情报检索语言与智能信息处理丛书

丛书主编／侯汉清

自然语言叙词表自动构建研究

杜慧平
仲云云／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(丛书主编　侯汉清)

自然语言叙词表自动构建研究

杜慧平　仲云云　著

东南大学出版社

·南京·

图书在版编目(CIP)数据

自然语言叙词表自动构建研究 / 杜慧平,仲云云著.
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.自… Ⅱ.①杜… ②仲… Ⅲ.自然语言—叙词表—研究 Ⅳ.G254.24 G356

中国版本图书馆 CIP 数据核字(2009)第 200907 号

情报检索语言与智能信息处理丛书(侯汉清主编)
自然语言叙词表自动构建研究

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 5.75 印张)
总 字 数 1 310 千字(本册 150 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　言

将检索语言中的词汇控制方法引入文本检索，是改进网络环境下自然语言检索系统性能的重要手段。如何结合自然语言处理技术，通过人机结合的方式进行词间关系的处理，则是其中的关键环节之一。本书正是针对这一迫切需要解决的问题进行的研究和探索。显然，本书中所说的自然语言叙词法已经不是传统意义上的叙词法，而是指以文本中数量巨大的自然语言词汇为对象，采用传统叙词法词间关系的控制形式，包括按等级、等同和相关关系等建立的词汇系统，以便据此在网络环境下的文本检索中结合使用。这使得本书的研究更加具有普遍意义。

作为国内第一本系统探索叙词表自动编制理论和方法的专著，本书做了大量开创性的工作。首先，本书对国内外自动编制的理论方法做了较好的梳理，并且详细标注了相关引用参考文献，可供研究者在此基础上进一步扩大阅读；其次，书中采用了实证研究的方式，以财税领域词表构建为例，提出中文叙词表自动构建方案，对词表构建各个阶段的理论、方法和相关技术系统进行阐述，言之有据，便于理解；此外，书中的许多探索，从自然语言叙词表概念的提出，到基于词聚类的等级关系识别方法的探索等，均颇具新意。尽管从目前应用的角度看，书中提出的一些新的理念与其实现方式，以及技术方法方面的试验改进，有待在今后的实践中进一步发展和完善，但它们对这一领域探索中的启发作用是不言而喻的。

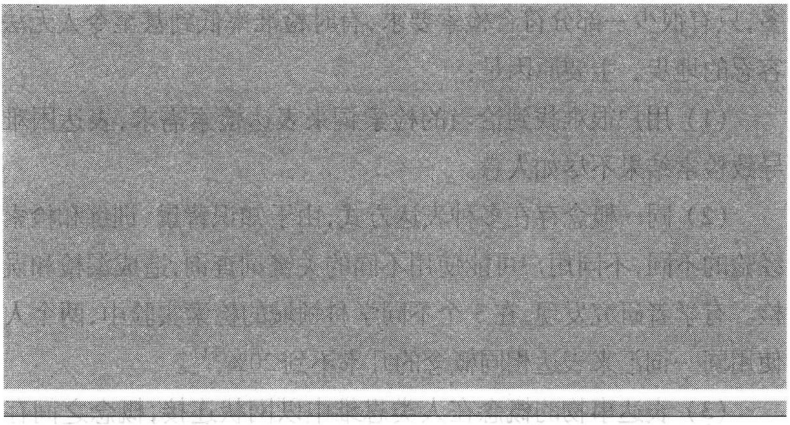
网络环境下的信息资源组织和检索改进，需要多个领域专业

人员的合作和持久探索。在计算机界、出版界、图书馆界专业人员合作的基础上制定的元数据规范就是一个典型的例子。相信这一点也同样会在知识结构的引入和词汇控制应用的实践中再次得到证实。长期以来，侯汉清老师一直致力于这一领域的研究和探讨，成果累累，同时培养出了一批高素质的专业人才，本著作就是这类成果的典型例子之一。杜慧平、仲云云两位作者在硕士学习阶段都是侯汉清教授的学生，早在她们完成硕士论文阶段，其研究达到的水准就给我留下了深刻印象，现在看到她们的成果得到进一步完善改进，即将作为专著出版，我感到十分欣喜。在将文献领域词汇控制及其应用的经验与网络环境下的应用相结合，并据以进行探索和拓展方面，青年学者的知识结构更加理想，具有一定的优势。相信本书中的研究一定会得到相关领域研究者的重视，对这一方向的研究产生积极的作用。

马张华
2009 年 7 月

目　次

图表目次



第1章
绪　论

人类进入网络时代后，信息资源爆炸性的增长趋势使人们意识到被“淹没”在信息的海洋中，如何从海量信息中有效获取所需信息成为亟待解决的问题。把叙词表等知识组织系统嵌入到网络信息检索系统中，能够有效提高检索效率，但应用环境和使用对象的改变使传统叙词表在编制和应用中面临一定困难，所以研究如何自动构建叙词表具有重要的理论意义及广阔的应用前景。

1.1　网络信息检索现状

目前，网络信息检索工具如搜索引擎等大多采用基于关键词的全文检索方式。这种方法建库简单，查找方便，但返回信息过

多,只有很少一部分符合检索要求,有时检准率低到甚至令人无法容忍的地步。主要原因是:

(1) 用户很难找到恰当的检索词来表达检索需求,表达困难导致检索结果不尽如人意。

(2) 同一概念存在多种表达方式,由于知识背景、训练和检索经验的不同,不同用户可能使用不同的关键词查询,造成漏检和误检。有学者研究发现,在5个不同学科领域的检索实验中,两个人使用同一词汇来表达相同概念的几率不到20%[1]。

(3) 表达事物的概念在人类思维中以网状连接,概念之间存在各种联系。表现在检索行为中,用户往往希望不仅得到含有检索词的文档,还能得到与检索概念相关的其他信息,虽然大多时候用户并没有明显地表达出这种愿望。

要解决网络信息检索中存在的问题,提高以搜索引擎为代表的纯自然语言检索系统的性能,需要把叙词表等控制机制引入检索系统中,实现概念检索和检索导航。所以编制适用于网络环境中信息检索所需要的叙词表成为当务之急。

1.2 叙词表编制及应用面临的主要问题

目前,计算机辅助编制叙词表技术已经较为成熟,计算机能够完成编表过程中诸如文字编辑处理、排序、打印以及词表维护等事务性操作,节省了大量人力和时间。但叙词表的词间关系仍然需要编表专家人工识别,无法克服知识获取瓶颈[2]:即对标引员或领域专家具有创建叙词表的认知要求,因此叙词表编制仍是一项知识密集型劳动。

同时,现有叙词表直接应用到网络信息检索中也存在一定问题:

(1) 现有叙词表通常反映较为通用的主题概念,将其应用到网络检索系统中,需要进行大量增删和修改,而且大多数信息检索

系统所支持的词表查询和词语转换技术会带来“嵌入迷失问题”(词表过大,使用户迷失方向)和“艺术博物馆现象”(用户花了很多时间却没找到任何详细信息)[3]。

(2) 国外一些检索试验表明,通用叙词表(General-purpose thesaurus)应用到特定领域的文献检索上,并不能明显改进检索效率[4],只有当叙词表的收词与文献中用词密切相关时,叙词表才会起作用。因此,根据领域文献本身特征,有针对性地自动及时地构建领域叙词表的方法是非常值得研究的课题。

1.3 本文研究的主要内容及意义

要实现对网络信息的有效组织,达到概念检索和智能检索,必须采用情报检索语言的基本原理——词汇控制,把检索语言和自然语言二者结合起来,探索新一代的知识组织工具。实际上,以张琪玉、侯汉清等为代表的学者早就主张情报检索语言发展的趋势是走人工语言和自然语言相结合的道路[5][6]。自然语言和人工语言在检索效率方面具有天然的互补性,自然语言词表扬弃二者的优缺点,并使二者有机结合,是一种新型情报检索语言。根据张琪玉教授的定义,“自然语言词表”指有自然语言成分的各种词表,或者说是自然语言应用于情报检索所需的各种词表[5]。

因此,针对目前词表编制和应用中存在的问题,本文旨在研究如何实现自动构建适用于网络信息组织和检索用领域叙词表。“自动构建”叙词表是指主要通过模式匹配、同现分析和聚类分析等自然语言处理技术自动识别词汇之间的等级、等同和相关关系。该词表在本领域的自然语料库基础上构建,保留了自然语言成分,并能实现对自然语言词汇的有效控制,又称作自然语言叙词表。

在国内仍以纯手工或机器辅助编制词表的现状下,自动构建领域叙词表技术在信息检索领域具有重要意义:

（1）编表速度快，费用低，时效性强。以前，叙词表完全靠手工编制，虽能够精确把握词间关系，结构复杂可靠，但是需要投入大量具有特殊要求的专业人员，成本高，构建速度慢，不易维护，而且无法克服知识获取瓶颈，即对标引员或领域专家具有创建叙词表的认知要求，因此词表编制是一项知识密集型劳动，工作量大，历时长。网络时代信息增长迅速，更新快，新词不断涌现，单纯靠手工编制词表是不现实的，这是叙词表在网络时代得不到推广使用的重要原因之一。自动构建叙词表，采用统计方法和自然语言处理技术，主要通过知识挖掘理论和方法识别潜在于语料库中的词汇语义关系，减轻编表人员的智力负担，基本保证词表质量，也能及时收录专业领域内的新概念和术语，用于词表更新。它弥补了手工编制词表的不足，具有良好的发展前景。

（2）词汇直接来源于本领域文本语料库，能客观真实地反映该领域的知识框架，提高专业领域信息检索的效率。自动构建的自然语言叙词表能够克服现有叙词表应用在网络信息检索系统中所遇到的困难。其收词一般直接来源于领域自然语言语料库，更能代表本领域的知识框架，概念更专指，能有效克服“嵌入迷失问题”（词表过大导致用户迷失了方向）和“艺术博物馆现象”（用户花了很多时间却没有找到任何有用信息）[3]。所以用自动生成叙词表的方法，有针对性地构建专业领域叙词表，是改进信息检索效果的有效途径。另外自动构建叙词表对于新兴领域尤其适用。

（3）符合网络环境中普通大众的检索习惯。随着网络的普及，信息载体逐渐从纸质文献向网页转移，人们逐渐习惯于上网搜索自己所需的知识。检索者从专门的信息服务人员转变为最终的信息需求者，信息使用者已经不再是以前图书馆中专门帮助用户查找信息的人员，而转变为来自不同行业、具有不同教育背景的普通老百姓，甚至有些人没有任何检索经验，他们更倾向于用自然语

言表达检索需求，这要求情报检索语言必须提供自然语言接口，具有自然语言与人工语言之间的自动转换功能。自动构建的自然语言叙词表遵循了文献保障原则和用户保障原则，能够把用户的自然语言检索词汇自动转换为受控语言进行匹配和检索，或向用户推荐检索词汇，减轻用户的智力负担。

（4）可用于自动标引、检索词提示与检索式扩展。自动生成的叙词表存储在数据库中，将之嵌入到信息检索系统，易于实现自动标引，对网络信息进行组织管理。对于网络用户来说，他们不熟悉词表收词范围和使用规则，给检索造成一定困难。较为理想的办法是，当用户输入表达检索需求的自然语言词汇时，检索系统能够提示与之对应的合适主题词供用户选择使用或直接转换成主题词进行检索查询。把自动构建的叙词表嵌入网络检索工具中或者作为检索系统的一个可调用接口，为用户检索策略的构造提供提示和导航，这样就提高了网络检索效率，真正体现了词表在网络检索中的应用价值。

本书内容安排如下：第1章分析目前网络信息检索存在的问题以及传统叙词表应用遇到的困难，提出本文的研究内容及研究意义；第2章概述当前叙词表在网络环境中的应用现状，简述纯人工编表、机器辅助编表和自动构建叙词表三种方式及其特点，介绍了国外优秀的词表管理软件；第3章介绍目前国内外叙词表编制技术研究进展，并分析了中文叙词表自动构建的可行性；第4章探讨叙词表自动构建的基础理论，包括理论依据、构建原则、研究方法和相关技术的介绍；第5章以构建财税领域词表为例，主要讲述自然语言叙词表自动构建方案，包括词表模式、收词选词、词间关系自动识别、词表存储和更新维护等阶段所采用的方法和技术；第6章着重研究基于词聚类的等级关系识别方法；第7章介绍自然语言叙词表自动构建系统总体设计思路及流程，描述试验用数据，