国家科技支撑计划子项“面向企业创新应用链的
知识管理体系建设总体框架及标准规范研究(2012BAH34F01)”
资助出版

专利知识挖掘关键技术研究

翟东升◎著

知识产权出版社
全国百佳图书出版单位

国家科技支撑计划子项“面向企业创新应用链的知识管理体系建设总体框架及标准规范研究（2012BAH34F01）”资助出版

专利知识挖掘关键技术研究

翟东升　著

知识产权出版社
全国百佳图书出版单位

内容提要

随着科学技术的高速发展，专利数据分析与挖掘变得越来越重要。本书通过深入研究专利数据特征以及各种专利分析方法，设计一套完整的专利知识信息处理模型，通过对专利数据库中专利信息的抽取和采集，最终实现对专利数据的深度分析和挖掘。

本书研究设计的专利知识信息处理模型主要包括：基于多 Agent 系统的专利信息采集原型系统和基于分布式多 Agent 系统的二层专利数据库信息获取原型系统实现从专利数据库中抽取和采集专利数据；面向主题的专利分析原型系统实现面向主题的专利数据仓库和数据集市的构建；基于语义与语境的专利信息查询扩展的原型系统实现专利知识库的创建；中文专利侵权检索原型系统实现对中文专利权利侵权判定。本书设计的原型系统分析深入，设计合理，具有较强的使用价值。

本书可作为从事专利情报信息研究人员、使用专利信息分析人员和相关科技信息研究人员的参考书，也可以作为从事专利情报研究分析的高等院校师生教学用书。

责任编辑：于晓菲

图书在版编目（CIP）数据

专利知识挖掘关键技术研究/翟东升著. —北京：知识产权出版社，2012.11
ISBN 978-7-5130-1686-5

Ⅰ.①专… Ⅱ.①翟… Ⅲ.①专利—情报检索 Ⅳ.①G252.7

中国版本图书馆 CIP 数据核字（2012）第 260955 号

专利知识挖掘关键技术研究

ZHUANLI ZHISHI WAJUE GUANJIAN JISHU YANJIU

翟东升　著

出版发行：知识产权出版社

社　　址：北京市海淀区马甸南村 1 号	**邮　　编**：100088
网　　址：http://www.ipph.cn	**邮　　箱**：bjb@cnipr.com
发行电话：010－82000893 转 8101	**传　　真**：010－82005070/82000893
责编电话：010－82000860 转 8363	**责编邮箱**：yuxiaofei@cnipr.com
印　　刷：知识产权出版社电子制印中心	**经　　销**：新华书店及相关销售网点
开　　本：787mm×1092mm　1/16	**印　　张**：19.25
版　　次：2013 年 1 月第 1 版	**印　　次**：2013 年 1 月第 1 次印刷
字　　数：286 千字	**定　　价**：49.00 元

ISBN 978-7-5130-1686-5/G·535（4540）

序　言

知识产权是现代化生产保护和促进科技进步的重要手段。在全球化的背景下，未来国际市场竞争主要是知识产权的竞争。知识产权代表了国家和企业的技术地位和核心竞争力。作为知识产权的重要组成部分的专利中蕴含着大量技术、商业和法律方面的信息，是一座知识挖掘的富矿。

根据世界知识产权组织统计，研发成果90%以上包含在专利文献中，科学有效地挖掘和利用专利信息，可缩短60%的研发时间，并可节约40%的研发经费，专利文献已成为国家和企业获取和挖掘技术、商业和法律知识的重要优质信息源之一。

如何在海量专利文献中高效采集相关专利并进行深度挖掘分析、识别机会威胁、及时监测预警、保护核心技术、规避法律纠纷、促进技术创新并抢占技术竞争制高点已经成为国家和企业迫切需要解决的问题，也是当前专利知识挖掘研究的热点和难点。

多年来，作者在专利知识挖掘领域进行了认真探索，将多代理、数据仓库与多维分析、数据挖掘、知识库等先进的信息技术应用于专利知识挖掘领域，取得了较好的研究成果，这本专著就是作者近年来部分研究成果的整理和总结。

我相信本书的出版，将促进专利信息资源的深度开发，并将进一步促进对本领域的研究，为我国的科技进步和经济发展做出更多的贡献，对提高我国在专利知识挖掘方面的研究水平有十分积极的意义。

中国工程院院士

2012 年 11 月 16 日

前　言

本书是作者在北京工业大学承担的国家科技支撑计划子项“面向企业创新应用链的知识管理体系建设总体框架及标准规范研究（2012BAH34F01）”的支持下，对近年来在专利知识挖掘研究方面取得的一些成果进行的梳理与总结。

近年来，专利分析技术的主要发展趋势是把专利数据采集、专利数据清洗、专利数据仓库、数据挖掘、文本挖掘、引用网络分析、信息可视化及智能信息检索技术引入专利分析中，形成专利知识挖掘这一新的研究方向，其主要方法是应用聚类、分类、关联、网络等挖掘算法用于海量专利文献的分析中，希望找出隐含的技术发展规律，例如，通过对特定主题专利文献的挖掘分析，洞察该技术领域可能形成的技术热点、前沿、技术机会、可能的技术演进路径、潜在的技术标准、新产品新技术特征、竞争对手的技术资源状况、整体的技术发展趋势等知识，从而辅助企业技术战略的制定，优化技术创新过程。

全书共包含5章，按照异构专利信息源场景下的专利数据采集与集成，专利数据清洗、加载与转化，专利数据仓库构建与应用，专利知识库建设与应用，文本挖掘在专利侵权检测中应用的思路来组织。其中，第1章主要介绍研究背景和意义、国内外研究现状、研究内容及本书整体框架安排等内容。第2章是专利信息采集方面的相关内容，包括研究中涉及的部分专利信息源特点简介，专利采集主要理论和技术，基于多代理技术的专利信息采集系统设计原理、算法与原型系统等。第3章是专利数据仓库构建与多维分析方面的相关内容，包括数据清洗、专利数据仓库构建、主题数据集市、技术生命周期分析、构建企业创新仪表盘等。第4章是基于本体的专利知识库构建内容，包括语义网络技术、文本挖掘技术、语料库技术、本体技术简

介，专利领域语料库设计与构建，基于本体的专利知识库构建，基于领域概念的专利信息检索等。第 5 章是专利侵权检索方面的相关内容，包括专利侵权及其判定、文本预处理、专利侵权检索模型及专利侵权检索模型等。

本书能够顺利完成，首先感谢赵京教授、阮平南教授，没有他们的大力支持与鼓励，本书不可能完成；感谢多年来在专利知识挖掘方向通力合作的张杰副教授、雷东升副教授；感谢参加专利知识挖掘方向研究的已经毕业的和在读的研究生们，他们是：刘晨、杨洋、潘虹、王立轻、常雅楠、柴立静、禾文汇、张帆、康宁、袁昕、陈蕾、马文珊、陈晨、徐颖、张欣琦、李倩、蔡万江等同学，正是他们的不懈努力和取得的良好成绩，为本研究方向做出很好的贡献，我为他们深感自豪；我还要感谢我的家人，感谢她们多年辛苦无怨的付出；感谢本书的编辑知识产权出版社于晓菲女士，正是她的积极鼓励和协调才促成本书的出版。

我期望本书的出版能为专利知识挖掘研究方向做出应有的贡献。由于作者水平所限，书中难免不足之处，恳请读者指正。

翟东升

2012 年 11 月

目　录

第1章 绪 论

1.1 研究背景意义

专利文献作为科研成果和科学知识的有形载体，不仅能够反映成果的研究内容，而且还蕴藏着许多表现科学活动的信息。与学术论文、学术报告、著作等文献载体相比，专利文献有更高的知识含量，具有及时性、启发性、可靠性和准确性。随着近年来技术的迅速发展，专利文献数量增长迅速，专利文献已成为获取技术、商业和法律竞争情报的主要战略性信息来源。通过对专利信息进行统计分析与知识挖掘，能够较好地了解国内外技术发展现状、动态趋势和特征，进而挖掘技术机会，帮助解决技术难题，为国家和企业制定决策提供可靠的依据。专利文献中还会提供一些不得不向公众透露的关键技术知识，从中可以挖掘出竞争对手的战略计划、市场策略、知识产权等多方面的竞争情报信息。根据世界知识产权组织（WIPO）统计，科学有效地挖掘和利用专利知识，可缩短60%的研发时间，并节约40%的研发经费，专利知识挖掘已经成为辅助国家进行技术决策、辅助企业进行技术创新研发的重要分析方法。

面对大量复杂的专利信息，如何快速挖掘出我们所需的专利知识就显得尤为重要。近年来，国内外已有许多机构和公司在专利信息抽取、采集和分析方面投入了大量的人力和物力，也产生了不少专利分析方法和专利分析软件，实现专利信息的分类、检索、管理、统计及引用分析等功能，主要针对专利文献的结构化外部特征项进行分析，在非结构和半结构化的专利信息多维分析、内容分析以及面向主题分析等知识挖掘方面虽然也进行了大量研究，但都还不够深入，处于探

索阶段。专利文献的深度分析和知识挖掘的困难主要在于：一是处理的专利文献数据量庞大，现今全世界每年约有160万条专利发表在专利文献上，目前仍缺乏很好的自动化处理这些海量数据的工具；二是缺少用于对专利文献进行深度知识挖掘的算法与模型。

本研究的目的是通过深入研究专利数据特征以及专利知识挖掘方法，设计一套专利知识挖掘模型和系统，初步实现专利信息的采集、专利数据的清洗、专利数据仓库和专利知识库的构建、专利信息多维分析、专利侵权知识挖掘等功能。

1.2 专利信息采集研究现状

1.2.1 Deep Web 研究综述

随着Web数据库的广泛应用，Web正在加速地“深化”[1]。Internet上有大量页面是由后台数据库动态产生，现有的搜索引擎不能索引这部分页面信息，使得这部分信息对用户来说是隐藏的，称之为Deep Web（又称为Invisible Web，Hidden Web）[2]。Deep Web是一个与Surface Web相对应的概念，最初由Dr. Jill Ellsworth于1994年提出，指那些由普通搜索引擎难以发现其信息内容的Web页面[3]。2001年，Christ Sherman，Gary Price将Deep Web定义为：虽然通过互联网可以获取，但普通搜索引擎由于受技术限制而不能或不作索引的那些文本页、文件或其他通常是高质量、权威的信息[4]。文献[5]对Deep Web定义为：那些大部分内容是不能通过静态链接获取的，特别是大部分隐藏在搜索表单后的，只有用户键入一系列关键词才可获得的页面。

2000年Bright Planet公司针对Deep Web作了一个详细的调查[6]，得出结论：Deep web是互联网上最大、发展最快的新型信息资源；Deep Web站点比一般站点涉及范围较小，内容更为精深；Deep Web包含的有效高质内容总量至少是Surface Web的1000～2000倍；Deep Web的信息内容与所有的信息需求、市场和领域高度相关；超过一

半的 Deep Web 内容都保存在专业领域数据库中；95% 的 Deep Web 信息都可被免费访问。据统计[7]，我国共有 Deep Web 站点数 24000 个，Web 数据库有 28000 个，Deep Web 站点大概占全部站点的 10.6%。

在 Deep Web 信息搜索领域，已有的研究工作主要集中这些问题上：Deep Web 爬行器与爬行策略，数据库入口、查询接口发现，接口特征分析以及表单自动填写等方面。

1.2.2 多 Agent 系统应用于信息获取领域的研究

自 20 世纪 90 年代以来，关于 Agent 和多 Agent 系统（MAS，Multi – Agent System）的研究逐渐引起重视并形成人工智能研究的热点。由于表达能力强，因此适用于动态开放环境的问题求解。近年来，Agent 及其相关技术被越来越多的应用在信息检索与信息获取领域。MAS 用于信息检索与信息获取领域按照应用的不同可以分为以下几个方向的研究。MAS 应用于元搜索引擎方面的研究 、应用于个性化信息搜索方面的研究 、MAS 应用于主题搜索方面的研究。

Multi – Agent 技术具有结构开放，模块化、并行，彼此协商共同完成复杂任务的特点，易于在分布式的环境下解决多目标问题。Multi – Agent 技术应用于信息获取领域，可以发挥 Agent 之间协调、推理、分布等特点。

多 Agent 系统调度问题研究 在基于 MAS 的系统中，存在多个功能相同或不同的 Agent，调度主要解决的问题是，如何根据各 Agent 的解题能力及任务解题需求，为每个 Agent 安排任务，在兼顾任务间的顺序关系及 Agent 的负载情况的同时，为任务确定开始时间且使其调度长度最短、任务分配均衡[8]。

对于多 Agent 调度问题涉及各个 Agent 分配任务，在分布式环境下多 Agent 进行任务均衡的并行计算，均衡推进任务从而降低整体的时空开销，降低系统负载、利用 Agent 间的协商高效快速完成任务等方面。

多 Agent 系统任务划分方面的研究 Rong Xie 等研究了移动 Agent 的多任务分配与调度算法，用简单的有向无回路图 DAG 表述移

动 Agent 的多任务调度模型[9]，模型综合了主机集合、主机间的数据传输速率、各任务在各主机上的执行时间、任务数据量以及任务的前继后续关系等因素，模型的 Agent 调度目标是将各任务映射到合适的主机上，寻找符合约束关系且调度时间最短的调度；Shervin N 用合同网协议（Contract Net Protocol，简称 CNP）来对 MAS 分配任务[10]。Agent 被动态分配为管理者和合作者两种角色，管理者 Agent 负责任务的分配。其他 Agent 是合作者，对当前任务进行投标，表达对于该任务的能力和意图。管理者据所有投标者的承诺，将任务分配给最适合的投标者；黄崇本等研究了 MAS 相关任务调度问题，从时间和空间两方面考虑，提出了一种多 Agent 相关任务的并行调度算法[11]：多 Agent 相关任务关联矩阵调度算法，利用可变的关联矩阵表示任务的时间需求与 Agent 的局部存储空间的关系及任务分配的状态。该算法具有最短或较短的调度长度，并且具有较好的时间均衡性和空间协调性。

多 Agent 系统中各 Agent 协调与调度研究　Yen 提出了一种基于 Agent 的交互基础结构用来解决不同的参与者在分布式网络调度系统中的合作和交流[12]。他提出了分布式调度系统的一种 Agent 模型，这种 Agent 依靠一种 Agent 交流语言实现彼此之间的交流与合作；Liu 等提出基于多 Agent 技术开发了一种合作式的多项目计划和调度系统[13]。在该混合系统中他们创建了六种智能体可以在项目执行时动态部署在每个指定位置，并且提出了一种协商机制使计划调度系统可以协调这些分布的 Agent；韩国栋等介绍了一种面向移动 Agent 的并行计算模型[14]，该模型允许多个计算任务在异构主机构成的分布式环境下同时进行计算，并且通过算法优化，降低移动 Agent 之间的通信成本，减少网络流量。文章用矩阵三角分解对任务进行粒度划分，并在此基础上运用十标度算法策略对任务进行描述，使计算任务得以初步量化，依据进程序列标度值大小进行排序，采用满射、贪心算法策略进行任务映射，使各主机的计算任务保持相对均衡。在利用移动 Agent 进行任务动态分配和协同计算时，能够减少网络通信流量，提高计算效率；刘爱珍等提出一种可覆盖全部解空间的移动 Agent 多任务分配与调度混合遗传算法[15]，给出问题模型及染色体表示方法，

采用禁忌表加随机算法生成初始种群，设计新的交叉机制保证交叉进化解的合法性。为促进算法的收敛，变异个体使用禁忌及任务均衡启发变异算子。还采用保持解的不降性的最佳个体保留策略。结果表明该算法进化的最优解较标准遗传算法有37.1%的平均改进量。

综合以上文献，多Agent系统中各Agent具有一定的独立解决问题能力，它们通过彼此之间的协商共同完成比较复杂的任务。MAS既可以处理单一目标的问题，也能处理多目标问题。将MAS应用于信息获取领域后，多Agent机制解决了并行、分布式高效信息收集、更新及信息检索的问题，可以有效降低负载，提高系统效率。但是目前将MAS应用于Deep Web的研究较少，本书第二章将研究基于MAS的Deep Web专利信息采集系统，充分利用MAS的协调性与灵活性较强的特点，通过一定的调度机制，实现对Deep Web中专利信息的采集。

1.2.3 分布式系统负载均衡机制研究

在一个分布式网络环境中，任务被随机分派到各个服务器节点上。在整个分布式网络中，各个服务器处理工作的能力存在一定差异，因此在运行时，存在超载与轻载现象。为了能够将整个系统资源的最大化利用，需要将任务从负载较重的节点转移出去，并由负载较轻的节点接收，实现资源的最大化利用。这个过程叫做负载均衡，通过负载均衡，可以将整个系统的资源有效利用，达到资源的最优化配置[16~17]。

按照调度方式来区分，有静态和动态两种类型。静态负载均衡通过以往的经验或是系统的静态数据，衡量负载，达到静态意义上的均衡。这种算法较为简单，但由于没有涉及到动态指标，因此往往在负载均衡过程中准确性存在缺陷。动态均衡涉及到系统的动态状态，其负载状态取决于分布式系统各个节点的运行状态，通过动态计算调整各个服务器执行任务的数量，将任务在节点之间迁移，达到整个系统的均衡，提高整个系统的效率[18]。

负载估计指标 对于主机上负载的估计指标，陈志刚等提出，通

常可以从 CPU 队列长度、可用内存大小等几方面衡量[19]。T. Kunz[20]认为，简单的指标往往比负载指标更能表示系统负载状态，效率更高。因此，其使用 CPU 队列长度作为评价指标，来表示各个节点的负载状态。

马雪梅[21]认为，负载均衡的计算需要考虑进程优先级问题，进程优先级对系统负载率会产生较大影响。在此基础上，提出了将系统资源的状态作为评价指标进行分布式系统负载均衡的计算指标，这样就可以解决任务优先级对负载率造成的干扰。其中涉及到 CPU 利用率、I/O 利用率、带宽利用率、内存利用率平均值五个参数指标。

负载均衡的调度算法 Wolf M E. 和 Lam M S.[22]通过研究提出，可以通过相对迁移和门限两种策略来确定某一节点在网络负载均衡中的地位。相对迁移策略较为简单，核心思想就是将系统中负载重的节点的任务转移给负载轻的节点来完成。门限策略的核心思想是为每个节点个性化制定负载上下限阀值。当负载超过上限时，则将其任务迁移；当负载低于下限时，则使该节点接收其他超载节点的任务。以此达到整个分布式系统的负载均衡。相比而言，第二种调度策略所花费的代价更小，效率更高，能够更有效的达到系统负载均衡状态，并被大多数算法所采纳。

静态调度算法主要包括了循环算法、目的地散列算法、最小链接算法等。这些算法从静态角度出发，采用固定的分配方式，通过经验或是系统状态对负载均衡进行静态调度，这种方法简单易行，但准确性较低[23,25]。

王春娟[24]提出了动态反馈负载均衡算法。其核心思想是：分布式系统中各个节点服务器，每隔一段时间向反馈调度器上报其相关负载指标参数。调度器会将其指标与阀值对比。若超越阀值，则将该服务器挂起，从任务接受队列中将其移除。若没有超出阀值，则根据模型计算其负载率 $Load_i$，并将其与负载率阀值对比，若其负载率过高，同样从接受任务队列中移除该节点。最后，对满足上述两个条件的节点，根据其负载率 $Load_i$ 与性能参数 C_i 计算负载状况 P_i，若 P_i 处于系统可接受负载区间内，则根据其负载状况分发任务。

Kim J. 和 Lee H.[26~27]提出了两阶段负载调度算法。其核心思想

是：在第一阶段，根据标准的算法对主进程进行分配，使用启发式贪婪算法将负载大的进程分派到轻载节点。在第二阶段，对节点发生故障情况进行估算，并以此为基础计算出故障情况下所带来的负载增加值，之后再次分配进程，达到负载均衡状态。

分布式系统的负载均衡调度模型主要有如下几种：

（1）链式模型。整个系统为一个链式结构，有头尾节点。节点只可以与相邻的节点进行信息交互，各个节点只需要保存与其相邻节点的地址信息，就可通过指针与其进行通信[28]。

（2）网络模型。由于链式模型中，节点只能与前后节点进行通信，限制了整个模型的通信能力。因此，又开发了网络模型。网络模型中，每一个节点都有若干个节点与其相邻。当某一节点超载时，将触发广播事件，负载状态小于下限阀值的节点接收到广播时，触发其接受超载节点的任务[29]。

（3）链网模型。从宏观上看，某一个分布式系统可能由若干个子网组成。可以将这些子网成为局域网，每一个局域网都是有网络模式构成的，并且任何一个节点只能属于一个局域网，每个局域网内部实现均衡，各个局域网之间只通过一个节点采用链状模型进行联通，这样的模型称为链网模型[29]。当局域网整体超载时，将触发任务迁移事件，实现整个分布式系统内部的负载均衡。

这些算法给出了一般分布式系统的负载均衡调度算法（如分布式网络服务器等），但是在专利信息抽取系统中，由于专利信息抽取工作具有自身特点（如同一IP每天的信息抽取限制），因此，在研究中需要进行分布式多Agent专利信息抽取系统的负载均衡机制的重新定制。

1.2.4 网页信息抽取技术研究

如果将网页信息抽取视为一个黑盒系统，那么其外在就可以形式化表现为信息输入与信息输出。输入为包含了相关信息的网页，这些网页可以是静态的，也可以是动态的，而输出的内容则是指经过信息抽取模块转换后的、便于分析处理的、具有高度结构化的信息