Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

面向多种信息源的专利分析系统建设*

□ 张静 / 中国科学技术信息研究所 北京万方数据股份有限公司 北京 100038
赵蕴华 霍翠婷 刘润生 / 中国科学技术信息研究所 北京 100038
张迎新 / 北京工商大学计算机与信息工程学院 北京 100048

摘要：专利作为一种非孤立型的信息资源，与期刊论文、标准、交易、诉讼等多种信息紧密相关，因此，专利分析的过程需要整合各种信息资源才能得到全面有效的分析结果。文章以专利信息为核心，面向多种与专利相关的信息资源，分析了面向多信息源的专利分析在各个环节需要解决的问题，提出了可解决这些问题的面向多信息源的专利分析系统整体架构。

关键词：多信息源，专利分析，专利分析系统架构
DOI：10.3772/j.issn.1673—2286.2010.09.007

1 引言

创新是企业乃至国家持续高速发展所不可或缺的动力，从宏观层面上看，国家已经把建设创新型国家列为重要发展战略；从微观层面上看，企业作为国家科技创新的主体和应用前沿，近年来也提高了对科技创新重要性的认识。

专利作为一种创新的载体和重要的科技文献，对国家和企业科技创新战略的实施有重要的影响意义，其价值越来越受到各界的重视。专利作为一种复合型的信息资源，集技术信息、法律法规信息、经济信息于一体，又与期刊论文、专利交易信息、国家宏观政策等信息都具有密切的关系。因此，专利分析并不是孤立的，其过程中需要引入多个信息源的信息进行综合，才能得出合理、准确的分析结果。

创新战略的制定需要其制定者全面了解领域的研究现状、市场现状等信息，深刻洞悉市场长期需求、竞争对手分布及其优劣势，寻找创新的技术突破口。然而，创新战略的制定者囿于自身的信息收集能力、信息整合能力、信息分析能力的限制，仅仅依靠自身的力量，往往不能做出正确的判断，其对来自数据、工具、方法三个层次的专利系统分析支撑平台的需求日益迫切。

2 国内外现有专利分析工具及系统

目前，国内外的主要专利信息服务提供商Thomson Scientific、台湾连颖、东方灵盾等已开发了一些较为成熟的专利分析工具和系统，包括Aureka、Delphion、Thomson Data Analyzer[1]、East Linden Door、保定大为、北京彼速、Patent Guider[2]等。这些工具主要提供的功能可概括为基于专利信息的基本统计分析、引证分析、聚类分析三种，其中各工具的基本统计分析功能差异不大，但大部分工具只能做简单的引证分析，国内专利分析软件则普遍不具备聚类分析的功能。

总体而言，目前分析系统还普遍存在以下不足：

1）分析数据源来源单一。虽然多数工具的检索数据源很全面，但很多分析工具仅能针对美国专利进行分析，并且不提供对来自不同专利数据源的专利数据的整合服务。并且上述工具都以专利数据为单一信息源，没有考虑到引入其他非专利信息进行辅助分析。

2）分析前数据预处理的功能有待加强。在进行专利分析前，将

* 本文系“科技基础性工作专项项目——主要国家重点研发领域及主要科技计划和重大专项发展现状的调查与监测平台建设”（2009FY240100）的研究成果之一。

检索结果按照一定规则去清洗、整合是非常必要的。但目前绝大多数工具都不提供此功能。

3）专利分析的灵活性有待提高。大多数专利分析工具和系统允许用户根据一定的指标体系对专利信息进行分析，但在为分析人员提供自定义的分析角度方面，灵活性尚存在欠缺。

4）聚类分析功能有待进一步挖掘。提供聚类分析功能的专利分析工具和系统非常少，且聚类结果并不十分理想。

3 面向多信源的专利分析系统

针对用户需求和国内外专利分析系统的发展现状，万方数据技术研究院专利小组与中国科学技术信息研究所合作，从2009初开始着手设计开发面向多信息源的专利分析系统。该系统的总体设计充分考虑到专利分析从信息采集、信息清洗、信息聚合到信息分析、挖掘的整个流程，尤其在多信源信息的引入整合、专利信息Raw Data的预处理等其他专利分析系统有所欠缺的环节进行了细致的设计，该系统生成的专利信息数据仓库可以支持用户对专利信息进行多角度不同粒度的深入分析。以下分别就构建系统相关的多种信息源的引入、面向深度分析的专利信息源预处理、面向创新服务的深入分析，及面向多种信息源的专利分析系统总体架构等研究着重进行介绍。

3.1 多种信息源的引入

对于专利分析，多信息源有着双重的含义：

广义地讲，专利文献的价值与引文信息、法律状态信息、专利交易信息、诉讼信息、学术研究信息、企业机构信息、政策法规信息息息相关。目前较为普及的专利分析系统都仅以单一信息源，即专利文献信息为分析依据，而未能考虑其他专利相关信息，这种信息缺失在一定程度上影响了专利分析结果的客观性和准确性。面向多信息源的专利信息系统充分考虑到了专利分析信息源的多样性，针对不同信息源设计不同的信息库，并在后期的数据抽取、转换、装载（Extaction Transformation Load，简称ETL）环节以及分析挖掘环节考虑各不同信息源的特点及其相互关系，做到多种信息的有效融合[3]。

狭义地讲，仅专利信息本身就具有多信息源的特点，不同的专利数据库提供的数据都有其特点和自身的局限性。仅仅依靠某个专利数据库的数据作为分析依据，往往只能管中窥豹，不能提供所关注技术发展、布局的全貌。

由于专利的地域性特点，不同国家的知识产权机构往往只收录本国的专利信息。个别专利服务机构或专利信息商业提供商也提供全球范围内的专利信息服务，如欧洲专利局（Europe Patent Office）的Worldwide Patent Database 提供全球90多个国家的专利文献信息；汤姆森路透（Thomson Reuters）集团的Derwent Innovation Index数据库则收录全球100个国家的专利文献信息[4]。但是这种提供跨国专利信息的数据库由于数据处理的缘故，其数据更新往往存在一定的时间滞后，尤其是非英语国家的专利文献，其更新更是比本国专利数据库滞后很多[5]。

此外，综合性的专利数据库往往在数据处理的时候还引入了不少错误，专利信息的细节也存在丢失。以DII为例，该数据库不提供专利权人的地址信息，并且其每条记录以专利家族为单位，专利家族所含的各个专利的信息存在不同程度的缺失，其中很多信息对专利分析至关重要。

鉴于各国专利局提供的专利信息更新快、覆盖范围小、准确性高，而综合性专利数据库覆盖范围广、更新慢、误差多的特点，我们在进行专利分析系统设计的时候，综合考虑了这两种类型的数据源，针对不同的需求提供不同的服务。在系统的数据层，我们同时建立针对不同国家的独立专利数据库以及面向综合专利数据源的统一库。其中，独立库面向对实时性要求较高的专利分析；而统一库则以综合性的专利数据源为基础，同时以各独立专利数据库的信息作为补充和修正依据，面向对全面性要求较高的宏观分析需求。

如图1所示，我们设计的面向多信息源的专利信息分析系统在信息采集的阶段，具有以专利信息为主、其他相关信息为辅、多个专利信息数据源相互补充的特点，力图在数据基础上避免使用单一信息资源所带来的不足。

3.2 面向深度分析的专利信息源预处理

准确、恰当的基础数据是信息分析结果正确性的基础，通常在数据分析、挖掘的实施过程中，数据预处理工作量往往占据整个分析挖掘流程的70%，数据预处理的重要性可见一斑。在专利分析领域，信

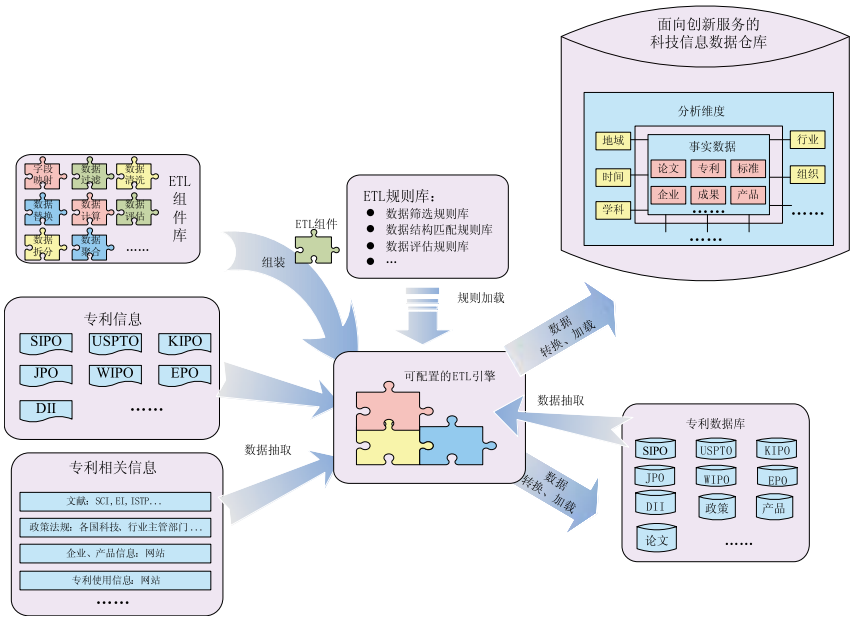

图1 多种信息源的引入

息源的预处理同样至关重要[6]。

仅以专利信息为例，直接从各国专利局以及专利信息提供商处获取的单一信息源原始数据一般都存在以下问题：1）数据的粒度与分析粒度不符；2）存在重复信息；3）专利权人缺乏权威控制等。而考虑到从多个专利数据库获取信息的情况，数据异构、语言差异等问题将为专利分析带来更多的障碍。更进一步，政策法规、专利交易、专利诉讼、学术论文等异质数据的引入，在丰富专利分析基础信息的同时，也为信息整合与预处理带来了更大的挑战。

如图2所示，在专利分析信息源预处理环节，我们的设计以可配置的ETL引擎为核心，通过其对ETL规则库和ETL组件库的调用提供针对异构异质信息源的预处理。其中，ETL组件库包括数据拆分、过滤、替换、聚合等数据预处理通用的功能模块，保证了专利分析信息源预处理的可操作性；ETL规则库中则包含ETL组件库中各功能模块执行需要依据的各个规则库，这些规则库可以根据信息源的变化进行更新，保证了专利分析信息源预处理的灵活性。清洗整合后的专利分析信息将以更适合深度分析的粒度上载到对应的数据库和数据仓库中（见图1）。

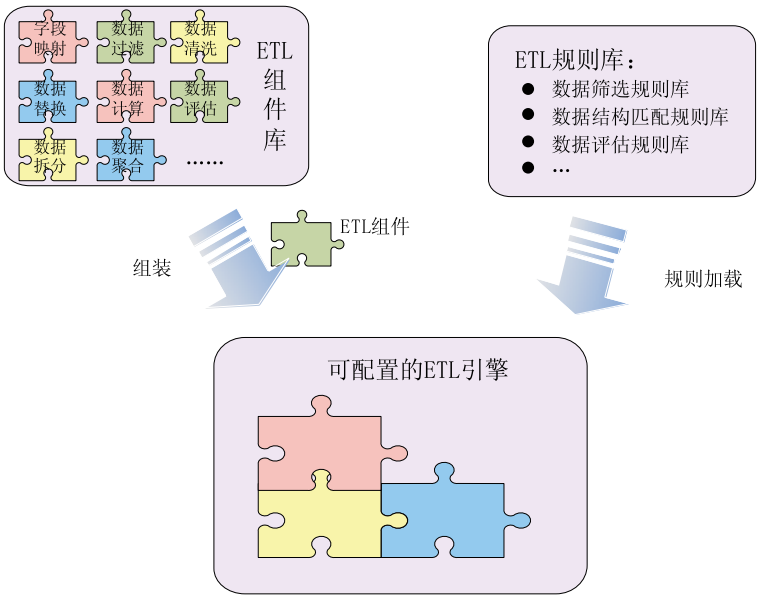

图2 面向深度分析的专利分析信息源预处理

以专利权人信息的清洗为例，ETL引擎可通过调用组件库中的拆分、替换、合并等功能模块，依据专利权人分类规则库、专利权人名称规范库、专利权人权威控制库等ETL规则库中的信息对专利权人进行规范化和权威控制，并生成服务于不同分析粒度的相关聚合信息。

3.3 面向创新服务的深入分析

信息分析模块是面向多种信息源的专利分析系统的核心，前期的多信息源性引入和信息预处理都是为更好地实现分析功能所做的铺垫。由于ETL引擎已经将原始信息转化为适于分析的最小粒度，这使得信息模块对数据仓库中的信息进行多维度的分析成为可能。

通过对现有专利分析软件以及企业需求的调研，我们将专利信息分解为地域、时间、行业、技术、领域等多个维度，通过OLAP[7]技术，可以在这些维度上进行不同粒度、层次的分析（Roll Up、Drill

Down），也可以对多个维度进行组合分析（Slice、Dice、Pivot）。

就分析的深入程度而言，我们在设计信息分析模块时将其划分为三个层次：

- 首先是基本统计层次，主要用以实现一些基本统计指标的生成，如专利地域分布、时间分布、技术分布、活动年期、技术生命周期等；
- 在此之上是复杂分析层次，主要用于实现专利权人合作关系、发明人合作关系、技术相关度、引用关系、专利 、论文主题相关度、同族专利分析、国际合作关系分析等较为复杂的分析功能；
- 最后是信息挖掘层次，借助文本聚类、复杂网络、时间序列等数据挖掘方法，实现技术聚类、热点监测、预测等功能。

3.4 面向多种信息源的专利分析系统总体架构

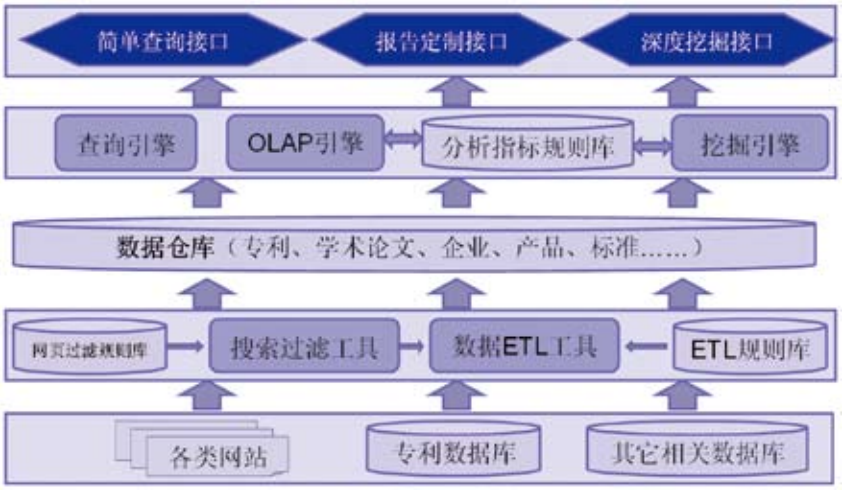

图3 面向多种信息源的专利分析系统总体架构

分析系统首先以各专利数据库和学术论文、政策法规、专利交易、诉讼等其他相关专利数据库作为信息源，同时通过搜索过滤工具根据网页过滤规则库的规则和算法从各类网站采集专利相关信息；然后通过ETL工具实行对信息源的清洗、转换、筛选、聚合等预处理，将聚合后适合分析的相关信息上载到数据仓库中。用户通过简单查询接口可对数据仓库中的信息进行检索查询；通过报表定制接口可以调用OLAP引擎，根据分析指标规则库中的规则生成用户定制的报告；通过深度挖掘接口调用数据挖掘引擎可对专利相关信息进行复杂的分析挖掘。

目前，面向多种信息源的专利分析系统中的多信息源引入模块和ETL模块的基本功能已经完成，分析模块中的简单统计和复杂分析层次的功能也已基本形成。

4 结语

中国科学技术信息研究所战略研究中心依托该系统推出的“重点科技领域检测与深度分析”系列报告获得了不错的反响，这也从一个侧面印证了面向多种信息源的专利分析系统的设计思路和功能。在系统的进一步研究和开发工作中，我们会进一步吸收新的有效的分析方法，更深入全面地考虑不同层面用户的需求，以期更好地为科技创新增添助力。

参考文献

[1] 刘佳佳,董旻,方曙. 国外专利分析工具的比较研究[J]. 现代图书情报技术, 2007(2).
[2] 张静,刘细文,柯贤能,等. 国内外专利分析工具功能比较研究[J]. 情报理论与实践,2008(1).
[3] LI X, HU D, DANG Y, et al. Nano Mapper: An Internet Knowledge Mapping System for Nanotechnology Development[J]. Journal of Nanoparticle Research, 2008(10).
[4] 顾震宇,林鹤. 网络环境下国外专利的有偿、无偿信息源的比较研究[J]. 情报科学,2004(3).
[5] 陆萍. 专利数据库USPTO、esp@cenet、DII的比较分析[J]. 情报科学,2006(9).
[6] RAHM E, DO H H. Data Cleaning: Problems and Current Approaches[C]// IEEE Data Engineering Bulletin,2000,23(30).
[7] CHAUDURY S, DAYAL U. An Overview of Data Warehousing and OLAP Technology[C]// ACM SIGMOD Record, 1997,26(1):65-74.

作者简介

张静（1975-），博士，研究方向为数据挖掘、商业智能、信息分析；发表相关文章十余篇。目前研究方向为专利信息挖掘、开放获取、知识组织。通讯地址：北京市海淀区复兴路15号，中国科学技术信息研究所 100038。E-mail: jane.zh.t@gmail.com
赵蕴华(1967-），硕士，副研究馆员，研究方向：信息咨询和信息资源服务研究。 通讯地址同上。
霍翠婷（1984-）加拿大温莎大学硕士，研究方向：专利分析、专利数据挖掘、信息资源服务研究等。通讯地址同上。E-mail: huoct@wanfangdata.com.cn
刘润生（1982-），硕士，中国科学技术信息研究所战略研究中心助理研究员，研究方向为专利分析、新能源与低碳发展研究。通讯地址同上。
张迎新（1967-），副教授，研究方向：数据库。通讯地址：北京工商大学计算机与信息工程学院，海淀区阜成路11号甲2楼528 100048。

Building a Patent Analysis System Oriented toward Multi-Information Resources

Zhang Jing / Institute of Scientific and Technical Information of China, Wanfang Data Co., Ltd, Beijing, 100038
Zhao Yunhua, Huo Cuiting, Liu Runsheng / Institute of Scientific and Technical Information of China, Beijing, 100038
Zhang Yingxin / BTBU, Institute of Computer and Information Engineering, Beijing, 100048

Abstract: As a non-isolated information resource, patent is closely related to academic publications, technical standards, trades, legal proceedings. Therefore, patent analysis process needs various information resources to make comprehensive and effective decision. In this paper, focusing on patent information, and taking a variety of patent-related information resources into consideration, we address the problems of each patent analysis phase, and propose a patent analysis system oriented toward multi-information resources for solving these problems.

Keywords: Multi-information resources, Patent analysis, Patent analysis system architecture

（收稿日期：2010-08-15）

业界动态

国内33家图书馆联合与国外出版商“斗法”

9月3日下午，国家科技图书文献中心、国家图书馆、中科院国家科学图书馆、北京大学图书馆等33家图书馆的代表约见媒体记者，公开了他们致中国科技文献读者的公开信和致国际出版商的公开信，并向媒体介绍了国内图书馆集体反对个别国外科技期刊出版商在全文数据库大幅涨价的态度。在致国内读者的公开信中，中国图书馆界呼吁担任国际出版商学术期刊编委、审稿专家、顾问的中国专家学者，积极向国际出版商施加影响，要求国外出版商“不要以过高价格或减少内容等手段来对中国广大用户获取和利用国外科技文献设置障碍”；呼吁广大教育科研人员积极支持中国图书馆界抵制个别国外出版商的大幅度涨价。

来源：http://www.bjkp.gov.cn/bjkpzc/kpxx/313820.shtml（查询时间：2010-09-06）