●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[1]

摘　要　目前,国内外信息领域都在致力于情报检索语言的互操作研究。介绍了基于词表结构的自动匹配和基于同义词表的语词匹配两种互操作技术。以教育类数据为例,选取了《中国分类主题词表》、《教育主题词表》、《社会科学检索词表》等叙词表,采用构建集成词库的方法实现不同叙词表间的兼容。图 3。表 4。参考文献 11。

关键词　叙词表　互操作　集成词库　《中国分类主题词表》　主题检索语言

分类号　G254.2

ABSTRACT　In this paper, the authors introduce the studies in interoperation of information retrieval languages in foreign countries and the two interoperation technologies of automatic matching based on thesaurus structures and word matching based on synonym thesauri. Taking data in the education class as an example, the authors use the method of constructing an integrated vocabulary to realize different thesauri, such as Classified Chinese Thesaurus, Educational Thesaurus and Thesaurus for Social Science. 3 figs. 4 tabs. 11 refs.

KEY WORDS　Thesaurus.　Interoperation.　Integrated vocabulary.　Classified Chinese Thesaurus.　Subject retrieval language.

CLASS NUMBER　G254.2

1　研究背景

从 20 世纪 60 年代起我国开始研究主题检索语言,到目前为止,已编叙词表约 100 多种,这些无规律、不受任何控制地增长的众多词表,给文献检索增添了新的困难和麻烦,于是,出现了叙词表兼容化的趋势。由于各个叙词表之间存在着较大差异,要实现它们之间的转换是一项非常复杂的工作。近年来,国内外学者一直在探讨检索语言的互操作问题,提出了多种解决方法,主要有自动匹配转换、中介词典、集成词表、叙词词库、映射、翻译[2]等,并在此基础上完成了很多互操作的项目[3],为用户的信息检索带来了很大方便。

我国从 20 世纪 90 年代初提出要实现统一的词表兼容体系。1991 年,傅兰生撰文分析了我国叙词兼容的两大方案——大词表方案和词库方案,论证了词库方案是现实可行的方案[4],并最终在 1989 年由国家科委立项,开始了“国家叙词库”构建项目。计划将国内几十部叙词表用集成词表方式汇合成一个国家叙词库,以促进全国情报检索语言的互操作。此后,朱岩、洪漪等都先后撰文[5-7],对“国家叙词库”的构建进行了具体的设计与分析。中国农科院科技文献信息中心成功地利用计算机建立了农业叙词库[8]。由于当时的技术水平有限加之经费短缺,项目只实施了第一期计划就中断了。但是该项目为我们提出了建立集成词库、实现不同检索语言之间互操作的思想,值得我们研究借鉴。侯汉清早在 1998 年就提出了建立以《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[9]。

基于以上经验和思想,本文采用构建集成词库的方法实现不同词表间的兼容。所谓集成词库,就是将某一特定主题领域的若干叙词表或分类表融合在一起,在各源词表的基础上建立一个含全部词条及相关参照的母表。可以通过识别等价词及准等价词建立一个词汇转换系统,从而实现词表间的互操作。

具体是建立一个以《中分表》为核心的兼容体系,即集成词库。这个集成词库包括:《中图法》与国内外分类法的互操作,《汉语主题词表》(以下简称《汉表》)与专业叙词表的互操作,以及受控语言与自然语言的互操作。本文主要研究叙词表的互操作技术,以教育类数据为例,选取了《中分表》、《教育主题词表》(以下简称《教词表》)、《社会科学检索词表》(以下简称《社科表》)等叙词表,具体介绍了基于词表结构的自动匹配和基于同义词表的语词匹配两种

* 本文为国家社科基金项目“基于知识组织系统的电子政务信息资源管理”(编号:05BTQ021)的研究成果之一。

叙词表互操作技术的实现过程。

2 基于词表结构的自动匹配

2.1 基本原理

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,在实现不同主题词表间的互操作时,也要考虑这些参照项内容,最终在表达同一概念的不同表达形式的主题词之间建立映射关系。

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由计算机自动进行匹配转换。通常情况下,两词表的结构越相似,学科覆盖重合率越高,则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。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。张雪英在她的硕士论文中曾利用这种方法设计了经济叙词表——叙词表的转换系统,证明了该方法的可行性[10]。

2.2 实验过程

本文也利用这种思想,通过词表本身的结构形式和款目参照关系,进行自动匹配,实现词表中部分词汇的转换。选用的《中分表》、《教词表》和《社科表》,都属于分面叙词表,款目格式几乎完全相同,所以利用其结构形式进行词表间的自动匹配是完全可行的,下面具体说明该方法的可行性。

(1)词表数据格式转换:主题词表对应的记录字段包括主题词、分类号、用(Y)、代(D)、属(S)、分(F)、参(C)等,即将每个词汇及其所有参照项组成一条记录,记录格式如表1所示。通常情况下,人们习惯用这种方式存储词表,以尽可能与原表保持一致。但在联机集成词表中,这样的存储格式既不便于词表数据的读取,也不便于词表的动态维护。所以,利用程序进行转换,将每个词汇对应的参照项逐条显示,形成词、关系、词及其词表来源的格式,具体格式如表2所示。

表1 词表记录格式

叙词	分类号	代(D)	属(S)	分(F)	参(C)	词表来源
本科	G64	本科教育;	高等教育;		大专;	中分表
家长教育	BD925.92	父母教育;			家长工作;家长学校;	教词表
勤工俭学	Q093.555	勤工助学;			半工半读;	社科表

表2 词汇关系表

word1	relation	word2	source
本科	D	本科教育	中分表
本科	S	高等教育	中分表
本科	C	大专	中分表

(2)反参照的生成。对实验过程(1)中生成的对应词汇,利用计算机程序生成对应的反参照条目,"D"对应"Y","S"对应"F","C"对应"C"。经过上述处理后,将所有叙词、非叙词及其对应参照项集成在一起,方便数据的读取。

(3)自动匹配。通过上述转换后,将词表中所有叙词、非叙词逐一列出,然后分别实现由《教词表》、《社科表》到《中分表》的自动匹配转换,匹配类型分为三种:① 完全匹配:指将两词表的叙词进行完全匹配,即《教词表》或《社科表》中的叙词在《中分表》中可以找到完全对应的叙词,则直接将其对应到该词下。② 同义词匹配:此处的同义词指在某一词表中具有"用"、"代"关系的词汇。可以根据各表提供的不同参照将这些词汇集中起来,即将两词表的叙词与非叙词、非叙词与非叙词进行匹配。③ 组配匹配:因为《中分表》主题词中很多是采用组配方式构成的主题词串,而《教词表》、《社科表》中都是独立的叙词,在这种情况下,将独立叙词也同样采用组代形式构成词串,进而与《中分表》的主题词串对应。匹配模式见图1:

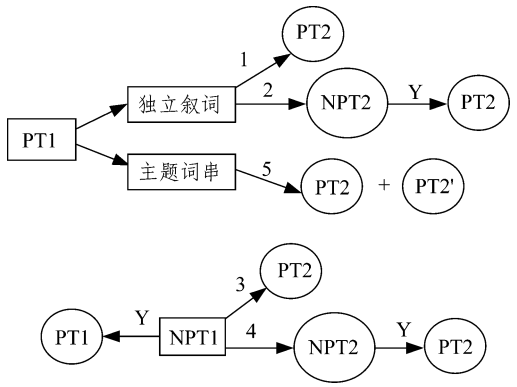

图1 自动匹配模式

其中,PT1 表示源词表中的叙词,NPT1 表示源词表中的非叙词; PT2 表示兼容词表中的叙词,NPT2 表示兼容词表中的非叙词; 1 表示叙词和叙词匹配,即 PT1 和 PT2 直接匹配; 2 表示叙词和非叙词匹配,即 PT1 和 NPT2 匹配,又 NPT2 Y PT2,则 PT1 和 PT2 匹配;3 表示非叙词和叙词匹配,即 NPT1 和 PT2 匹配,又 NPT1 Y PT1,则 PTI 和 PT2 匹配; 4 表示非叙词和非叙词匹配,即 NPT1 和 NPT2 匹配,又 NPT1 Y PT1,NPT2 Y PT2,则 PTI 和 PT2 匹配。5 表示如果 PT1 是主题词串,则用 PT2 + PT2′组配匹配。匹配过程按 1、2、3、4、5 的顺序进行,最终在各词表的正式叙词之间建立关系。

2.3　结果分析

通过上述方法,分别完成《教词表》和《社科表》到《中分表》的匹配,匹配结果见表 3。

表 3　自动匹配结果统计

匹配类型 / 兼容词表	完全匹配(A)	同义词匹配(B)			去重总计(A+B)	叙词总数	百分比(A+B)	组配匹配(C)	主题词串	百分比(C)
	$\frac{PT1}{PT2}$	$\frac{PT1}{NPT2}$	$\frac{NPT1}{PT2}$	$\frac{NPT1}{NPT2}$						
教词表	446	48	102	76	574	915	62.73%	424	736	57.6%
社科表	216	24	56	31	299	915	32.68%	206	736	28%

由表 3 看出,采用自动匹配转换实现叙词表之间的互操作是完全可行的,可以在各词表中完全相同或同义词相同的叙词之间实现兼容,尤其是《教词表》,其覆盖度达到了 62.73%,因为该词表是教育专业词表,本身收录有《中图法》的教育类目数据,所以转换率比较高;而《社科表》的覆盖度偏低,因为它是一部综合词表,教育类只是其中的一个小类,参与转换的数据就少。另外,通过组配也可以实现《中分表》中主题词串的部分匹配。

3　基于同义词表的语词匹配

3.1　基本原理

对于不能完全兼容的语词需要采用其他方法来匹配。基本方法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。针对此不足,本文考虑在该算法中引入同义词表,以提高计算的准确度。

首先编制一部语义精良的同义词表,该同义词表包括受控词、非受控词、表达完整概念的语词以及不可再切分的词素等;然后将匹配词采用自动分词技术,基于上述同义词表进行切分,成为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,先将同义词表中出现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。

3.2　实验过程

具体实现步骤如下:

(1)同义词表的编制。收集教育类语词,编制同义词表。为计算方便,为每个词赋予一个 ID 号,则将同义词用相同的 ID 号标注。主要有以下途径完成同义词表的编制:①主要是采用陆勇提出的基于模式匹配的汉语同义词自动识别方法[11],对不同的语料库定义不同的模式,从而提取和挖掘相应的同义词。本系统分别从《教育大辞典》、Web 网页和期刊论文中进行了同义词提取。②现有词表中具有“用”、“代”关系的词汇,直接作为同义词对,赋予其相同的 ID 号,加入到同义词表中。此处包括《中分表》、《教词表》及《社科表》的所有具有“用”“代”关系的叙词和非叙词。③利用《同义词词林》中的同义词对。《同义词词林》是一部对汉语词汇按语义全面分类的词典,对不同的词汇进行分类,并赋予相应的语义编码,具有相同语义编码的词即为同义词。根据上文已建立的词素词典,从《同义词词林》中抽取这些词素及其对应编码,并将同一语义编码下的词作为该词素的同义词。

经过上述处理,将收集到的三种同义词数据整合去重,加入适当的人工识别,删除明显错误的记录,基本的同义词表编制完成。每个词对应一个ID号,同一ID号下的词为同义词。因为此处的同义词表还将作为分词词典使用,对于没有同义词的词素也要添加在内,供下面分词使用,这些词素可以不赋ID号。

(2)对匹配词进行切分,将其切分成多个词或词素的集合。采用最大正向匹配算法(MM法)对匹配词进行分词。基于步骤1中建立的同义词表进行分词,分为以下三种情况:如果匹配词可以直接在同义词表中找到,则直接用对应的ID号来标识该词;如果匹配词被切分的词素有对应的ID号,也将其用词素对应的ID号标识;如果词素没有对应ID号,则直接用对应的词素对其进行标记。

(3)利用相似度计算公式,计算两个词之间的相似度。经过上述分词后,将每个词用词素或ID号进行标识,然后采用基于词素的相似度计算公式来进行计算。此处的计算单位可以是ID号也可以是词素,通过计算两个词所含的相同ID号或词素个数以及它们在各词中的位置来计算其相似度。

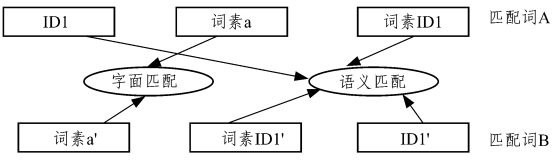

图2　同义词相似匹配算法示意图

(4)确定上述计算公式后,然后利用程序进行计算。实际计算过程中,仍然要考虑不同词表的不同表达形式,分别计算匹配词表和源词表中的叙词与叙词、叙词与非叙词以及非叙词与非叙词之间的相似度,需要进行四次交叉匹配,这里用PT表示叙词,NPT表示非叙词,则需要进行如下匹配:PT1/PT2、PT1/NPT2、NTP1/PT2、NPT1/NPT2。

(5)确定阈值,筛选主题词。通过上述方法可以计算出不同词表中每两个词之间的相似度,但不可能都作为映射结果,需要设定一个标准,筛选出合理的对应主题词,即可为《中分表》每个主题词提供一系列对应的主题词数据,供用户进一步选择。系统通过一定的实验后,确定阈值为0.6,即提取出结果大于等于0.6的词,认为这两个词相似,从而完成两个词表的对应。

整个过程可用以下流程图表示:

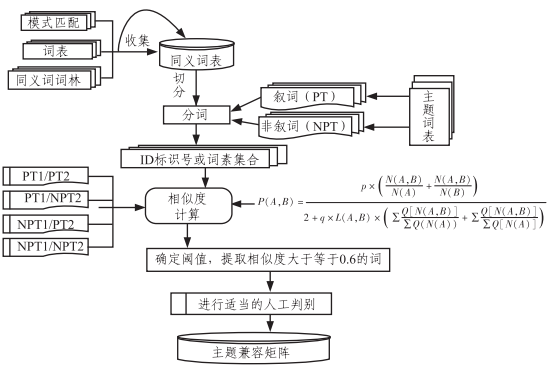

图3　基于同义词表的语词匹配流程图

3.3　结果分析

通过上述方法,分别完成《教词表》和《社科表》到《中分表》的匹配,匹配结果如表4所示:

表4　相似度匹配结果统计

匹配类型 / 兼容词表	叙词匹配	非叙词匹配	组配匹配	去重总计	叙词总数	百分比
教词表	608	302	524	1274	1651	77.17%
社科表	451	203	376	970	1651	58.75%

由上表数据可以看出,采用基于同义词表的语词匹配算法是完全可行的,加入同义词表,明显提高了相似度计算的准确性,从而使匹配覆盖度有了明显的提高。《教词表》的覆盖度达到了77.17%,也就是说通过这种方法可以使大部分的叙词实现转换,证明了该种方法的可行性。

4　教育集成词库的结构

教育集成词库是以《中分表》为核心,由若干词表构成的一个兼容体系,实现了各种分类语言、主题语言等之间的互操作。这个集成词库可以由多种形式、多种结构的众多兼容工具组成,该系统采用两种主要兼容结构形式:①字顺兼容矩阵。以每个主题概念为款目词纵向展示,即将《汉表》中的每个主题词或主题词串按字顺方式显示,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《汉表》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。②分类兼容矩阵。以《中图法》分类号为主干竖向展示,即按分类号顺序显示《中图法》类目,并列出其对应的《汉

表》中的专指主题词和附属主题词,把其他参与兼容的分类法横向展示。将参与兼容的分类表类号与《中图法》的类号相对照,列出其等值兼容或近似兼容的概念。最终形成一个以《中分表》为核心的兼容体系,实现不同词表之间的兼容互换。

参考文献:

[1] 刘华梅. 基于情报检索语言互操作技术的集成词库构建研究——以教育词库为例[D]. 南京:南京农业大学信息管理系,2006.

[2] Marcia Lei Zeng, Lois Mai Chan. Trends and issues in establishing interoperability among knowledge organization systems[J]. Journal of the American Society for Information Science and Technology, 2004,55(5): 377 - 395.

[3] 刘华梅,侯汉清. 近十年情报检索语言互操作研究进展[J]. 图书馆理论与实践,2006 (4):31 - 33.

[4] 傅兰生. 我国叙词兼容两大方案的分析——兼论国家级叙词兼容词库的建立[J]. 情报学报,1991,10(4):257 - 264.

[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] 朱岩. 对建立国家叙词库的认识与思考[J]. 科技情报工作,1991(2):15 - 17.

[7] 洪漪. 我国国家叙词库建设中几个问题的探讨[J]. 情报学刊,1991,14(3):209 - 211.

[8] 方陆明. 利用电子计算机建立农业叙词库及其管理系统——兼谈机编叙词表的几个问题[J]. 农业图书情报学刊,1989(1):61 - 65.

[9] 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.

[10] 张雪英,侯汉清. 叙词表词汇转换系统的设计[J]. 情报学报,2000,19(5):451 - 457.

[11] 陆勇. 面向信息检索的汉语同义词自动识别[D]. 南京:南京农业大学信息管理系,2005.

刘华梅　南京农业大学信息管理系硕士研究生毕业,现在国家图书馆工作。通讯地址:北京中关村南大街33号。邮编100081。

侯汉清　南京农业大学教授,博士生导师。通讯地址:南京农业大学信息科技学院。邮编210095。

(收稿日期:2007-11-20)

(上接第94页)

[20] Witten I H, Paynter G W, Frank E, Gutwin C, Nevill-Manning C G. KEA: Practical Automatic Keyphrase Extraction[C]. In: Proceedings of the 4th ACM Conference on Digital Library (DL'99), Berkeley, CA, USA, 1999: 254 - 256.

[21] 韩客松, 王永成. 中文全文标引的主题词标引和主题概念标引方法[J]. 情报学报, 2001, 20(2): 212 - 216.

[22] Keith Humphreys J B. Phraserate: An Html Keyphrase Extractor[R]. Technical Report, University of California, Riverside, 2002.

[23] 侯汉清, 章成志, 郑红. Web概念挖掘中标引源加权方案初探[J]. 情报学报, 24(1): 87 - 92.

[24] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segementing and Labeling Sequence Data[C]. In: Proceedings of the 18th International Conference on Machine Learning (ICML01), Williamstown, MA, USA, 2001: 282 - 289.

[25] 周俊生, 戴新宇, 尹存燕, 陈家骏. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006, 34(5): 804 - 809.

[26] 侯汉清, 马张华. 主题法导论[M]. 北京: 北京大学出版社, 1991: 199.

[27] CRF + +: Yet Another CRF toolkit[OL]. [2005-12-20]. http://chasen.org/~taku/software/CRF + +.

[28] 人大报刊复印资料[OL]. [2007-12-01]. http://art.zlzx.org.

[29] 中文自然语言处理开放平台[OL]. [2005-12-25]. http://www.nlp.org.cn.

[30] Vapnik V. The Nature of Statistical Learning Theory[J]. New York: Springer-Verlag, 1995: 1 - 175.

[31] Zeng H J, He Q, Chen Z, Ma W Y, Ma J. Learning to Cluster Web Search Results[C]. In: Proceedings of 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR04), Sheffield, 2004: 210 - 217.

[32] SVMlight[OL]. [2005-12-20]. http://svmlight.joachims.org.

章成志　南京理工大学信息管理系讲师,博士,中国科技信息研究所博士后。通讯地址:南京理工大学信息管理系。邮编210094。

苏新宁　南京大学信息管理系教授,博士生导师。通讯地址:南京大学信息管理系。邮编210093。

(收稿日期:2008-01-10)