

·理论探索·

基于文本挖掘的专利分析工具的比较研究

葛小培¹ 孙涌^{1,2} 马玉龙³ 陈祥荣¹
(1. 苏州大学计算机科学与技术学院, 江苏 苏州 215006;
2. 江苏省计算机信息处理技术重点实验室, 江苏 苏州 215006;
3. 空军第一航空学院计算机教研室, 河南 信阳 464000)

〔摘要〕本文根据文本挖掘技术、可视化技术、优势及局限性等特征, 对国外 12 种典型的专利分析工具进行了比较研究, 同时对这些工具的适用性进行了评价并提议哪些用户组可以从中受益。本文所讨论的工具实例包括成熟的文本挖掘软件包, 一些较简单的全文搜索工具, 以及少数集成了更多成熟的软件包和全文搜索功能的数据可视化工具。希望能为专利分析人员选取适当的工具及软件公司改进专利分析工具提供参考。

〔关键词〕专利分析工具; 文本挖掘; 聚类; 统计分析; 语义分析
〔中图分类号〕G255.53 〔文献标识码〕A 〔文章编号〕1008-0821(2010)04-0008-04

Study on the Foreign Patent Analysis Tools in Text Mining

Ge Xiaopei¹ Sun Yong^{1,2} Ma Yulong³ Chen Xiangrong¹
(1. School of Computer Science & Technology, Soochow University, Suzhou 215006, China;
2. Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou 215006, China;
3. Office of Computer Teaching and Researching, First Aeronautical Institute of Air Force, Xinyang 464000, China)

〔Abstract〕Twelve typical foreign patent analysis tools were presented in this paper to provide a comparison of text mining capabilities visualization technology, perceived strengths potential limitations output of results etc, while the applicability of these tools were evaluated and suggestions as to which user groups may benefit from it. Summarized some valuable information for patent information analysts to choose proper tools and for software analysis tool provider to improve these tools.

〔Key words〕patent information analysis tool; text mining; clustering; statistical analysis; semantic analysis

1 专利分析工具比较

本文选取了 12 种国外主要的专利分析工具或专利分析技术提供商, 分别是 ClearForest^[9]、Goldfire Innovator^[7]、In-sight^[8]、TEMIS^[9]、Quosa^[10]、RefViz^[11]、STN AnaVist^[12]、VantagePoint^[13]、Aureka^[14]、M - CAM Doors^[15]、Wisdomain^[16]、Derwent Analytics^[17]。根据其处理不同结构的数据将这些工具分为三类: 第一类主要处理非结构化数据, 第二类主要处理结构化文本, 第三类用于处理混合数据(包括结构化和非结构化)。并从工具的类型、性能、数据源、输出结果、优势及局限性等方面具体讨论比较。

1.1 第一类工具

此类工具主要用于处理如专利全文、邮件、新闻、期刊论文、网页等非结构化数据, 包括: ClearForest、Goldfire Innovator、In-sight、TEMIS。

1.1.1 ClearForest

ClearForest 是一个专利分析工具, 具有文本挖掘功能。应用时从非结构化文本(如新闻报道、网络调查和 HTML 文档)中抽取相关特征项, 然后自动分类并导入到分类系统。该工具实现了从分类系统搜集到信息抽取之间关系的可视化, 这样可以通过把大量基于全文内容的信息转化为积极的商业情报来做出更好的商业决策。

其优点是在文本分析过程中使用配套的工具将大量的

收稿日期: 2010-01-20
基金项目: 国家自然科学基金项目(项目编号: 60873116); 教育部科学技术研究重点项目(项目编号: 207040)
作者简介: 葛小培(1982-), 女, 硕士研究生, 研究方向: 智能信息处理技术, 专利分析。

非结构化数据转化为可操作利用的结构化数据。

1.1.2 Goldfire Innovator

Goldfire Innovator 是一个文本挖掘工具, 使用复杂的语义分析技术, 可以将非结构化的文本转换为可检索的语义索引。还提供基于关键字概念的静态分类、动态文件摘要, 能够方便地进行知识跟踪、竞争情报分析、技术趋势分析和引文分析。

该工具操作复杂且费用极高, 为此限制了很多人对它的使用。

1.1.3 Inxight

Inxight 是一家专注于数据可视化、信息检索及自然语言处理的软件公司, 企业数据大约有 80% 都是非结构化数据, 主要是非结构化的文本、元数据和实体, 使用 NLP 进行全文抽取, 文件则根据预先确定好的层次分类标准进行分析。

其优点是同时提供多个在线数据库检索, 且拥有该领域最强大的语言算法, 能提供 32 种语言^[2]。但在定义分类时要求太高, 一般很难 (如医疗、科学和新闻) 被实际使用。

1.1.4 TEMIS

TEMIS^[2] 采用语义分析的方案, 其功能包括抽取、分类和聚类。该工具整合了 Insight Discover Extractor、Insight Discover Categorizer、Insight Discover Cluster、a multilingual engine XeLDATM and Skill Cartridges。文本首先经过多语种引擎 XeLDA 处理后转换成基于上下文的单词, 然后 Discover Extractor 使用 Skill Cartridges 对这些单词进行抽取。该工具可以对任何格式的文件做分析操作, 其输出结果包括聚类、排序、引用文献列表等。另外, 还提供了一个化学文档浏览器, 此浏览器有专门抽取模块, 可以把化学名称转换成化学结构。

TEMIS 不足之处是缺乏多样化的可视化选择方法, 成本高可能也是限制其使用另外一个因素。

1.2 第二类工具

这类工具主要处理数据库中存储的专利扉页和参考文献信息等结构化数据, 包括 Quosa、RefViz、STN-AnaVist、VantagePoint。

1.2.1 Quosa

QUOSA 是查询 (Query)、组织 (Organize)、储存和共享 (Save & Share) 及分析 (Analysis) 这几个单词首字母的组合, 是一个基于概念抽取和聚类的文本挖掘工具, 提供全文分析, 其统计分析包括特征项抽取和词频排序, 概念抽取使用由麻省理工学院和哈佛大学的数学家改进的动态抽取算法。该工具的管理界面划分为 3 个区: 检索浏览区 (Browser Pane), 检索结果区 (Results Pane), 文献管理区 (Organizer Pane)。其优点是全文搜索、批量下载大量的期刊文章和专利全文。文本还可下载到 EndNote 作进一步研究。但其

在专利信息源方面的应用不如非专利信息源有优势。

1.2.2 RefViz

RefViz 是 OmniViz 公司的产品, 但只有 OmniViz 的部分功能, 是专门为参考文献的引用而设计的。该软件用于统计和语言分析, 只适用于结构化数据, 采用 K-means 聚类算法进行有结构数据聚类, 选用 Galaxy 视图输出结果^[3-4]。其主要特征是引用 Retriever (可用于多个在线资源同时进行检索)。

其优点是对参考文献信息的后期处理, 文件按主题内容为索引组织分类。

1.2.3 STN AnaVist

STN AnaVist 根据频率进行分析, 提供多种分析方式的可视化。它遵循 K-means 等常规聚类算法^[2], 采用 Research Landscape 视图来表示聚类结果。

目前的 STN AnaVist (2.0 版) 仅用于处理 CAPLUS 数据库、PCTFULL 全文数据库、USPATFULL 全文数据库, 近年来更多的是处理 Derwent DWPI 下载的参考文献。虽然它处理参考文献信息的结果很好, 但对全文信息的文本挖掘功能并不显著。

1.2.4 Vantage Point

Vantage Point 主要通过模型匹配、基础规则和自然语言处理 (NLP) 针对题录数据库数据进行文本挖掘, 该工具提供数据清洗和整理, 允许用户创建用户管理词典。其最佳工作环境是几百条数据, 当然也可以对几千条数据加以分析, 并提供多维分析。

Vantage Point 在提取数据后采用 factor map 和 cross-correlation map 模型进行聚类^[3-4], 并允许用户只标记出文档中的一部分用于聚类。该工具支持对 VBscript 的扩展, 它添加了 5 个 VB 类以及 50 余个 VB 方法, 用户可编辑 VBscript 脚本, 实现一些特殊的定制分析功能^[4]。

1.3 第三类工具

此类工具主要处理混合数据, 包括 Aureka、MCAM Doors、Wisdomain、PatAnalyst 和 Derwent Analytics, 都提供本地化专利数据库。

1.3.1 Aureka

Aureka 是专利文本挖掘和可视化领域的先驱, 允许组织和知识产权管理。该文本挖掘功能是基于关键字和统计分析。该平台提供本地专利数据库, 数据范围包括 US、EP、PCT、JP (英文), 专利数据定期更新^[5]。

平台中的 ThemeSape software 文本分析软件依据专利 a 样本中的相关词频, 应用聚类分析生成词汇地图来描述专利技术主题分布情况; 引用树软件 (Aureka citation tree) 利用专利引证信息构建双向引证树, 形象化的显示出研究对象引用在先专利和被其后专利引证的信息; 报告工具 (Aureka Reporting Tool) 可提供 3 种类型的报告: 一是信息摘要; 二是详细文本式报告; 三是图表式报告。

采用ThemeScape 绘制技术地图时，用户不能加入同义词词组或希望关注的主题概念是平台的一个缺陷。此外，引用树软件只显示直接引用关系，无法显示整个引用链，且只适用于美国专利，及系统使用费十分昂贵都制约了它的使用和发展。

1.3.2 M—CAM DOORS

M—CAM DOORS^[2]采用语义分析和引证分析相结合的方法进行技术分析，最大特色在于语义分析方面。系统不仅能检索相同主题的专利文献，还能借助于系统自建的同义词库自动搜索与主题相似的专利文献。用户可通过 M—CAM 网站进行专利分析，寻找在先技术获得专利情报。从整体上看，它主要侧重于专利和技术的侵权分析。

分析用的数据仅限于美国专利数据，同时图表的选择范围不多，且需要同义词库来支持，这增加了运行成本和用户使用成本。

1.3.3 Wisdmain

Wisdomain 公司研发了 Focust 和 PatentLab—II 等产品进行专利检索和分析，数据范围包括 US、EP、JP、PCT 和 IN-PADOC 的专利数据^[1]。

PatentLab—II 为客户提供专利数据可视化分析，其主要功能如下：提供几种类型的 HTML 报告格式、Analysis Wizard 为用户提供操作简单功能强大的专利分析、Patent Viewer 方便用户随时查看专利全文、Charts&graphs 增强了用户可视化能力，此外还提供在线的 IPC/UPC 进行对照显示。

FOCUST 由检索模块、引文模块和分析模块组成。检索模块有灵活的检索入口，用户在一个界面下方方便的检索 US、EP、JP、PCT 和 WIPO 的专利数据；引文模块向用户展

示了被分析专利的所有引证信息，同时显示引用专利的历史，它提供了 3 种引用树图形，即向前引用、向后引用和间接引用^[1]；分析模块提供了诸如文本挖掘分析、高级可视化技术分析及灵活多样的专利文件管理。

Wisdomain 使用自己收集的专利进行分析似乎不允许其它数据源中的数据集链接到该软件。与 Aureka 一样，Wisdomain 引文分析仅限于美国专利。

1.3.4 Derwent Analytics

Derwent 公司是目前世界上集专利信息的收集、加工、整理和分析，并提供全球性服务的大型专利信息供应商之一^[3]。

Derwent Analytics 采用自然语言加工技术，对 Derwent 标引专家加工的专利文献标题中最经常出现的名词词组进行提取和分组，并结合关键技术领域专利活动活跃的专利权人、国家和时间等因素，组成各种图表，来反映某技术领域关键技术的变化、R&D 投入状况以及新的技术空白点等。其依托 VantagePoint 作为技术支持，用户界面友好、直观，用户可以方便地从 Derwent 世界专利索引数据库中提取有价值的信息和重要的商业情报。

Derwent Analytics 可为情报专家、法律专家、科研人员和人力资源专家等提供可靠的决策依据，同时为企业提供相关的竞争情报。

2 专利分析工具的比较结果

根据前文的分析介绍，将 12 种专利分析工具的比较结果总结如表 1、表 2 所示：

表 1 12 种工具比较结果

工具名称	类 型			性 能						潜在用户				输出结果						
	文本挖掘	数据可视化	自建数据库	语义分析	金融风险分析	自然语言处理	统计分析	关键字分析	概念抽取	多语言分析	商业情报	科研人员	情报专家	法律或专利部门	列表直方图饼图	矩阵图	摘要	分类图	聚类图	引证图
ClearForest	✓	✓		✓				✓		✓				✓	✓		✓	✓		
GoldfireInnovator	✓	✓	✓	✓					✓			✓					✓	✓		✓
Inxight	✓	✓		✓		✓		✓	✓	✓	✓							✓		
TEMIS	✓	✓		✓		✓		✓			✓	✓			✓			✓	✓	✓
Quosa	✓	✓					✓		✓	✓		✓								✓
RefViz	✓	✓					✓			✓			✓			✓			✓	✓
STN AnaVist		✓	✓				✓				✓	✓	✓		✓	✓	✓		✓	
Vantage Point	✓	✓				✓					✓		✓		✓	✓	✓	✓	✓	
Aureka	✓	✓	✓				✓	✓	✓		✓	✓	✓	✓	✓				✓	✓
M—CAM Doors		✓	✓	✓	✓				✓	✓	✓		✓	✓						✓
Wisdomain	✓	✓	✓				✓	✓	✓			✓	✓		✓					✓
Derwent Analytics	✓	✓	✓			✓		✓			✓		✓	✓	✓				✓	

从表 1 中可以看出，由于第一类工具主要处理非结构化数据，所以在语义分析、自然语言处理、概念抽取及分

类分析方面突显了其明显的优势，主要是为企业和科研人员提供服务；第二类工具主要处理结构化数据，其优势在

与其强大的统计分析和聚类分析功能，主要是为科研人员和情报专家的研究提供服务；第三类工具可以处理混合数据，其优势主要集中在关键字分析和概念抽取方面，该类工具拥有自建数据库，这使得分析的功能更强大、更全面，在性能和输出结果列中充分显示了这一点。例如，能

为法律或专利部门提供专门的服务，这是前两类工具所不具备的独特功能。另外，引证分析也是其一大特色，从引证分析的深度来看 Wisdomain 要优于 Aureka，不过都只能对美国专利进行引证分析。

表 2 12 种工具比较结果

工具名称	数 据 源	优 势	局限性
ClearForest	网站，专利文献等的非结构化数据	抽取模型	预期风险投资
Goldfire Innovator	超过1 500万项的全球专利库、9 000个科学效应库、3 000个专业技术网站以及语义索引的企业和个人知识库等的非结构化数据	高级语义分析	需要内部培训，高成本
Inight	网站、内部数据库等的非结构化数据及抽取和识别预处理文件中的元数据和实体类型	抽取和联合检索	生命科学领域实际应用时需要输入更具体的分类
TEMIS	网站、专利文献、临床试验、邮件、生物信息学等的结构化和非结构化数据	采用专门的 Skill Cartridges 工具进行抽取	缺乏多样化的可视化选择方法，高成本
Quosa	PubMed、Ovid、NCBI、Google Scholar and USPAT-FULL 及专利摘要等的结构化和非结构化数据	全文检索和管理	主要应用与非专利文献源，使用其它数据源需要添加配置
RefViz	Web of Science、PubMed、Ovid、OCLC、Library of Congress and Purdue University 等的结构化数据	文献资料的后期处理	主要应用与非专利文献源，用其它数据源时需要添加配置
STN AnaVist	CAPLUS、USPatful、PCTFull、DWPI 等的结构化数据	标准化的词汇库，同义词词典	提供的数据源相对少
Vantage Point	文献著录项目领域的结构化数据	应用于技术和公司评估的分析工具模块	大量数据集的清理列表难以实现
Aureka	MicroPatent 数据库	专利地图，聚类，引证分析	识别 Themescape 符号有困难，引证分析仅用于美国专利
M－CAM Doors	88 个官方专利局网站及期刊	专利单值性和实施性分析	只专注于专利信息，高成本
Wisdomain	US、EP、PCT、PAJ、CN 和韩国的摘要，国际专利文献中心的法律状态	高相关性引证分析	引证分析仅用于美国专利，不提供数据导入功能
Derwent Analytics	世界专利数据库（收录了全球 40 多个专利机构的3 000多万个专利）	引文链接，快速建立用户自定义的词表，功能全面	不提供引文分析，限制访问权限，高成本

表 2 主要针对这 12 种工具的数据源、优势及潜在的局限性方面进行了比较。这 3 类工具大都可以分析专利文献和非专利文献，都具有其优势和不同之处，不能相互替代。比如，有侧重于非专利文献的分析工具，如 Quosa 和 RefViz；有侧重于专利文献的分析工具，如 M－CAM Doors；有侧重于生命科学领域分析工具，如 Inight 和 TEMIS；也有侧重于全行业的专利分析工具，如 Aureka、Wisdomain、Derwent Analytics、Goldfire Innovator 等。

总体而言，这 3 类工具都具有文本挖掘和可视化功能，是诸多专利分析工具中的佼佼者。虽然这些工具都具有很强大的功能，分析很深入，但没有一种工具能包含表 1 所列出的所有功能，聚类分析的准确性、可视化程度及引证分析的深度有待于进一步提高。

3 结束语

根据调研可知，第一类工具在处理广泛的数据源和使用先进的语义技术方面最灵活，但需要内部培训或预投入大量的时间、精力和成本。第二类工具主要是处理结构化数据，因此，可用于优化参考文献信息。第三类工具用于处理混合数据，功能更强大、更全面，但其局限性是依赖性，例如，只提供美国专利的引文分析。

由于每种工具是针对不同的用户群体而设计的，提供不同方面的文本挖掘和可视化。所以合并两个或更多的这些工具将会是比较理想的选择，可以利用到他们最佳的功能，如基于语义分析数据可视化。

另外，目前为了拓展专利分析工具的功能，越来越多的开发商将自然语言处理、人工智能等领（下转第 161 页）

果他们的法律知识不过关的话，那么农民的知情权将直接成为大打折扣的权利；三是通过农民夜校、职校和农民电视、电台频道进行农村法制教育，使农民懂得法律赋予每个公民的知情权是什么，如何行使它，权利受到侵害后，救济的办法和保障在哪里。从而在农村大力营造学法、守法，依法办事、依法行政的良好氛围。

3.4 其他保障

除了要由政府、传媒和社会 3 个层面来对农民知情权进行保障之外，还可以从农村这个层面通过村务公开和村委会会议公开等来保障农民的知情权，如（1）在村中的大道旁树立公示牌，公示牌上公示的是国家相关政策，自然也就成了农民手中的“盾牌”^[15]；定期免费向村民发放明白纸，在明白纸上公布村财务、村集体经济合同、宅基地申报情况、“一事一议”筹资酬劳情况、种粮直接补贴审批结果、村级民主决策等情况^[16]。（2）村委会设立一块村务公开栏，把村中的重要事务包括财务情况、村民需求和村委会会议等全部内容公布在上面，而且要定期更新。

参 考 文 献

[1] 杨耕身. 农民民主权利和主体地位是新农 [EB/OL]. <http://www.ayxdj.gov.cn/ReadNews.asp?NewsID=8327>, 2009-10-22.

[2] 陈丽平. 保障农民知情权的成功实践 [N]. 法制日报, 2007-12-17 (1).

[3] 黄琼枝. 论公务员之守密义务与国民的知情权 [A]. 法学研究所硕士论文 [C]. 台北: 中兴大学, 1986.

(上接第 11 页)

域的创新应用于专利分析。本计算机信息处理技术重点实验室也于 2005 年开始从事基于 deep web、本体、语义及集群计算技术等于一体的生物医药领域专利信息服务平台的研究，2010 年预期投入试用，由于篇幅有限，将另文详述。

参 考 文 献

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

[2] YunYun Yang, Lucy Akers Thomas Klose. Text mining and visualization tools [J]. World Patent Information, 2008, (30): 280-293.

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

[4] 刘佳佳, 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): 67-74.

[5] 张静, 刘细文, 柯贤能, 等. 国内外专利分析工具功能比较研

[4] Right to know—Wikipedia [EB/OL]. http://en.wikipedia.org/wiki/Right_to_know, 2009-07-15.

[5] 姚小林. 知情权的法理分析 [EB/OL]. <http://www.cngdsz.net>, 2009-10-20.

[6] 朴再玉. 关于构建我国公民知情权保障体系之若干设想 [D]. 延吉: 延边大学, 2005.

[7] 蒋永福. 信息获取自由及其限度 [M]. 北京: 社会科学文献出版社, 2007: 103.

[8] 全国文化信息资源共享工程介绍 [EB/OI]. <http://www.ccnl.com.cn/html/gxgc/index2.htm?file=01-01>, 2009-06-26.

[9] 赵炳林. 农民的知情权应由司法保障 [N]. 农民日报, 2005-04-09 (3).

[10] 杨静, 周志新. 和谐社会构建中公民知情权的制度保障 [J]. 时代经贸, 2008 (2): 75.

[11] 汪习根, 陈焱光. 论知情权 [J]. 法制与社会发展, 2003, (2): 62-74.

[12] 陈叶兰. 农民的环境知情权、参与权和监督权 [J]. 中国地质大学学报: 社会科学版, 2008, (6): 18-22.

[13] 安泉林. 当前我国农民权利缺失问题及保障措施创新研究 [D]. 石家庄: 河北师范大学, 2007.

[14] 左群, 张晓楠. 试论社会主义新农村建设进程中农民权利保障的几个问题 [J]. 特区经济, 2008, (12): 162-164.

[15] 朱国亮. 公示牌还给农民知情权 [EB/OL]. <http://www.gs.xinhuanet.com/old/gansu/news-shehui/sh-2427.htm>, 2009-10-21.

[16] 刘丽娜. 公开什么内容农民说了算 [N]. 青岛日报, 2009-10-04 (1).

究 [J]. 情报理论与实践, 2008, (1): 141-145.

[6] <http://www.clearforest.com/Technology/TechnologyOverview.asp> [EB]. 2009-10-12.

[7] <http://www.invention-machine.com> [EB]. 2009-10-12.

[8] <http://www.inxight.com/products/sd-es> [EB]. 2009-10-16.

[9] <http://www.temis.com/index.php?id=59&sel=1> [EB]. 2009-10-16.

[10] <http://www.quosa.com/solutions.html> [EB]. 2009-10-20.

[11] <http://www.refviz.com/rvinfo.asp> [EB]. 2009-10-22.

[12] <http://www.stn-international.de/stninterfaces/stnnavist/stn-anavist.html> [EB]. 2009-10-24.

[13] <http://www.thevantagepoint.com> [EB]. 2009-10-28.

[14] http://aureka.micropat.com/7w/html/7w_default.asp [EB]. 2009-10-28.

[15] <http://www.m-cam.com/doors> [EB]. 2009-11-01.

[16] <http://www.wisdomain.com/Overview.htm> [EB]. 2009-11-01.

[17] <http://scientific.thomson.com/support/faq/analytics> [EB]. 2009-11-01.