信息管理
新视野论丛

JI YU ZHUAN LI WEN XIAN DE

基于专利文献的本体构建及应用

BEN TI GOU JIAN JI YING YONG

谷 俊 著

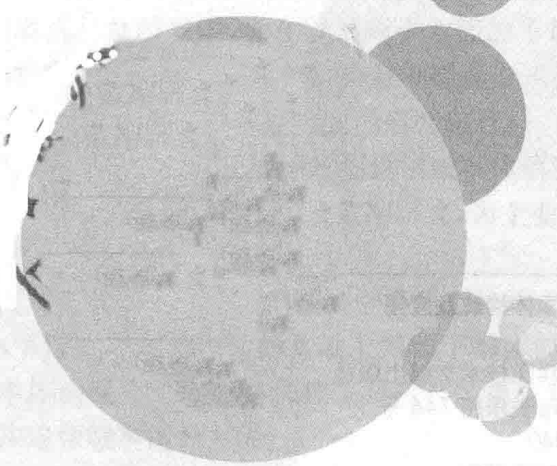
上海科学技术文献出版社
Shanghai Scientific and Technological Literature Press

JI YU ZHUAN LI WEN XIAN DE

基于专利文献的本体构建及应用

BEN TI GOU JIAN JI YING YONG

谷 俊 著



上海科学技术文献出版社
Shanghai Scientific and Technological Literature Press

图书在版编目（CIP）数据

基于专利文献的本体构建及应用 / 谷俊著．—上海：上海科学技术文献出版社，2017
ISBN 978-7-5439-7243-8

Ⅰ.① 基… Ⅱ.①谷… Ⅲ.①专利文献—研究 Ⅳ.① G306.4

中国版本图书馆 CIP 数据核字 (2016) 第 281865 号

责任编辑：徐　静
封面设计：徐　炜

基于专利文献的本体构建及应用
谷　俊　著
出版发行：上海科学技术文献出版社
地　　址：上海市长乐路 746 号
邮政编码：200040
经　　销：全国新华书店
印　　刷：常熟市华顺印刷有限公司
开　　本：787×1092　1/16
印　　张：11.5
字　　数：279 000
版　　次：2017 年 5 月第 1 版　2017 年 5 月第 1 次印刷
书　　号：ISBN 978-7-5439-7243-8
定　　价：58.00 元
http://www.sstlp.com

前 言

本体是一种有效的知识组织方式，被纳入语义网体系，因其具有明确性、形式性和共享性三大特征，可以在网络资源上融入计算机理解的信息，达到资源的语义理解，是语义层面上网络信息的交换与共享的基础。它将 WEB 资源通过语义的方式组织起来，使得互联网的资源获取更加便捷，是在因特网上提供高效服务的先决条件。目前，本体在人工智能、信息检索、知识工程、数据挖掘等学科领域中被广泛研究和应用。

科学技术日益发展的今天，专利文献作为一种披露最新技术的公开信息来源，被越来越多的科研工作者所认识和利用。据世界知识产权组织统计，专利信息包含了世界上 90%～95%的技术信息，并且技术信息的公开要比其他载体早 1～2 年；在所有研发活动中，专利是最有力的工具；在世界研发平均产出中，与其它活动相比，专利经济价值超过了 90%。

在海量的专利文献中获取需要的专利主要依赖于专利检索系统，但是就目前已有的专利检索系统来看，大多数基于结构化数据库、基于单纯关键词匹配的检索方式，不能有效地满足用户快速、准确获取知识的需求。总的来说，存在以下两个方面的不足：①检索方式简单化，仅提供基于关键词匹配的检索方式，无法提供推理，导航等检索方式；②难以获取隐含知识，主题之间的关联揭示不充分，无法准确把握当前行业的研究热点。

针对专利情报工作中遇到的这些问题，本书尝试通过构建领域本体模型，以达到提高检索效率的目的，并在此基础上，实现基于专利文献的相关监控研究，为企业的研发战略和知识产权战略提供服务。

(1) 本体构建相关理论研究　领域本体的构建一直是学界研究的热点问题。为了了解前人在本体构建方面的做法，本书首先在文献调研的基础上回顾了本体构建的基本理论方法，并对已有的本体描述语言、本体构建方法和本体构建工具进行了简要对比分析，尝试找出最适用于利用中文专利文献构建领域本体的方法。

(2) 本体半自动构建方法研究　为了对本体半自动构建方法进行深入研究，以国际专利分类号为 C21 的中文专利实验数据，提出了本体半自动构建相关技术，包括文本中术语的抽取技术、术语间关系获取技术、本体形式化等技术。

利用专利文献构建本体的第一步便是从专利文献中抽取相应的技术术语。首先尝试在 ICTCLAS 词典分词的基础上，利用串频最大匹配算法从中文专利文本中抽取候选术语，再利用 TFIDF 算法得到相关特征项的权重，经过筛选后得到最终概念术语，完成对专利文献中技术术语的抽取，为后续的本体构建奠定基础。

在抽取术语结果的基础上，为了实现本体中术语分类关系的构建，针对 K－Means 聚类

算法难以确定初始类中心点的问题，利用蚁群聚类方法进行初始聚类，确定初始类中心点，再通过 K－Means 聚类算法对初始聚类的结果进行进一步分层聚类，并结合术语综合相似度计算的方式提取每个类的标签，实现了术语层次关系的构建和本体分类关系的初步搭建。

此外，简单分类关系的构建无法实现最终本体的搭建，因此，在术语抽取结果的基础上，通过关联规则来获取术语间的非分类关系。首先利用基于上下文的术语相似度获取方法得到术语间的相似度权重；再通过加入谓语动词的关联规则算法计算，结合搜索引擎技术得到候选关系对集合，通过置信度和支持度的对比分析，抽取非分类关系结果。

为了补充关联规则抽取非分类关系的结果，本书还在分析专利文献句法特点的基础上，提出利用规则匹配和 CRFs 算法结合的非分类关系抽取方法。首先对专利摘要进行拆分，形成独立子句，经过人工规则学习的结果与实验数据进行匹配，得到语料句式库；最后利用 CRFs 学习和测试模型对专利文献中的非分类关系进行抽取，经过进一步的关系抽取，完善了本体的非分类关系，从而完成最终本体的构建。

(3) 基于本体的专利检索及监控方法研究　本体构建的目的是为了提高专利情报工作的效率，提高检索和分析的准确性，进而提高专利监控的效果。因此，在分析专利文献描述特点的基础上，提出了适用于本体模型的专利检索、分析和监控技术。

首先，为了提高专利检索结果的查全率，扩大检索结果的范围，利用本体对检索词进行语义扩展，并将扩展后的检索词集合提交给 Lucene 搜索引擎进行检索，提高了检索的查全率。除了检索之外，传统的专利情报研究还需要对检索结果进行定量分析，因此，本书结合专利分析模型，实现了诸如专利权人排名、发明人排名、申请日排序、公开日排序、申请人技术发展对比、申请人技术领域对比、发明人技术领域对比等若干计算机自动分析方法，提高了分析结果的准确性。

专利分析和专利监控，除了对技术发展的宏观态势进行分析之外，还需要对专利技术本身进行监控和识别。为帮助政府和企业识别专利侵权问题，提出了相似专利检测方法：结合本体模型，利用 TFIDF 算法和 VSM 找出相似专利；再通过有序最长公共子序列匹配算法找出专利文献中相似的部分，并将其提取出来供情报人员参考。这一做法能够提高专利侵权判别的效率和准确性。

此外，为帮助企业寻找技术领域中的关键专利，提出了适用于计算机自动处理的重要专利评价体系和评价方法。评价体系中主要包括国际专利分类号数量、权利要求项数量、法律状态、专利存活期、专利家族规模、合作者数量、发明人数量等指标，经层次分析法和德尔菲调查法形成相应权重体系，通过对专利语义检索的结果进行逐条评价打分，达到重要专利自动抽取的目的。

为帮助企业情报部门监控行业最新技术进展，提出了新技术术语的识别方法。利用 ICTCLAS 分词系统和停用词表抽取文档词元，通过改进的 TFIDF 模型计算词元权重并筛选出热点词元；再通过词间距测算对热点词元按顺序进行组配，经权重计算和阈值筛选后得到候选新技术术语集，提供给领域专家验证，加快新技术术语识别的速度，有利于政府和企业监控行业发展最新动向。

最后，利用 Java 语言对专利检索及监控相关技术的相关方法进行了实现，设计并开发出专利监控原型系统，并为政府相关部门和企业提出了部分专利文献的本体构建和应用工作的意见和建议。

目 录

第 1 章

绪　论

1.1　现状及问题

随着中国经济的飞速发展，越来越多的企业认识到仅仅依靠低成本的劳动密集型运作方式已经不能适应新的需求，而通过技术创新提高企业的核心竞争力则逐渐成为企业发展的主导。专利文献作为技术信息最有效的载体，囊括了全球 90%以上的最新技术情报，相比一般技术刊物所提供的信息早 5～6 年[1]，而且 70%～80%发明创造只通过专利文献公开，并不见诸于其他科技文献；相对于其他文献形式，专利更具有新颖、实用的特征。因此，更多的企业把目光转向了潜在价值巨大的专利文献，希望通过专利文献的检索与分析工作，配合企业的战略和研发部门，提高研发起点，节约研发成本，预测技术发展动向，监控竞争对手，增加核心技术的储备，从而提高自身的竞争力。

1.1.1　专利检索技术及存在的问题

专利文献可以通过各个国家的知识产权部门免费检索获得，如中国知识产权局、日本专利特许厅、韩国知识产权局、美国专利商标局、欧洲知识产权局等。上述来源虽然可以免费获取专利文献，但是需要检索者熟悉各种专利数据库的检索方法，而且无法同时对多个数据库进行检索，检索起来比较费时费力。因此，一些数据供应商将世界上大部分国家公开的专利文献进行了整合，同时将不同语种的专利文献按同一语种进行了翻译，并提供统一的检索入口（如 Thomson Innovation[2]、Dialog 系统[3]、Total Patent[4] 等），提供收费服务。这些数据库拥有统一的检索语法和数据源，并按自己的规则对数据进行了标准化，大大提高了检索者的检索效率，得到了广泛应用。

专利文献的检索方式主要可分为基本检索、表格检索、布尔检索和号码检索：①基本检索，提供了一键式检索的环境，用户在单一的检索框中输入任意检索词进行检索；②表格检索，可以根据不同的著录项目（如专利名称、发明人、专利权人等）及其组合进行检索；③布尔检索，支持逻辑连接符（AND，OR，NOT 等）和通配符检索；④号码检索，支持各种专利号码

---

〔1〕邵波. 企业竞争与反竞争情报中的专利分析研究. 情报科学，2006，24(2)：235－238.

〔2〕Thomson Renters. Thomson Innovation [EB/OL]. [2011－05－17]. http://www.thomsoninnovation.com.

〔3〕ProQuest. Dialog 系统[EB/OL]. [2011－05－17]. http://www.dialog.com/products/dialogweb/.

〔4〕Lerisnexis. Totzl Patent [EB/OL]. [2011－05－17]. http://www.lexisnexis.com/totalpatent/.

(例如公开号、申请号、优先权号、国际专利分类号等)的检索。

以 Thomson Innovation 专利数据库为例。该数据库收录了全球主要国家的专利数据(见表 1-1),由领域专家对不同语种的专利题录信息进行了专业的英文翻译和改写,并辅以特有的 Derwent 手工代码进行标注,以提高专利文献的检索效果和可读性。

表 1-1 Thomson Innovation 专利数据库收录范围[5]

数据库名称	说 明	语 种	时间跨度
德文特世界专利索引库(DWPI)	由领域专家重新加工后的专利数据,收录了涵盖全球 44 个专利公开国,约 2 000 万个专利家族,超过 4 200 万件专利数据。	英语	1963 以来
世界知识产权组织专利数据库(WIPO)	收录了在世界知识产权组织申请公开的所有专利数据。	70% 英语 15% 德语 5% 法语 1% 西班牙语	1978 以来
美国专利数据库(US)	美国申请公开和授权公开专利数据库	英语	1836 以来
欧洲专利数据库(EP)	涵盖欧洲 31 个国家的申请公开和授权公开专利数据库	60% 英语 30% 德语 10% 法语	1978 以来
英国专利数据库(GB)	英国申请公开专利数据库	英语	1916 以来
法国专利数据库(FR)	法国申请公开专利数据库	法语	1971 以来
德国专利数据库(DE)	德国申请公开和授权公开专利数据库	德语	1983 以来
中国专利数据库(CN)	中国申请公开和授权公开专利数据库	英语	2007 以来
日本专利数据库(JP)	日本申请公开和授权公开专利数据库	英语	1971 以来
韩国专利数据库(KR)	韩国申请公开和授权公开专利数据库	英语	1978 以来
DocDB(INPADOC)	全球超过 60 个国家专利的法律状态信息	英语	—

但是专利的检索依然依赖于关键词匹配等传统检索方式,检索结果的好坏完全依赖于检索者个人的知识掌握程度和对问题分析的透彻程度,无法深度剖析专利内部知识以及知识之间的关系,因而无法满足检索者查全率和查准率的要求。总的看来,信息检索存在以下三个方面无法回避的问题[6,7]:

①“忠实表达”问题 检索者很难用简单的检索词或者检索词的组合来表达自己所需要检索的真正内容,导致检索困难。现在虽然有很多检索工具提供了“相似检索”的功能(例如 SooPAT 专利检索[8]),但是这些工具仍然停留在关键词匹配的简单功能上,无法实现对于用户检索词的语义理解。

②“表达差异”问题 人类的自然语言中,随着时间、地点和环境的变化,对于同一事物的描述是不一致的,这就导致了专利信息中对于同一技术或者概念的描述会发生一定的变化(例如,对于“钢管”的描述,在专利中会有“焊管”、“无缝钢管”、“油井管”、“OCTG”、

[5] Thornson Rellters. Patent Data [EB/OL]. [2011-05-17]. http://www.thomsoninnovation.com/ti/contentsets/patents/.

[6] 董慧. 基于本体论和数字图书馆的信息检索[J]. 情报学报,2003(6):648-652.

[7] 杜文华. 本体构建及其在数字图书馆的应用研究[D]. 武汉:武汉大学,2005.

[8] Soopat. SooPAT 专利搜索[EB/OL]. [2011-05-17]. http://www.soopat.com/.

“ERW”、“特殊扣”、“UOE”、“套管”等多种表达方式);同时,有些专利为了保护自己的利益,会在专利说明书中尽可能多地使用晦涩难懂的法律术语(例如“梯子”被描述成“攀登的工具”等),在这种情况下,检索者如果利用自己理解的关键词去进行检索,很可能遗漏了其他描述形式的内容,导致检索失败。

③ “词汇孤岛”问题 人们在检索时,除了希望获得与自己输入的检索词完全匹配的信息外,还希望能够获得与检索词相关的其他信息;但是传统的专利检索无法满足检索者的这一要求。其特点决定了检索结果一定是包含检索词的文档和信息,而用户的检索词无法得到扩展,造成了“词汇孤岛”现象,从而降低了查全率。

语义检索则可以把人工智能技术(AI)和自然语言处理技术(NLP)运用到专利检索中,通过构建相关领域本体,将领域概念和推理机制融入检索过程中,从语义层面理解用户的检索请求,并利用概念间的关系和推理规则进行辅助检索,力图从根本上解决传统检索中遇到的一系列问题。

1.1.2 专利分析技术及存在的问题

专利分析是在大量专利文献的基础上,通过定量和定性的分析,将零碎的专利信息转换为系统的认识,从而有助于企业制定策略、做出决策[9]。专利分析可以分为定性分析和定量分析,定性分析是指通过专利说明书、权利要求项、图纸等来识别专利,按技术特征来归并相关专利并使其有序化,一般用来获得技术发展路线、企业研发重点以及特定的权利要求等方面的情报。

定量分析是通过对专利文献外部特征(公开日期、专利权人、专利公开号等)进行统计统计分析,识别竞争对手的技术特点,揭示其技术优势、技术热点以及预测对手的技术发展方向等。例如,通过作者排名分析可以得出所关注领域的技术专家;通过国际专利分类号(IPC)的排名分析可以得出关注领域的技术热点;通过机构与年份的交叉分析,可以得出竞争对手的历年的技术研发投入程度和科研进展情况;通过机构的聚类分析,能够得到竞争对手的研发相似程度等。在上述分析的基础上,可以通过进一步的分析,得出最终结论。定性与定量分析是相辅相成的,在实际工作中往往结合使用[10]。

定性与定量分析的结合一直是专利分析工作者研究的课题,而专利地图的出现为两者的结合提供了纽带。专利地图由日本专利特许厅((Japan Patent Office, JPO)最先发明并使用。它为专利情报研究提供了新的方法和表现形式,对专利信息中所蕴含的科技、经济、法律等情报等进行进一步分析,并通过各种可视化的图表形式反映隐藏在专利文献中的各种信息,分析技术分布态势,从而预测技术发展路线,为研发决策提供更直观的支持[11]。专利分析流程见图 1-1。

专利信息挖掘和隐性知识发现是专利工作的根本目的。从现有的专利分析技术来看,面对大量的专利文献,大部分依赖计算机辅助的分析技术都停留在浅层次的基于专利文献外部属性的分析(例如年代分析、专利权人分析、IPC 分析等),更深一层的专利知识分析和挖掘在很大程度上仍然需要人的介入,而人工分析因其存在费时费力、精度差、易受主观因素

---

〔9〕彭爱东. 一种重要竞争情报——专利情报的分析研究[J]. 情报理论与实践,2000,23(3):196-199.

〔10〕云明向. 技术竞争情报理论与方法研究综述[J]. 情报科学,2010(1):154-160.

〔11〕肖国华. 专利地图研究与应用[D]. 成都:四川大学,2006.

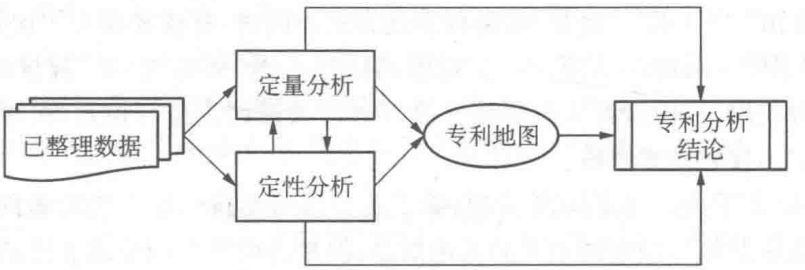


图 1-1　专利分析和专利地图

影响等缺点，往往成为专利信息挖掘和专利情报发现的瓶颈。如何将隐含在专利文献中的专利情报发掘出来，从而达到监控竞争对手和行业技术发展、为企业提供战略发展提供支撑的目的，一直是专利工作者研究的重点。对于专利分析来说，主要存在下面两个方面的问题：

（1）数据预处理效果不佳　情报部门采集到的专利数据必须经过一系列的预处理工作才能进行分析。以“技术领域分析”为例，一般技术领域的判定以国际专利分类号（IPC）为依据，此类领域划分方式对于企业实际的专利分析应用来说意义不大（见图 1-2），情报部门必须根据公司自身的特点和发展战略重新划分技术领域。因此，基于自建技术领域的数据分类和标引显得较为重要（见图 1-3）。

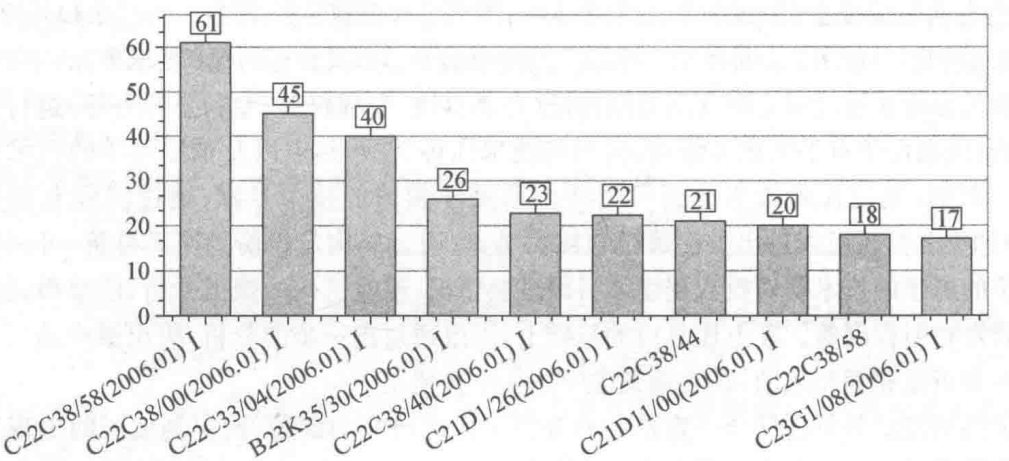


图 1-2　国际专利分类号（IPC）分析结果举例

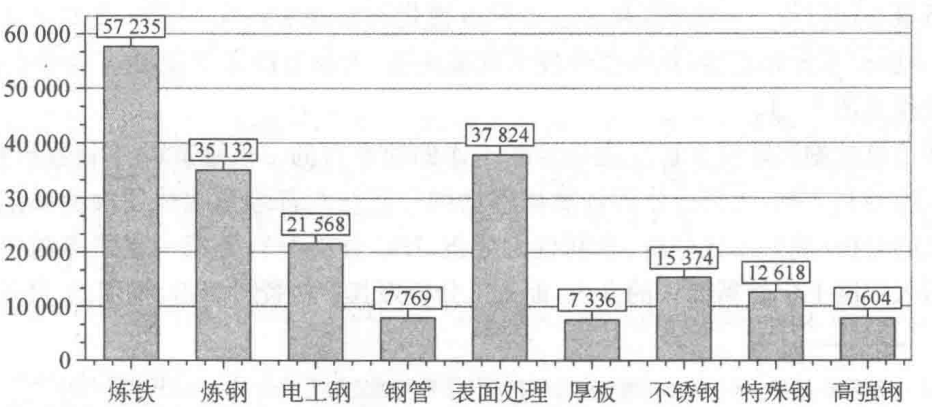


图 1-3　自建分类分析结果举例

为了确保标引结果的准确性，情报部门一般会根据领域技术专家给定的意见，编制特定的检索词组合，并将其提交至自建库中进行检索，最终得到分类标引结果。因其检索过程与传统检索相同，“忠实表达”、“表达差异”和“词汇孤岛”三类问题依然存在，因此，检索词选取的结果直接影响到数据标引的质量，从而影响最终的分析效果。虽然情报人员会不断调整检索词以确保达到最佳检索效果，但是这一过程是由人工完成的，不仅需要花费大量的时间对检索词进行调整和维护，调整过程还会受到人为主观因素的制约，成为数据预处理的瓶颈。而融入语义检索的数据分类标引则能够最大限度地提高标引效果，从而提高分析结果的质量。

(2) 计算机辅助分析精度不够　目前国内外出现了一些辅助专利分析的软件，这些软件试图通过先进的计算机处理技术，辅助人们对专利的内容进行分析挖掘，从而减轻人的负担。其包括 Thomson Data Analyzer[12]，Vantage Point[13]，BizInt Smart Charts for Patents[14]，SciFinder[15]，STN Express with Discover[16]，Vivisimo[17]，M - CAM DOORS[18] 等。这些专利分析工具对专利挖掘技术的发展起了很大的推动作用，但是由于技术的限制，本身并不完善。

例如 Thomson Data Analyzer 软件是汤森路透公司开发的一款优秀的专利分析和挖掘软件，具有数据导入与转换、数据清理、概念分组[19]、列表/直方图、比较矩阵、自然语言处理和聚类等功能，因其功能全面、处理速度快受到了专利分析人员的欢迎；但其自然语言处理和聚类部分的功能仅依赖于后台的简单词库，软件对文献内容进行分词处理后仅仅通过统计手段得到分析结果，而对于文献中的同义词、近义词以及同一概念的多种描述并未加以整合，导致分析结果精确程度下降。另外，由于字符编码问题和中文本身的复杂性，该软件并不支持中文专利文献分析。其他的专利分析软件也存在类似的问题。因此，迫切需要一种能够深入挖掘文献内容、揭示概念间语义关系的专利文献分析软件，以提高分析结果的准确性。

1.2 研究思路

本书以钢铁领域为例，以某大型钢铁公司开发的“钢铁技术情报支撑系统”为基础，抽取系统内中文专利数据库中国际专利分类号(IPC)为 C21(铁的冶炼)的发明专利摘要数据作为实验数据，探索在非结构化文本的基础上构建领域本体方法以及基于本体的语义检索和专利预警原型系统的设计与实现。研究思路见图 1 - 4。

---

[12] Thomson Data Analyzer. [EB/OL]. [2011 - 05 - 18]. http://science.thomsonreuters.com.cn/productsservices/TDA/.

[13] VantagePoint. [EB/OL]. [2011 - 05 - 18]. http://www.thevantagepoint.com/.

[14] BizInt Smart Charts for Patents. [EB/OL]. [2011 - 05 - 18]. http://www.bizcharts.com/.

[15] SciFinder. [EB/OL]. [2011 - 05 - 18]. http://www.cas.org/products/scifindr/index.html.

[16] STN Express with Discover. [EB/OL]. [2011 - 05 - 18]. http://www.stn-international.de/stn_express.html.

[17] Vivisimo | Information Optimization. [EB/OL]. [2011 - 05 - 18]. http://vivisimo.com/.

[18] M - CAM DOORS. [EB/OL]. [2011 - 05 - 18]. http://www.m-cam.com/.

[19] Trippe A. Patinformaties: Tasks to Tool [J]. World Patent Inform, 2003(25): 211 - 221.

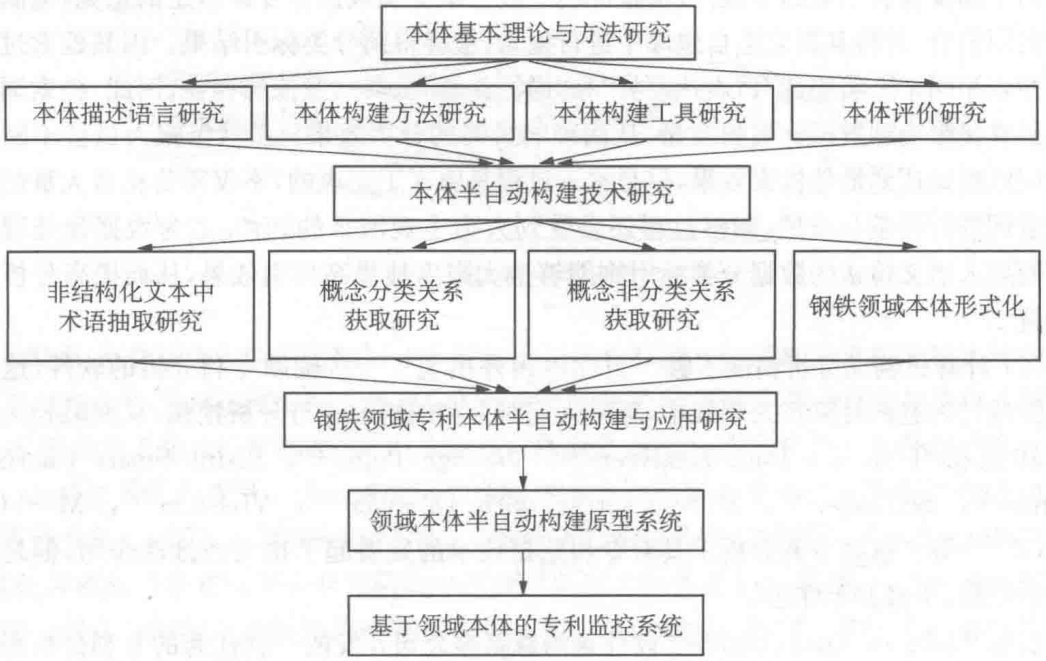


图 1-4　基于中文专利的本体半自动构建及应用研究思路

1.3　研究内容

(1) 专利文献应用技术现状分析　企业对专利情报的重视，专利文献的数字化，为专利应用技术提供了良好的发展土壤，推动了专利文献检索和分析技术的进步。而传统的专利文献应用方法，已经制约了专利情报工作的进一步开展。本书在文献调研和实践研究的基础上，从专利检索技术和专利分析技术两个方面对目前的专利情报获取技术进行剖析，指出其中不足，提出基于本体的专利数据组织方案，将专利文献资源的管理由信息层面提升到语义层面。

(2) 本体相关理论方法研究　自 Tim Berners-Lee 提出语义网以来，本体技术作为语义网的核心部分，得到了长足的发展。本书在文献调研的基础上，从本体描述语言、本体构建方法、本体构建工具和本体评价四个方面对本体相关理论方法进行分析和比较，从而得到指导钢铁领域本体半自动构建的最佳方法。

(3) 钢铁本体半自动构建技术研究　专利文献需要从文本中进行本体的学习。本书以机器学习和自然语言处理技术为基础，从术语抽取、分类关系获取、非分类关系获取和本体形式化等四个方面实现钢铁领域本体的半自动构建，以提高本体构建的效率。

(4) 基于钢铁领域本体的专利预警机制研究　本书从语义检索、技术发展趋势、专利侵权和突破性技术等方面探讨专利自动预警的机制与方法，结合钢铁领域本体，设计实现专利信息的自动预警系统原型。

1.4 研究方法

(1) 文献研究法 文献研究法是根据一定的研究目的或课题,通过调查文献来获得资料,从而全面地、正确地了解掌握所要研究问题的一种方法。通过文献研究法,调研本体构建与应用的相关文献,了解本体构建与应用的历史与现状,为钢铁领域的本体构建寻找理论支持和方法依据。

(2) 对比分析法 按照特定的指标系将客观事物加以比较,以达到认识事物的本质和规律并做出正确的评价。通过对比研究,分析现有的本体构建方法和技术路线,比较各自的优劣势,为钢铁本体半自动构建技术路线提供有力支持;通过对实验结果与前人的研究进行对比分析,揭示本书所提出方法的优势和不足,为将来的进一步改造优化提供依据。

(3) 定量分析法 通过对钢铁相关专利数据库中的词频进行统计分析,抽取相关术语,通过术语间的共现分析,获取术语间语义关系,为钢铁领域本体的构建提供数据基础。

(4) 实验法 抽取部分实验结果数据作为样本,对生成的本体模型进行检验。

(5) 系统构建法 使用 C# 和 Java 语言,开发领域本体半自动构建系统和专利预警系统。

1.5 研究意义

本书的研究意义在于:①通过对专利文献中标题、摘要等字段的深入分析,利用机器学习的方法半自动构建出领域本体,为利用中文文本构建领域本体提供借鉴;②通过对专利中所包含的技术内容和法律内容进行挖掘,初步总结出利用计算机技术,结合领域本体实现企业专利监控的方法,并构建出相应的原型系统,对企业搭建专利监控系统来说,具有一定的指导意义。

第2章 文献综述

本书主要涉及本体构建方法以及本体在专利检索和专利分析中的应用，因此，为了了解业界对于这两个方面的最新研究进展，从本体构建研究和基于专利文献的本体应用研究两个角度进行了相关文献调研。

2.1 本体构建研究概况

本体是一种有效的知识组织方式，被纳入语义网体系，成为语义网发展的基础。本体的构建研究一直是学界研究的热点，然而，本体在计算机中的应用起步较晚，属于探索阶段，因此，在学界一直没有一个统一的构建标准。目前被广泛认可的构建原则来自于 Gruber 在 1995 年提出的清晰(clarity)、一致(coherence)、可扩展性(extendibility)、最小编码偏好程度(minimal encoding bias)和最小本体约束(minimal ontological commitment)五条构建原则[1]。在本体构建方法上，大致分为两类：一是对现有的叙词表或者分类表进行改造后生成本体；二是利用现有的信息资源，通过统计和数据挖掘等手段，辅以人工半自动的生成本体。

2.1.1 利用叙词表/分类表改造成本体

(1) 国外研究进展　国外对于叙词表/分类表的本体化改造起步较早，而且也取得了一定的成果。早在 1983 年，Brachman, R J[2] 就讨论过 Semantic Network 中的分类问题，对“is-a”的概念和内涵进行了详细的描述，并为下一代知识描述语言(本体)提出了构建建议。Brachman 的这篇文章对后来的本体构建研究产生了深远的影响，此后的学者们在此基础上提出了利用已有的叙词表进行本体的构建。

在 ICES - KIS 的 MIA 项目中，荷兰阿姆斯特丹大学的 Wielinga 等发现，虽然叙词表能够为工艺品提供结构化的标准词汇表述，但是如果对工艺品对象进行语义描述的话，叙词表就显得捉襟见肘。因此，他们尝试对 AAT(工艺品叙词表)进行改造，构建仿古家具领域的本体，并用 RDFs 进行形式化表示。这是利用叙词表改造生成本体的一次有益尝试，

[1] TR Gruber. Toward principles for the design of ontologies used for knowledge sharing [J]. International Journal of Human and Computer Studies, 1995(43):907 - 928.

[2] Brachman R J. What IS - A Is and Isn't: An Analysis of Taxonomic Links in Semantic Networks [J]. Computer, 1983,16(10):30 - 36.

Wielinga 认为，用于构建本体的叙词表必须满足以下三个条件：①叙词表必须有一个严格完整的层次体系；②叙词表中概念的表示必须唯一；③叙词表的表述格式应该符合当前的 Web 标准[3]。

美国马里兰大学的 Golbeck J 等人[4]，通过 NCI 叙词表建立了一个 OWL（Ontology Web Larguage）本体。他们主要以人工方式进行叙词的抽取和本体加工，以控制本体质量，并使用 Apelon 公司的术语开发软件（Terminology Development Environment）和工作流管理工具（Workflow Manager）加以辅助，以确保本体加工过程标准化和加工结果的准确性（见图 2-1）。生成的本体使用 Apelon's Ontylog XML 格式保存，最后用 OWL 语言对 XML 数据文件进行转换，得到最终本体。

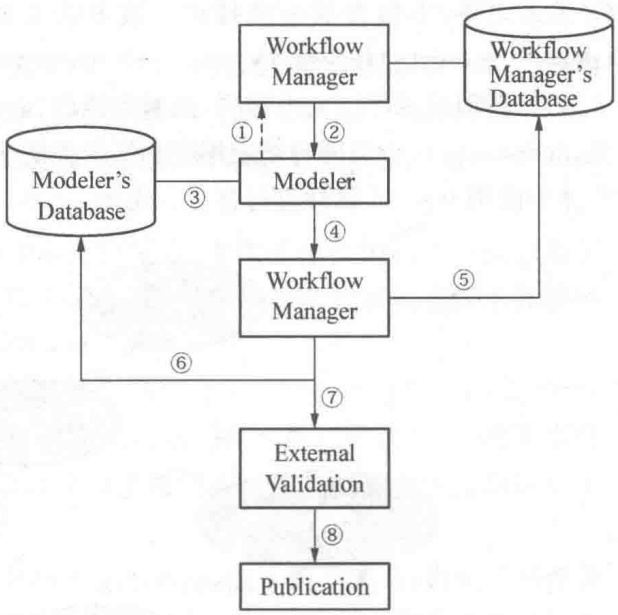


图 2-1 NCI 本体加工流程

英国中心实验室研究理事会（CCLRC）的 Matthews B 等人，以 CRIS 词表的转换为例，详细介绍了叙词表的发展和组织结构，总结了从叙词表到本体转换的两种基本方法：①基于术语的构建方法，先抽取代表概念的术语作为首选术语类，其他术语作为该术语的子类，使用 Broader，Narrower，Related 等属性与首选术语类成员关联；②基于概念的构建方法，叙词表中的一类术语代表一个概念，概念与术语集关联，从而以概念为基准进行构建。Matthews B 等对上述两种方法进行了比较，并结合实际应用，认为 RDFs 比 OWL 更适合叙词表到本体的转换。

德国卡尔斯鲁厄大学的 Steffen Lamparter 等人[5]介绍了从分类架构（文件目录、WEB 目录等）到本体的半自动构建方法。具体步骤是：①确定概念和实体；②词义消歧；③分类构建；④非分类关系定义；⑤本体生成。Steffen 等为此开发了一个原型系统，在没有人工干预的情况下，转换准确率在 70%～80%之间，达到了一定的效果[6]。

佛罗里达海湾海岸大学的 Hepp 对产品和服务标准（PSCS）进行了本体转换研究。他分析了现有产品和服务本体（UNSPSC）的不足，提出了一种能够正确反映 PSCS 分类关系的本体转换方法，并开发了与现有 OWL 语言版本兼容的转换机制，最终生成了一个产品和服

---

[3] Wielinga B J, Schreiber A T, Wielemaker J, et al. From thesaurus to ontology [C]. ACM, 2001.

[4] Golbeck J, Fragoso G, Hartel F, et al. The national cancer institute's thesaurus and ontology [J]. Journal of Web Semantics, 2003,1(1):75-80.

[5] Matthews B, Miles A, Wilson M. CRISs, Thesauri and the Semantic Web [C]. 7th Internatinal Conference on Current Research Infromation System, 2004:113-124.

[6] Lamparter S, Ehrig M, Tempich C. Knowledge extraction from classification schemas [C]. On the Move to Meaningful Internet Systems Coopis Don & Odbase, 2004,3290:618-636.

务标准领域的本体[7]。随后，他又对 PSCS 的本体转换方法(GEN/TAX)进行了详细的描述[8~10]，该方法以“分类体系中每个类目可以都分解成一个通用概念和一个分类概念”思想为依据，其优点是可以直接与 RDFs 或者 OWL 等本体描述语言交互，而且除了 sub Class Of 关系之外，不包含其他的推理。该方法主要包括三个方面：①将概念集分为通用类(generic category)和层级类(taxonomy category)；②层级类使用及其子类(subClassOf)的关系，按照原有的等级关系排列，而通用概念(generic concept)只负责对类命名；③增加注解类(annotation)，该类同时是通用类和层级类的子类(sub Class Of)。这样，任何资源都会在本体中使用 type-of 属性进行标注。见图 2-2。

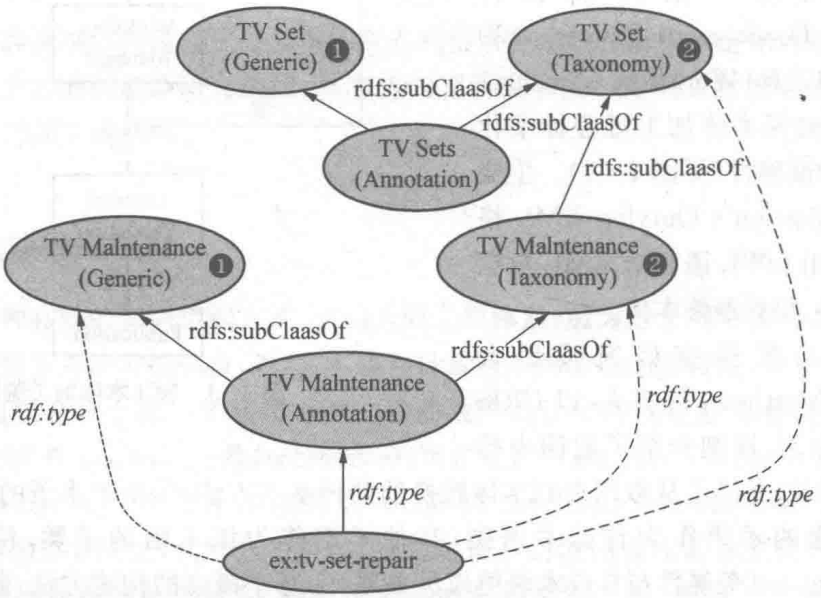


图 2-2 使用“GEN/TAX”方法构建的本体进行产品标注

荷兰阿姆斯特丹自由大学的 Mark van Assem 等人[11]在综合了前人本体转换方法的基础上，提出了 SKOS 本体转换的方法。其步骤包括：①叙词表分析；②构建对应关系；③利用算法实现转换；④评估。此外，Mark 等人还利用该方法分别实现了 IPSV、GTAA 和 MeSH 的本体转换。

美国马里兰大学的 Soergel 等人[12]，介绍了支持从叙词表改造成本体转换框架和方法，

---

[7] Hepp M. A methodology for deriving OWL ontologies from products and services categorization standards [C]. Regensburg, Germany: 2005.

[8] Hepp M. Representing the hierarchy of industrial taxonomies in owl: The gen/tax approach [C]. Citeseer, 2005.

[9] Hepp M. Products and services ontologies: a methodology for deriving OWL ontologies from industrial categorization standards [J]. International Journal on Semantic Web and Information Systems, 2006,(1):72-99.

[10] Hepp M, De Bruijn J. GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies [J]. The Semantic Web: Research and Applications, 2007:129-144.

[11] Van Assem M, Malaisé V, Miles A, et al. A Method to Convert Thesauri to SKOS[J]. The Semantic Web: Research and Applications, 2006,4011:95-109.

[12] Soergel D, Lauser B, Liang A, et al. Reengineering Thesauri for New Applications: the AGROVOC Example [J]. Journal of Digital Information, 2006,4(4): 36-52