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

面向深度分析的领域专利信息特色资源服务平台建设*

□ 桂婕 乔晓东 朱礼军 张兆锋 李鹏 / 中国科学技术信息研究所 北京 100038

摘要：专利信息是集技术、经济、法律信息于一体的重要情报分析数据来源，也是支持技术创新管理的重要决策依据。但目前，面向科技领域专利深度分析过程中存在着信息整合程度低、深层信息揭示力度小、先进IT技术方法应用不足而不能快速响应决策支持等问题。基于此，文章提出了以重点科技领域战略研究为服务目标、建设面向深度分析的领域专利信息特色资源服务平台的研究思路，并介绍了该平台实现的流程与方法。

关键词：专利信息，专利资源服务平台，科技决策支持
DOI：10.3772/j.issn.1673—2286.2010.07.008

1 引言

研究表明，在几大科技知识资源（主要包括专著、专利、科技论文和技术报告等）中，新知识含量最高的是专利[1]。同时，专利文献是集技术信息、经济信息、法律信息于一体的最重要的情报源之一，对于技术创新管理具有重要意义。近年来情报学和管理学研究领域都将其视为重要的研究数据来源，专利信息分析已成为一种重要的研究手段，为科研管理、技术研发等提供重要的决策信息支持。在这种日益高涨的应用需要驱动下，相关专利信息分析工具的研究及分析系统的研制开发成为当前科技信息服务业的热点。

比较分析国内外主要专利信息分析系统，国外专利信息分析系统主要包括TDA[2]、AUREKA[2]、Delphion[3]、INAS[4]、VantagePoint[5]、Focust[6]、STN AnaVist[7]、M-CAM DOORS[8]、OmniViz[9]等。上述专利信息分析系统的主要特点为：处理数据对象为英文专利，其中大多数都与特定专利数据库绑定，并可提供专利数据的检索、基本清洗和二维、三维图表的统计展示、引文分析、文本聚类、关联共现分析、专利地图等功能。国内专利信息分析系统主要包括中国知识产权信息中心[10]、恒和顿[11]、大为[12]、东方灵盾[13]等机构的专利信息分析系统，国内专利分析系统的处理对象包括了中英文专利，提供的分析指标偏重专利著录项数据的定量分析及二维、三维图表的统计，但其处理深度文本分析的能力落后于国外专利分析系统[14]。

本文作者在重点领域科技监测研究的开展过程中发现，面向管理决策的科技领域专利分析具有数据量大、数据清洗准确率要求高、定量分析与文本内容分析整合、分析指标个性化多样化等特点，而现有专利分析系统在数据筛选、清洗、分析等功能模块上，其质量、效率及灵活性等方面还无法完全满足深度领域专利分析工作的开展，需要大量人工干预且耗费时间长。

基于此，本文提出了建设“面向深度分析的领域专利信息特色资源服务平台”（以下简称专利特色资源服务平台）的研究工作思路，在选定专利数据库来源的基础上，通过整合专利数据的下载与抽取、专利权人清洗、基本专利指标分析、专利文本知识抽取、面向分析

* 本文得到国家科技部“十一五”科技支撑计划（项目编号：2006BAH03B03）、中国科学技术信息研究所重点工作项目（项目编号：2009KP01—7—1）、中国科学技术信息研究所预研基金项目“领域专利知识组织实现的关键问题研究”等项目的资助。

的数据检索与导出等功能模块，开发建设面向领域专利分析、灵活易用的在线专利信息服务平台。

2 专利特色资源服务平台总体设计

2.1 平台系统架构设计[15-17]

专利特色资源服务平台的设计从两个方面着手：一是数据层构建，二是涵盖了数据获取、数据存储与处理和数据分析的工作流程的技术实现。

2.1.1 数据层构建

按照专利特色资源服务平台建设中数据获取、数据存储与处理和数据分析的工作流程，数据层的建设也相应分为三个层次，数据层的设计如图1所示。

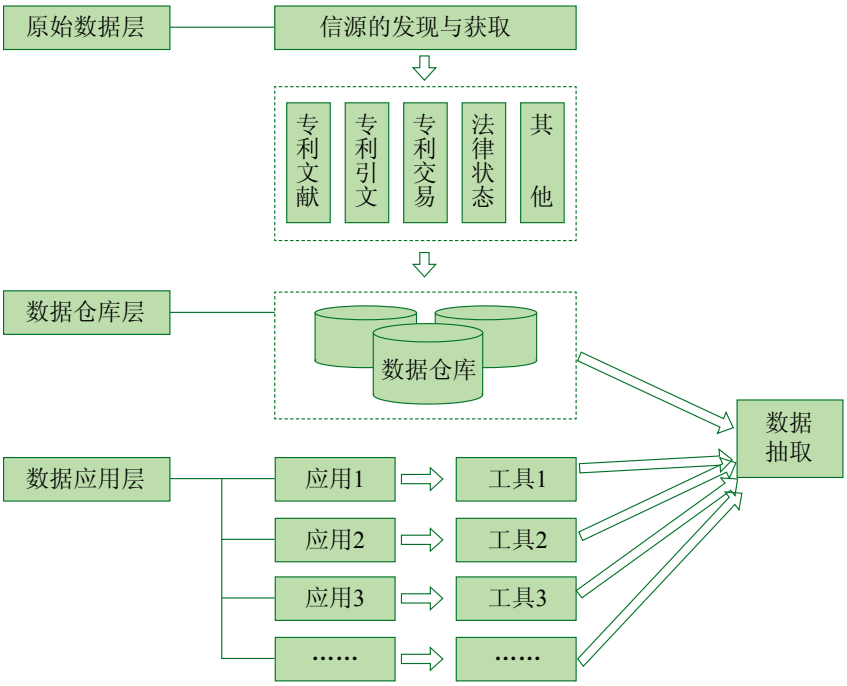

图1 专利特色资源服务平台数据层设计

（1）原始数据层：主要存储不同来源的数据，包括特定领域的专利文献数据（专利文献的外部著录项字段、文摘、权利要求项等），专利引文数据（主要是美国专利的引用与被引数据），专利交易的新闻报道与相关数据，专利的法律状态信息等。

信源的发现与获取工作主要针对特定领域的专利信息获取需求，以各类专利数据库和网络科技新闻为数据的获取来源，使用数据自动捕获与下载的工具与技术将所需数据抓取到本地服务器上，并按一定规则进行数据清洗工作。其中，专利法律状态信息将考虑采取实时使用、实时获取的方式获得，确保法律状态信息的及时性。

（2）数据仓库层：专利数据仓库的建设遵循保持数据事实的原则。经清洗规范后的各类数据按照应用需求以数据仓库的体系架构进行存储与管理，在应用阶段进行改造和重复使用数据仓库中的各种粒度的数据。平台设计重点侧重于数据仓库的结构设计及异构数据的融合。

（3）数据应用层：数据应用层以表现数据特征为基础，以特定专利分析应用触发相应工具的开发，并驱动对数据仓库中数据的抽取与利用。应用包括数据检索、指标分析和可视化展示等三大模块，对应每个模块对数据需求的不同特点，开发对应的专利聚类、趋势图、矩阵图、关联图谱等工具。

2.1.2 基于专利分析流程的平台架构设计[18-20]

基于专利分析流程的平台架构设计如图2。平台架构包括了数据获取、数据处理与转换和指标分析与展示等三部分。

（1）数据获取

数据获取有网络专利信息下载和本地数据导入两种方式。

① 网络专利信息下载。通过集合专家意见和参考主题词表制定该领域专利数据检索策略，存入领域检索表中。启动平台上的下载程序，可自动通过爬取网页的方式获得网络专利数据，抽取后存入原始数据库。下载引擎可以自动设定爬虫程序定期去网站下载更新的数据，保证了系统数据的及时更新。此种方式优点是可以获得免费数据，但受到速度和下载条数的限制。

② 本地数据导入。对于已经购买了定制专利数据库的使用者，可以通过数据导入程序直接将购买的本地数据导入系统。此种方法优点是速度快，缺点是需要人工更新及

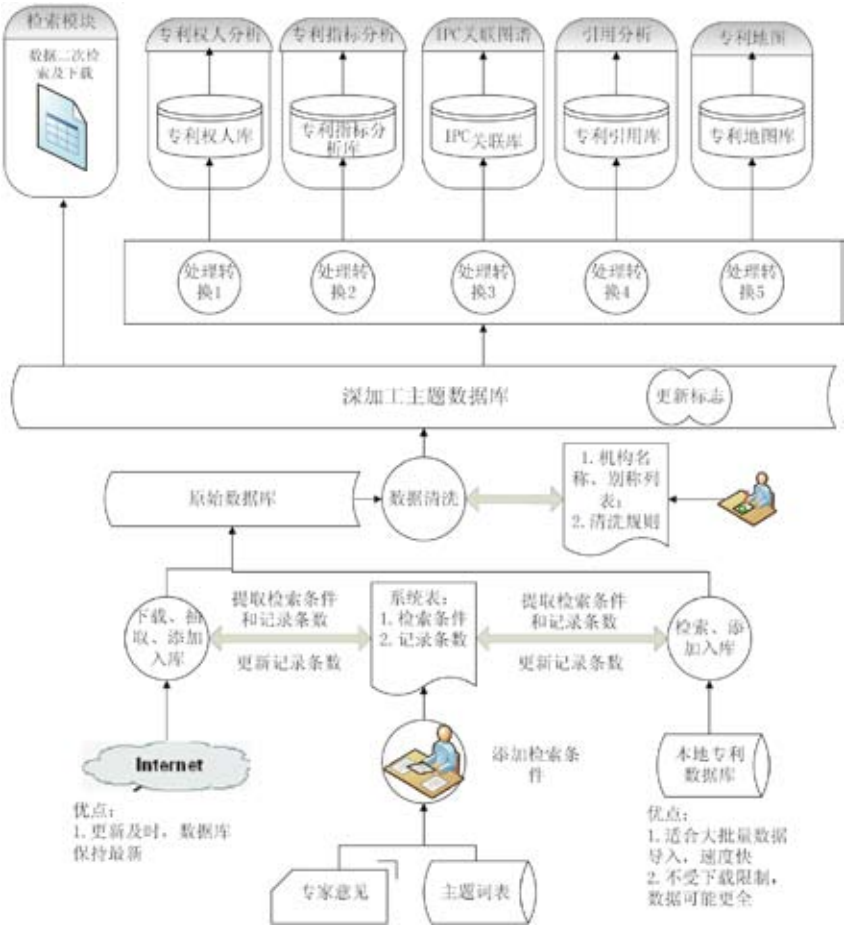

图2 基于专利分析流程的平台架构设计

耗费较多的人力。

（2）数据处理与转换

① 数据处理。平台的数据处理以自动处理与人工处理相结合的方式进行。系统数据库内置有专利数据的清洗规则表和专利权人权威控制表。系统根据这两个表可自动对专利数据进行清洗，及对机构名称进行规范统一。同时，可人工对两个表进行添加修改等维护工作，使得对数据处理的准确度不断提高。处理后的数据存入深加工主题数据库。

② 数据转换。深加工主题数据库存储的是最小粒度的专利数据，基于此库可以生成面向指标分析和图形展示的专项库。如：专利权人库、专利指标库、IPC关联库、专利引用库和专利地图库。专项库支持各类指标分析的快速展示。同时，平台可也提供深加工主题数据库的数据导出功能，将需要的数据导出到Excel表中进行深入分析。

（3）指标分析与展示

针对生成的专项库生成各种类型的展示图，如柱状图、饼形图和折线图等；基于引文数据生成引用关系图；基于对专利数据的深度挖掘而生成专利地图。系统提供的友好界面可以对生成图的参数进行调整，以最大限度满足分析需求。

2.2 平台设计的“特色”定位

本文的专利特色资源服务平台主要解决专利分析流程中的数据自动获取与处理、数据自动深度清洗、文本内容提取等功能的软件工具实现问题，最大限度地提高数据获取与分析工作的完成效率。平台设计的特色定位主要包括数据来源选择、数据拆分与清洗规范的制定及实现、基于专利著录项的指标分析与数据导出的集成和专利文摘知识抽取等。

2.2.1 数据来源选择

区别于集成了数据源与分析工具的专利分析系统，本平台只提供面向选定数据来源的专利处理与分析工具。由于重点科技领域专利分析工作的数据分析对象一般为世界各国专利（以技术领先国家专利分析及与中国专利的对比分析研究为研究重点），综合考虑专利数据库的数据质量与检索功能的全面性，本平台选择了Derwent专利数据库、美国专利商标局（USPTO）专利数据库和中国知识产权局专利数据库等作为数据来源。其中，Derwent不仅收录了世界各国专利的英文版本，还单独建立了Derwent专利的应用领域分类、技术领域分类、机构代码表等，从检索功能上有效弥补了关键词检索和IPC分类号检索的不足；而美国USPTO专利数据库，因具有授权专利质量高、专利信息丰富（除专利文献数据外，还可获取美国专利的被引信息、缴费信息、法律状态信息等）等特点，一直是专利分析领域研究者的重要研究数据来源。

2.2.2 数据拆分与清洗规范的制定与实现

数据拆分与清洗质量的好坏直接影响分析结果的准确程度。因此，本平台在原始数据拆分设计上以数据粒度最小化为原则，使拆分后的数据可直接用于指标分析。同时，专门针对专利权人清洗设计开发了清洗规范及配套软件工具。在专利数据清洗过程中，除专利权人外，其他专利著录项信息都可通过软件工具实现自动的拆分与清洗，而专利权人的清洗则需要软件工具与人工参与的方式来完成。

国外高质量专利数据库和领先专利分析系统都非常重视专利权人的清洗与规范工作。如Derwent专利数据库专门自建了专利权人机构代码表，每条专利都加入了机构代码信息；TDA、VantagePoint等都提供专利权人机构的清洗与合并功能。因此，本平台的专利权人自动清洗工具及辅助人工清洗工具也将是研究重点之一。

2.2.3 基于专利著录项的指标分析与数据导出的集成

平台的分析功能设计上，除提供基本专利指标分析结果展示外，还将支持分析结果与原始分析数据链接的功能，使用户可将感兴趣的相关数据内容导出到本地进行深入分析，且此导出结果将包含未拆分的原始数据和拆分后的最小粒度数据两类。

2.2.4 专利文摘知识抽取[21,22]

在专利文本分析领域，并没有专门明确定义专利知识抽取的内涵，而是将知识抽取过程中非结构化自由文本的知识内容获取的相关方法、技术融合于专利分类/聚类、文本挖掘、可视化、主题识别、趋势发现等各类研究中，专利文本的知识抽取研究是专利内容分析的重要基础。

本文专利知识抽取研究将在分析专利结构特征与内容特征（领域特征、技术特征、功效特征）基础上，构建专利知识抽取模型，研究规则与统计相结合、多种方法集成的混合专利知识抽取技术，并将工具集成在平台上，为研究者开展专利分类/聚类、文本挖掘等研究提供工具支撑。

3 专利特色资源服务平台主要模块的实现

3.1 专利数据自动抽取

在本文专利特色资源服务平台设计上，用户可选择Derwent、美国USPTO和中国SIPO等三类来源的专利数据进行领域专利分析工作。本文以Derwent专利数据为例，介绍平台专利数据自动抽取模块的实现。

Derwent专利数据库提供了比较全面的检索功能，用户在Derwent数据库中检索得到所需领域专利后，可下载获取专利的著录项信息和摘要信息等，其网站上提供了三种数据下载格式，分别为html格式、纯文本格式和制表符分隔的格式。考虑机器抽取对象的易读性和数据完整性，本平台选择html格式的专利数据作为抽取处理对象。

在技术实现上，数据自动抽取分为两个步骤：

（1）采用基于正则表达式的抽取技术，从html格式的原始信息中提取专利的基本信息，形成基本信息表；

（2）使用SQL存储过程，将基本信息表中的数据拆分为细粒度的分析数据表。

3.2 专利权人清洗

专利权人主要被用来进行专利申请所属机构分析，如机构专利申请量排名、机构类型分布等，因此专利权人清洗的准确率对分析结果有着重要的影响，这也是专利数据处理中最为耗费人力的环节之一。本平台在制定专利权人清洗规则基础上，开发了自动专利权人清洗工具，以最大限度减轻人工清洗工作的强度。专利权人清洗分为两个步骤：

（1）利用机构名称特征表，自动标注该机构的类型，类型如表1所示。其中，类型3、4、5的机构类型未完全明确，需要人工干预进行二次清洗。

（2）建立专利权人权威控制表，自动处理专利权人信息中拼写不一致、名称不规范等问题；同时，将专利权人的清洗分为机构独立法人名称和所属上级集团名称两层，不仅弥补了Derwent机构代码信息不全面的问题，也使专利权人

表1 专利权人所属机构类型

类型序号	专利权人所属机构类型
1	科研机构
2	企业
3	推荐为科研机构
4	推荐为企业
5	无

的分析维度更全面和深入。另外，平台还为专利权人权威控制表提供了人工辅助的更新维护功能。专利权人权威控制清洗及示例如表2。

3.3 专利分析指标

表2 专利权人权威控制清洗示例

机构名称原始信息	专利数量	二级权控（独立法人）	一级权控（所属集团）	专利总量
MANDO MACHINERY CORP LTD	29	Mando Machinery Corporation	Mando	72
MANDO MACHINERY CO LTD	21			
MANDO MACHINERY CORP	3			
MANDO CORP	18	Mando Corporation		
MANDO KIKAI KK	1	MANDO KIKAI KK		

本文平台上已实现的专利分析指标分为基本指标分析和复杂网络分析两类，如图3。其中基本指标分析主要包括特定技术领域专利的整体趋势分析、技术发展趋势分析及申请机构分析等，可满足领域专利分析的基本需求。

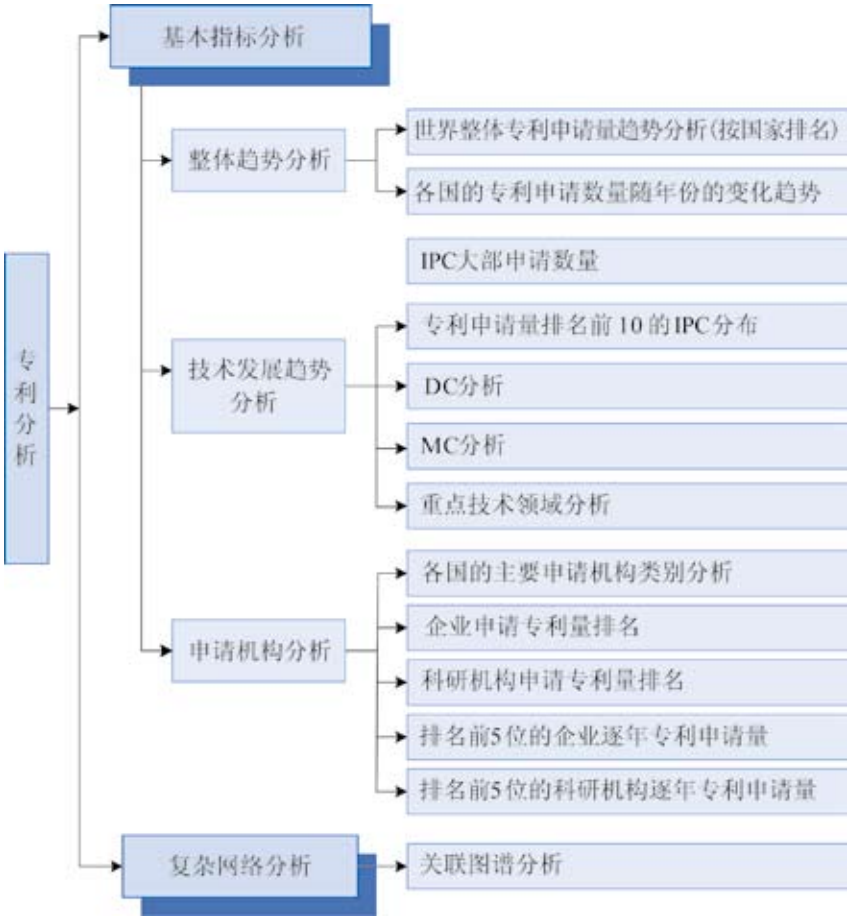

图3 平台专利分析指标内容

4 实证研究：以新能源汽车技术领域为例

本研究在进行了平台架构设计与功能模块的代码开发后，选择了新能源汽车技术领域作为实证研究领域，以测试平台的可用性。图4为开发完成的“专利特色资源服务平台”主界面。

4.1 数据来源

新能源汽车技术领域专利数据采集来源为Derwent数据库，数据时间范围为入库时间为1963-2009年的专利数据，检索日期为2009年4月。数据采用了关键词与Derwent手工分类代码结合的检索方式得到。

在使用本平台的数据获取、清洗与分析等功能集成上，以中国新能源汽车技术领域分析为例，演示平台对专利分析的支持功能。图5为新能源汽车技术领域的数据获取界面。

4.2 分析结果展示

图6-图8中显示的数据图表都在本文的专利特色资源服务平台上生成，分别显示了新能源汽车技术近10年在各国的专利申请情况、中国新能源汽车技术近10年IPC8大部类专利申请情况、中国新能源汽车技术领域近10年各类专利权人申请专利比例等分析结果。由图显示，在平台分析界面上，用户不仅可通过自主选择国别、年份等参数快速获得目标分析结果，还可在同一界面上完成与分析原始数据的链接、

图4 “专利特色资源服务平台”主界面

图5 新能源汽车技术领域的数据获取界面

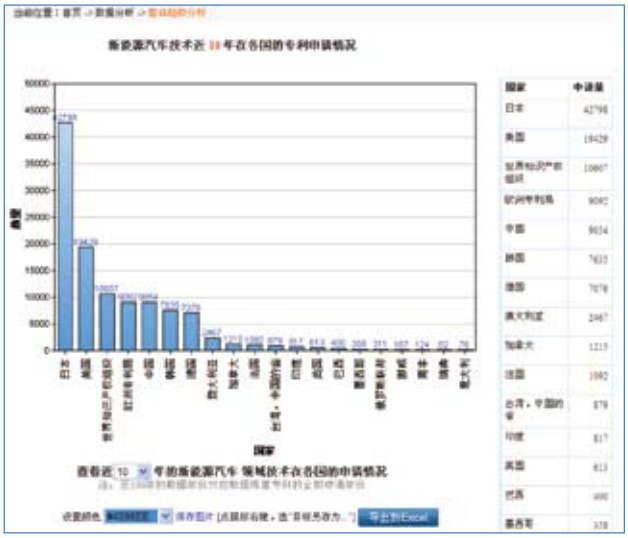

图6 新能源汽车技术近10年在各国的专利申请情况

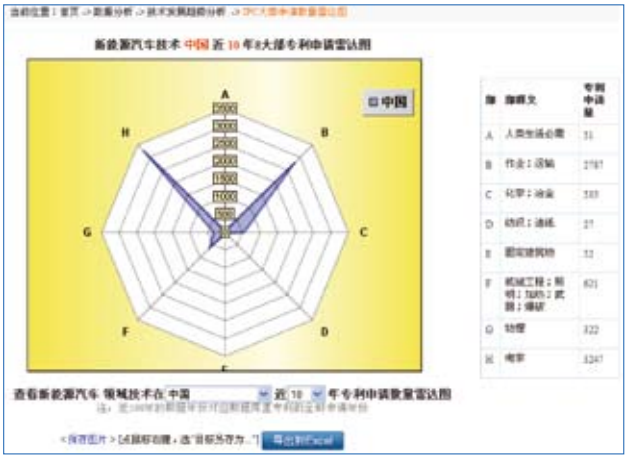

图7 中国新能源汽车技术近10年IPC8大部类专利申请情况

图8 中国新能源汽车技术领域近10年各类专利权人申请专利比例

分析结果的文本数据下载、分析图形下载等工作。

对新能源汽车技术领域专利分析的实证研究表明，专利特色资源服务平台在专利获取速度、自动清洗效果、分析结果展示及相关数据下载等方面都可有效地快速支持领域专利分析工作的开展。

5 总结

本文“面向深度分析的领域专利信息特色资源服务平台”的研究工作探索了“专利信息+分析工具+

服务平台”的科技管理决策服务模式。在下一步的研究中，将重点研究专利文本内容的处理和语义表示方法与技术，并将相关研究成果以软件工具形式集成到平台上，不断扩充其分析功能。

参考文献

[1] 王朝霞. 专利知识获取及其支持概念创新设计的方法研究[D]. 浙江大学,2009.
[2] Science - Thomson Reuters [EB/OL]. [2010-04-05]. http://scientific.thomson.com.
[3] Delphion Research intellectual property network - international and US patent search database [EB/OL]. [2010-04-05]. http://www.delphion.com.
[4] WinsLab [EB/OL]. [2010-04-05]. http://www.winslab.com.
[5] VantagePoint - Text Mining software for Technology Management - Search Technology, Inc. [EB/OL]. [2010-04-05]. http://www.thevantagepoint.com.
[6] Wisdomain [EB/OL]. [2010-04-05]. http://www.wisdomain.com.
[7] STN International: STN AnaVist [EB/OL]. [2010-04-05]. http://www.stn-international.de/stn_anavist.html.
[8] M•CAM, Inc. [EB/OL]. [2010-04-05]. www.m-cam.com.
[9] BioWisdom Ltd. [EB/OL]. [2010-04-05]. www.biowisdom.com.
[10] 中国知识产权网[EB/OL]. [2010-04-05]. http://www.cnipr.com.
[11] 北京恒和顿创新科技有限公司网站[EB/OL]. [2010-04-05]. http://www.all-patent.com.
[12] 保定市大为计算机软件开发有限公司[EB/OL]. [2010-04-05]. http://www.daweisoft.com.
[13] 北京东方灵盾科技有限公司[EB/OL]. [2010-04-05]. http://www.eastlinden.com.
[14] 张静,刘细文,柯贤能,黎江. 国内外专利分析工具功能比较研究[J]. 情报理论与实践,2008,31(1):141-145.
[15] 翟东升,王明吉,余旸. 基于专利地图和Multi-Agent思想的专利分析系统构建[J]. 情报学报,2006(3):316-321.
[16] 王曰芬,张旭,邬尚君. 在线专利分析软件的总体架构[J]. 现代图书情报技术,2008(10):48-53.
[17] 颜端武,张秀梅,邬尚君. 在线专利分析软件的应用:企业技术创新性与竞争性分析[J]. 现代图书情报技术,2008(12):66-72.
[18] WANNER L, et al. Towards content-oriented patent document processing [J]. World Patent Information, 2008 (30): 21-33.
[19] SHIH M-J, et al. Discovering competitive intelligence by mining changes in patent trends[J]. Expert Systems with Applications, 2010 (37): 2882-2890.
[20] KIM Y G,SUH J H, PARK S C. Visualization of Patent Analysis for Emerging Technology[J]. Expert Systems with Applications, 2008 (34): 1804-1812.
[21] 张运良,桂婕,朱礼军,乔晓东. 中文专利深度内容标引规范研制[J]. 数字图书馆论坛,2008(11):18-21.
[22] 赵蕴华,桂婕,张运良,朱礼军,姜彩红. 基于深度标引的专利文本挖掘框架研究[J]. 数字图书馆论坛,2008(11):1-5.

作者简介

桂婕（1976-），博士，助理研究员，研究方向：专利分析和科技创新管理。通讯地址：北京市复兴路15号信息技术支持中心 100038。E-mail: guij@istic.ac.cn
乔晓东（1965-），硕士，研究员，研究方向：信息服务和信息资源管理。通讯地址同上。E-mail: qiaox@istic.ac.cn
朱礼军（1973-），博士，副研究员，研究方向：知识组织。通讯地址同上。E-mail: zhulj@istic.ac.cn
张兆锋（1979-），硕士，工程师，研究方向：信息系统和信息可视化。通讯地址同上。E-mail: zhangzf@istic.ac.cn
李鹏（1979-），硕士，助理研究员，研究方向：智能信息处理。通讯地址同上。E-mail: lipeng_cn@istic.ac.cn

Construction of Domain Patent Information System for Deep Patent Analysis

Gui Jie, Qiao Xiaodong, Zhu Lijun, Zhang Zhaofeng, Li Peng / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Patent literatures are regarded as the important information source of technology, economy and law, also serve for decision-making in technological innovation management. However, three key problems limit the development of patent analysis, which are involved in integration of patent information, exposure of patent text contents and usage of advanced IT technologies. Facing the problems, we propose the idea of constructing the domain patent resource system based on deep patent analysis, which would serve as decision-making in key scientific and technological fields. The paper explores the basic methods and processes to realize the system.

Keywords: Integration of patent information, Patent database, Decision-supporting

（收稿日期：2010-04-09）