基于受控词表互操作的集成词库构建研究*

刘华梅　侯汉清

摘　要　目前,国内外信息领域都在致力于受控词表的互操作研究。本文以教育类数据为例,通过对检索语言互操作技术的研究,借鉴国内外实现不同受控词表之间互操作的经验和方法,建立一个以《中分表》为核心的兼容体系,即建立一个可以不断扩充的集成词库。为了便于浏览和使用词库数据,采用单机模式、XML 格式及本体构建工具对词库兼容数据进行可视化显示,进而为用户提供各种服务。图 6。参考文献 11。

关键词　分类法　主题法　互操作　集成词库　《中分表》
分类号　G254

ABSTRACT　At present, information professionals have been working on the issue of the interoperability among different indexing languages. Through studies of interoperability technique between indexing languages, the authors construct a compatible system, which takes the class Education as a test example and regards *Classified Chinese Thesaurus* (CCT) as core and extensible basis. In order to browse and use the compatible data, the system adopts stand-alone mode, XML and ontology tools for display, and can provide various services to users. 6 figs. 11 refs.

KEY WORDS　Library classification. Subject indexing. Interoperability. Integrated vocabulary. *Classified Chinese Thesaurus*.
CLASS NUMBER　G254

1　研究背景

叙词表、分类表作为我国传统的知识组织工具,不仅广泛应用于图书、文献的分类主题标引,也适用于网络信息资源的组织和检索。但由于存在主题词表、叙词表、分类表和网络分类法等多种信息组织方式,使得用户在检索同一学科或主题的文献时,需要使用不同的检索标识,在用户不熟悉各种词表的情况下,检索变得尤为困难。由此可见,实现检索语言的标准化和兼容化势在必行。

关于受控词表的互操作问题,早在 20 世纪 60 年代就已提出,一直是图书情报学界的研究热点,国内外很多学者都致力于这方面的研究,也提出了很多解决方法和意见。实现互操作的方法有很多,如中介词典、宏观/微观词表、集成词表等等。

国外一直致力于多种分类法、主题法及自然语言之间的互操作,在这方面开展了很多积极有效的研究计划,如 CAMed(Complementary and Alternative Medicine)项目,由哥伦比亚大学与肯特州立大学主持,是对医学资源的补充和选择,是一个集英、法、中、日等国叙词表的集合词表管理系统和跨词表检索系统[1]。MACS(Multilingual Access to Subject)项目由法国、德国、英国、瑞士的国家图书馆共同完成,在 SWD(德语)、RAMEAU(法语)、LCSH(英语)这三个主题标题表中概念相等的标题词之间建立相等关系[2]。HILT(High-Level Thesaurus Project)项目主要研究用不同知识组织系统组织的网络信息资源的检索和浏览问题,通过建立一个转换机制在不同知识组织系统之间提供服务,实现跨库、跨领域的信息检索和浏览[3]。UMLS(Unified Medical Language System,一体化医学语言系统)是美国国家医学图书馆(NLM)自 1986 年开始研制的项目,其是计算机化的受控词表集成系统,它不仅是语言翻译、自然语言处理及语言规范化的工具,而且是实现跨数据库检索的词汇转换系统[4]。

* 本文为硕士学位论文的部分研究成果。

综观国外这些项目,都是以集合词表或词库的模式来实现各种检索语言之间的互操作,并已应用到具体的信息组织与知识服务中。而国内在这方面的研究还很少,20 世纪 90 年代初提出的构建“国家叙词库”项目也因故未能完成[5],医学领域正在研制的“统一的中国医学语言系统”和“中医药一体化语言系统”可以说是这方面的具体实例[6],但该项目的实施更多地依赖手工操作完成,不具有可推广性。

早在 1998 年侯汉清教授就曾撰文提出建立以《中国分类主题词表》(以下简称《中分表》)为核心的检索语言兼容体系[7],本文正是在这一思想的基础上进行构思,利用计算机技术、自然语言处理技术、数学方法等来解决同义词识别、语词匹配、映射关系发现等问题,实现不同分类法、主题法之间的互操作,进而构建一个以《中分表》为核心的教育集成词库,并在此词库的基础上开展各种服务。

采用集成词库的模式,不仅可以将各种词表或类表全部收集、利用起来,而且不需要对原词表和类表进行加工,主要通过发挥计算机联机显示、统计及转换等功能,实现互操作;也无须建立完全等价的对应关系,只要表现出所有相关及同义关系即可;还可以随时添加新的兼容词表,不断扩充词库。另外,《中分表》作为我国最主要的分类主题一体化词表,在国内有着独特的地位,有着最广泛的影响和最众多的用户,现在大多数的图书馆都采用它来类分、标引图书。再者,《中分表》实际上已经起到在不同程度上兼容各种专业分类表和叙词表的作用。《中分表》本身就属于一种将分类语言和主题语言融为一体的集成词表,它的兼容模式易于扩充和发展,而且适应性广。所以本系统选取《中分表》作为核心,实现其他分类法、主题法到它的兼容。

2 集成词库的设计

集成词库主要是将某一特定主题领域的若干叙词表或分类法汇编在一起,通过识别等价词及准等价词建立一个词汇转换系统,用于在联合分类或标引活动中实现分类法和叙词表之间的互操作。

本文以教育类为实验对象,收集现有的各种词表(含分类表、主题词表、关键词等)的教育数据,以《中分表》为核心,通过一定的方法实现其与分类语言、主题语言之间的互操作,从而构建一个教育集成词库。集成词库的具体构建过程包括以下几个步骤:

(1)选择语料库,获取实验数据。选择中外著名的教育主题词表、分类表以及综合性词表或分类表中的教育大类。对于有电子版本的词表,通过预处理,规范其格式,将其数据导入数据库直接使用;对于没有电子版本的词表,采用扫描识别或手工录入的方法来获取数据。本文选取了《中分表》(教育类)、《科图法》(教育类)、《杜威十进分类法》(教育类)、《教育主题词表》、《社会科学检索词表》以及从教育大辞典、图书馆编目数据中抽取的教育类关键词或关键词串等数据。

(2)利用计算机技术、自然语言处理技术、数学方法等实现不同词表到《中分表》的互操作,包括各种分类法的互操作、主题法的互操作。

(3)集成词库的存储。本文采用兼容矩阵的形式来存储词库数据,具体包括两种形式。

- 字顺兼容矩阵:将《中分表》的主题词或主题词串按字顺方式纵向显示,并标明其相应的《中图法》分类号,把其他参与兼容的主题词表横向展示,与《中分表》的主题词或主题词串相对照,列出其等值兼容或近似兼容的一个或多个主题词。
- 分类兼容矩阵:将《中图法》的分类号按顺序纵向显示,并列出其对应的主题词,把其他参与兼容的分类表的类号与《中图法》的类号相对照,列出其等值兼容或近似兼容的概念,并将与《中图法》类号相对应的关键词列出。

(4)建立词库管理和应用系统,将互操作结果可视化显示。为了用户更好地浏览和利用词库中的词汇数据,实现词库的信息交换和信息再利用,采用数据库的单机模式、XML 元数据格式以及基于本体构建工具对词库数据进行可视

化显示。

综上所述,集成词库构建的主要流程如图 1 所示:

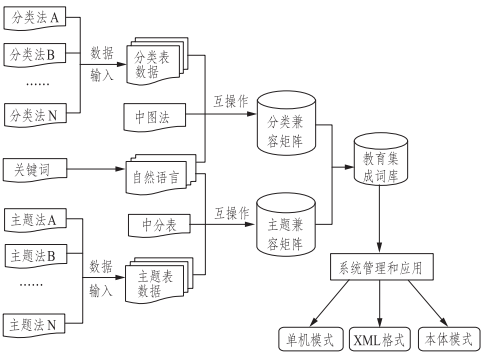

图 1　教育集成词库的构建流程

3　集成词库构建系统模块

根据集成词库构建的基本原理和流程,构建系统模块分为分类法互操作、主题法互操作、数据显示等模块,如图 2 所示:

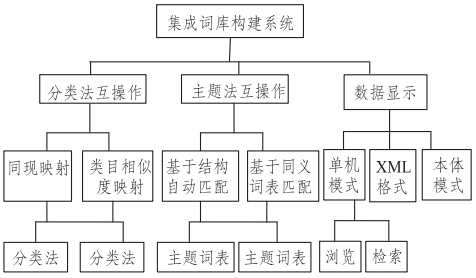

图 2　集成词库构建系统模块结构

3.1　分类法互操作模块

该模块主要是完成不同分类法与《中分表》分类法部分的互操作,具体技术包括同现映射技术和基于类目相似度的映射技术两种[8]。

(1)基于同现信息的互操作

分类法的实质就是表达一系列文献情报内容概念及其相关关系的号码标识系统,可以用不同分类法的分类号来标识同一文献或图书。与之相应,标识同一文献或图书的不同分类号之间必定具有一定的关联,所以本文采用基于不同分类法的同现信息来确定类目的兼容关系。如果两部分类法的分类号经常出现标识同一文献或图书的情况,则认为这两个分类号是可以互相兼容转换的。目前的图书馆普遍采用 MARC 机读目录格式存储数据,对同一图书可以采用多种分类法进行分类,用不同的字段来标识,如 690 字段是中图法分类号,692 字段是科图法分类号,所以可以从图书馆的机读目录中下载一批同现数据。另外,对同一图书,在不同的系统中可能用不同的分类法标识,也可以通过查找不同系统,找到同一图书的不同对应分类号。

首先从图书馆书目数据中收集共现信息,并利用程序分别统计出各类号单独出现的频次和同时出现的频次。然而,由于不同分类法的体系结构不同、类目含义不同、标引不一致等,会造成一部分类法的某一类号与另一部分类法中多个类号兼容的现象,这其中势必存在着错误的标引结果,从而影响互操作的准确性。本文采用基于统计学的相关度算法——Dice 测度方法计算两类号的相关度,对同时出现的不同分类号进行约束,Dice 测度公式表达形式如下:

$$\mathrm{Dice} = \frac{P(A,B)}{\frac{1}{2}[P(A)+P(B)]}$$

通过计算得到两种类号的 Dice 测度值,并按值从大到小排序,根据多次试验结果分析确定一个阈值,在阈值范围内的作为备选兼容类号,最后再进行适当人工干预,实现不同分类法间的互操作。

(2)基于类目的直接映射

等级体系分类法一般是通过类目和注释来表达各种复杂概念的,并且对类目名称进行了一定的规范化,要求类目名称用词确切,能反映类目的实际内容范围,采用比较通行的科学名词;当类目名称不能确切表达类目的实际内容时,使用注释加以补充、说明[9]。所以本文选择类目对应的类名词及其注释词进行映射计算。

首先对类目概念和注释进行规范化处理,包括去掉停用词及特殊符号,将多主题概念进

行分解,从注释中提取能表达主题概念的语词等等,经过处理后将每个类目表示为多个能表达完整概念的语词,然后将其进行格式转换,形成词—分类号的对应形式。最后利用词素相似度计算方法分别计算类名词、注释词的相似度,如计算 A、B 两词的相似度,则计算公式如下:

$$P(A,B)=p\times\frac{\left(\frac{N(A,B)}{N(A)}+\frac{N(A,B)}{N(B)}\right)}{2}+q\times L(A,B)\times\frac{\left(\sum\frac{Q[N(A,B)]}{\sum Q[N(A)]}+\sum\frac{Q[N(A,B)]}{\sum Q[N(B)]}\right)}{2}$$

经过上述公式计算后,得到各词两两之间的相似度,根据多次试验结果分析确定一定的阈值,在阈值范围内的词作为对应词。对于一个词存在大量对应词的情况,按其相似度选择排在前面的词,进而通过语词的对应关系确定类目的对应关系。

3.2 主题法互操作模块

该模块主要是完成不同主题法与《中分表》主题词部分的互操作,具体技术包括基于词表结构的自动匹配技术和基于同义词表的语词相似度匹配技术[10]。

(1)基于词表结构的自动匹配

自动匹配转换实质是借助各词表本身结构的兼容性,当词汇以机器可读形式存在时,使两词表相互对应的词可由计算机自动进行匹配转换[11]。通常情况下,两词表的结构越相似,学科覆盖重合率越高,可自动转换的词就越多。对词表兼容性影响较大的主要是词表的微观显示结构,即每一条叙词款目的构成。如果两表的显示结构越相似,数据处理就越容易,二者的兼容转换就越容易实现。

主题词表都具有规范的结构形式,除主题词本身外,还包括代、属、分、参等参照项内容。本文通过一定的格式转换,将词表中的所有叙词、非叙词逐一列出,然后通过计算机自动实现完全匹配、同义词匹配、组配匹配等过程,在各词表中完全相同的叙词或非叙词之间以及组配的主题词串之间实现兼容。整个匹配模式见图 3。

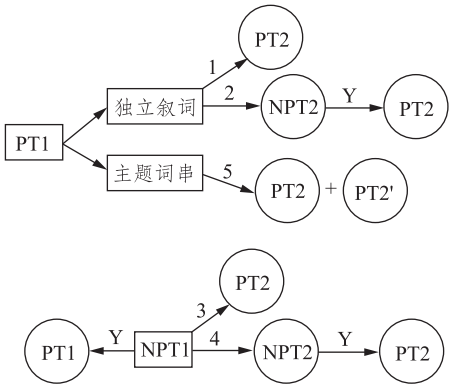

图 3 自动匹配模式

(2)基于同义词表的语词匹配

该种技术的基本思想是识别出不同词表中的同义词,将其进行匹配,此处的同义词包括意义完全相等的词以及意义相近或相关的词。由于汉语构词特点,大部分意义相同或相近的语词大多包含有相同的字,所以基于单汉字或词素的字面相似度算法是比较常用的一种方法;但该方法最大的不足是对于字面不相似的异形同义词不能很好地识别。

针对上述不足,本文考虑在该算法中引入同义词表,以此提高计算的准确度。基本思想是:首先利用现有的各种技术编制一部语义精良的同义词表,该同义词表可以包括受控词、非受控词、表达完整概念的语词以及不可再切分的词素等;然后基于上述同义词表采用自动分词技术将匹配词切分成一系列词或词素的集合;再根据切分的词或词素设计算法计算相似度,在设计算法时先将同义词表中有的词或词素进行语义匹配,对无法采用语义匹配方法的词采用基于词素的字面匹配;最后提取相似度在一定阈值范围内的词作为同义词或相关词,匹配到对应《中分表》主题词下。匹配过程如图 4 所示。

通过采用以上不同的方法基本完成了分类语言和主题语言之间的兼容转换,建立了分类兼容矩阵和主题兼容矩阵两个集成词库,但互操作结果以后台数据库形式存储,不方便用户查看和使用,需要通过一定的可视化界面将其显示给用

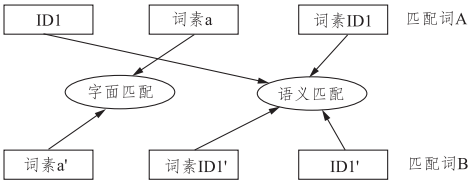

图 4　同义词相似匹配算法示意

户。本文设计了基于数据库的单机模式、XML 格式和基于本体构建工具对词库兼容数据进行可视化显示,进而为用户提供各种服务。

3.3　数据显示模块

(1)基于数据库的单机模式显示

设计词库应用和管理系统,根据需要读取后台数据库的数据,以系统界面的模式将结果返回给用户,供用户浏览和选择。浏览功能主要是通过《中图法》分类号按树形结构展开,可以查看其对应的其他兼容分类表的信息,还可以查看该分类号对应下的主题词与其他主题词表的互操作信息。检索功能提供了分类号、主题词、关键词三种检索途径。关键词可以是用户输入的任意词,系统为其转换到对应或相关的《中分表》分类号或主题词,进而再查看其他对应词表的相关信息。单机模式数据查看界面如图 5 所示。

图 5　单机模式数据查看界面

(2)基于 XML 格式显示

采用系统界面模式显示数据,需要多次读取数据库,并且数据不能一次完全显示,只能单条浏览、检索,不便于用户使用。所以为了使其更广泛、有效地得到利用,实现词库的信息交换和信息再利用,现将其转换为 XML 文档,进而可以在 XML 文档上进行数据浏览和使用。基本思想是:以《中图法》分类号为起始节点,类目为概念节点,然后依次显示该类目在其他分类法、主题法中的对应信息。分为两种情况:一种是主题格式(subject. xml),即对应结果只显示类目在其他词表中的对应概念词;另一种是主题描述格式(subject_description. xml),是要具体显示各对应词表的详细信息,如分类表显示其对应分类号、类名、注释,主题词表显示对应主题词及其分类号、英译名、参照关系(代、属、分、参)等内容。

(3)基于本体构建工具的可视化显示

基于 XML 格式存储和显示数据,虽然可以将词库数据进行整合,一次显示多条记录,但只能以树状层次结构展开,可读性差,不便于用户理解。本体可以将某个知识体系的信息资源进行结构化组织,反映词汇之间的语义关系,并且可以利用本体构建工具,将知识体系以图形化的界面形式提供给用户。所以本文利用 Protégé2000 本体构建工具,将词库中的数据进行可视化显示。基本思想是:首先以《中分表》概念主题作为首级类,包括其分类号、类名词及主题词;子类包括类目对应的其他兼容分类法以及主题词对应的其他兼容主题词表。然后确定出属性,包括类号和词之间的对应属性,对应的分类表、主题表属性,以及一些词间关系属性,如用、代、属、分、参等。最后加入实例,如分

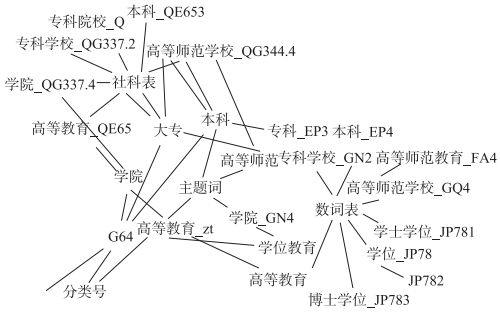

图 6　基于本体构建工具的可视化显示

类号“G64”,对应类名词“高等教育”,对应主题词“本科、大专、学院、高等师范学院”等,类名词和主题词再分别对应其他词表中的兼容信息,如图6所示。这种方式可以从不同的角度对词库信息进行多维度的揭示,把有语义联系的事物都连通起来。

4 总结

本文是国内在采用集成词库模式解决受控词表互操作方面的首次尝试,将会对我国词表互操作和智能信息检索的发展有一定的推动作用。具体可以应用于下列方面:

(1)可以为用户提供对应的分类号、主题词或关键词,进行检索服务,进而可以开展网络检索的提问扩展,实现概念检索,为构建本体、语义网、主题图等知识组织系统提供语义来源等术语服务。

(2)实现分类法、主题法之间的互操作,可以减少图书馆员、情报工作者等的工作强度,节省标引、分类、编目时间;一次概念判断,可以同时赋予多个分类号或主题词。

(3)可以使采用不同分类号类分的图书集中收藏,便于读者查询和借阅;还可以与国外分类法兼容,从而更好地实现中、英文文献的共享。

(4)把现在“闲置”的各种词表全部收集、利用起来,发挥这些术语资源或语义工具的作用。

(5)可以用于修订、更新各种词表,还可用于编制各种专业叙词表、电子政务叙词表或多语种叙词表。在此基础上还可以建设专业术语数据库、同义词词典等。

当然,本系统是一个实验性系统,还存在一些不足,很多方面有待改善和扩充:①分类表映射过程中没有考虑复分、仿分,这在推广使用中会有很多障碍,以后可以考虑解决;②映射方法上还需要创新,提出新的方法实现情报检索语言之间的互操作,提高转换的准确度,另外计算机参与的自动化程度还有待于提高;③基于词库、XML的Web应用及本体可视化,上传下载、扩展检索等一些术语服务有待开发。

参考文献:

[1] Thesaurus construction and maintenance [EB/OL]. [2010-01-15]. http://circe.slis.kent.edu/mzeng/bin/tmstools.exe? DBN = TMS.

[2] MACS. Multilingual Access to Subjects [OL]. [2010-01-15]. http://infolab.uvt.nl/pub/hoppenbrouwersj-2001-23.pdf.

[3] HILT: High-level thesaurus project proposal [EB/OL]. [2010-01-18]. http://hilt.cdlr.strath.ac.uk/abouthilt/proposal.html.

[4] Unified medical language system [EB/OL]. [2010-01-28]. http://www.nlm.nih.gov/research/umls/about_umls.html.

[5] 朱岩. “国家叙词库”建库设计与分析[J]. 情报理论与实践,1991(4):28-30.

[6] 张晓梅,李丹亚,胡铁军. 一体化医学语言系统与本体论研究[J]. 医学信息学杂志,2006(2):89-92.

[7] 侯汉清. 建立以《中国分类主题词表》为核心的检索语言兼容体系[J]. 北京图书馆馆刊,1998(4):35-39.

[8] 刘华梅,侯汉清. 基于类目相关度和相似度算法的分类法互操作技术研究[G]// 中国科学技术信息研究所,全国信息与文献标准化技术委员会. 信息资源组织及其标准规范学术研讨会论文集. 北京:2008:35-42.

[9] 戴剑波,侯汉清. 文献分类法自动映射系统的构建:以《中国图书馆分类法》与《杜威法》为例[J]. 情报学报,2006(5):595-596.

[10] 刘华梅,侯汉清. 叙词表互操作技术研究:教育集成词库的试验[J]. 中国图书馆学报,2008(5):95-98.

[11] 张雪英,侯汉清. 叙词表词汇转换系统的设计[J]. 情报学报,2000(5):451-457.

刘华梅 国家图书馆馆员。通讯地址:北京市海淀区中关村南大街33号。邮编:100081。

侯汉清 南京农业大学信息管理系教授,博士生导师。通讯地址:南京。邮编:210095。

(收稿日期:2009-05-22;
最后修回日期:2009-08-14)