情报检索语言与智能信息处理丛书

丛书主编／侯汉清

情报检索语言的兼容转换

张雪英／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)

情报检索语言的兼容转换

张雪英　著

东南大学出版社
·南京·

图书在版编目(CIP)数据

情报检索语言的兼容转换/张雪英著. —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.情… Ⅱ.张… Ⅲ.情报检索—检索语言—兼容性 Ⅳ.G254.0

中国版本图书馆 CIP 数据核字(2009)第 200914 号

情报检索语言与智能信息处理丛书(侯汉清主编)
情报检索语言的兼容转换

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 6.0 印张)
总 字 数	1 310 千字(本册 155 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

随着计算机、通信和互联网技术的快速发展，信息检索面临着信息量激增、用户大众化和服务创新等多重挑战。目前，自然语言处理、模糊数学、机器学习等先进技术，已经成功应用于语音识别、电子政务、机器翻译和工业控制等领域。因此，如何应用人工智能技术推进信息检索的智能化和大众化，无疑成为当前信息服务的主要议题。

情报检索语言是信息检索的语言保障工具，其兼容性问题成为信息共享和服务的最大障碍。本书较为全面地阐述了情报检索语言兼容转换的基本理论和国内外研究进展，讨论了多种检索语言的兼容转换方法，包括基于集成词表的叙词表兼容转换方法、基于字面相似度的分类兼容转换方法、基于最大似然法的分类表与叙词表兼容转换方法、基于粗糙集的检索语言兼容方法，提出了基于 N-Gram 的中文文本关键词自动抽取方法，分析了“中图法”知识库的构建理论与方法。

本书内容组织独具匠心，既有详细的理论基础，也有具体的方法和系统。从技术角度看，首先讨论传统的方法，逐步过渡到统计方法和机器学习方法；从应用角度看，首先介绍相关技术在文献信

息检索领域的应用，然后拓展到其他领域。通过本书，读者既可以较为系统地了解情报检索语言兼容转换的相关技术，也可扩展学术视野，触类旁通。

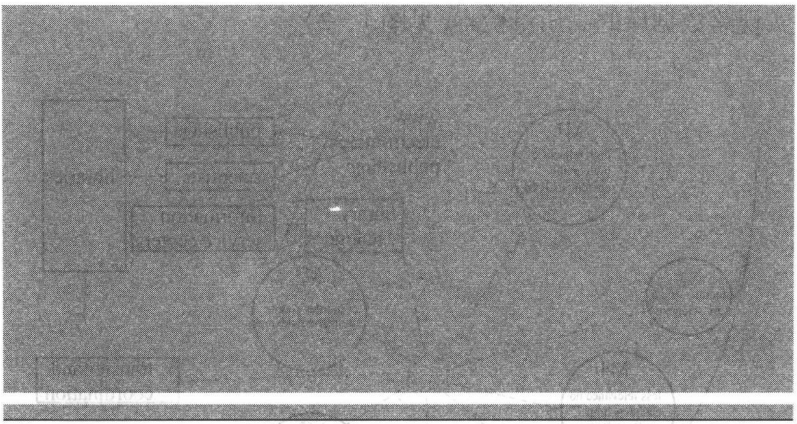
在我指导的博士生当中，张雪英勤奋好学和吃苦耐劳的精神给我留下了很深的印象。在德国期间，通过她的不懈努力，使得南京理工大学和波恩大学签署了学术合作协议。在一些繁琐的事务处理中，她的组织能力、应变能力和语言交流能力获得了中德专家的好评。本书是她多年来刻苦努力的写照和研究成果的积累。我相信，读者一定会从中获得不少有益的启发。

刘凤玉

2009 年 7 月于南京理工大学

目　次

图表目次



第1章
绪　论

1.1　多元信息空间

与四十年前相比，信息检索环境发生了巨大的改变。信息提供者和信息用户都呈现出了多元化的发展趋势[1]。多元化信息空间使得用户可以轻易地跨越空间距离，访问世界各地的信息资源（见图1-1）。

但是，跨越信息空间进行有效的信息检索却不是轻而易举的事情。其根本原因在于：信息检索系统采用的检索语言之间存在严重的语义兼容性问题，导致信息检索系统的共享能力和服务效果大大降低。事实上，对于普遍用户而言，最理想的模式是通过一个统一的检索界面，使用一种检索语言（最好是自然语言）就可以

实现跨数据库的高质量检索(见图1-2)。

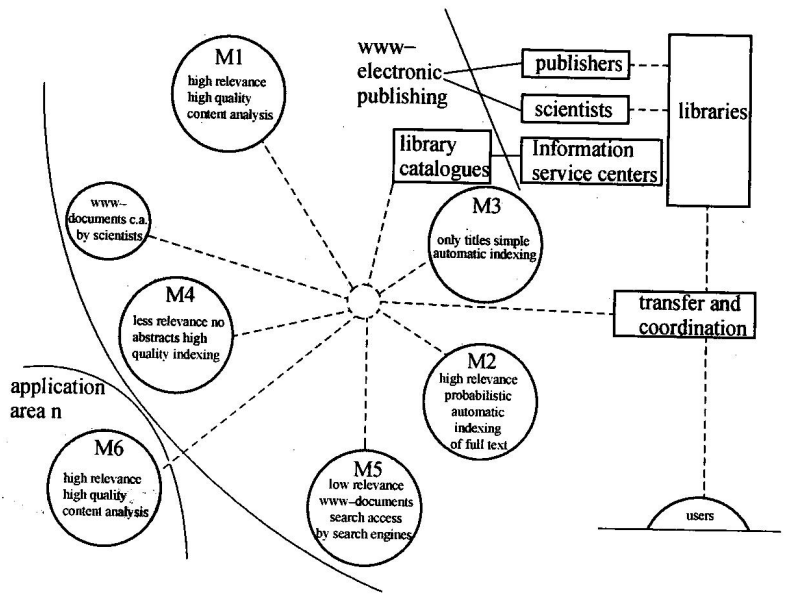


图1-1　多元化信息空间

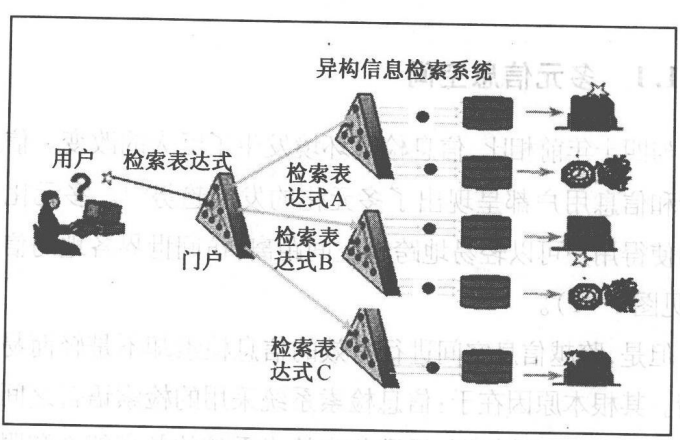


图1-2　联网环境下的信息检索模式

1.2 情报检索语言的兼容性

情报检索语言是表达一系列概括文献信息内容的概念及其相互关系的概念标识系统，包括分类、主题和代码三大语系。在文献信息检索系统中，情报检索语言是检索服务质量控制的语言保障工具。

情报检索语言的兼容性是指用某种词表的词汇或代码及其构造的检索式，可以直接适用于、或通过交换适用于多个情报检索系统[2]。事实上，各个检索语言在基本原理、原则和方法之间是相互联系的，联系的基础和实质是主题概念和逻辑关系。实现情报检索语言的兼容，就是要找到一种方法，使具有不同标识、结构、物质载体的类表和词表成分，能够在语义上互相关联起来，消除情报检索语言之间的语义异构性。因此，情报检索语言的兼容性处理，不仅有利于用户查询各种系统中的文档资料，而且有利于文献的集中处理，为文献检索网络化、集成化和共享服务提供支撑[3]。

根据情报检索语言的类型及语种，可以将情报检索语言的兼容性定义为以下几种类型[4]：

- 不同类型检索语言的兼容。比如，《中国分类主题词表》（简称《中分表》）就是我国第一部大型的、综合性的、分类与主题兼容、先组式检索语言与后组式检索语言兼容的工具书。
- 相同类型检索语言的兼容。比如，叙词表之间、分类法之间的兼容。
- 综合性检索语言与专业检索语言的兼容。比如，《汉语主题词表》与专业词表的兼容，《中图法》与专业分类表的兼容。
- 中外文检索语言的兼容。不同语种的检索语言之间兼容，实现起来较为困难，但是意义重大。
- 受控语言与自然语言的兼容。随着 Internet 的普及，自然