专利文献研究2016

主　编◎甘绍宁

副主编◎张　鹏

知识产权出版社

全国百佳图书出版单位

专利文献研究

2016

主　编◎甘绍宁

副主编◎张　鹏

知识产权出版社

图书在版编目（CIP）数据

专利文献研究. 2016/甘绍宁主编. —北京：知识产权出版社，2016.9

ISBN 978-7-5130-4427-1

Ⅰ.①专… Ⅱ.①甘… Ⅲ.①专利—文集 Ⅳ.①G306-53

中国版本图书馆 CIP 数据核字（2016）第 209901 号

内容提要

本书从文献基础研究和产业促进服务两个维度筛选、梳理了五十余篇文章编纂形成《专利文献研究（2016）》。

责任编辑：龚　卫　　责任校对：董志英

装帧设计：张　冀　　责任出版：刘译文

专利文献研究（2016）

Zhuanli Wenxian Yanjiu（2016）

甘绍宁　主　编

张　鹏　副主编

出版发行：知识产权出版社有限责任公司　　网　　址：http://www.ipph.cn

社　　址：北京市海淀区西外太平庄 55 号　　邮　　编：100081

责编电话：010-82000860 转 8120　　责编邮箱：gongwei@cnipr.com

发行电话：010-82000860 转 8101/8102　　发行传真：010-82000893/82005070/82000270

印　　刷：北京科信印刷有限公司　　经　　销：各大网上书店、新华书店及相关专业书店

开　　本：787mm×1092mm　1/16　　印　　张：36

版　　次：2016 年 9 月第 1 版　　印　　次：2016 年 9 月第 1 次印刷

字　　数：780 千字　　定　　价：100.00 元

ISBN 978-7-5130-4427-1

《专利文献研究（2016）》编委会

《专利文献研究（2016）》编辑部

出版说明

《专利文献研究》系列丛书聚焦专利文献最新研究成果，深度挖掘专利文献的技术和经济价值，以期有效发挥专利文献在专利与产业之间的纽带作用。2015 年 12 月发布的《国务院关于新形势下加快知识产权强国建设的若干意见》明确提出，要“促进新技术、新产业、新业态蓬勃发展，提升产业国际化发展水平，保障和激励大众创业、万众创新”。毋庸置疑的是，专利文献中所承载的技术、法律、经济和战略信息，无论是在创新型小微企业的孵化培育、传统企业的产业转型，还是新兴产业和互联网等新型企业的发展壮大过程中，都正在并将愈加发挥重要的引导和支撑作用。

自 2010 年首刊，《专利文献研究》迄今已出版五期，所选近三百余篇文章均由业内人士撰写，在一定程度上及时呈现了该领域的最新研究动向和成果。本年度，编者延续了往年的选题角度，从文献基础研究和产业促进服务两个维度筛选、梳理了五十余篇文章编纂形成《专利文献研究（2016）》。为了突出专利文献对产业发展的引导和支撑作用，本书着重收集了基于专利文献展现产业创新现状和发展趋势的分析类文章，对新兴和热点技术领域的专利信息动态进行重点跟踪和研究，且所涉技术领域较往年更为多样化。同时，为了呈现近年来业界对专利文献基础研究的智力成果，本书围绕数据加工、分析方法、国外检索资源等收录了若干篇专业性高、指导性强的文章以供参鉴。再者，为了更好地显现专利文献所蕴含的战略价值，本书亦收录了基于专利文献的专利竞争情报分析类文章，从技术、申请者、专利布局等层面分析了相关热点技术领域的竞争动态。

衷心希望本书的出版能够促进专利文献研究成果的传播和利用，更好地引导公众基于专利文献挖掘所需信息，同时也鞭策专利文献工作者在专利对经济社会发展的促进作用日益显现、社会公众对专利信息的需求日渐增长的新形势下，更好地履行汇聚智慧、传播信息、支撑创新的历史责任。

《专利文献研究（2016）》编辑部

2016 年 9 月

目　录

基础研究

产业服务

JICHU
YANJIU

基础研究

基于语义的专利文本数据加工*

王亚利　侯金霞**　张正阳　翟佳雯　姜春涛

摘　要　本研究借助自然语言处理、本体工程和机器学习等技术从专利文本数据中自动或半自动地挖掘有用的专利语义信息。本研究通过不同的维度对专利文本进行了针对性的标注，标注包含了词法和句法标注、特征标注及功能标注三个维度。其中词法和句法标注借助现有的自然语言处理工具来实现；特征标注则通过术语和术语之间的关系来提取；功能标注是对前两层标注结果与其他技术耦合后进行的更高层级的标注。研究中利用了开源工具对中文专利文本进行分词、词性标注和句法关系标注；通过机器学习和依存关系树等方法设计了多种算法原型；利用本体编辑工具建立了本体，同时利用该本体对该领域的文本进行了语义标注，并进行了实证研究。

关键词　语义　本体　专利　数据加工

一、数据挖掘和本体论在专利分析中的应用原理

（一）文本挖掘

文本挖掘是指利用数据挖掘技术从大量无结构的文本信息中发现潜在的和可能的数据模式、内在联系、规律、发展趋势等，抽取有效、新颖、有用、可理解、散布在文本中的有价值的知识，并利用这些知识更好地组织信息的过程。

因为专利数量的巨大，以及不同创新主体对同一技术描述的差异，大量专利文档中普遍存在内容与已知的技术和表述存在很大差异。所以对于专利文献而言，传统的信息检索技术已不适应其日益增加的大量文本数据处理的需要。若不清楚专利文献中描述的真实内容，很难从专利文献的文摘、权利要求项或说明书全文中找到合适的专利文献，更难得到正确的分析结论并获取有价值的信息。借助文本挖掘工具对不同专利文本进行对比整理，按专利文献重要性和相关性组合排列，可为创新主体检索并利用专利信息提

* 本文改编自国家知识产权局专利局专利文献部2015年专利信息人才专项研究课题“大数据下专利数据库建设的理论、方法及实践研究”，课题负责人：吕阳红，统稿人：张正阳，主要执笔人：王亚利、龚跃鹏、翟佳雯、姜春涛、张正阳、左勇刚、徐琦，其他参与人员：顾东蕾、高劼、王勇、陆敏、黄俊。

** 等同第一作者。

供有益的帮助。

（二）专利文本标注框架简介

为构建有效的专利文本挖掘系统，帮助用户在专利检索时获得更精确的结果，对获取的专利进行自动分类和排序、技术趋势预测。本研究选取了中文专利，在专利数据层和专利内容挖掘层之间，增加专利数据标注层，方便从专利文本中自动或半自动地提取有用的专利语义信息。利用这些语义信息可以建立专利关键词库、构建专利相似性矩阵、专利自动分类或聚类、专利可视化研究等一系列专利内容挖掘任务的语义特征。本研究设计的专利数据标注分为：（1）词法和句法标注，（2）特征标注，（3）功能标注，其结构如图 1 所示。

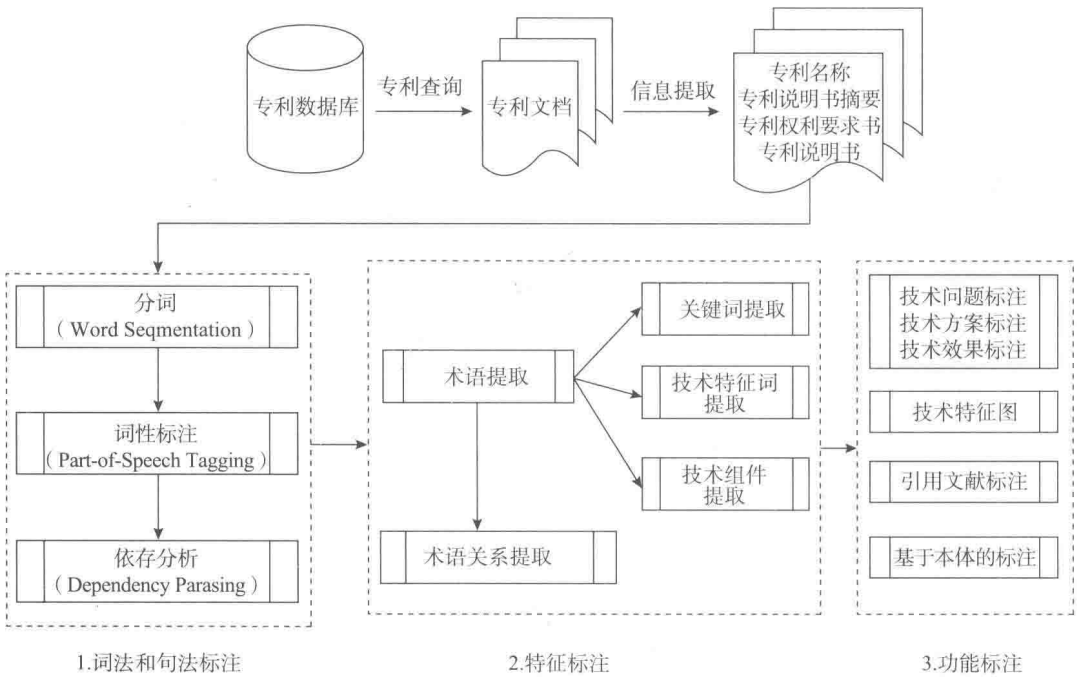

图 1　专利数据标注

图 1 三个标注子层从左至右、由低到高地说明了其标注层级的递进，即功能标注依赖于特征标注的结果，而词法和句法标注则是实现特征标注的先决条件。

（三）本体

本体是由一个正式的概念（类）和属性的描述、关系、约束和行为组成的。本体为每个特定的领域提供了一个通用的词汇表，使得自动搜索相同的概念成为可能。使用本体可以映射不同的“词语”到相同的“概念”，从而解决一词多义（polysemy）和同音异义（homonymy）的问题。

1. 本体的表达——语义网络

本体通常被显示为互相关联的概念结点的语义网络。如图 2 所示的有关交通工具和

其部件的概念模型的片段网络。

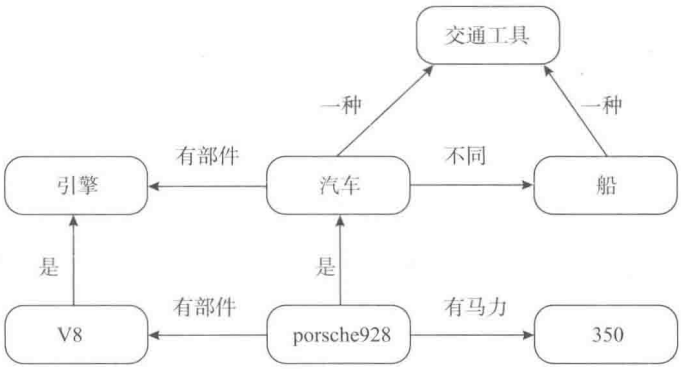

图2　语义网络实例

2. 本体学习

本体学习特指从各种类型的数据源中自动或半自动生成本体，或者说从无结构的自然语言文本中获取本体知识，其实质是由词法或语言学所激发的本体学习与基于逻辑的方法。本体学习包含获取、概念化、评价和形式化领域依存的知识。

3. 本体语言

（1）RDF（Resource Description Framework）。作为致力于元数据的标准化的 RDF 和本体语言 RDFS（RDF Schema）出现在 W3C 所倡议的语义网的背景下。RDF（S）作为一种本体语言，通过在资源类和属性上建立层次结构，已经成为具有完善体系和被广泛认可的适用于元数据和 Web 基础本体的编码标准。

（2）OWL（Web Ontology Language）。OWL 建立在 RDF 语义基础之上，用来提供更高表述水平的语言。除了从 RDF（S）继承的类和包含关系，OWL 还提供通过逻辑表达式的方式从简单类关系构建复杂类关系。

4. 本体编辑工具

（1）Protégé。Protégé 是一个开放源码的本体编辑器，具有可扩展的知识模型、可输出为多种格式、支持本体查询和可视化和支持数据库存储等多个特性。

（2）KAON。KAON 是一个支持增强型的 RDF 本体，它拥有图形化的本体编辑工具 OIModele 和基于 KAON 的 KAON Server，以及从自由文本中获取本体的学习工具。

5. 本体填充

本体填充是对现存本体加入概念和关系实例的过程，它被认为是一个概念或概念之间的关系以及它们的实例集合。本体填充需要填入的是最初本体和一个实例提取引擎，实例提取引擎负责定位文本集中的概念和关系的实例。

（四）基于领域本体的文本语义标注原理

领域文本是含有特定领域知识的文本集合，包括领域专业术语，在领域本体的帮助

下，领域文本中的概念、属性和实例可以被准确地识别出来。语义标注可被认为是在文本中标注所有提及领域本体的概念，即类、实例、属性及关系。应用经过语义标注的领域文档就是基于领域本体的语义检索模型，图3和图4介绍了一种基于领域本体的语义标注流程和经过语义标注的基于领域本体的语义检索系统。

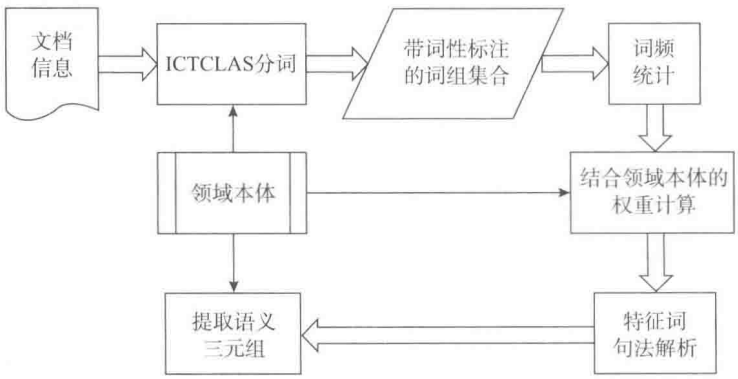

图3　基于领域本体的语义标注流程

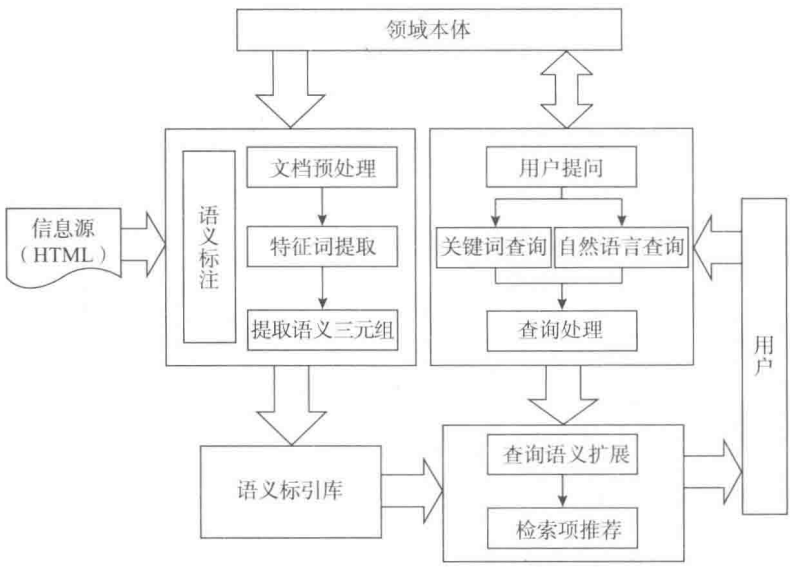

图4　基于领域本体的语义检索系统

语义标注系统根据需要既可以标注文本还可以填充实施本体，语义标注有人工、自动或半自动等不同的模式。所谓半自动执行，即首先借助自动系统完成一些标注，这些标注可由人工标注者进行后续编辑和改正。绝大多数情况下，人工标注是用于检测自动标注方法的质量和估计半自动标注方法所需消耗的人力和财力。

1. 信息提取——自动或半自动语义标注

信息提取（Information Extraction，IE）是执行自动或半自动语义标注的通用技术之

一。IE 执行的主要任务包括识别文本中出现的不同类型实体名称的命名实体识别、判断论述中的两个语言表达式是否指代同一个实体的指代消解和识别文本中出现的实体之间关系的关系提取。IE 使用的主要自然语言处理工具如词性标注、构词分析(morphological analyzer)、命名实体识别、句法分析器等。信息提取的方法主要是基于规则和基于机器学习两类。基于规则的方法是直接利用人类的直觉和领域知识，由语言学工程师设计的用于提取词典信息的规则。基于规则的方法不需要训练数据集，但需要语言处理技能。基于机器学习的方法是由经过人工标注的训练语料库训练而得到的机器学习模型（也称为有监督的机器学习），或不使用训练语料库的方法（也称作无监督的机器学习）进行相关信息的提取任务。不同信息提取方法的选择取决于目标领域的应用、语义标注的复杂性、人力资源的可获得性等因素。

2. 语义相似度计算

语义相似度计算主要分为基于 WordNet 的相似度的计算、基于非 WordNet 的相似度的计算及基于维基语义相似度的三种方式。其中，基于 WordNet 的相似度计算是使用共享信息量的方法对命名实体进行标注，而基于非 WordNet 的相似度计算是使用 Google 距离的方法或编辑距离的方法，基于维基语义相似度的计算使用词汇间的语义相关度来比较领域文档中词汇和领域本体中概念属性的相似程度，并对识别出的实例进行标注。基于非 WordNet 的相似度计算使用了英语相似度的计算来作语义相似度的计算，基于维基语义相似度的计算对英文的标注是利用编辑距离和维基语义相似度来进行的，而对中文的标注则是利用维基语义相似度和百度距离来实行的。

二、专利数据加工与挖掘模型研究

（一）词法、句法标注

1. 专利文本的预处理

专利数据的存储形式多种多样。有存储在 Oracle SQL Database，并以 XML 文件格式输出的，也有从 Microsoft Excel 文件中读取专利数据的。在进行专利文本内容分析时，用户通常希望浏览专利文件特定部分的文本，这样使得自动从多种格式的专利数据中提取特定部分的文本成为必要。此外，为了利用自然语言处理工具对中文专利文本进行语义分析，所提取的专利文本数据还需要进行清洗、分词、词性标注、关键词识别等预处理，以提高语义分析的精确度。课题研究团队将在下面分别介绍这几部分的处理过程。

（1）提取专利文本。

专利文件的文本主要包括以下四项：①专利名称，②说明书摘要，③权利要求书，④说明书。为了便于进行专利文本标注，本研究设计一专利文本提取工具，自动提取这

四项文本（包括其中一项或任意几项的组合）。此外根据专利法要求，说明书必须包括：①技术领域，②背景技术，③发明内容，④附图说明，⑤具体实施方式五部分（附图说明部分是和图有关的，与本研究的文本处理不相关，故忽略此部分内容）。因为对说明书各部分的文本进行独立的标注分析，要更具有针对性，本研究编写了一说明书文本分解处理工具，能够自动批量提取技术领域、背景技术、发明内容、及具体实施方式这四部分任意组合的文本内容。

课题研究团队使用 Apache POI 开源工具读取以 Microsoft Excel 格式存储的专利数据，使用 JavaStax API（内存需求量小）和 Java DOM API（内存需求量大）读取以 XML 格式存储的专利数据。

（2）分词。

经过特定部分专利文本的提取后，要利用自然语言词法/句法分析工具，分词是必不可少的一步，分词结果的精确度决定了后续语义分析的可靠性。经过对一系列中文分词工具的比较与分析，本研究选用 Stanford Segmenter，对中文文本进行分词处理。例如，对于从发明专利 CN200910031044. X 中提取的句子："本发明目的是提供一种安装和拆卸方便，能够提高自动扶梯承载能力的自动扶梯支架组件"经过分词处理的结果为："本发明目的是提供一种安装和拆卸方便，能够提高自动扶梯承载能力的自动扶梯支架组件。"

（3）词性标注。

词性标注是对经过分词处理的文本中每个词进行词性的标记，课题研究团队应用 Stanford Part - of - Speach Tagger，对中文文本进行词性标注。如继续上节的例子，对经过分词处理的文本，进行词性标注的结果为："本［DT］发明［NN］目的［NN］是［VC］提供［VV］一［CD］种［M］安装［NN］和［CC］拆卸［NN］方便［VA］，［PU］能够［VV］提高［VV］自动［JJ］扶梯［NN］承载［VV］能力［NN］的［DEC］自动［JJ］扶梯［NN］支架［NN］组件［NN］。"其中，每个词相邻的括号中的标记为该词的词性。这些标记由 Chinese Penn Treebank Standard 标记集所定义。

（4）关键词自动提取。

运用"TextRank"算法使用图的结构来表达文本，其中词汇单元作为图的结点、两边则由词汇单元之间的关系来决定。这样，通过对所构建的文本表达图、实施图的结点排序算法（经过改进的 PageRank 算法）来选择结点值高的结点作为关键词。课题研究团队使用 Java，通过修改复旦开源自然语言处理工具的源代码，实现了 TextRank 算法的代码，并用该算法实现从经过词性标注的专利文本中自动提取关键词。该算法的一个重要特性是它不需要深层次的语言知识，也不需要特定的经过标注的语料库，因此从理论上可以适用任何领域、任何语言。