情报检索语言与智能信息处理丛书

丛书主编／侯汉清

网络环境中知识组织系统构建与应用研究

薛春香／著

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

情报检索语言与智能信息处理丛书(丛书主编　侯汉清)

网络环境中知识组织系统构建与应用研究

薛春香　著

东南大学出版社
·南京·

图书在版编目(CIP)数据

网络环境中知识组织系统构建与应用研究/ 薛春香著.
—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.网… Ⅱ.薛… Ⅲ.计算机网络—情报检索—检索语言—研究 Ⅳ.G254.0 G354.4

中国版本图书馆 CIP 数据核字(2009)第 200934 号

情报检索语言与智能信息处理丛书(侯汉清主编)
网络环境中知识组织系统构建与应用研究

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 7.875 印张)
总 字 数	1 310 千字(本册 205 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)

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

2001 年 7～8 月《中图法》编委会在京对《中国分类主题词表》第 2 版初稿进行综合审定，南京农业大学、中山大学、上海空军政治学院等信息管理系的老师带领部分研究生参加了审定工作。薛春香就是这批学生中的佼佼者之一。她敏捷严谨的思维、沉稳扎实的作风以及深究勤问的性格给我留下深刻的印象。此后，她攻读博士、做博士后，始终在信息组织这块沃土上辛勤耕耘，由此也与《中图法》、《中国分类主题词表》结下了不解之缘，成为《中图法》编委会的常客。

八年后，薛春香的《网络环境中知识组织系统构建与应用研究》问世，这是她在博士论文的基础上从更广阔的情报语言领域进行深入研究后完成的，论证的结构更加完善、内容也更加丰富。这本书对网络环境中的知识组织系统及研究现状、知识组织系统构建与描述标准等作了全面的概述，深入对网络环境中知识组织系统的设计、传统知识组织系统自动构建以及知识组织系统的互操作与整合进行了研究，并对网络环境中知识组织系统的应用做了分析，是目前国内一本较全面地研究知识组织系统构建与应用的著作。

网络信息组织研究现已成为信息组织中极为重要的领域。这

不仅是因为因特网成为有史以来最大的信息载体，也因为它拥有世界上最多的信息用户，并且这种趋势还将持续发展。传统的检索语言、检索工具只有面向和融入新的社会需求才会发展，才会保持自己的生命力。然而，我们在这方面反应有点迟钝，研究的方向和切入点也不够清晰，很长一段时间是以评论员的角色面对网络信息及其组织的大潮，更像一个观潮者而不是弄潮儿。

不论信息存储在什么载体上、以什么形式传播，其组织的目的和基本原理是一样的，只不过信息的计算机化、网络化、数字化使信息组织的手段和形式更加多样化，对信息组织的规范化、互操作性要求更高，同时也给信息组织带来空前的便利。情报语言学理论是我们这“一亩三分地”里的核心理论之一，正在被诸多相关领域重视并得到应用，“搜索引擎”就是一个例证。而图书情报界自身的研究却不尽如人意。

受国内学术界一些浮躁和急功近利倾向的影响，目前我国网络信息组织领域分量重、影响大的成果不多。除了体制上的深层原因（如机构设置、经费使用不合理）外，研究的目的性也是重要的原因。图书情报界论文的总量是相当大的，但与其相应的知识生产量、创新性成果却是很不匹配。为项目、为资金、为会议、为指标、为职称、为学业等而进行研究或撰写论文的现象很普遍，这种状况使研究人员很难在一个方向上集中精力进行长期、艰苦的研究，很难产生高水平的成果。

我常感慨于一个外国科学家甚至是研究生可以为某一研究在荒野丛林一呆就是数年乃至数十年，这种探求事物真知的科学精神与执著令人钦佩。在网络信息组织研究领域，像元数据、XML、RDF、语义 Web、SKOS、Ontology、OWL、MARC21 等的创新均来自

国外，我国还处于介绍、学习、吸收、改进阶段，一些基础性研究（如“知识组织体系描述规范研究”）尚停留在翻译、整理水平。即便这样，从概念到概念的泛泛之谈还为数不少。

不论是传统的检索语言应用于网络信息组织，还是网络信息组织及其系统，都有很多新课题等待我们去探索，比如传统的分类法及主题词表面向网络信息组织的系统改造、当主题词的同义词同时又是其他主题词的常用词素时如何处理才能满足网络组织的需求、严格的逻辑组配标引对网络信息检索效率的影响及改进方法；又如 Folksonomy、Wiki、Blog、SNS、RSS 中信息自组织的规律及在受控信息系统中的应用等等。不论是大项目还是小题目，只有钻进去坚持不懈的努力，才有可能有所创新。创新研究需要动力，社会需求、使命感、个人兴趣等都会成为重要的动力。做学问是件苦差事，除了严谨的科学精神、科学方法，还要脚踏实地，锲而不舍，耐得住寂寞，把过程和结果看得同等重要。

张琪玉老师在本丛书总序说的“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”在薛春香身上也得到了体现。信息组织是她的主要研究方向，大量的实践活动和“用于中文信息自动分类的《中图法》知识库的构建”、“基于语料和基于标引经验的自动分类模式比较”、“农史知识组织系统构建与应用研究”等成果为她这本书的写作奠定了基础。她在这本书里以农史知识为对象，运用知识组织系统的理论提出农史知识组织系统设计方案，通过实验构建了专业分类表、领域词表、农史特色词表，从而完成农史文献题录库、全文库、农史工具书的整合，形成农史研究的知识库：一个小而独立的网络信息组织系统。其中不少方法、细节具有创新性，展示出她的情报语言学和信息组织的功底。

尽管本书还有一些研究尚待深入或不够成熟，如知识组织系统互操作一节略显薄弱，但不论从宏观上了解知识组织系统构建还是从微观上把握构建知识组织系统的技术，这都是一本值得阅读的著作。

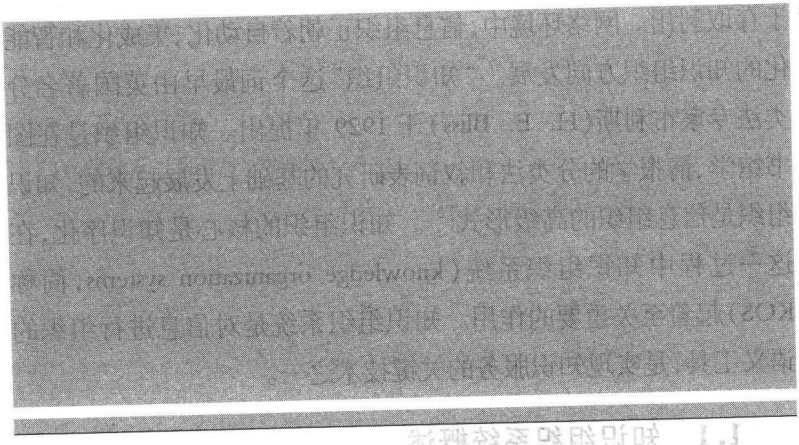
最近读了一段薛春香的博文："我只想做点自己喜欢做的东西，不用为发文、写申请、写报告而烦恼！""但这个浮躁喧哗的时代，除非你成了疯子，否则你的精神世界是不可能清静了！再积极入世，再积极超然！人生本来就是一个螺旋叠加的上升！"我想，只要这样积极地面对人生，笑迎学习、工作和生活中的各种挑战，才会活得更潇洒，才会在事业上有所成就。

祝愿薛春香像她期待的那样："2009，面向大海，春暖花开！"

陈树年
2009 年 9 月于沪

目　次

图表目次



第1章
网络环境中的知识组织系统概述

这是一个信息极大丰富的时代，海量信息给人们的工作、学习、日常生活都带来了快捷便利。但是，由于海量信息处于分散、混乱、无序的状态，导致人们无法充分享受资源丰富所带来的快速获取所需知识的便利。“我们淹没在信息中，但是却渴求知识”[1]，就是因为“失去控制和无组织的信息在信息社会里并不构成资源，相反，它成为信息工作者的敌人”[1]，因此，有效组织信息，建立数字世界的秩序，提供全面、快速、准确的信息检索服务，已成为网络时代亟待解决的问题。这是人们充分有效地利用信息资源的前提和保障，也是网络环境中信息组织研究的重点和方向。

信息组织，简而言之，就是组织信息，将信息予以序化，从而便

于存取利用。网络环境中，信息组织正朝着自动化、集成化和智能化的知识组织方向发展。“知识组织”这个词最早由英国著名分类法专家布利斯(H. E. Bliss)于1929年提出。知识组织是在图书馆学、情报学的分类法和叙词表研究的基础上发展起来的，知识组织是信息组织的高级形式[2]。知识组织的核心是知识序化，在这一过程中知识组织系统(knowledge organization systems，简称KOS)起着至关重要的作用。知识组织系统是对信息进行组织的语义工具，是实现知识服务的关键技术之一。

1.1 知识组织系统概述

1.1.1 知识组织系统的概念

知识组织系统也称知识组织体系，是对各种人类知识结构进行表达和有组织阐述的语义工具的统称，包括分类法、叙词表、语义网络、本体以及更泛指的情报检索语言、标引语言[3]。“知识组织系统”这一概念试图囊括所有组织信息和促进知识管理的模式和方法，作为这一含义而言的“知识组织系统”这个词最早由网络环境中的知识组织系统/服务(Networked Knowledge Organization Systems/Services，简称NKOS)研究小组在美国计算机学会(Association for Computing Machinery，简称ACM)1998年的数字图书馆会议上提出[4]。而实际上，知识组织系统由来已久，传统的文献分类法、主题词表、术语表等都是一种知识组织系统，只是未用“知识组织系统”来统称这些知识组织工具。本文所论述的网络环境中的知识组织系统(NKOS)既包括电子化、网络化的传统知识组织系统，也包括在网络环境尤其是在语义网环境中产生的新型知识组织工具。

术语“知识组织系统”的产生基于网络环境这样一个大背景：

(1) 随着信息环境、信息数量、信息需求的变化，越来越多的

知识组织工具不断出现，如语义网络、概念地图、本体等，名称越来越多，而区别越来越不明显。

（2）随着信息资源的激增，人们为了能够获取到真正想要的知识，对信息组织揭示程度要求越来越高，已从文献组织层面上升到知识组织层面。人们发现，单独使用任何一种工具都难以组织好信息，必须多个知识组织工具配合使用，才能发挥组织的最佳效能。这从国内外对分类法、主题词表之间互操作研究的热衷和关注程度可见一斑。

（3）超链接、标记语言、可视化等计算机网络技术，被大量引入传统检索语言之中。二者的结合，使以往以分类表、词表等形式出现的检索语言名不副实，不得不改名。

（4）网络环境既对多种知识组织工具的集成使用提出需求，也为其创建及应用创造条件，知识组织系统的发展离不开网络环境以及网络环境中发展起来的各种信息技术。

“系统”一词在《现代汉语词典》中定义为：“同类事物按照一定的关系组成的整体。”因此，提出“知识组织系统”一词来统称各种信息组织工具，反映了网络环境中多种知识组织工具集成使用的趋势，某种知识组织工具独行天下或各自为政的局面逐渐消亡。

1.1.2　知识组织系统的分类

关于知识组织系统的界定和分类，先后有过多种不同的看法。

武汉大学学者马费成从知识工程的角度，按照知识组织系统的演进顺序，将其分为基于文献单元的知识组织系统、基于数据单元的知识组织系统和基于人工智能的知识组织系统三种类型[5]。这里的知识组织系统既包含了知识组织的工具，也包含了知识资源库，是一种广义的知识组织系统。

台湾学者吴万钧认为，知识组织并不都是借助于文献来完成