情报检索语言与智能信息处理丛书

丛书主编／侯汉清

受控词表的互操作研究

戴剑波
刘华梅／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）

受控词表的互操作研究

戴剑波　刘华梅　著

东南大学出版社
·南京·

图书在版编目(CIP)数据

受控词表的互操作研究/ 戴剑波,刘华梅著. —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.受… Ⅱ.①戴… ②刘… Ⅲ.检索语言—研究 Ⅳ.G254.0

中国版本图书馆 CIP 数据核字(2009)第 200908 号

情报检索语言与智能信息处理丛书(侯汉清主编)

受控词表的互操作研究

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 6.625 印张)
总 字 数	1 310 千字(本册 173 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

《受控词表的互操作研究》一书是戴剑波、刘华梅在侯汉清教授指导的同主题硕士论文的基础上，经过反复修改增补后出版示人的一部重要论著。该书从情报检索语言的互操作模式和方法研究出发，在分析国内外情报检索语言互操作项目的基础上，研发《中图法》与《杜威十进分类法》(DDC)的互操作系统、《中分表》为基础的教育科学的集成词库系统。虽然两个系统均为试验研究系统，但基于 Borland Delphi7.0、SQL2000 等软件工具和自动匹配及映射技术开发的互操作系统在国内均应为原型和首创系统，其研究成果意义深远。

近年来，随着网络应用技术的发展，知识服务的深入研究，跨语言、跨结构的知识组织系统互操作问题，越来越成为国际社会文献信息资源共建共享亟须解决的问题。该问题不仅是各种国际性会议和各国学科门户网站研究、合作和报道的焦点，而且也是有关知识组织方面的英美国家标准、国际标准修订补充的主要内容，这些都充分说明了戴剑波、刘华梅基于中国检索语言特点开展其互操作研究的重要性。

这部论著立意明确，结构清晰，材料翔实，有理有据。尤其对《中图法》与 DDC 类目的自动映射方法和技术，对集成系统的构建设计，语词、结构的自动匹配及类目映射，互操作的显示及可视化技术等做了大量试验研究，为汉语同义词自动识别和挖掘技术、相似度计算等积累了很多试验素材和分析方法。这些研究成果将对我国核心的知识组织系统《中图法》、《中分表》与其他国内外分

类法、叙词表等知识组织系统的互操作项目的实际开展起到技术指导和支持作用，对国内开展知识组织系统互操作项目研究有重要的参考价值。

早在20世纪90年代，国家图书馆中外文图书曾希望开展拟标引或转换DDC类号工作，《中图法》编委会办公室为此拟立项研究《中图法》与DDC等国外主要分类法的对照转换系统，并从理论和对照实践上，对其类目体系比较分析，提出立项报告，但由于多方面因素影响，该项目未被批准。十几年过去了，刘华梅已经成为《中图法》编委会办公室的一员，希望他们的这部论著能在《中图法》、《中分表》实践基地上“开花结果”，转换为知识组织系统互操作的实际产品，扩大《中图法》、《中分表》的应用范围。

在本书即将出版之际，侯汉清教授委托我为其学生的论著写序。侯教授一直是我的良师益友，借此序表达我对侯老师的仰慕和敬佩之情，也为图书情报界推出新秀。借此书出版问世之机，祝贺侯教授学子满园，桃李芬芳！祝贺戴剑波、刘华梅学业有成，更上一层！

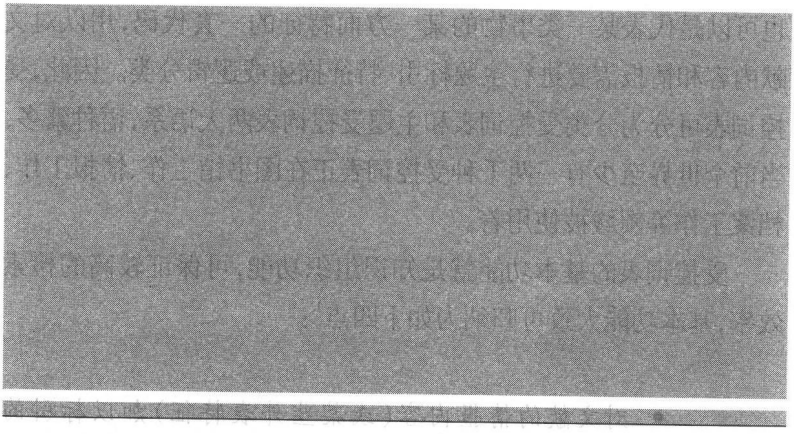
卜书庆

2009年7月

目　　次

第 1 章　受控词表的互操作 …… 1
1.1　受控词表概况 …… 1
1.2　受控词表的互操作问题 …… 7
1.3　受控词表互操作的模式和方法 …… 13
第 2 章　受控词表互操作研究进展 …… 23
2.1　国外受控词表互操作研究状况 …… 23
2.2　国内受控词表互操作研究状况 …… 40
第 3 章　《中图法》与《杜威法》映射系统研究 …… 63
3.1　分类法之间的映射原理 …… 63
3.2　分类法之间人工映射方法 …… 74
3.3　分类法自动映射系统的总体设计 …… 79
3.4　分类法自动映射系统的构建 …… 83
3.5　自动映射系统的使用与测试评价 …… 108
第 4 章　教育集成词库的构建系统研究 …… 121
4.1　集成词库的设计 …… 122
4.2　分类语言互操作技术 …… 130
4.3　主题语言互操作技术 …… 143

4.4 自然语言与受控词表的互操作 …… 154
4.5 受控词表互操作的显示及可视化 …… 160
4.6 结语 …… 176
附 录
附录1:类目与关键词主题词对应数据(样例) …… 180
附录2:CLC 与 DDC 类目批处理映射结果(样例) …… 182
附录3:分类兼容矩阵结果(样例) …… 184
附录4:主题兼容矩阵结果(样例) …… 186
名称索引 …… 189
主题索引 …… 194
后 记 …… 200



第1章
受控词表的互操作

1.1 受控词表概况

受控词表，又称受控语言或标引语言，是根据情报检索的需要而创制的人工语言，专门用于各种手工的和计算机化的情报检索系统，表达文献主题概念和检索课题概念。它作为提供文献内容检索途径的情报检索系统的一个构成因素，在其中起着语言保证作用。

受控词表的实质是表达一系列概括文献情报内容的概念及其相互关系的概念标识系统。它可以是从自然语言中精选出来并加以规范化的一套词汇，可以是代表某种分类体系的一套分类号码，

也可以是代表某一类事物的某一方面特征的一套代码，用以对文献内容和情报需要进行主题标引、特征描述或逻辑分类。因此，受控词表可分为分类受控词表和主题受控词表两大语系，语种繁多，当前全世界至少有一两千种受控词表正在图书馆工作、情报工作、档案工作等领域被使用着。

受控词表的基本功能就是知识组织功能，可保证较高的检索效率，基本功能大致可归纳为如下四点[1]：

- 对文献的情报内容（或某些外表特征）加以标引的功能；
- 对内容相同及相关的情报加以集中或揭示其相关性的功能；
- 对大量情报加以系统化或组织化的功能；
- 便于将标引用语和检索用语进行相符性比较的功能。

1.1.1 分类法概况

自1876年世界上第一部现代意义上的分类法面世以来，分类法在信息组织管理中一直占据非常重要的位置。在其100多年的发展历史中，分类法发展成了种类繁多的信息组织工具，它广义上可以分为：

- 国际通用分类法，例如《杜威十进分类法》（Dewey Decimal Classification，DDC），《美国国会图书馆分类法》（Library of Congress Classification，LCC），《国际十进分类法》（Universal Decimal Classification，UDC）等等。
- 国家通用分类法，例如《中国图书馆分类法》（简称

《中图法》，缩写 CLC)，瑞士国家图书分类法(SAB)等等。

- 专业图书分类法，例如《美国国家医学图书馆分类法》(National Library of Medicine Classification)等等。
- 自立分类体系，例如 YAHOO 等等。

就 DDC 而言，目前已经产生了 22 个版本，基本平均 6 年就产生一个新的版本。种类繁多，版本不断地更新，足以说明分类法在信息组织管理方面旺盛的生命力。

在我国第一部叙词表产生前，分类法在我国的科技情报界、图书馆界信息资源的组织管理中起着绝对的作用。1917 年，由沈祖荣、胡庆生合编的《仿杜威书目十类法》可以说是我国近代真正意义上的文献分类法。在其影响下，民国期间就先后共出现了 30 多部类似的文献分类法，形成了“仿杜”、“改杜”、“补杜”等流派。我国现代图书分类普遍认为从 20 世纪 50 年代开始，先后诞生了《中国人民大学图书馆图书分类法》、《中国科学院图书馆图书分类法》、《中国图书馆图书分类法》、《中国图书资料分类法》、《中国档案分类法》等多部大型分类法[2]。经过 50 多年的发展，我国图书分类法目前已经形成以《中图法》为主的局面，全国图书情报部门九成以上的单位都使用它。

传统分类法按照其编制方式可以分为等级列举式、分面组配式、列举组配式三种。等级列举式分类法也称枚举式分类法，是将所有类目组织成一个等级体系，并尽量列举，这种分类法有类目表达的概念较复杂、先组程度高等特点。分面组配分类法不采用详尽列举的方式，而是通过主题概念的范畴划分为不同的组面，每个组面中的概念为单元概念，任何复杂的主题都可以分解成相应的单元概念，也可以通过简单概念(单元概念)组成复杂的类目。列举组配式是一种半分面的分类法。

目前正在使用的分类法一般来说都有专门的权威机构维护更新;分类法具有按学科内容进行浏览的功能,并且具有很好的层次性和系统性;多数分类法都采用符号标记,不受限于专门的语言,可以较好地实现多语种信息检索;分类法应用于浏览时,具有较好的上下文环境,用户可以在明确的语义环境中浏览,利用分类法体系中的类目可以很好地实现信息的扩检和缩检;另外,分类法适应于非文本信息资源的组织管理[3]。正是这些优点使分类法在图书馆、情报界得到长期的使用,形成了非常广泛的用户基础。

90 年代以来,伴随着信息资源网络化、数字化的发展,网络信息资源从数量到内容都有了突破性的增长,呈现出多类型、多媒体、非规范、跨时间、跨地区、跨语种等特征,给用户查询和利用信息带来了很多困难。鉴于分类法在组织信息方面的优点,图书馆界人士开始了用分类法组织网络信息资源的研究。利用分类法进行网络信息资源的管理,进行了多种模式的探索,主要包括以下几种:

(1) 传统分类法直接应用于网络资源的管理。如同传统的信息资源一样,图书馆管理员利用分类法直接编目网络信息资源。目前,应用国际通用分类法来组织网络信息资源的试验系统中,应用 DDC 来组织网络信息资源的就达 30 多个,如 Canadian Information by Subject, Blue Web'n Content Categories, The Internet Resource Subject List in Classification Order 等[4];用 UDC 组织的如 WWW Subject Tree of WAIS Databases, BUBL, GERHARD 等共计逾 10 个;The WWW Virtual Library, Cardinal Stritch College Library 等近 20 个采用 LCC 来组织网络信息资源[5]。依据专业类表建立的专业性网络检索系统比较典型的如:瑞典技术大学图书馆建立的"瑞典工程电子图书馆"(Engineering E-Library, Sweden, EELS)。

(2) 对传统分类法进行改造,满足网络资源的自动化标引和

检索要求。OCLC(Online Computer Library Center，联机计算机图书馆中心)对于 DDC 应用于网络信息资源的组织和挖掘作了大量的研究工作，如 Scorpion 项目，主要研究电子文献的索引和编目，但重点是构架自动主题识别工具的研究；旨在增强 DDC 主题处理能力的 Dewey ETC Trees 项目；为增强 DDC 自然语言处理能力的 Wordsmith 项目等等。目前在广泛使用 DDC 的视窗版和 WEB 版，就是改造传统分类法应用于网络信息资源组织非常成功的实例。

(3) 重新编制分类法。传统分类法不适合处理网络信息资源，需要重新编制适应自身需要的分类法。例 YAHOO 等搜索引擎就采用了自编的分类体系，本质上也可以视为一种分类法。另外，将传统分类法与网络搜索引擎相结合而成的网络分类搜索引擎，吸收了网络搜索引擎的长处，并能改善传统分类法不能反映网络信息新主题，检索途径单一，用户服务面窄等不足[6]，具有很大的发展前景。

1.1.2　主题法概况

主题法是各种主题受控词表的一个统称，是用自然语言语词或受控的自然语言语词直接表达主题概念，按语词字顺排列主题概念，并用参照系统显示概念之间关系的受控词表[7]。

主题法真正发展历史已有 100 多年。1895 年美国图书馆协会根据克特的思想，编制、出版的《字典式目录使用的标题表》，即《美国图协标题表》是世界上第一部大型的标题表，也可视为世界上第一部真正意义上的主题法[8]。进入 20 世纪 50、60 年代，又在标题法的基础上陆续发展出元词法、关键词法和叙词法。我国主题法的研究起步较晚，1964 年，我国航空部编印《航空科技资料主题表》，是我国建国后编制的第一部主题词表。1979 年，由中国科

学技术信息研究所参与组织编制出版的《汉语主题词表》是我国叙词语言发展的重要里程碑，从理论和实践上都为促进我国主题词表的进一步发展起到了极其重要的作用[9]。之后相继出版了百余部主题词表，几乎覆盖了各专业和文献类型，构成了我国叙词语言体系。

这种完全建立在自然语言基础上的主题法，由于能直接以事物为中心集中文献信息，以直观的语词表达信息检索要求，采用字顺方式组织信息，符合用户在获取信息时的方便性和易用性要求，很好地满足了用户特性检索的需要，因而，一度成为信息组织的主流方法，同分类法一起构成了信息组织与检索的两种主要方法。

主题语言用主题词组织与揭示信息具有直接和直观的特点，而且其标识基本上是独立完整的事物概念，满足人们对特定事物、特定主题检索的需要，因而在网络环境中也得到广泛的应用。主题法在网络信息组织中的使用主要表现为两种方式，一是使用现有词表（叙词表、标题表）组织网络信息。目前，使用现有词表组织网络信息的还不多，主要是《美国国会图书馆标题表》（Library of Congress Subject Headings，LCSH）和《医学主题词表》（Medical Subject Headings，MeSH）被一些网络信息检索系统采用。采用LCSH的系统有：Clinic Web Browse，Alphabetical List of NLM Sections等。二是广泛采用关键词法。由于关键词法具有种种优点，关键词的抽取完全可以自动化，因此关键词检索在网络中的应用相当广泛。目前，大部分搜索引擎的索引数据库几乎都采用关键词法进行信息组织，如Ahavista是关键词搜索引擎的典型代表[10]。

数以千计的分类表、叙词表以及各种自然语言词表纷纷问世，在满足信息组织和信息检索不同需求的同时，也给信息检索，尤其是跨学科、跨数据库、跨语种等检索带来种种困难。而解决此种困难的对策之一，就是研究受控词表的兼容性，建立不同受控词表之

间的互操作系统。

1.2 受控词表的互操作问题

1.2.1 受控词表的兼容化

任何一种受控词表,无论它的体系多么完善,所采用的方法多么先进,一般说来它都不可能适用于一切检索系统,满足一切检索要求,于是就提出了受控词表的兼容化问题。

所谓兼容,是指两个实体结合起来工作的能力。具体地说,受控词表的兼容性是指不同词表、类表之间可以实现兼容与互换,即用某种词表的词汇及其构造的检索式(或标引记录),可以直接适用于、或通过交换适用于多个情报检索系统[11]。联网环境下受控词表的兼容,是指用户只用一种受控词表或者不用任何受控词表(即直接使用自然语言)就可以实现联网环境下的跨数据库检索。也就是说,用户只要使用一个检索式就可以直接检索多个相关领域的数据库,而不需要每检索一个数据库就重新构造一个检索式。

实现受控词表之间的兼容,就是要找到一种方法,使具有不同标识、结构、载体的分类表或主题词表的成分互相联系起来[12]。这不仅有利于人们对各种不同文档的查询,而且有利于文献的集中处理,为文献检索网络化的实现提供可靠保证。总之,兼容化是为了提高网络资源检索效率、实现信息资源共享提出的新要求,它是受控词表发展的重要趋势。

受控词表兼容性的范围包括[13]:

- 不同类型受控词表的兼容,如分类法与主题法的兼容,也称分类主题一体化。如《中国分类主题词表》(简称《中分表》)是我国第一部大型的、综合性的,分类与主题兼容、先

组式受控词表与后组式受控词表兼容的工具书。

- 同一类型受控词表的兼容，如各种叙词表之间、各种分类法之间的兼容。

- 综合性受控词表与专业受控词表的兼容，如《汉语主题词表》与专业主题词表的兼容，《中图法》与专业分类表的兼容。

- 中外文受控词表的兼容，这关系到受控词表的国际通用性，实现起来较为困难，但却意义重大。

- 规范化语言与自然语言的兼容，这是在自然语言的应用越来越广泛的情况下提出的，随着计算机的普及并向网络化方向发展，这一问题已经受到越来越多的重视。

在两个词表之间进行转换时，影响其转换效果的，主要有以下几个因素[14]：

- 词表所覆盖主题领域的重叠程度。重叠越多，转换越易；重叠越少，转换越难。如将一个工程方面的词表转换成一个医学方面的词表，效果不可能理想，两个词表中共同的概念范畴和共同的词汇都很少，因此很难进行对应转换。

- 词表结构化的程度。显然，结构相似的两个词表进行词汇的对应转换时，找到对应词比较容易，结构完全不同的两个词表找到对应词比较困难。

- 词汇专指度。专指性高的词易转换成专指性低的词，反之，专指性低的词不易转换成专指性高的词。前者转换可通过多对一的方法实现，后者的转换则很难。

- 词汇的先组程度。两个词表的先组程度越相似，转换概念越容易。若两个词表中的词汇先组程度相差很多，则由