情报检索语言与智能信息处理丛书

丛书主编／侯汉清

领域本体的半自动构建及检索研究

何　琳／著

東南大學出版社
SOUTHEAST UNIVERSITY PRESS

情报检索语言与智能信息处理丛书(侯汉清　主编)

领域本体的半自动构建及检索研究

何　琳　著

东南大学出版社
·南京·

图书在版编目(CIP)数据

领域本体的半自动构建及检索研究/ 何琳著. —南京:东南大学出版社,2009.12

(情报检索语言与智能信息处理丛书/ 侯汉清主编)

ISBN 978-7-5641-1913-3

Ⅰ. 领… Ⅱ. 何… Ⅲ. 计算机网络—情报检索—研究 Ⅳ. G354.4

中国版本图书馆 CIP 数据核字(2009)第 200928 号

情报检索语言与智能信息处理丛书(侯汉清主编)

领域本体的半自动构建及检索研究

出版发行 东南大学出版社
出 版 人 江 汉
社 址 南京市四牌楼 2 号(邮编:210096)
印 刷 南京玉河印刷厂
责任编辑 李 正
(电话:025-83790887;E-mail:leezheng1978@sina.com)
经 销 新华书店
开 本 880 mm×1 230 mm 1/32
总 印 张 50.625(本册 7.375 印张)
总 字 数 1 310 千字(本册 190 千字)
版 次 2009 年 12 月第 1 版 2009 年 12 月第 1 次印刷
总 定 价 200.00 元(共 8 本)

---

*东大版图书若有印装质量问题,请与读者服务部联系,电话:025-83792328

丛书总序

这部丛书包括下列八本专著：

（1）薛春香著《网络环境中知识组织系统构建与应用研究》；

（2）陆勇著《面向信息检索的汉语同义词自动识别》；

（3）杜慧平、仲云云著《自然语言叙词表自动构建研究》；

（4）章成志、白振田著《文本自动标引与自动分类研究》；

（5）张雪英著《情报检索语言的兼容转换》；

（6）刘华梅、戴剑波著《受控词表的互操作研究》；

（7）何琳著《领域本体的半自动构建及检索研究》；

（8）李运景著《基于引文分析可视化的知识图谱构建研究》。

这八本专著是侯汉清教授多年来指导博士生、硕士生们进行科学研究（有些是同他们合作研究）的具体成果的一部分。这些著作的主题内容，可以归结为“情报检索语言的自动化”和“自然语言检索”两个相关的问题，或者更概括地说，就是“信息检索自动化的升级问题”，属于当前信息检索学术研究的前沿课题。

这些专著，如果将其分散来看，或许不觉得分量之重；但如果把八本专著放到一起，就可以看出其成果之丰硕。侯汉清教授在带研究生中看准一个方向不断开拓、持之以恒的精神，可以出大成果，值得我们效法。南京农业大学在侯汉清教授领导下进行的有

益的研究工作，我想一定会成为我国信息检索自动化发展史册之中浓浓的一笔。

这一类项目，本质上都是情报语言学的研究课题。所以，在研究中必须遵循情报语言学的理论，吸取情报语言学的已有成果，其结论应切合情报语言学的要求。它们只是利用计算机技术作为方法手段来达到研究目的而已，不能过分强调网络环境的特殊性而置情报语言学关于检索效率的基本要求于不顾。计算机技术应当与情报语言学密切结合。侯汉清教授和他的弟子们同时具备这两方面的知识，是顺利地较好地完成这些研究项目的关键。

这八个研究项目，大多采取实验研究法，故其成果具有较大的可信度和易理解性。其中有些项目，难度较大，甚至极难，专著只是作了认真、有益的探索；有些项目，虽然尚有一些不足，但作为中间成果，可在当前信息检索工作中推广应用，在应用中进一步完善。

信息检索自动化的初级阶段已在我国普遍实现。但要晋升一级，扩大自动化过程的范围和提高自动化的水平，当前的研究还属起步，发表的科研成果尚少见，学术研究有待扩大和深入。这部丛书起了很好的开拓作用，为继续研究打下了基础，是研究者很好的学习和参考用书，希望对此感兴趣的读者能从中获益。

张琪玉

2009 年 7 月

序　言

自从半路出家转入信息组织这一行当，就认真学习过侯汉清教授关于主题法分类法等方面的著作。近几年来，侯教授及其众多弟子在图书情报学领域核心期刊上发表的论文，也是我经常阅读和引用的文献，伴随着我在信息组织领域的成长。2009 年 6 月，在上海参加“全国第五次情报检索语言发展方向研讨会”期间，得知侯教授及其弟子将集结出版情报检索语言与智能信息处理丛书，今有幸提前拜读丛书中何琳博士的《领域本体的半自动构建及检索研究》一书，写一点个人阅读体会，权当学习心得与广大读者分享。

我曾经使用重庆维普数据，以“ontology”作检索词，使用高级检索方法，分类范畴大类选择为“图书情报”，期刊范围为“全部期刊”，检索入口为“题名”或“关键词”，查阅每年发表论文数量。2001 年以前，没有查到任何文献；从 2001 年到 2009 年，每年的论文数分别是 1、1、5、8、24、35、14、14、3，发现从 2007 年开始，与本体相关的论文数在减少。通过本人 2009 年完成本体构建相关的国家自然科学基金项目所得出的结论，以及近年来为部分图书情报学核心期刊本体相关论文审稿的体会，感觉到本体构建已经成为图书情报学领域本体研究的瓶颈。如果没有实用的通用本体、领

域本体，本体的美好应用将永远只是前景。因此，很高兴看到有本体构建和应用的相关论著面世，尤其是半自动构建方面的成果；同时，期望通过本著作的出现，使本体研究和应用重新步入快速增长的轨道。

本书开篇从语义 Web 入手，介绍了语义 Web 的提出过程、结构和应用前景等背景知识，明确了本体是语义 Web 结构的一个重要组成部分。在语义 Web 的召唤下，分析了目前的网络信息资源组织方式，包括基于关键词的网络搜索引擎，分类或主题词表的信息组织方式，以及数字图书馆和信息门户，所有这些网络信息资源组织中，如果引入本体技术，计算机就可以为用户提供语义级别的信息服务。在结构和应用的铺垫下，第三章对本体相关知识进行详细的解释和分析，包括本体的分类、本体构建与开发工具介绍、本体构建概况、本体应用概况，对本体基本知识进行了清晰的梳理。随后用一章的篇幅明确了古农学本体构建的困难之处、采用的技术路线、研究工作的意义等，同时针对将要进行的研究工作，详尽分析介绍了更多直接相关的信息技术，如自动分词、词性标注、命名实体、自动标引、自动聚类、句法分析等大量相关信息技术，同时也提供了大量的古农学方面的基础知识，例如古代农业耕作制度、耕作技术、育种栽培方法、农机具、农作物、古农书等相关知识。第五章是本著作的主体和主要创新部分，在信息技术和古农学专业知识的基础上，设计了相应的算法和规则，进行了概念的获取、等级关系及非等级关系的获取、数据的形式化等工作，并且开发了本体半自动构建工具，详细介绍了古农学本体的半自动构建过程，为读者提供了系统的半自动构建方法和知识。应用研究章节内容丰富，既有理论分析、试验框架设计，又有应用实例，而且通过具体系统编程实现了一些检索功能，具有重要参考价值。最后一章是本体评价方面的探讨，关于本体的评价研究目前还很少，

著者在此方面开了一个好头。

几乎是在一个周末就读完了这十几万字的著作，心情愉悦，感觉良好。著者从本体时代背景分析到结构组成，从基础知识介绍到存在问题讨论，从本体半自动构建研究到检索应用研究，最后再加上评价讨论。著作总体逻辑结构合理、层次清晰、丝丝入扣；研究内容方面，在作者之前，还没有看到与古农学本体相关的、如此内容丰富的本体理论与实践研究；尤其在本体的自动与半自动研究方面，作者示范了详尽的研究与实践过程，本体的半自动构建将是图书情报学领域今后一个时期在本体构建方面的一个重要发展方向。总之，这部著作是本体研究与应用领域的一本好书，将对广大本体研究工作者具有重要参考价值和启发作用。

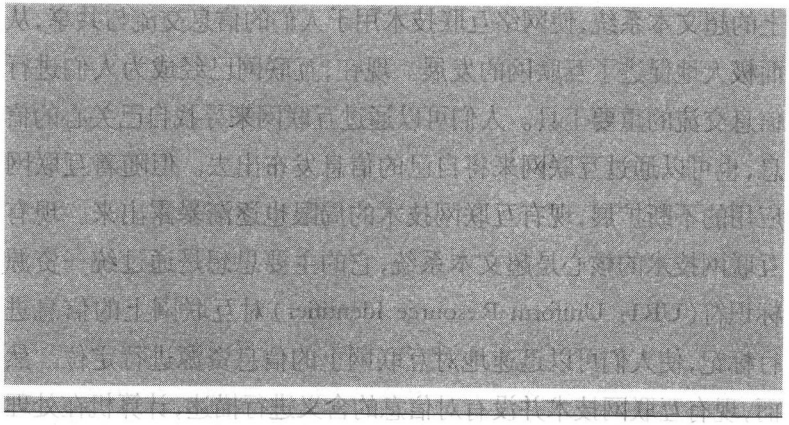
常 春

中国科学技术信息研究所 博士、研究馆员

2009 年 8 月

目　次

图表目次



第1章
语义网概述

计算机技术、通信技术和网络技术的发展为史学研究的交流和传播提供了更为便利的条件，遍布全球的学术资源通过网络得到了有机的整合。然而信息资源爆炸性的增长趋势，使得人们意识到了被“淹没”在数据的海洋中，如何更为有效地从海量数据中获取有用的信息是目前亟待解决的问题。

1.1 语义网的产生

1.1.1 语义网的提出

1990 年，蒂姆·伯纳斯·李(Tim Berners-Lee)发明了互联网

上的超文本系统，使网络互联技术用于人们的信息交流与共享，从而极大地促进了互联网的发展。现在，互联网已经成为人们进行信息交流的重要工具。人们可以通过互联网来寻找自己关心的信息，也可以通过互联网来将自己的信息发布出去。但随着互联网应用的不断扩展，现有互联网技术的局限也逐渐暴露出来。现有互联网技术的核心是超文本系统，它的主要思想是通过统一资源标识符（URI：Uniform Resource Identifier）对互联网上的信息进行标记，使人们可以迅速地对互联网上的信息资源进行定位。然而，现有互联网技术并没有对信息的含义进行描述，计算机在处理信息时只是按照 URI 来定位信息，但对信息的内容并不关心。而人们真正关心的是信息的内容，也就是互联网上的文本、图片等资源所包含的意义。由于现有互联网技术的局限，互联网上信息处理的自动化、智能化程度是很低的，计算机处理器的强大功能也没有得到有效利用。

互联网技术的研究者正在研究新的技术以改变这种状况，而其中最令人瞩目的就是语义 Web 技术。语义 Web 是互联网研究者对下一代互联网的称谓，通过扩展现有互联网，在信息中加入表示其含义的内容，使计算机可以自动与人协同工作。也就是说，语义 Web 中的各种资源不再只是各种相连的信息，还包括其信息的真正含义，从而提高计算机处理信息的自动化和智能化。而计算机并不具有真正的智能，语义 Web 的建立需要研究者们对信息进行有效的表示，制定统一的标准，使计算机可以对信息进行有效的自动处理。

在 2000 年的世界 XML（Extensible Markup Language）大会上，万维网创始人蒂姆·伯纳斯·李做了题为 Semantic Web 的演讲，对语义 Web 的概念进行了解释，并提出了语义 Web 的体系结构。2001 年 5 月，Scientific American 封面文章发表了蒂姆·伯纳

斯·李的“The Semantic Web”一文，描绘了语义 Web 的美好前景，并对其中的主要技术进行了简明的介绍。语义 Web 也被网格研究者们纳入信息服务网格的研究范围。据美国《福布斯》杂志预测，网格技术将在 2004—2005 年出现一个高峰，推动信息产业市场的持续高速发展，在 2020 年产生一个产值为 200 000 亿美元的大工业。而语义 Web 正是网格技术中信息服务网格技术的基础，在网格技术的研究中占有极其重要的地位。

鉴于语义 Web 研究的重要价值，国外的很多大学、研究机构、大公司都成立了专门的项目组来推动这项技术的发展，W3C (World Wide Web Consortium)组织也成立了专门的工作组来推动语义 Web 技术的发展。2001 年 7 月 30 日，在斯坦福大学召开了题为 Infrastructure and Applications for the SemanticWeb 的学术会议。2002 年 7 月 9 日，在意大利召开了 1st International Semantic Web Conference 会议。国内这方面的研究刚刚起步，2002 年，中国的 863 计划将语义 Web 技术列为重点支持项目。

当前，语义 Web 作为信息技术的一个热点，得到了研究者们极大的关注，也得到了许多政府、科研机构及商业部门的投入，近几年必将得到较大的发展。

1.1.2　什么是语义网

所谓“语义”就是文本的含义。语义需要理解文本的意思和结构，而与显示方式无关。语义网就是能够根据语义进行判断的网络。目前在万维网中，网页仅仅是一个单调的内容显示，电脑只负责将一个网页链接到另一个网页，网络不能按照用户的要求自动搜寻和检索网页，直至找到所需要的内容。而语义网则是希望计算机能“看懂”网页的内容，使计算机成为“智能”的导航工具。当然语义网还并不仅仅能完成这个功能，它比这还要“聪明”得

多。简单地说，语义网是一种能理解人类语言的智能网络，它不但能够理解人类的语言，而且还可以使人与电脑之间的交流变得像人与人之间的交流一样轻松。

Tim 等人所倡导的语义 Web 定义为："语义 Web 不是产生一种新的 Web，而是针对现有 Web 的扩展，Web 中的信息语义被良好定义，使人与人、人与计算机之间能够更好地协同工作。" 语义 Web 不仅仅是为了共享资源，进一步希望人们共享知识，这种网络环境下的知识共享能够发挥比个人拥有知识更大的作用；其次，目前海量的网页数据中，有近 80% 的内容来源于数据库中的数据，因此在万维网中共享的不是网页内容，而是数据内容的集成和共享。因此在语义 Web 中，处理的对象不再是以超级链接连接的页面文本，而是具有一定语义超链接的数据内容；进而，万维网成为一个语义关联的数据网络，人们能够像使用数据库那样在万维网上获取到所需的知识。

语义网的基本思想是提供基于机器可处理的数据语义，并应用这些元数据的启发式进行自动化的信息访问[3]。数据语义的显性表示和领域理论（本体）将使得 Web 提供一种全新质量的服务。其最终目标是将人类知识编织成一个巨大的网络，并以机器处理的方式来实现它。各种自动化服务将帮助用户以机器可理解的格式访问和提供信息，并使得计算机自动化处理过程和 Web 信息集成更为方便。

在 2001 年斯坦福大学召开的语义 Web 国际研讨会之后，有了每年一度的语义 Web 国际会议（ISWC）。此后，欧洲语义 Web 国际会议（ESWC）和亚洲语义 Web 国际会议（ASWC）也相继在 2004 年和 2006 年开始组织，并受到研究者的广泛关注和参与。值得一提的是，在 Web 领域最高的国际会议 WWW 上，语义 Web 研究也逐渐显示出其重要地位。在 2007 的 WWW 会议中，语义