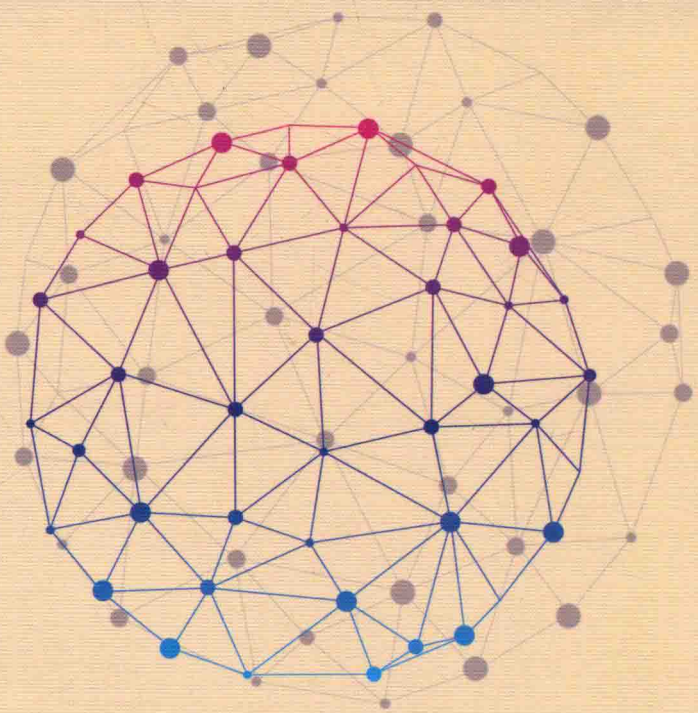
面向专利知识服务的
汉语本体学习研究

王昊　吴志祥　著

MIANXIANG ZHUANLIZHISHI FUWUDE HANYU BENTIXUEXI YANJIU

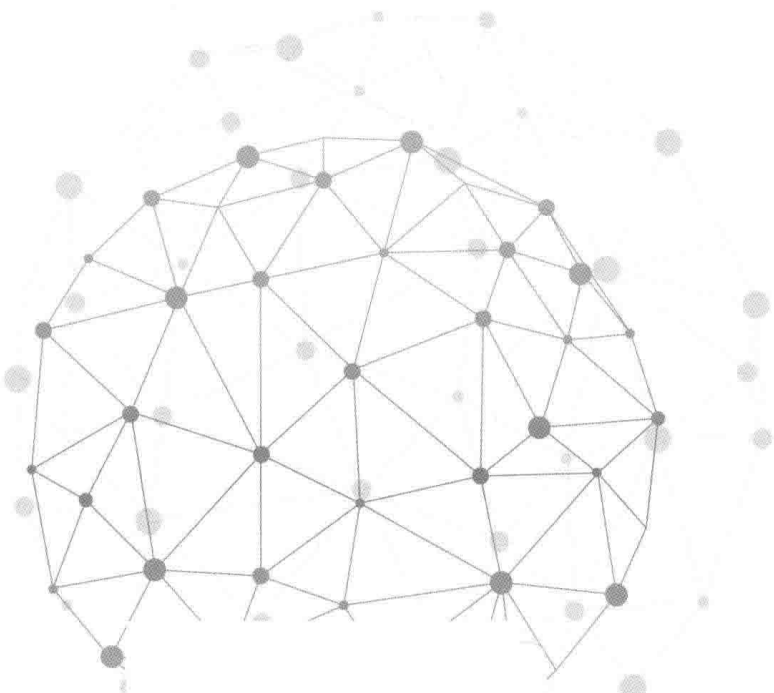


中国社会科学出版社

面向专利知识服务的
汉语本体学习研究

王昊　吴志祥　著

MIANXIANG ZHUANLIZHISHI FUWUDE HANYU BENTIXUEXI YANJIU



中国社会科学出版社

图书在版编目（CIP）数据

面向专利知识服务的汉语本体学习研究／王昊，吴志祥著．—北京：中国社会科学出版社，2018.11

ISBN 978-7-5203-3251-4

Ⅰ.①面… Ⅱ.①王… ②吴… Ⅲ.①专利—管理—汉语—语言学—研究 Ⅳ.①H1

中国版本图书馆 CIP 数据核字(2018)第 233286 号

出 版 人 赵剑英
责任编辑 赵 丽
责任校对 王秀珍
责任印制 王 超

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京明恒达印务有限公司
装 订 廊坊市广阳区广增装订厂
版 次 2018 年 11 月第 1 版
印 次 2018 年 11 月第 1 次印刷

开 本 710×1000 1/16
印 张 19
插 页 2
字 数 293 千字
定 价 79.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

本书获南京大学“双一流”建设经费资助

目　录

第一章

汉语专利本体及本体学习概述

第一节　本体、专利本体和汉语专利本体

本体（Ontology）原本是哲学领域中的概念，是对客观存在的系统解释，用来描述现实的抽象本质[①]。在20世纪90年代中期，本体被引入知识工程领域，用于描述知识的内涵，表达知识的语义。到目前为止，知识工程领域尚未对本体形成统一定义，研究者在不同实践中应用本体，不断赋予本体以新的内涵。一般认为，本体是共享概念模型的形式化规范说明，它包含四方面的含义：概念模型、明确性、形式化和共享性[②]。本体也是一种元数据，它提供丰富的原语来描述领域的概念模型，澄清领域知识的结构[③]，具有知识表示的能力。图1－1为“钢铁冶金”领域的本体片段示例，图中显示了该领域部分中文术语之间的层次分类关系，如“钢铁冶金”领域术语可分为“方法”“工艺”“装置”“原料”等四大类目，而其中“方法”又可分为“喷吹”“轧钢”“干馏”等子目，“高炉”是一种“装置”，“矿石”是一种“原料”等。

---

① 王向前、张宝隆、李慧宗：《本体研究综述》，《情报杂志》2016年第6期。

② Studer R., Benjamins V. R. and Fensel D., “Knowledge engineering: principles and methods”, *Data & Knowledge Engineering*, Vol. 25, No. 1－2, 1998.

③ 张云中、徐宝祥：《基于形式概念分析的领域本体构建方法优化研究》，《图书情报工作》2010年第8期。

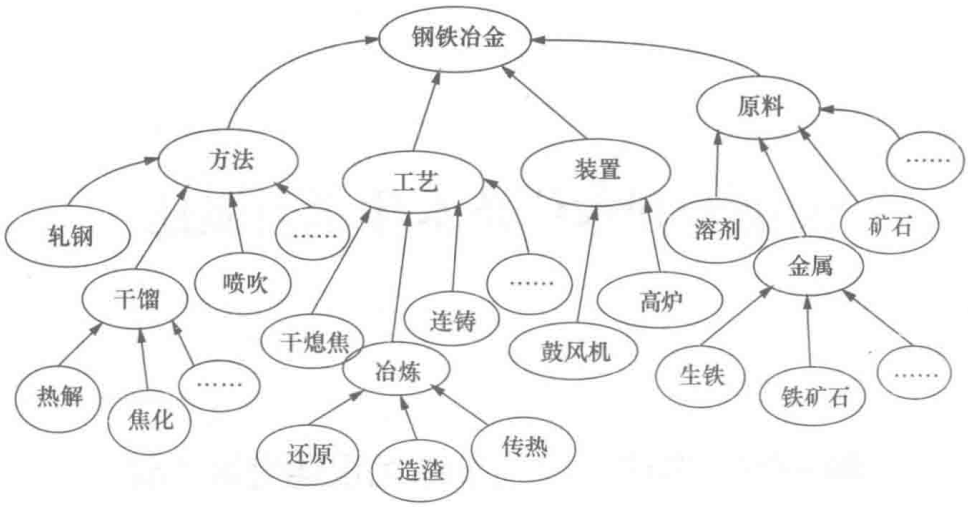


图1-1 “钢铁冶金”领域的本体片段

本体可重用，避免了重复的领域知识分析；本体提供了大量受约束的、明确定义的、机器可处理的统一术语和概念，可以构建完整的“术语表”来定义网络中的数据，使知识共享成为可能；本体还能够对知识进行推理和验证。

本体机制的广泛应用主要得益于语义网（Semantic Web）概念的提出。1998年Berners-Lee首次提出“语义网”[①] 这一全新概念，并阐释了语义网的七层体系结构（如图1-2所示）。“语义网”是指建立一个使用能够表达语义（或机器可处理）的元素来描述信息，以满足智能软件代理对异构、分布信息的有效访问、合理交换、语义处理和准确检索等要求的公开环境。而本体机制作为语义网的信息组织层面最为核心的技术，得到了前所未有的重视，还被广泛应用于除语义网外的其他各个研究领域，如智能信息检索、搜索引擎、电子商务、自然语言处理、软件工程、知识管理、数据挖掘、多代理系统、机器学习、信息分类、地球信息科学和数字图书馆等。

① Tim B. L.,“Semantic Web Road Map”, 2017.9, Semantic Web (http://www.w3.org/DesignIssues/Semantic.html).

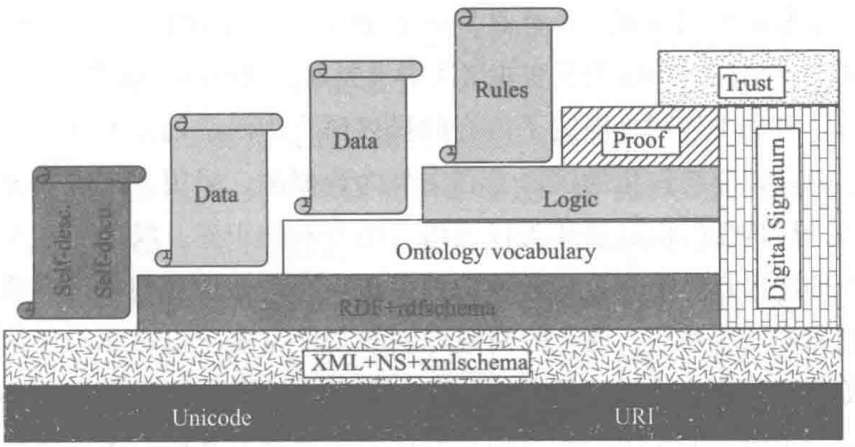


图 1－2　语义网体系结构

根据本体实际应用领域的差异，研究人员对本体的探索也各有侧重。这种涉及某一特定的学科或应用领域的本体，被称为领域本体（Domain Ontology）①。具体领域的专利本体就是一种典型的领域本体，图 1－1 中的术语层次结构片段即可被认为是“钢铁冶金”领域的专利本体。本书所指的专利本体是指一种利用专利文献的内容资源所构建的，为专利检索和开发提供服务的知识库。专利文献本身就是一种重要的知识来源，然而由于其缺乏结构性和形式化，在使用过程中极其不便。随着本体理念的兴起，将文本形式的专利说明转化为具有语义的知识本体以有效利用专利资源成了研究人员的共识②③。于是，在专利本体的语义化知识框架下，开发专利预警平台以提供知识地图、语义检索和创新开发等知识

① 李军锋：《专利领域本体学习方法研究》，博士学位论文，北京信息科技大学，2015 年，第 5—70 页。

② Taduri S., Lau G. T., Law K. H. et al., “An ontology to integrate multiple information domains in the patent system”, IEEE International Symposium on Technology and Society, 2013.

③ Bermudez-Edo M., Noguera M., Hurtado-Torres N et al., “Analyzing a firm's international portfolio of technological knowledge: A declarative ontology-based OWL approach for patent documents”, *Advanced Engineering Informatics*, Vol. 27, No. 3, 2013.

服务逐渐演化为专利应用的典型模式①②③。从形式化的角度来说，专利本体就是领域专利知识库；从具体内涵来说，专利本体包含了具体领域的术语、术语间的层次关系和非层次语义关系，以及推理规则。

汉语专利本体则是限定了内容的描述语言为汉语的专利本体。具体来说，就是以非结构化的汉语专利文献为数据源，采用人工或机器自动化手段，所构建的以汉语作为描述语言的专利知识库。汉语专利本体研究是当前汉语自然语言处理领域的前沿热点，是知识管理得以实现和推广的技术基础。然而，现有的领域专利知识库基本上以英语、德语、波斯语等字符语言作为描述语言，基于汉语文本抽取本体元素的方法体系和面向知识服务的汉语专利知识库都还没有出现，其研究大多都停留在构建本体知识库的关键步骤的技术攻关上。汉语本体知识库难以构建的一个重要因素就是汉语等象形语言具有语法特征复杂性和规则多样性等特点，在自然语言处理中存在较大的技术瓶颈。

第二节 本体学习

本体学习（Ontology Learning，OL）是指利用语言分析、机器学习和数学统计算法等，通过计算机自动或半自动地从已有的数据资源中发现潜在的概念、概念间关系和公理等本体元素的方法体系④⑤。OL 在“快速

---

① Zhai D. , Chen L. and Wang L. ,“Research on Patent Information Retrieval Based on Ontology and its Application”, *IEEE*, Vol. 1, 2009.

② Wu C. Y. , Trappey A. J. C. and Trappey C. V. , *Using Patent Ontology Engineering for Intellectual Property Defense Support System*, Switzerland: Springer, 2013, pp. 207 - 217.

③ 谷俊：《本体构建技术研究及其应用——以专利预警为例》，博士学位论文，南京大学，2012 年，第 11—100 页。

④ Yu M. , Wang J. and Zhao X. ,“A PAM-based ontology Concept and Hierarchy Learning Method”, *Journal of Information Science*, Vol. 40, No. 1, 2014.

⑤ Maio C. D. , Fenza G. , Loia V. et al. ,“Hierarchical Web Resources Retrieval by exploiting Fuzzy Formal Concept Analysis”, *Information Processing & Management*, Vol. 48, No. 3, 2012.

开发知识本体以适用于语义网"[①] 的背景下产生并得到迅速发展，其实质是信息抽取[②]在知识层面上的进一步延伸。根据抽取任务的不同，OL 可分解为如图 1－3 所示的复杂度逐层上升的 6 个层次。Terms 是指由自然语言描述的在领域中有意义的词语或词组；Synonyms 是指具有相同语义的术语集合；Concepts 是对术语的抽象描述，由 Intension、Extension 和 Lexical realizations 三部分构成；Concepts' Hierarchies 则表现为术语之间的包含关系；Semantic Relations 是指除层次关系以外的所有语义关系的总称，一般通过对象—属性的方式加以描述；本体中最复杂的是 Rules 元素，是本体概念之间的一种约束，而 Axioms 是指永真规则。

在 OL 的层次体系中，自下而上其知识的复杂度逐渐上升，知识抽取特别是从非结构化文本中抽取的难度也逐层提高。对于英语等字符语言，面向文本资源的 OL 研究和应用已经覆盖全部层次，但在公理和约束规则的抽取上仍然涉及较少[③④]；而对于中文等象形语言，由于其语法特征的复杂性和规则的多样性，目前研究主要停留在第 4 层次，很少涉及非层次语义关系特别是公理的抽取研究[⑤]，而且面向文本资源概念或实体层次关系的抽取也多聚焦于实验论证阶段，基本上还没有可进行实际应用的工具或方法出现。

为此，本书将借鉴面向字符语言的本体学习系统的功能组成和学习流程，来改变传统的手工构建模式，将各种机器学习模型、数据挖掘算法和语义学分析方法引入中文本体的自动构建研究中，建立面向中文文本资源的本体学习系统，进而以非结构化的专利文献作为学习对象，基

---

① Nanda J., Simpson T. W., Kumara SRT et al., "A Methodology for Product Family Ontology Development using Formal Concept Analysis and Web Ontology Language", *Journal of Computing & Information Science in Engineering*, Vol. 6, No. 2, 2006.

② 王昊、邓三鸿：《HMM 和 CRFs 在信息抽取应用中的比较研究》，《现代图书情报技术》2007 年第 2 卷第 12 期。

③ Terrientes L. D. V., Moreno A. and Sánchez D., *Discovery of Relation Axioms from the Web*, Philadelphia: Engineering and Management, 2010, pp. 222－233.

④ Shamsfard M. and Barforoush A. A., "Learning ontologies from natural language texts", *International Journal of Human-Computer Studies*, Vol. 60, No. 1, 2004.

⑤ 谷俊、严明、王昊：《基于改进关联规则的本体关系获取研究》，《情报理论与实践》2011 年第 34 卷第 12 期。

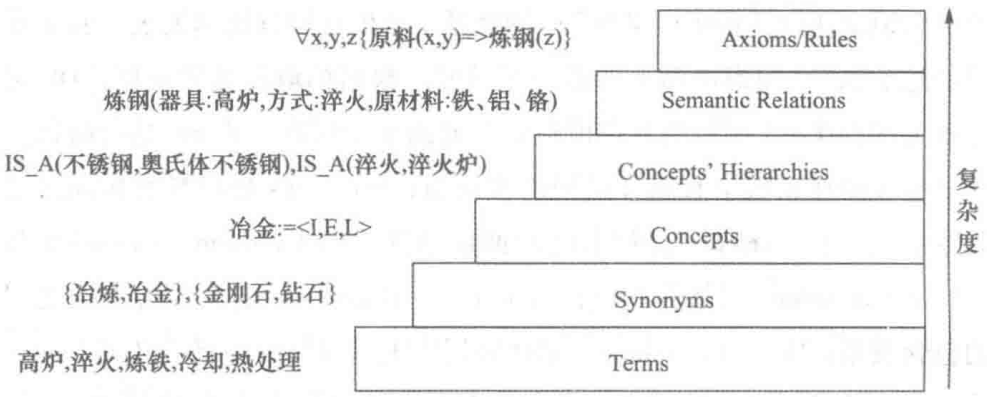


图 1-3 本体学习的层次体系

于本体学习自动构建特定领域的知识本体库，一方面通过实践应用论证本体学习理论模型的正确性和合理性、系统平台的实用性和有效性；另一方面则为进一步实现基于本体的专利语义网准备知识结构基础，为特定领域的专利管理提供全面而有效的知识服务，包括基于知识地图浏览的专利分类、基于语义匹配检索的专利预警以及基于逻辑蕴含推理的专利创新等。

第三节 汉语专利本体的自动构建及应用概述

随着领域本体研究的发展，构建领域本体已经成为该领域研究的核心和关键。目前，领域本体的构建方法主要有：手工方式、半手工（半自动化）方式和自动化方式。然而，由于领域本体的规范化要求高以及自动化技术的不成熟，当前多停留在领域专家手工或半手工构建阶段①，领域本体难以大范围和快速化生成，极大地限制了领域知识的交流和传承。随着领域本体研究的发展，一系列用于指导本体构建过程的方法不断涌现和发展，其中较为典型的有 METHONTOLOGY、KACTUS、骨架法、

① 张文秀、朱庆华：《领域本体的构建方法研究》，《图书与情报》2011 年第 1 期。

SENSUS、IDEF－5、七步法等[①]。这些方法遵循不同行业标准，适用的应用领域也各不相同，如 TOVE 法[②]专用于构建 TOVE 本体（企业建模过程本体），而 METHONTOLOGY 法[③]却是在构建化学本体的过程中产生的。

当前，在汉语领域还缺乏行之有效且影响深远的本体构建解决方案，自动化构建汉语本体的相关研究更是少之又少。究其原因，多是受限于象形语言语法的复杂性和汉语文本内容的多样性，因此，汉语本体的自动化构建方法都基本停留在理论探索和实验阶段。总的来说，根据数据来源，汉语本体的构建方法主要有三类：一是基于已有词典或术语表自动构建本体，这类方法所构建的本体知识库需依赖于现有词典或术语表；二是从文本中抽取概念及语义关系构造本体，该类方法的重点是进行术语间层次关系和非层次关系的解析，而汉语语法的复杂性和多样性导致该方法一直难以实现完全自动化和规模化的实施；三是根据需求将已有本体集成构造新的本体[④]。由于针对汉语专利本体构建的研究大多集中在基于结构化的专利数据的实例抽取与映射方面，而本书所探讨的汉语专利文本是非结构化的数据资源，因此，需采用上述第二种方法来对汉语专利文本进行术语抽取和语义关系分析等研究。

由于新颖性、可靠性和权威性，专利文献通常被认为是一种重要的知识来源[⑤]，而自动化和规模化的汉语专利本体构建，是实现高效利用专利文献中所包含领域知识的基础。专利文献在广义上包含了发明创造过程和成果的所有原始文献；狭义上仅指专利说明书[⑥]。专利文献具有外部特征和内部特征，外部特征一般指专利文献的篇名、作者姓名、出版者、

---

① 刘宇松：《本体构建方法和开发工具研究》，《现代情报》2009 年第 29 卷第 9 期。

② Gruninger M.，"Designing and Evaluating Generic Ontologies"，*European Conference of Artificial Intelligence*，1996.

③ Fernndez L. M. and Informtica F. D.，"Overview of Methodologies for Building Ontologies"，*Proceedings of Ijcai－99s Workshop on Ontologies & Problem Solving Methods Lessons Learned & Future Trends*，1999.

④ 李蓉蓉：《面向复杂语义的专利本体构建方法研究》，博士学位论文，武汉大学，2014 年，第 5—91 页。

⑤ 曹琴仙、于森：《基于内容分析法的专利文献应用研究》，《现代情报》2007 年第 27 卷第 12 期。

⑥ 王朝晖：《专利文献的特点及其利用》，《现代情报》2008 年第 28 卷第 9 期。

报告号、专利号等，容易组织为结构化数据，再加以检索和利用，又被称为结构特征；内部特征一般指专利文献的具体内容描述，包括题名、关键词、摘要、正文等，又被称为内容特征，其中包含了大量非结构化的文本数据，蕴含丰富的可挖掘的知识，但难以实现有效的组织和利用，这要归咎于汉语专利文本语义关系的复杂性和汉语自然语言处理的技术难关。目前，有关汉语专利本体的构建大多是基于专利文献的外部特征进行的，这些研究以结构化的汉语专利数据为处理对象，基本采用 OWL 语言来描述本体模型[①②]，但在构建过程中依然需要进行大量手工标注的工作；而基于内容特征来构建本体的相关研究更是寥寥无几，而且由于专利文献内容的复杂性和汉语语法的多样性，相关研究基本上都停留在理论或实验室讨论的小体量阶段，只能采用手工或半手工方式进行小范围的汉语专利本体构建[③]。对此，本书将汉语专利文献中的内容特征专指为汉语专利文本，进行汉语专利本体的自动构建研究，进而为实现有效的知识检索和知识创新提供全面的知识服务。

第四节 钢铁冶金领域专利文本概述

为实现建立面向知识服务的汉语专利知识库这一最终目标，本书将从特定的学科领域——“钢铁冶金”（Iron and Steel Metallurgy，I&SM）领域切入，收集该领域的汉语专利文本，从中抽取相关的专利术语，并解析术语间层次关系和非层次关系，进而构建 I&SM 领域的专利知识库。为此，笔者从中国国家知识产权局（State Intellectual Property Office of the P. R. C.，SIPO）专利检索平台[④]中下载了所有国际专利分类号（International Patent Classification，IPC）为 C21（铁的冶金）和 C22（冶金等）

① 谷俊：《冶金行业专利本体模型的构建研究》，《情报杂志》2012 年第 31 卷第 3 期。

② 翟东升、张欣琦、张杰：《Derwent 专利本体设计与构建》，《情报科学》2013 年第 12 期。

③ 谷俊：《中文专利本体半自动构建系统设计》，《图书情报工作》2013 年第 57 卷第 3 期。

④ SIPO 检索及分析入口（http：//www. pss-system. gov. cn/）。

的专利记录共计7597件，并以其题名和摘要作为实验的汉语专利文本，具体形式和内容如图1-4所示。

示例文本1

编号：CN1858266

题名：SAE1008脱氧工艺

摘要：一种SAE1008脱氧工艺，包括如下步骤：1. 出钢前，先将10kg电石预脱氧剂加到钢包底部，出钢9—12秒后根据钢水中的终点碳含量情况再分批次加入电石预脱氧剂，每批次加入量控制在9—12kg/t，每两批次之间的时间间隔不少于5秒，电石预脱氧剂加入完毕后再加入铝—锰—铁终脱氧剂，该终脱氧剂的加入量是0.9—1.1kg/t；2. 出钢后期，先向钢包内同时加入硅—钙—钡终脱氧合金以及硅钙，其中硅—钙—钡终脱氧合金的加入量是1.9—2.1kg/t、硅钙的加入量是0.9—1.1kg/t，然后根据钢水中的终点碳含量情况向钢包内一次性加入铝—锰—铁终脱氧剂10—40kg/炉。本发明的脱氧工艺不仅操作步骤简单，易于操作，而且可以有效地降低钢中的氧含量，能够极大地提高钢水纯净度以及最终的钢材质量。

示例文本2

编号：CN1861812

题名：高速钢冷轧辊端部热处理方法

摘要：一种高速钢冷轧辊端部热处理方法，它包括下述步骤：a. 把高速钢冷轧辊置入1160±5℃盐浴炉加热16—24分钟；b. 取出后再置入240℃—280℃硝盐炉中冷却25—35分钟；c. 取出后把高速钢冷轧辊端部置入800℃—850℃盐浴炉中加热5—10分钟后；d. 从盐浴炉取出后再置入240℃—280℃硝盐炉中等温20—40分钟；e. 从硝盐炉取出后空冷到50℃—80℃；f. 把高速钢冷轧辊置入560℃—570℃硝盐炉加热60分钟以上；g. 从硝盐炉取出空冷到室温；h. 再把高速钢冷轧辊置入560℃—570℃硝盐炉加热60分钟以上从硝盐炉取出并冷却到室温，把步骤f与步骤g再重复两次。本高速钢冷轧辊端部热处理方法轧辊端部不开裂并且硬度不低于60HRC，同时不影响轧辊辊身的硬度。

图1-4　“钢铁冶金”（I&SM）领域专利文本示例

在构建本体之前，需要深入剖析I&SM领域中的汉语专利文献的文本特征，通过对专利文献及其领域术语的抽样调研，并结合上述所收集的

I&SM 领域的汉语专利文本为例，本书将从宏观和微观两个角度来总结该领域专利文本的特征。

从宏观角度来看，I&SM 领域的专利文本具有以下特征：①反映了当前 I&SM 领域的最新科技信息，具有较强的新颖性。新颖性是获得专利权必须具备的首要的实质性条件，专利文本所记载的往往是当前时代最新颖、最前沿的内容，因此针对专利文本的研究能够保证所获取的知识是 I&SM 领域的学科热点和发展动态。②专利文本的内容具有高度唯一性，相同内容的发明专利权只授予最先提出申请的发明，因此专利文本中不存在重复性内容。同样，由此获取的 I&SM 领域知识不是人云亦云的、无价值的、重复性的内容。③专利文本中所记载的信息详细公开。专利说明书表述清楚、规范，达到所属技术领域的普通专业人员能够理解和实施的程度。按照各国专利法的规定，专利权人公开的技术信息必须使所属技术领域的技术人员能够实现。而本书所侧重的是专利文本的题目和摘要，这两部分内容则包含了专利文本的概要内容，信息量所占比重最大。④专利文本中所记载的内容具有较强的可靠性。专利说明书一般均需受过专门训练的专利代理人同发明人一起撰写，同时还要通过严格的专利审查制度，以及知识产权局的审批程序，因此专利信息的技术内容是准确可靠的。综合上述 I&SM 领域的特征，研究 I&SM 领域的专利文本对于把握当前冶金行业研究热点及深入地了解该领域具有非常重要的意义。

从微观角度来看，以图 1 - 4 中为例，总结 I&SM 领域专利文本的特征如下：①I&SM 领域专利文本的题目和摘要中含有大量的术语，如编号为“CN1858266”的专利文本中，题名就是由“SAE1008”“脱氧”“工艺”这三个术语组成的，摘要中则有近半的词是该领域的专有术语，如“预脱氧剂”“终脱氧剂”“终点碳含量”“钢水”“钢材”等。这一特征使得从术语角度分析 I&SM 领域的专利文本成为可能。②I&SM 领域专利文本的题名和摘要中所包含的术语之间存在着丰富的潜在关系。一般题名中所包含的是主要术语，摘要中所出现的则是与主要术语相关的次要术语，如编号为“CN1858266”的专利文本中，题名中所出现的主要术语是“脱氧”和“工艺”，而摘要中所出现的“预脱氧剂”“终脱氧剂”

“终点碳含量”等术语，均是对标题术语“脱氧”和“工艺”的解释、延展和扩充。这一特点使得从分析术语间层次关系的角度来获取 I&SM 领域专利文本中的知识成为可能。③I&SM 领域专利文本中的术语存在共现关系。如编号为“CN1858266”的专利文本中，“预脱氧剂”和“终脱氧剂”这两个术语之间就存在共现关系；编号为“CN1861812”的专利文本中，“开裂”和“硬度”这两个术语之间就存在共现关系。这一特征的存在对于研究 I&SM 领域的专利文本所隐含的内容关联和知识关联具有重要作用。

综合 I&SM 领域专利文本的上述特征可以发现，解析 I&SM 领域专利文本中术语之间的层次和非层次关系，可以更加深入地了解该领域的结构体系和研究重点，还可以进一步挖掘该领域的潜在知识，从而提供更有效的知识服务。对此，本书将根据 I&SM 领域文本的上述特征选取合适的本体学习理论，将机器学习、数据挖掘、潜在语义分析、形式概念分析等方法引入 I&SM 领域专利文本的术语抽取、语义关系的层次分析和非层次分析中，采用自动化的方法构建 I&SM 领域的专利本体知识库。