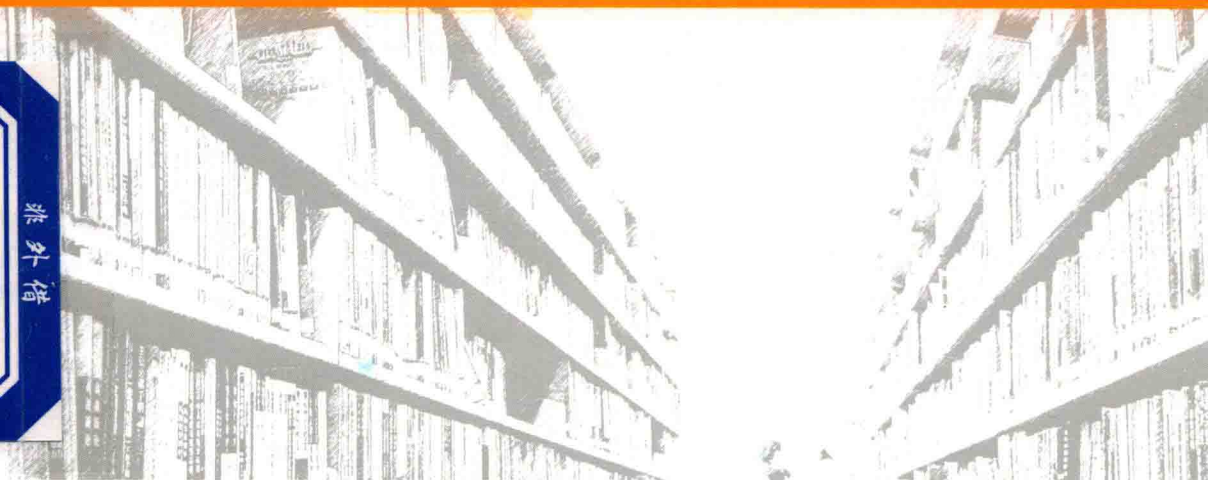




图书馆资源组织语义化 理论及方法研究

刘 耀 ◎ 著



科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

国家社会科学基金项目“图书馆资源组织语义化理论及方法研究”
(12BTQ006) 资助

图书馆资源组织语义化 理论及方法研究

刘 耀 著



科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

· 北 京 ·

图书在版编目 (CIP) 数据

图书馆资源组织语义化理论及方法研究 / 刘耀著. —北京: 科学技术文献出版社, 2018.2

ISBN 978-7-5189-3642-7

I. ①图… II. ①刘… III. ①图书馆工作—文献资源—馆藏管理—研究
IV. ①G253.5

中国版本图书馆 CIP 数据核字 (2017) 第 289068 号

图书馆资源组织语义化理论及方法研究

策划编辑: 周国臻 责任编辑: 周国臻 白建刚 责任校对: 文 浩 责任出版: 张志平

出 版 者 科学技术文献出版社
地 址 北京市复兴路15号 邮编 100038
编 务 部 (010) 58882938, 58882087 (传真)
发 行 部 (010) 58882868, 58882874 (传真)
邮 购 部 (010) 58882873
官 方 网 址 www.stdp.com.cn
发 行 者 科学技术文献出版社发行 全国各地新华书店经销
印 刷 者 北京教图印刷有限公司
版 次 2018年2月第1版 2018年2月第1次印刷
开 本 710×1000 1/16
字 数 195千
印 张 12.75
书 号 ISBN 978-7-5189-3642-7
定 价 58.00元



版权所有 违法必究

购买本社图书, 凡字迹不清、缺页、倒页、脱页者, 本社发行部负责调换

前 言

光阴荏苒，我在中国科学技术信息研究所工作已有 10 年光景。2007 年，我从北京大学计算语言学研究所博士后出站，之后一直从事自然语言处理、知识组织与知识工程相关的工作。多年在一线进行研究与实践工作，一路走来积累了丰富的经验，在不断迭代和升华中，逐渐形成了一套完整的覆盖资源获取、资源加工到知识服务全流程的思想理论。在该理论思想的指导下，围绕着技术目标，我领导团队逐步开展工程实践研究，对理论方法进行验证。同时，根据工程实践中的具体问题，进行关键技术攻关，对其中的一些流程、步骤、细节等进行了微调和补充，不断升级与完善 PYROIS 系统。PYROIS 系统初建于 2007 年，2011 年年初，PYROIS 系统 1.0 版本上线，目前该系统 3.0 版本已研发完成。

我在很早之前就有想要写这本书的想法，分享研究实践中的一些想法和体会，为从事相关研究的人员提供一些思路、方法和参考，但一直没有时间。其间，也陆陆续续地发表了一些研究论文，但没有进行系统性阐述。PYROIS 系统 3.0 版上线之后，空闲时间较以前也多了些，我便抓紧时间完成了早已有的写书的想法。

本书的主要思想是利用自然语言处理技术和机器学习方法对已有的公认领域知识进行重构并加以利用；在构建领域本体的基础上，对相关文献进行语义标注；并在大量内容相对语义化的基础上，结合传统组织资源，通过机器学习等方法，生成初始语义元数据；然后，在辅助平台的帮助下，实现图书馆资源组织语义

化过程与语义元数据体系构建同步进行,将语义标注文献与语义索引分别存储,实现图书馆资源组织语义化。

在思想理论指导及技术目标的驱动下,相关研究及实践工作体现在以下几个方面。

知识元数据库构建技术。从充分开发和利用百科全书资源的角度出发,利用自然语言处理技术和语言分析工具分析百科全书资源,发现其中隐含的知识点及其之间的内在关联,将大量的、不断出现的知识点结构化地组织和关联起来,构成智能连接的网状图,建立知识元数据库。

语义元数据构建与标注一体化技术。将图书馆资源组织语义化看作图书馆资源语义标注的浅层标注,是内容语义化与形式语义化交互实现的过程,基于 NLP 理论与方法,结合传统图书馆资源组织方式与领域本体构建技术,搭建辅助构建及标注一体化平台,实现语义元数据体系构建与资源组织语义化过程同步实施。

领域本体自动构建技术。通过领域本体构建、语义标注等信息技术的融合与集成,将行业领域知识合理分类,构建以知识点为基本单元的图书、期刊、专利数据库,以智能检索、知识语义导航、可视化等方式为科学研究、技术开发、工程设计、工程应用的开展提供知识服务,实现知识的共享与重用。

一体化爬虫技术。针对语义爬虫存在的不足,实现从一个概念出发,自动生成语义结构,用该语义结构指导爬虫抓取数据资源;同时,在爬虫爬行过程中,不断从数据库中抽取相关的概念及其之间的关系,并填充进语义结构,实现语义结构进化和爬虫爬行迭代一体化。

资源融合精准解析技术。以科技信息资源和服务体系为基础,

对专利、图书、论文、科技报告等资源进行精准解析与深度融合，突破海量知识数据的存储、组织、索引和检索的效率与性能，构建领域知识关联网络，以检索、分析、评价及可视化的方式提供服务，为政府决策、企业创新提供信息支撑。

受限领域语义资源构建技术。在领域资源匮乏的情况下，根据已知线索从多种渠道获取和利用外部资源，进行结构化、知识化处理，最大限度地完成资源内部知识体系的构建与资源的语义化标注处理，为行业领域知识服务应用场景提供数据支撑。

技术寻源技术。连接全球的创新企业、机构、专家、个人用户，汇集各方需求，实时匹配一流的技术解决方案，从传统的“靠自己”的研发模式转型为用户、资源全流程交互的开放研发模式，帮助用户获得产品的优先体验权和资源获得优先供货权，并助力小微企业快速实现产品迭代升级。

现在这个阶段主要围绕基础技术、行业共性技术及部分个性技术进行研究，并对思想理论进行验证。后续将吸收与结合大数据、深度学习等一些先进理论、方法和技术，进一步深入开展相关研究，以面向行业领域提供一站式的知识服务综合解决方案。

本书是国家社会科学基金项目“图书馆资源组织语义化理论及方法研究”（项目编号：12BTQ006）的最终研究成果。中国科学技术信息研究所的领导及科学技术文献出版社的编辑对本书的出版给予了大力支持，在此深表感谢。

在本书的编写过程中，作者参阅了大量的专业图书和文献，汲取了很多精髓，特别是引用了部分图表、数据等，在此，向有关作者表示诚挚的感谢。我的研究生郭志军、郑德举、张子渊、龚幸伟、王睿佳、孙丽君、帅远华、王明程、肖铮等人不同程度

地参与了前期的研究工作，研究生黄毅在资料整理方面做了大量的工作，在此一并表示感谢。同时，也向开发团队人员表示感谢。

由于水平和经验有限，加之书中涉及内容广泛，难免有不足之处，恳请同行专家和读者批评指正，以便在再版时进一步修订完善。

刘 耀

2018年2月

目 录

1 绪论	1
1.1 研究意义	1
1.2 研究思路	3
1.3 研究内容	5
1.4 关键问题	5
1.5 创新之处	6
1.6 撰写思路	7
1.7 本章小结	7
2 多资源融合	8
2.1 资源获取	8
2.1.1 网络资源获取	8
2.1.2 数据库资源获取	17
2.1.3 本地资源获取	17
2.2 资源解析	18
2.2.1 多种资源类型解析	18
2.2.2 多模态资源的解析	29
2.3 数据交换	45
2.3.1 数据交换技术	46
2.3.2 标准化体系建设	53
2.4 资源存储	61
2.5 本章小结	63
3 资源加工与结构化	65
3.1 结构化语料加工	65

3.1.1	词性标注及消歧	67
3.1.2	词性标注及消歧的主要功能	68
3.1.3	句法自动标注	68
3.1.4	语义角色自动标注	68
3.1.5	文本分割	71
3.1.6	句间及段落关系标注	76
3.2	结构化词表构建	81
3.2.1	专业词典构建	81
3.2.2	语义词典构建	82
3.2.3	概念词表构建	86
3.2.4	同义词表构建	88
3.3	定制化处理	90
3.3.1	基本原理	91
3.3.2	模型构建	91
3.3.3	实验与结果分析	92
3.4	本章小结	98
4	本体构建技术	99
4.1	相关理论	99
4.2	总体框架	103
4.2.1	基础流程	103
4.2.2	扩充流程	103
4.3	模型构建	104
4.3.1	树状结构到多层嵌套网状结构	106
4.3.2	文献检索到专家系统	107
4.3.3	自然语言描述到主题词描述	108
4.4	概念获取	110
4.4.1	基本思想	111
4.4.2	技术实现	111
4.4.3	实验与结果分析	115
4.5	属性获取	117
4.5.1	基本思想	118

4.5.2	技术实现	119
4.5.3	实验与结果分析	122
4.6	关系获取	127
4.6.1	基本思想	128
4.6.2	技术实现	128
4.6.3	实验与结果分析	130
4.7	本章小结	131
5	语义资源生成与标注一体化	132
5.1	语义资源生成	132
5.1.1	基于种子文件	133
5.1.2	基于本体结构与语料	137
5.2	语义标注	141
5.2.1	技术框架与思路	141
5.2.2	语义标注算法	142
5.2.3	实验与结果分析	146
5.3	语义资源评价	148
5.3.1	概念覆盖程度评价	149
5.3.2	属性完整性评价	149
5.3.3	语义关系复杂度评价	151
5.4	本章小结	151
6	应用案例研究	152
6.1	雷达语义资源生成与标注一体化	152
6.1.1	目标与要求	152
6.1.2	分析与构建	152
6.1.3	构建结果	158
6.1.4	拓展应用	162
6.2	面向技术创新的铝行业资源组织语义化	166
6.2.1	目标与需求	166
6.2.2	分析与构建	166
6.2.3	构建结果	177

6.2.4 应用服务	178
6.3 本章小结	182
参考文献	183

图表目录

图 1-1	流程与架构	4
图 2-1	普通爬虫工作流程	9
图 2-2	语义爬虫工作流程	10
图 2-3	初始语义结构示意	11
图 2-4	第一层关联词提取流程	14
图 2-5	第二层关联词提取流程	15
图 2-6	不同阈值时的平均 F_1 值	17
图 2-7	FTP 工具	18
图 2-8	图书样例	19
图 2-9	论文样例	21
图 2-10	专利样例	23
图 2-11	词典样例	25
图 2-12	叙词表样例	27
图 2-13	表格线提取与字符识别的具体步骤	30
图 2-14	表格检索平台	31
图 2-15	表格检索结果	31
图 2-16	表格详情	32
图 2-17	研究方案及整体技术路线	33
图 2-18	待解析的文件	40
图 2-19	公式识别结果	40
图 2-20	公式相似度匹配实验方法	41
图 2-21	平方和公式的 MathML 描述	41
图 2-22	检索平方和公式的界面	42
图 2-23	检索结果界面	42
图 2-24	详细检索结果	43
图 2-25	检索结果公式的 MathML 描述	43

图 2-26	检索结果公式还原	44
图 2-27	图片检索界面	45
图 2-28	图片检索结果	45
图 2-29	SGML、HTML 和 XML 文件组成	47
图 2-30	XML 解析代码	47
图 2-31	XML 解析结果	48
图 2-32	WORD 解析代码	48
图 2-33	WORD 解析结果	49
图 2-34	Excel 解析代码	50
图 2-35	Excel 解析结果	50
图 2-36	PDF 解析代码	51
图 2-37	PDF 解析结果	51
图 2-38	资源上传流程	52
图 2-39	资源编辑流程	53
图 2-40	资源导出流程	54
图 2-41	研究框架结构	55
图 2-42	专利 Schema 结构模型	59
图 2-43	Schema 模板上传界面	60
图 2-44	Schema 模板列表页面	60
图 2-45	Schema 模板修改页面	61
图 2-46	MongoDB 安装界面 1	62
图 2-47	MongoDB 安装界面 2	62
图 2-48	启动 MongoDB	63
图 2-49	MongoDB 插入数据代码	63
图 3-1	语料库加工流程	66
图 3-2	词语切分和词性标注及校对工具	68
图 3-3	句法标注工具	69
图 3-4	语义角色自动标注工具	71
图 3-5	概念语义结构	73
图 3-6	篇章单位的层次结构	77
图 3-7	篇章标注工具	81
图 3-8	概念词表构建	88

图 3-9	同义词表构建	90
图 3-10	面向专利文献的定制化管线处理流程模型	92
图 3-11	插件管理器	93
图 3-12	MySegPOS 处理部件	93
图 3-13	创建和运行定制化管线应用	94
图 3-14	动词分类	94
图 3-15	专利文本示例	94
图 3-16	依存句法分析树	95
图 3-17	短语结构句法树	95
图 3-18	抽取规则 1	96
图 3-19	抽取规则 2	96
图 3-20	抽取规则 3	96
图 3-21	抽取规则 4	97
图 3-22	抽取规则 5	97
图 3-23	抽取结果	97
图 4-1	基础流程	103
图 4-2	扩充流程	104
图 4-3	医学类树状结构示意	104
图 4-4	解剖类的子类树状结构示意	105
图 4-5	有机体类的子类树状结构示意	105
图 4-6	疾病类的子类树状结构示意	106
图 4-7	疾病类知识的横向关联示意	107
图 4-8	疾病类知识元的临床属性描述框架	108
图 4-9	解剖类的临床属性描述框架	109
图 4-10	化学制品和药物类的临床属性描述框架	110
图 4-11	概念属性示意	110
图 4-12	按代码分层的医学主题词表	111
图 4-13	按 Tab 键分层的主题词表	112
图 4-14	按上下位关系分层的主题词表	113
图 4-15	词表转换整体流程	114
图 4-16	基于代码的结构化词表导入流程	115
图 4-17	层级关系转换流程	116

图 4-18	医学主题词表转换结果	117
图 4-19	冶金工业词表转换结果	118
图 4-20	属性提取流程	119
图 4-21	属性提取算法	121
图 4-22	结构化文本片段格式	123
图 4-23	结构化提取后格式	124
图 4-24	半结构化文本片段格式	124
图 4-25	新建本体属性抽取工程	125
图 4-26	属性提取结果	125
图 4-27	将属性更新到本体	126
图 4-28	单文档属性提取结果	126
图 4-29	多文档属性提取结果	127
图 4-30	知识关系获取整体流程	129
图 4-31	导入属性 1 流程	129
图 4-32	获取知识关系流程	130
图 5-1	流程与结构	132
图 5-2	本体进化整体流程	134
图 5-3	关键词获取流程	135
图 5-4	新知识添加流程	136
图 5-5	业务流程	138
图 5-6	语料文件	139
图 5-7	编码测试	140
图 5-8	语义标注总体框架	141
图 5-9	语义标注算法	144
图 5-10	语义索引目录结构	145
图 5-11	语义索引文件数据格式	145
图 6-1	概念词表	153
图 6-2	对象属性表	153
图 6-3	数据属性表	154
图 6-4	初始本体	157
图 6-5	语料示例	158
图 6-6	部分规则模板示例	159

图 6-7	雷达本体构建结果	160
图 6-8	实例清单	161
图 6-9	雷达与探测本体实例输出	162
图 6-10	实例导入 Protégé 可视化展示	162
图 6-11	雷达知识库服务平台首页	163
图 6-12	本体可视化	163
图 6-13	图书知识图谱	164
图 6-14	语义检索	165
图 6-15	自动撰写	165
图 6-16	中国铝业广西分公司平果铝厂主要设备说明	168
图 6-17	中国铝业广西分公司平果铝厂设备管理分类	168
图 6-18	设备概念模型的建立流程	169
图 6-19	设备概念新增及修改后的层级类目	169
图 6-20	资源岗位技能培训手册	171
图 6-21	岗位及安全概念模型建设流程	172
图 6-22	岗位概念的类层级结构	172
图 6-23	岗位知识要求概念的子类目	173
图 6-24	安全概念的类层级结构	174
图 6-25	工艺概念模型建立流程	175
图 6-26	《铝冶炼生产技术手册》中对于工艺过程的描述	176
图 6-27	调整后的工艺概念层级结构	176
图 6-28	铝行业资源组织语义化构建结果	177
图 6-29	语义检索页面	178
图 6-30	市场人员个人空间页面	179
图 6-31	市场人员信息推送流程	179
图 6-32	市场人员数据中心页面	180
图 6-33	机房管理员个人空间页面	180
图 6-34	机房管理员信息监控及故障方案链接	181
图 6-35	操作人员个人空间页面	181
图 6-36	操作人员数据中心页面	182
表 2-1	词间的主要语义关系表	11

表 2-2	用来构建初始语义结构的资源来源	12
表 2-3	各类资源对应的准确率、召回率及 F_1 值 (%)	16
表 2-4	各类资源权重	16
表 2-5	构建第二层语义结构时的各类资源权重	16
表 2-6	公式属性描述	33
表 2-7	公式要素描述	34
表 2-8	公式的语义标签	35
表 2-9	公式的基本标记元素标签	35
表 2-10	公式的总体布局元素标签	36
表 2-11	公式的标注与极限标签	36
表 2-12	公式的表格与矩阵标签	36
表 2-13	科技文献中的部分贝叶斯公式结构	38
表 2-14	元素属性描述	56
表 2-15	著录项目数据元素	57
表 3-1	文本分割实验结果对比	76
表 3-2	篇章关系标签集合	77
表 3-3	标注科学文献的内容标签示例	78
表 3-4	标注“皮肤病”语料的篇章内容标签	79
表 3-5	中医药语义词典示例	82
表 3-6	专利抽取结果	97
表 4-1	本体与词典、百科全书的关系	100
表 4-2	本体与数据库模式的关系	101
表 4-3	本体与分类法、主题法的关系	101
表 4-4	概念编码示例	114
表 4-5	指定词表结构	116
表 4-6	选取的特征模板	120
表 4-7	医学领域文本属性提取结果	127
表 4-8	知识关系获取结果 1	130
表 4-9	知识关系获取结果 2	131
表 5-1	互联网进化结果	136
表 5-2	互联网进化结果 F 值	137
表 5-3	搜狗搜索引擎与语义标注算法逆序数对比	146