国外专利文本挖掘可视化工具研究*

王　敏　李海存　许培扬
中国医学科学院医学信息研究所　北京 100020

[摘要] 首先简要介绍专利信息分析概念、专利分析的一般流程, 专利分析工具可实现的主要功能; 其次依据专利分析工具可分析的数据源, 将分析工具分为非结构化数据分析工具、结构化数据分析工具和混合型数据分析工具三大类, 并从分析工具类型、分析数据源、主要功能、结果呈现、用户群 5个方面对国外常用的 12种专利文本挖掘可视化分析工具进行系统介绍和比较; 最后对专利分析工具应用及其发展提出建议。

[关键词] 文本挖掘　可视化工具　专利分析　竞争情报
[分类号] G353.1　G306

Foreign Text Mining and Data Visualization Tools in Patent Information Analysis

Wang Min　Li Haicun　Xu Peiyang
Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020

[Abstract] This paper starts with an introduction of the definition of patent information analysis and its process, and then summarizes the main function of patent analysis tools, which can be classified into structured analysis tools, unstructured analysis tools and hybrid analysis tools according to different types of data sources. The main part in this paper is to provide a detailed overview and comparison of twelve foreign text mining and data visualization tools in its type, applicable data sources, achievable analysis methods, visualize output of results and intended users. Finally, the paper gives some suggestions on the application and development of patent information analysis tools.

[Keywords] text mining　data visualization tools　patent information analysis　competitive intelligence

随着知识经济全球化进程的加快, 专利文献作为反映科技发展, 特别是技术发展态势的重要情报源, 在科技战略制定中发挥着日益重要的作用, 如何对其开展有效分析, 辅助政府部门、科研机构、高新企业进行专利战略布局和专利技术研发, 成为情报机构开展情报分析、战略决策的重要方向。专利分析离不开高效分析工具的支持, 专利分析方法、分析工具的合理使用是决定信息分析水平、效率以及质量的重要因素。本文首先对专利信息分析进行简要概述, 并对国外常用专利分析工具进行系统调研, 以期为国内人员开展专利信息分析工作提供借鉴。

1　专利信息分析概述

专利信息分析是竞争情报分析的重要形式, 是在对专利文献进行筛选、鉴定、整理基础上, 利用文献计量学方法, 对其所含的各种信息要素进行统计、排序、对比、分析和研究, 从而揭示专利文献的深层动态特征, 了解技术、经济发展的历史及现状, 进行技术评价和技术预测[1]。

专利信息分析流程分为准备期、分析期和应用期三个阶段。准备期是保证专利信息分析达到目标的基础; 分析期是专利信息分析工作的主体, 主要包括数据采集和数据分析两个阶段; 应用期是分析工作的延伸, 是专利信息分析的价值体现, 各阶段具体包括的内容见图 1[2]。

2　专利分析工具的主要功能

随着信息技术飞速发展, 文本挖掘、信息可视化技术已被应用到专利分析领域, 众多专利分析工具应运而生, 尽管不同分析工具各有专长, 但总的来说专利分

* 本文系中国医学科学院医学信息研究所中央级公益性基本科研业务费专项资助课题/信息可视化在医学信息分析中的应用研究0 (项目编号: 08R0129)研究成果之一。
收稿日期: 2009-06-22　修回日期: 2009-09-18　本文起止页码: 86-90　本文责任编辑: 王传清

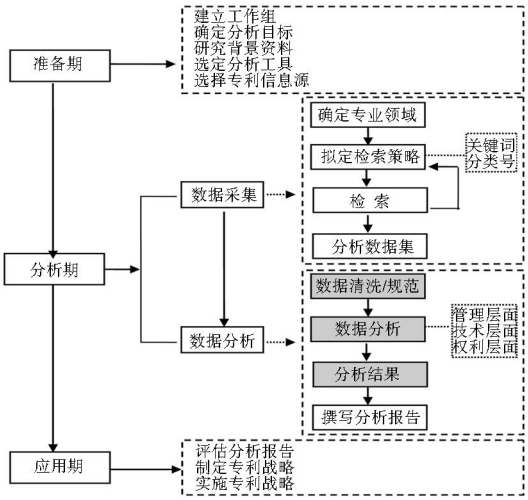

图 1 专利信息分析流程

析工具功能主要有以下两点:

2.1 数据清洗

数据清洗又称数据规范, 是影响专利分析效果至关重要的一步, 其作用是为专利分析提供准确的数据, 主要包括: ¹ 对检索到的专利文献进行相关性筛选, 将符合条件专利文献纳入分析数据集; º 同一概念不同写法进行规范, 以消除同一概念、同一事物不同写法造成的分析误差。

2.2 分析方法实现及其结果可视化显示

专利分析方法通常分为定性分析、定量分析和拟定量分析, 分析工具可实现的方法主要包括基本统计分析、共现分析、聚类分析和引证分析 4类:

基本统计分析: 是指依据专利文献标引项, 对专利申请时间、申请人、申请机构、申请国家、同族专利量等指标进行统计, 用于把握专利文献分布状况及其发展态势。分析结果通常以列表、直方图表形式展现。

共现分析: 是指相同或不同类型特征项信息共同出现的现象。通过对专利分类号、专利权人、申请时间、申请国、专利技术焦点等进行组配统计[3], 用于揭示专利信息的内容关联和特征项所隐含的知识。分析结果显示方式主要有共现矩阵和曲线图。

聚类分析: 是指利用聚类技术将同一数据集中的专利, 按照技术类别聚成不同的子类, 以揭示该特定技术领域内各个子领域的分布情况, 分析各主要竞争对手专利分布情况。目前专利聚类主要基于主题, 呈现结果可以按竞争对手和时间顺序进行浏览[4]。聚类结果展现方式主要有聚类地图、结构化数据聚类和非结构化数据聚类[5]。

引证分析: 是指对专利的引用和被引用情况进行分析。通过研究专利间的引用关系及其规律, 探求技术间的联系和发展规律, 跟踪不同技术专利网络, 反映特定技术领域的生命周期以及竞争对手间的技术依赖关系。分析结果呈现主要有引证表、引证树和引证地图[4]。

3 国外常用专利分析工具

本部分重点列举国外常用专利分析工具, 并依据分析工具可分析的数据源, 将其分为非结构化数据分析工具、结构化数据分析工具和混合型数据分析工具三大类[6]。

3.1 非结构化数据分析工具

非结构化数据分析工具是指用于分析专利全文、期刊论文、网页内容等非结构化数据的软件, 主要包括 ClearForest、OmniViz和 TEMIS。

3.1.1 ClearForest ClearForest是美国 Thomoson Reuters公司开发的具有强大功能的文本分析工具, 包括先进的文本标记抽取平台、分析平台以及开发环境。其最具特色功能是可以将非结构化数据转化为结构化数据, 如从论文、网页等非结构化文本中抽取相关词语生成结构化数据, 进而对结构化数据进行文本挖掘, 如分类、聚类, 生成列表、共现矩阵、聚类图等。此外该工具还提供文本分析可视化功能, 用于挖掘类间隐含关系和发现新知识[7]。

3.1.2 OmniViz OmniViz是英国 BioWisdom 公司开发的一款单机版可视化数据分析软件。该软件有两大特色: ¹ 分析数据类型广泛, 可对数值数据、分类数据、基因序列、化学结构以及专利、论文等多种数据类型进行分析; º 整合复杂的统计算法、文本算法对大规模数据进行分析生成可视化图谱辅助用户对数据的理解, 可视化图谱主要有 Galaxy图、CoMet图、ThemeMap和聚类图等[8]。

3.1.3 TEMIS TEMIS是美国 TEMIS公司开发的一款用于商业智能的文本挖掘工具。其特色功能是支持概念化检索, 依靠强大的数据算法、语言学算法将多语种、多种文本类型的非结构化数据转化为结构化数据, 对其进行数据提取、分类、聚类等[9]。TEMIS价格昂贵限制了在国内的应用。

3.2 结构化数据分析工具

结构化数据分析软件主要用于对数据库中的专利信息、文献题录信息进行分析, 主要包括 Thomson Data Analyzer(简称 TDA)、VantagePoint、Quosa、ReViz、STN

AnaVist和 Vxinsight。

3.2.1 Thomson Data Analyzer TDA是美国 Thomson Reuters与 Search Technology公司联合推出的数据挖掘和可视化分析工具，由 Search Technology公司的 VantagePoint引擎提供技术支持。TDA除支持德温特世界专利索引 (DII)、Web of Science和 Pubmed等常用数据库外，还支持 Excel数据的导入。TDA软件提供强大的数据清洗功能保证了数据分析的准确性，支持基本统计、共现分析、聚类分析，并可自动生成列表、矩阵、聚类图、报告等[10]。

VantagePoint软件与 TDA软件功能基本类似，在此不再赘述。

3.2.2 Quosa Quosa是美国 Quosa公司开发的一款集文献检索、全文下载、文献管理分析于一体的单机版文本挖掘工具。该软件支持 Ovid、PubMed、美国专利数据库等的直接搜索，可将 PDF全文下载到本地进行组织管理，并可对文献进行概念提取和聚类。其文献全文自动下载、最新进展追踪、PDF文献信息自动识别以及全文分析功能是同类文献管理软件所不具备的[11]，但其分析功能与专业类文献分析软件相比功能还较少。

3.2.3 RefViz RefViz是美国 Thomson Reuters公司开发的单机版文献信息分析可视化软件。该软件主要特色是具备强大的语义分析功能，可利用词库工具对数据进行清洗，分析结果可生成 Galaxy视图和二维矩阵视图。但是 RefViz仅能分析结构化数据，如来自文献数据库或文献管理软件的文献题名、摘要、主题词等信息，不能分析非结构化数据[12]。

3.2.4 STN AnaVist STN AnaVist是美国化学协会分支机构化学文摘服务社 (CAS) 与 FIZ Karlsruhe开发的一款科技文献、专利文献文本分析可视化软件。该软件可对化学文摘、DII、欧洲专利和美国专利等多个数据库进行分析，并可利用 CAS词表对机构、技术术语进行数据规范。其主要特色是可采用聚类技术生成研究内容全景分析图[13]。

3.2.5 Vxinsight Vxinsight是美国能源部桑地亚 (Sandia)国家实验室开发的一款免费的单机版文本分析可视化软件。该软件的主要特色是采用三维虚拟地图的形式来模拟聚类信息，以揭示科技文献、专利、蛋白、基因间的相关性[14-15]。

3.3 混合型数据分析工具

混合型数据分析软件是一类可分析结构化数据和非结构化数据的软件，主要包括 Aureka、M2CAM Doors、Wisdomain和 PatAnalyst，这些工具都整合了专利数据库检索功能。

3.3.1 Aureka Aureka是美国 Thomson Reuters公司开发的在线知识产权管理分析平台，提供专利检索、管理、分析 (专利引证分析、专利地图分析等)、预警等功能。在专利分析方面，ThemeScape提供聚类分析可生成专利地图，Citation Tree提供引文分析可生成引证树，揭示专利信息间的相互关联，为用户技术研发与自主创新、专利评价与评估、专利权保护、企业联营与合作或兼并等的生产经营决策活动提供帮助。该软件数据清洗功能较弱是其主要不足[16]。

3.3.2 Wisdomain Wisdomain是美国 Wisdomain公司开发的一个专利分析解决方案，整合 FOCUSTI、Patent2Magnet、PatentFamilyTree、PatentLab2II 4个工具，支持美国、欧洲、中国、日本、韩国、世界 PCT专利检索以及 Inpadoc法律状态检索，提供基本统计、共现分析和引证分析功能，分析结果可以列表、聚类图、引文图、二维或三维图形显示[17]。

3.3.3 Delphion专利信息平台 Delphion是美国 Thomson Reuters公司开发的专利信息服务平台，集成 Snapshot、Corporate Tree、PatentLab2II、Text Clustering、Citation Link 5个工具，分别提供在线分析、公司名称规范、列表和直方图等图表生成、文档聚类、引文分析功能。该平台收录范围广、整合分析工具多是其主要特色，但其按服务项目、专利下载数量收费的服务模式，使得一般用户难以承受其高昂的费用[18]。

3.4 专利分析工具比较

以上对国外常用的非结构化数据分析工具、结构化数据分析工具、混合型数据分析工具进行了简单介绍，下面将从分析工具类型、分析数据源、主要功能、结果呈现、用户群 5个方面，对 12个分析软件进行比较[6]，见表 2。

非结构化数据分析工具，主要基于语义分析技术，将非结构化数据转化为结构化数据，进而利用强大的分析功能对其进行分析。这三款软件中，ClearForest、TEMIS价格昂贵，限制了在国内的应用；Ominiviz为单机版软件使用便捷，除具有文本挖掘功能外还具有强大可视化功能，其可视化功能在众多软件中尤为出众。

结构化数据分析工具，主要用于分析结构化数据。TDA是目前国内科技文献、专利文献分析应用较多的软件，支持 20多种文献数据源，是目前已知文献信息

表 2 国外 12种专利文本挖掘可视化工具比较

工具名称	工具类型	分析数据源	主要功能						用户群
			数据清洗	分析方法				结果呈现	
				基本统计	共现分析	聚类分析	引证分析		
非结构化数据分析工具									
CleanForest	文本挖掘	结构化数据和非结构化数据	有	无	有	有	无	列表、矩阵、聚类图	商业智能
OminViz	文本挖掘/可视化	结构化数据和非结构化数据(数值数据、分类数据、基因序列、化学结构)	有	有	有	有	无	交互式可视化图谱(Galaxy图、CoMet图、ThemeMap和聚类图等)	研发人员
TEMIS	文本挖掘	结构化数据和非结构化数据	无	有	不详	有	无	列表、聚类图	研发人员/商业智能
结构化数据分析工具									
Quosa	文本挖掘/文献管理	结构化数据(PubMed, Ovid, USPTO等)	无	无	无	有	无	数据分组和注释	研发人员
RefViz	文本挖掘/可视化	结构化数据(SCI, PubMed, OCLC等)和来自参考文献管理软件的数据	有	有	有	有	无	Galaxy图和矩阵图	研发人员/信息管理人员
STN AnaVist	文本挖掘/数据库检索	结构化数据(CA Plus, USPTO, DII等)	有	有	有	有	无	列表、图表、研究景观图(research landscape)	信息管理人员/商业智能/研发人员
Thomson Data Analyzer	文本挖掘	结构化数据(SCI, PubMed, DII等)及Excel格式数据	有	有	有	有	无	列表、图表、矩阵、聚类图、专利报告	信息管理人员/商业智能
Vxinsight	文本挖掘/可视化	结构化数据(ODBC方式存取的多种数据类型)	无	有	无	有	无	聚类图(二维、三维)	研发人员/信息管理/混合型数据分析工具
混合型数据分析工具									
Aureka	文本挖掘/可视化/数据库检索	US, DE, EP, GB, JP(仅文摘)和PCT专利	有	有	有	有	有	ThemeMap、引文树、聚类图、专利报告	研发人员/信息管理人员/决策人员/商业智能
Wisdomain	文本挖掘/数据库检索	US, DE, EP, JP, PCT, 中国, 韩国, INPADOC	有	有	有	有	有	列表、图表、系统树、引文图	研发人员/信息管理
Delphion专利信息平台	文本挖掘/数据库检索	US, DE, EP, JP, PCT, INPADOC, DII	有	有	无	有	有	列表、引文树、聚类图	研发人员/信息管理/商业智能

分析工具中支持数据最为广泛的软件, 且支持 Excel文件 (含中文)的导入; 此外该软件具有强大的数据清洗功能、自动生成专利报告功能, 这些功能是其他软件无法比拟的; 但是 TDA在专利地图制作、文献结果可视化方面还存在不足, 在专利分析中需与其他专利分析工具联合应用。 Quosa和 Refviz主要用于期刊文献的管理和分析, 支持数据源较少; STN AnaVist自带技术术语、机构分析词表可用于专利文献数据清洗, 但对大规模数据库的清洗仍是该软件面临的巨大挑战。 Vx2 insight是本文介绍的分析工具中唯一一款免费的软件, 主要特色是可以生成二维、三维聚类地图用于揭示专利、文献间的关系, 但该软件在专利分析方面功能较弱。

混合型数据分析工具, 除提供专利分析功能外, 还提供专利文献检索、数据下载功能, 文中提到的三个分析工具分析功能完备, 均具有数据清洗功能, 提供基本统计、共现分析、聚类分析、引文分析 (仅对美国专利进行分析), 并可对分析结果进行可视化显示。但这三个工具在专利分析方面各有其优势与不足, 如 Aureka可采用聚类分析生成主题 (词汇)地形图, 用于专利技术主题分布研究, 而在专利国家、机构分析分析方面由于缺乏数据清洗功能, 分析结果准确性不足; Wisdomain仅能分析自带数据库检索结果, 不具有数据导入功能; Delphion主要用于专利数据检索, 在数据分析方面相比 Aureka和 Wisdomain功能较弱。

4 结 语

专利分析工具是顺利开展专利信息分析的重要保障, 分析工具的好坏将直接影响专利分析的效率和结果的准确性, 在应用专利分析工具时, 还应注意以下几点:

融会信息分析思维, 选择恰当分析工具。国外专利分析工具众多, 在开展专利分析工作时, 应根据不同的分析目的、拟解决的问题, 结合不同分析工具的主要功能, 选择恰当的分析工具。由于国内购买国外产品途径不畅、价格较高等原因, 目前国内应用较多的主要有 TDA、Aureka和 Ominiviz。

结合人工干预, 提高分析质量。高质量专利分析报告的完成离不开对专利文献的文本挖掘, 但是仅有文本挖掘工具或信息技术专家是不够的, 还需要具备专业知识背景专家的干预, 因此在专利分析工具使

用中, 从数据检索、数据规范、数据分析以及结果的解释都离不开人工的干预以及专家的支持。

分析工具尚不完善, 分析功能有待进一步提升。随着文本挖掘和信息可视化技术的应用, 专利分析工具中有了较大提升, 但仍存在一些不足, 如多数据源融合度低、数据清洗功能弱、知识挖掘程度浅等, 相信随着自然语言处理、人工智能创新技术的不断进步, 分析工具功能将不断完善。

参考文献:

[1] 骆云中, 陈蔚杰, 徐晓琳. 专利情报分析与利用. 上海: 华东理工大学出版社, 2007. 130.

[2] 陈燕, 黄迎燕, 方建国, 等编. 专利信息采集与分析. 北京: 清华大学出版社, 2006. 67.

[3] 暴海龙, 朱东华. 专利情报分析方法综述. 北京理工大学学报(社会科学版), 2002, 4(S1): 91-93.

[4] 张静, 刘细文, 柯贤能, 等. 国外专利分析工具功能比较研究. 情报理论与实践, 2008, 31(1): 141-145.

[5] Trippe A J. Patinformatics: Tasks to tools. World Patent Information, 2003(25): 211-221.

[6] Yang Y Y, Akers L, Klose T, et al. Text mining and visualization tools - Impressions of emerging capabilities. World Patent Information, 2008, 30(4): 280-293.

[7] Thomoson Reuters ClearForest. [2009-04-15]. http://www.clearforest.com.

[8] BioWisdom. OminViz. [2009-04-15]. http://www.biowisdom.com.

[9] TEMIS. TEMIS text intelligence. [2009-04-15]. http://www.temis.com.

[10] Thomson Reuters. Thomson Data Analyzer. [2009-04-15]. http://thomsonreuters.com/products_services/scientific/Thomson_Data_Analyzer.

[11] Quosa. Quosa. [2009-04-15]. http://www.quosa.com.

[12] Thomson Reuters. RefViz. [2009-04-15]. http://www.refviz.com.

[13] CAS. STN AnaVist. [2009-04-15]. http://www.cas.org/products/anavist.

[14] Sandia National Laboratories. Vxinsight. [2009-04-15]. http://ipal.sandia.gov/VxInsight.

[15] Boyack K W, Wylie B N, Davidson G S, et al. Analysis of patent databases using vxinsight. [2009-04-15]. http://www.cs.sandia.gov/projects/VxInsight/pubs/npivm00.pdf.

[16] Thomson Reuters. Aureka. [2009-04-15]. http://aureka.micropat.com.

[17] Wisdomain. Wisdomain. [2009-04-15]. http://www.wisdomain.com.

[18] Thomson Reuters. Delphion. [2009-04-15]. http://www.delphion.com.

1作者简介2王 敏, 女, 1979年生, 馆员, 发表论文 6篇; 李海存, 男, 1984年生, 硕士研究生; 许培扬, 男, 1953年生, 研究员, 研究室主任, 发表论文 60余篇。

(上接第 45页)

参考文献:

[1] Dretske F I. Knowledge and the flow of information. Oxford: Basil Blackwell, 1981.

[2] Tuominen K, Taj S, Savolainen R. Discourse, cognition, and Reality: Towards a social constructionist metatheory for library and inforation science// Bruce H, Fidel R, Ingwersen P, et al. Emerging Frameworks and Methods: COLIS 4: Proceedings of the 4th International Conference on Conceptions of Library and Information Science. Seatle, WA: Libraries Unlimited, Greenwood, CT, 2005: 271-283.

[3] Cole C. Activity of Understanding a problem during interaction with an / Enabling0 information retrieval system: Modeling information flow. Journal of the American Society for Information Science, 1999, 50(6): 544-552.

[4] Ingwersen P. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. Journal of Documentation, 1996, 52(1): 3-50.

[5] Toms E G. Information interaction: Providing a framework for information architecture. Journal of the American Society for Information Science, 2002, 53(10): 855-862.

[6] Cool C, Belkin N J. A classification of interactions with information// Bruce H, Fidel R, Ingwersen P, et al. Emerging frameworks and methods: COLIS4: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science. Greenwood Village, Colorado: Libraries Unlimited, 2002: 1-15.

[7] Spink A. Information science: A third feedback framework. Journal of the American Society for Information Science. 1997, 48(8): 728-740.

1作者简介2杨延铮, 男, 1972年生, 馆员, 发表论文 7篇; 刘秋让, 男, 1967年生, 副教授, 副馆长, 发表论文数篇; 师俏梅, 女, 1970年生, 副研究馆员, 发表论文 5篇, 参编专著 3部; 田苍林, 男, 1959年生, 研究馆员, 发表论文 20余篇, 出版著作 3部; 黄 辉, 男, 1978年生, 助理馆员, 发表论文 2篇。