和完整性需求出发, 对比分析来源于欧洲专利局和美国专利商标局的美国专利题录信息属性, 制定异构数据资源集成的加工流程, 并加以实现, 最终构建了数据规范和信息全面的美国专利数据库。

关键词: 异构数据; 美国专利; 数据集成; EPO; USPTO

中图分类号: G306; G250.73

DOI: 10.3772/j.issn.1673-2286.2015.09.004

在科学技术迅猛发展和经济全球化的时代背景下, 各个国家和企业越来越重视新技术的开发和利用。作为一种技术、经济、法律信息资源, 专利数据被广泛应用于技术创新、技术变革与技术管理研究中。专利信息情报不仅可用于了解竞争对手的研发状况, 还可以协助国家、科研机构、创新型企业确定研发战略、避免重复研究等^[1-2]。因此, 如何有效的发挥专利情报的作用, 成为当前研究的热点。

专利情报分析是指对来自专利说明书、专利公报中大量的、个别的专利信息进行加工及组合, 并利用统计方法或技术手段使这些信息具有纵览全局及预测的功能, 并且通过各种分析手段, 将原始的专利信息由普通的信息上升为企业经营活动中有价值的情报。在分析过程中, 主要利用专利申请信息, 如发明人、申请人、优先权、引证信息、专利家族、专利分类和专利摘要等^[3-4]。

专利情报的方法体系包括数据获取、数据预处理、数据集成、数据分析、信息情报五个层次^[5]。其中, 数据

集成 (data integration), 或称为数据整合, 是一种资源整合的理念和方式, 源于20世纪90年代初, 伴随信息系统集成的兴起而逐步引起重视。所谓数据集成, 是对异构资源系统中的异质、异类的的数据在逻辑上或物理上有机地集中, 提供统一的表示和操作, 以解决多种异构数据资源的互联与共享^[6]。

对于专利数据来说, 不同来源的专利数据缺乏统一的组织机制, 相同信息的表达方式不一致, 不能保证数据的完整性、可靠性, 难以支持数据检索、主题聚类、内容挖掘等专利情报分析的需求, 所以专利数据集成是亟需解决的问题。这一问题成为当前学术界研究的重点, 相关科研人员试图从不同角度进行尝试和探索, 提高信息资源的利用程度^[7-10]。

本文对比了不同来源的美国专利数据的属性特征, 分析了异构数据集成的可能性, 并制定了美国专利数据集成的处理逻辑, 在保持原有数据信息特色的基础上, 提供更为全面和规范的美国专利信息资源, 满足专利情报分析需求。

* 本研究得到科技部国家科技支撑计划“面向科技创新的专利信息加工与服务关键技术研究与应用示范” (编号: 2013BAH21B01) 资助。

1 专利数据资源集成需求

目前, 欧洲专利局 (European Patent Office, EPO) 收集了包含美国、中国、德国、日本在内的93个国家、地区或知识产权组织的专利数据资源^[11-12], 是目前世界上收集专利数据最多的专利组织。在数据应用过程中, EPO对异构异质的专利数据资源进行了加工、清洗和规范。

本文将来源于欧洲专利局和美国专利商标局的美国专利著录项信息进行资源集成, 以期满足如下数据集成的需求:

(1) 数据一致性。在集成不同来源、不同数据组织形式的专利信息资源的过程中, 建立统一的数据规范, 并以关系型数据库的形式予以组织和保存, 构建统一资源ISTIC美国专利数据库, 方便用户检索、分析美国专利数据资源。

(2) 数据获取的直接性。经过集成后的信息集成系统, 应是各异构数据源的有机集成。对一个查询请求, 可以直接获取全面的美国专利信息, 不需分别访问不同的资源系统。

(3) 数据完整性。从信息集成系统中获取的数据信息, 保留了不同来源的数据特色, 从不同角度揭示了美国专利信息资源的特征属性, 用户可最大限度地挖掘美国专利数据潜在的价值信息。

2 专利数据资源集成方法研究

专利信息资源集成使用的数据是欧洲专利局和美国专利商标局提供的美国专利题录数据。欧洲专利局提供的美国专利数据资源包括2001年至今的专利申请公开数据 (Published Applications) 和1790年至今的核准公布数据 (Issued Patent); 美国专利商标局提供的美国专利数据资源包括2001年至今的专利申请公开数据和1796年至今的核准公布数据, 1790年-1975年的数据只有图像全文说明书, USPTO仅提供检索服务^[13]。

考虑到美国专利商标局不提供1790年-1975年的专利数据下载, 因此本文试图集成1976年-2014年12月期间的核准公布数据, 以及2001年-2014年12月期间的专利申请公开数据。数据集成的思路是: 首先通过关联字段匹配两个来源的美国专利数据, 以保证数据的一致性; 经过宏观和微观的数据对比分析, 确定数据集成方案, 以保障数据获取的完整性; 最后通过重新组织和

数据存储, 实现数据获取的直接性。

2.1 关联字段的选择

美国专利申请号是循环编号, 每轮循环均从000001-999999以内连续编排, 周而复始, 为了区分不同循环周期的申请号, 美国专利商标局引入了申请号系列码 (Application Number Series Code)。如图1中所示的专利 (Utility Patent) 申请于2006年, 申请号编码进入了第11个循环周期, 因此USPTO提供的专利申请号是11/594928, EPO对美国专利申请号码进行了规范化处理, 申请号码包括申请年份流水号信息, 所以对于同一件专利, EPO经过规范加工后提供的申请号码是20060594928, 因此对于相同的专利, USPTO和EPO提供的申请号的表达方式不同。

文献号是各工业产权局在公布专利文献时编制的序号。专利申请一经受理, 随后便依审查制度和审批程序可以一次公布或多次公布, 从而导致一件申请有一个申请号和有多个文献号 (如申请公开号和核准公告号)。对于美国专利而言, 来源于EPO和USPTO的专利文献号基本一致, 因此本文使用专利文献号对不同来源的美国专利数据进行关联, 具体关联方式及示例如图1所示。

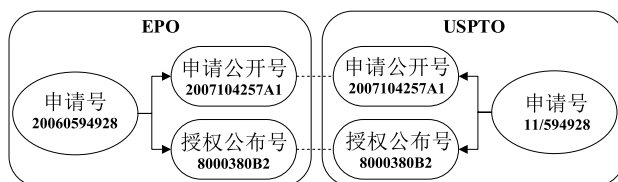


图1 两种来源美国专利数据关联示意图

通过使用专利文献号作为两个来源数据的关联字段, 去匹配、对接、筛选、规范欧洲专利局和美国专利商标局提供的美国专利信息资源, 并以专利申请号为单位组织专利信息的存储, 构建ISTIC美国专利数据库。

2.2 宏观数据对比

欧洲专利局获取各国家、地区、组织的专利数据资源后, 为了规范、统一的专利信息资源, 会对不同国家的专利著录项信息予以取舍和清洗。美国专利商标局提供的美国专利数据是最原始和官方的信息资源, 因此两者各有优势, 其信息可以相互补充, 表1是EPO和

USPTO提供的美国专利数据著录项信息对比表。

表1 两种来源美国专利题录信息宏观对比表

	EPO	USPTO	
		申请公开	授权公布
标题信息	✓	✓	✓
摘要信息	✓	✓	✓
国际专利分类信息	✓	✓	✓
欧洲专利分类信息	✓		
联合专利分类信息	✓		
美国专利分类信息	✓	✓	✓
洛迦诺分类信息			✓
植物专利分类信息			✓
引证信息	✓		✓
发明人信息	✓	✓	✓
申请人/专利权人信息	✓	✓	✓
代理机构/代理人信息		✓	✓
审查员信息			✓
优先权信息	✓	✓	✓
简单专利家族信息	✓		
分案、续案信息		✓	✓
相关PCT专利信息			✓

集成上述两种来源的美国专利数据，就是针对来源于EPO和USPTO的专利数据特点，为共有的信息和特有的信息建立不同的加工标准。对于共有的信息，如标题、摘要、国际专利分类信息等，通过对比两种来源数据的准确性和合理性，优先选取更为准确、全面的数据；对于特有的信息，如审查员信息、代理机构/代理人信息、欧洲专利分类信息、美国专利分类信息等，有选择地予以保留。根据上述规则，共有信息和特有信息经过进一步的数据清洗与规范，以关系型的数据库形式存储，构建信息全面、数据规范的ISTIC美国专利数据库。

2.3 微观数据对比

2.3.1 标题、摘要信息

一般情况下，欧洲专利局和美国专利商标局提供的专利标题和摘要是一致的。考虑到授权公布的专利经过了专利审查员的审查和申请人的修改，更为准确，

因此集成数据库中的专利标题和摘要优先选取授权公布的专利文献信息，即以最新公布文献中的专利标题和摘要信息为准。

2.3.2 发明人、申请人/专利权人信息

欧洲专利局对发明人、申请人/专利权人数据进行了统一的规范化处理，包括标点符号、英文字母大写转换等；相较而言，美国专利商标局的发明人、申请人/专利权人的信息更接近专利原文，且地址信息更为全面，尤其是对于申请人/专利权人居住地为美国，地址信息精确到“州”。以专利7594121（专利授权公布号）为例，两种来源提供的专利权人信息对比情况如表2所示。

表2 两个来源的申请人/专利权人信息示例

来源	序号	专利权人名称	国别	州	地址
USPTO	1	Sony Corporation	JP		Tokyo
	2	Sony Electronics Inc.	US	NJ	Park Ridge
EPO	1	SONY CORP	JP		
	2	SONY ELECTRONICS INC	US		

在专利信息分析过程中，规范的发明人姓名、申请人/专利权人名称以及详细的地址信息对于竞争对手的战略分析和地域分析都非常重要，因此，我们保留了EPO提供的发明人姓名、申请人/专利权人名称，以及USPTO提供的地址信息。如表2中的示例经过数据集成后的结果如表3所示。

表3 经过数据集成后的信息示例

序号	专利权人名称	国别	州	地址
1	SONY CORP	JP		Tokyo
2	SONY ELECTRONICS INC	US	NJ	Park Ridge

2.3.3 国际专利分类信息

国际专利分类（International Patent Classification, IPC）是目前世界上唯一通用的专利文献分类体系，对于海量专利文献的组织、管理和检索具有极其重要的作用^[14]。但是由于美国专利采用的是美国专利分类（U.S. Patent Classification, USPC）体系，USPTO只是用计算机系统将USPC分类号转换成IPC分类号，标注在专利文献的扉页，并且不进行任何后续的回溯和完

善; 而欧洲专利局自第8版IPC 推广使用以来, 按照第8版的分类标准, 陆续对1-7版的IPC 信息进行追溯和修正^[15]。因此就IPC信息而言, 欧洲专利局提供的IPC 信息更加准确。

2.3.4 其它专利分类信息

美国专利商标局一直沿用美国专利分类体系, 欧洲专利局之前同时使用IPC分类体系和欧洲专利分类体系(European Classification, ECLA)。2010年10月, 美国专利商标局和欧洲专利局宣布合作开发联合专利分类(Cooperative Patent Classification, CPC), 美国专利商标局承诺将在CPC实施之后放弃其近200年历史的USPC体系, 欧洲专利局也已经停止使用ECLA体系^[16]。目前上述两个机构均已开始推广联合专利分类体系, 并分别提供了CPC检索入口。因此, ISTIC美国专利数据库未保留USPC和ECLA分类信息, 而是选择了CPC分类信息。

2.3.5 优先权信息

欧洲专利局规范了专利的优先权信息, 尤其是优先专利申请号码, 并标引了优先专利申请与本专利申请的关系类型, 包括分案申请、续案申请、部分续案申请等^[12]。欧洲专利局在规范后的优先权信息的基础上构建了简单专利家族数据库。因此, 欧洲专利局提供的优先权信息更加准确。

2.3.6 其它信息

对于两个来源的独有信息(除欧洲专利分类信息和美国专利分类信息), 如欧洲专利局提供的简单专利家族信息, 美国专利商标局提供的审查员信息、代理人/代理机构信息、分案/续案信息等, 均予以保留并规范。

3 美国专利信息资源集成及实现

利用专利文献号, 对欧洲专利局和美国专利商标局提供的美国专利信息资源进行集成。数据集成流程如图2所示。

(1) 数据预处理, 主要是对EPO和USPTO提供的

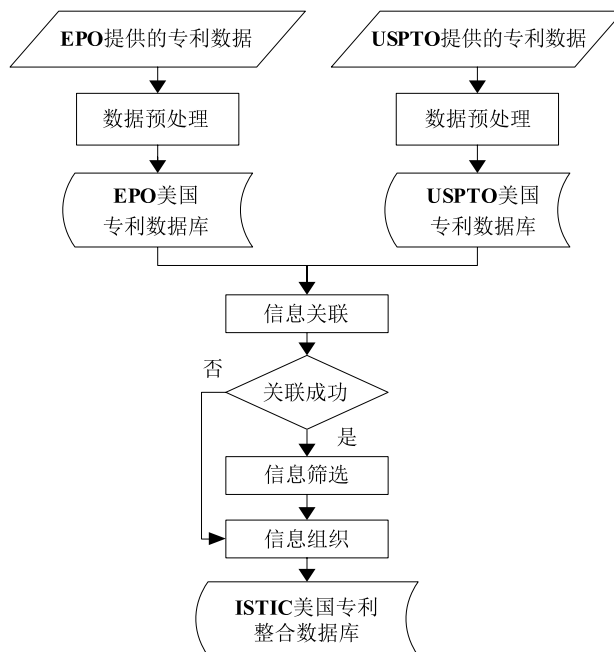


图2 美国专利信息资源集成流程图

美国专利原始数据进行初步的数据清洗, 同时将其转化成具有标准格式并能够存储于数据库的数据, 为数据资源集成提供基础数据。

(2) 信息关联, 根据同一件专利申请号相同, 文献号不同的原则, 确定同一件专利的多次公开/公布记录; 通过专利的文献号关联两种来源的数据, 实现来源于EPO和USPTO的美国专利数据的匹配。

(3) 信息筛选, 经过不同来源数据的宏观和微观比对, 确定数据集成加工方案, 以保障数据的完整性。

(4) 信息组织, 根据预先制定的数据集成加工方案, 对数据进行统一的清洗与规范, 并予以组织和存储, 构建ISTIC美国专利数据库。

ISTIC美国专利数据库以专利申请(Application)为核心组织信息资源, 包括多次公开/公布记录信息

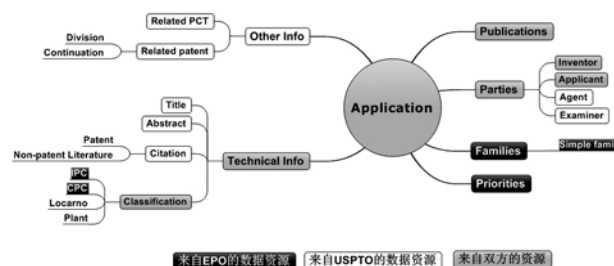


图3 ISTIC美国专利数据库概念图

(Publications)、技术信息(Technical Info)、相关人员信息(Parties)、优先权信息(Priorities)、简单专利家族信息(Families),以及其他相关信息(Other Info),其概念图如图3所示。

4 结论与不足

本文对比了来源于欧洲专利局和美国专利商标局的专利数据属性特征,对美国专利信息进行异构集成,实现了数据资源的互补,最终构建了ISTIC美国专利数据库,保证了美国专利数据的完整性和全面性,为专利情报分析提供数据支撑。本文对美国数据资源的集成仅限于著录项信息的集成,未涉及到法律信息的整合,而专利法律信息的分析是专利情报分析的重要组成部分。专利法律状态信息资源的集成以及其他国家专利数据资源的集成探索将是下一步的研究重点。

参考文献

- [1] 张巍,汪雪锋,郭颖,朱东华. 中国专利数据检索策略研究[J]. 科学学研究, 2011, 29(6): 833-839.
- [2] 赖院根,朱东华,刘玉琴. 专利申请信息与法律状态信息的集成分析研究[J]. 科学学研究, 2007, 25(6): 1083-1090.
- [3] 于潇,孙英隽. 专利情报分析对企业创新的影响研究[J]. 情报科学, 2007, 25(11): 1668-1671.
- [4] 刘红光,吕义超. 专利情报分析在特定竞争对手分析中的应用[J]. 情报杂志, 2010, 29(7): 35-39.
- [5] 黄圆圆,朱东华,任智军,张诚. 对比分析方法在专利情报分析中的应用研究[J]. 现代图书情报技术, 2006(10): 60-65.
- [6] 马文峰,杜小勇. 基于数据的资源集成[J]. 情报资料工作, 2007(1): 41-45.
- [7] 翟东升. 专利知识挖掘关键技术研究[M]. 北京:知识产权出版社, 2013: 9-10.
- [8] 强韶华. 不同语种专利信息资源互操作设计[J]. 情报科学, 2011, 29(3): 446-450.
- [9] 刘婕. 世界林业专利信息资源的集成与利用[J]. 世界林业研究, 2007, 20(3): 70-73.
- [10] 朱新超,霍翠婷,刘会景. 不同来源的中国专利信息资源集成研究[J]. 情报理论与实践, 2014, 37(11): 90-93.
- [11] Information on EPO data [EB/OL]. [2015-06-14]. <http://www.epo.org/searching/data/data/tables/weekly.html>.
- [12] EXCHANGE FORMAT EPO - Patent Information Resource [EB/OL]. [2015-06-14]. [http://documents.epo.org/projects/babylon/eponet.nsf/0/75d409ce227914c9c1257b8d003edda/\\$FILE/st36_user_documentation_v_2_5_1.pdf](http://documents.epo.org/projects/babylon/eponet.nsf/0/75d409ce227914c9c1257b8d003edda/$FILE/st36_user_documentation_v_2_5_1.pdf).
- [13] Electronic Bulk Data Products [EB/OL]. [2015-06-14]. <http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products>.
- [14] 陈立新,梁立明. 技术领域的集成与集成研究——基于美国专利IPC的关联分析[J]. 情报杂志, 2013, 32(1): 37-41, 72.
- [15] The IPC and EPO products [EB/OL]. [2015-06-14]. <http://www.epo.org/searching/essentials/classification/ipc-reform/impact.html>.
- [16] 朱新超,霍翠婷,刘会景. 合作专利分类系统(CPC)与传统专利分类系统的比较分析[J]. 数字图书馆论坛, 2013(9): 38-44.

作者简介

刘会景,女,1984年生,硕士,万方数据技术研究院,研究方向:专利数据挖掘、信息资源服务, E-mail: huijingliu@live.cn。
凌锋,男,1977年生,万方数据技术研究院,研究方向:知识组织与发现、数据挖掘。
朱新超,男,1990年生,万方数据技术研究院,研究方向:专利数据加工、信息资源服务。
齐萍,女,1986年生,万方数据技术研究院,研究方向:专利数据挖掘、专利分析算法设计。
杨冠灿,男,1981年生,博士,中国科学技术信息研究所助理研究员,研究方向:专利分析、技术竞争情报。

Study on Patent Data Integration from Heterogeneous Sources: A Case of US Patent Data

LIU HuiJing^{1,2}, LING Feng², ZHU XinChao², QI Ping², YANG GuanCan¹

(1. Institute of Scientific and Technical Information of China, Beijing 100038, China;

2. WANFANG Data Co., Ltd, Beijing 100038, China)

Abstract: The US patent bibliographic data from European Patent Office (EPO) and United States Patent and Trademark Office (USPTO) was compared and analyzed from both macroscopic and microcosmic views. Then the processing flow of data integration from heterogeneous sources mentioned above was discussed. The ISTIC-US patent database with consistent and integrate data was finally constructed.

Keywords: Heterogeneous Data Source; US Patent; Data Integration; EPO; USPTO

(收稿日期: 2015-06-20; 编辑: 王立学)