

利用篇名数据库自动完善后控词表

□张宇萌 马张华

摘要 通过论述后控词表的建立原理、完善方法,指出实现自动完善、自动维护是后控词表得到广泛应用的关键。针对已有的文献篇名数据,提出一套利用篇名数据自动完善后控词表的方法,并用计算机实现。

关键词 自然语言检索 后控词表 篇名数据库

随着信息技术的不断发展,自然语言检索系统的优势越来越明显,大有取代受控语言之势。但是,自然语言不受控制的致命缺陷意味着自然语言检索永远都离不开控制,因此自然语言与受控语言结合的后控词表将是信息检索语言的一个重要发展方向。

目前,我国的后控词表理论研究已经相当成熟,在国际上处于领先地位。但真正投入使用的却非常少,原因就是词汇的收集、词表的完善非常困难。如何利用一些现有的数据库,实现自动收集词汇、完善词表不失为一个“多快好省”的办法。在这里,笔者结合计算机编程,利用现有的篇名数据库实现自动完善后控词表,供大家参考。

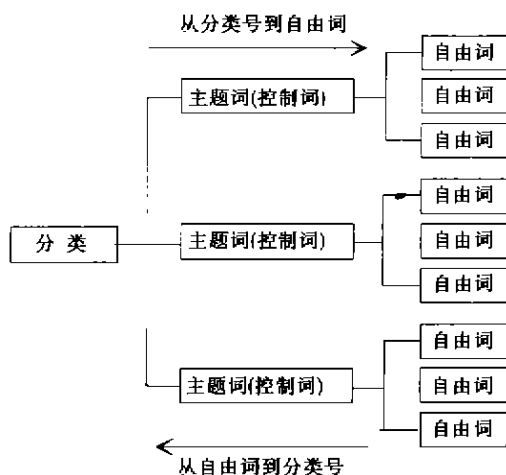


图 1

例如,《中国分类主题词表》第一卷 523 页有如下款目,每条对应款目分为左右栏两部分:

0321 线性振动	线性振动 共振;反共振;混合共振;偶合共振;共振频率;间谱振动;强迫频率;强迫振动;线性阻尼;自由振动
0322 非线性振动	非线性振动 冲荡振动;非线性 Landau 阻尼;非线性力学;非线性阻尼;非线性振动理论;马其诺方程;松弛振动;希尔方程

(1) 分类号、类名和类目注释,置于分类号—主题词对应表的左栏;

(2) 类目所对应主题词和主题词串以及对应注释,置于分类号—主题词对应表的右栏。

根据此框架,可以建立以下分类表款目:

0321 线性振动	0321 偶合共振
0321 共振	0321 共振频率
0321 反共振	0321 间谱振动
0321 混合共振	0321 强迫频率

1 后控词表的编制方式

张琪玉教授在《论后控技术》一文中指出,后控词表主要有四种编制方式,在这里,本系统采用的是第 4 种方式:“将自然语言检索标识与某种词表或分类表对应”。即以影响大、通用性强、用户比较熟悉的《中图法》的分类体系为骨架,并参考各类主题词表,对《中国分类主题词表》作必要的增补、修订,以《中国分类主题词表》中的主题词作为控制词,在控制词下建立一个自由词词群,并标示出包括所有主题词、自由词的代、属、分、参关系,形成内容完善的后控词系统。这样建立的分类表结构如图 1 所示。以这种结构编制的后控词表可以直接取代分类主题词表,只要把分类号对应的主题词(同义词)颠倒过来,就可改造成主题—分类号对应表(可把主题词转换成对应的分类号)。这样,这种后控词表就可与自动抽词词典联接。

然后从《中国分类主题词表》第二卷(主题词-分类号对应表)、《物理学汉语主题词表》找到各个主题词的相关词,统一置于各主题词下。

采用这种方式可以借鉴《中图法》、《汉表》、《中国分类主题词表》等重大检索语言成果,直接利用其中的分类体系和词汇,使后控词表具有分类主题一体化词表的性质和功能,为以后进行基于《中图法》的自动分类和基于《汉表》的后控词表的自动标引奠定了基础。这样编制的后控词表适应面广,易于推广,是一种简单易行的途径。它的款目形式见表1:

表1 后控词表的款目形式

KZH	KZH1	BSF	CLASSFYNO	CLASSFYNO1	WORD
548	1		03		比重
**	1	D	03		比重法
**	1	D	03		密度
553	10		03		轨道力学
**	10	C			月球地球飞行轨道
**	10	C			轨道计算
**	10	C			轨道摄动
**	10	Z			力学

注:KZH(控制号)中用数字表示的是控制词(正式词),“**”则表示该词是控制词的相关词;KZH1 相同的数字表示它们都是同一个控制词的相关词,其中也包括控制词(通过 KZH 来识别);BSF 用来表示非控制词与控制词的词间关系;CLASSFYNO, CLASSFYNO1 表示主题词和等同词对应的类目和交替类目;WORD 表示后控词表中的词汇,包括控制词和自由词。

后控词表的构建一般可分两个阶段:初级阶段和完善阶段。初级阶段是指词表组成、体系结构、显示方式的确定,并根据分类主题一体化词表建立一个分类表,这时后控词表只有主题词(控制词),极少、甚至没有后控词(自由词),词汇还很不丰富,虽然可以投入使用,但效率很低,后控词表自然语言词汇检索的功能不能充分实现;完善阶段则通过大量收录具有标引检索价值的词汇,建立起它们与表内词的联系,将丰富的自然语言词汇纳入控制系统,从而实现直接从自然语言词入手,达到词表扩检、缩检的目的。这一阶段是目前后控词表建设的关键,但由于后控词表涉及的词汇数量大、范围广、处理难度大,因此完善后控词表的工作应当充分利用现有的标引成果和现代技术手段,结合文献处理过程的特点进行。笔者认为,第二阶段的工作可以确定在科学的数据提取和数学方法的基础上,采用这种方法进行:利用现有文献数据库篇名数据自动完善后控词表。

2 数据的提取和数学方法

2.1 数据的提取

我国目前投入市场的中文文献数据库中大多数都包含了篇名、文摘、分类标引数据和主题(关键词)标引数据等,每条数据都对应于实际的文献,都是标引员根据他们的知识和经验,并对文献内容进行具体分析而形成的,具有广泛的文献保障,是一批自动构建后控词表宝贵的资源。有《中文社科报刊篇名数据库》、《中文科技期刊篇名数据库》(光盘版)、《复印报刊资料专题目录索引》等等。

这些文献数据库大都使用《中图法》进行分类标引,参照《中国分类主题词表》进行主题标引或关键词标引;有的数据库系统还作了一些词汇控制处理,如清华大学的《中国学术期刊(光盘版)》提供了“蕴涵检索”和“相关检索”功能,重庆维普公司的《中文科技期刊篇名数据库》在辅助工具中提供了“同义词表”和“分类主题词表”。这些现成的数据完全可以被利用,通过计算机统计、筛选有效数据来丰富后控词表。侯汉清教授就采用过概率统计的方法对这些数据进行提炼并对标引词进行归类。本文在此基础上,从另一个角度出发,采用一种更简洁的算法,并且是对基于自然语言的篇名数据进行提炼,抽取新词进行归类。

2.2 计算模型

文献中的分类号与标引词(自然语言词汇)是多对多的关系。从统计角度看,分类号与标引词(自然语言词汇)共现的频率越高,其关联程度也越高,该标引词(自然语言词汇)对应该分类号的概率就越高。

统计学上测定两个事件的关联程度最常用的就是条件概率模型:

$$P(A|B)=P(AB)/P(B)=(M/C)/(N/C)=M/N \quad (\text{公式1})$$

其中 $P(A|B)$ 表示标引词 B 对应分类号 A 的概率;

$P(AB)$ 表示标引词 B 与分类号 A 同时出现的概率;

$P(B)$ 表示标引词出现的概率;

C 表示样本库的总记录数;

M 表示标引词 B 对应分类号 A 的出现次数;

N 表示标引词 B 出现次数。

从公式1中可以看出,在样本库不变的情况下(N值为常数),标引词B对应分类号A的出现次数(M值)越大, $P(A|B)$ 值就越大,表明事件B对应事件A的概率越大。通过计算机统计计算,自动选取 $P(A|B)$ 值最大者为该标引词的最佳分类号。

后控词表是不断增长的词表,只有收集到高质量的新词,才可能构筑高质量的词表。自然语言是最为活跃、变化最快的一种语言,篇名、文摘数据中出现的新概念、新术语及它们的不同表现形式有时不能及时、完全反映并被标引出来。因此,利用篇名数据库给自然语言语词定类的方法(以下简称“篇名数据法”),则可以利用篇名、篇名的分类号数据,采用条件概率法对著者直接使用且表达更为自由的自然语言词汇归类。在这里,本文仍以《中文科技期刊篇名数据库(重庆)》中的“03”力学类的文献数据为样本进行探讨。

例如,文献篇名数据“Bingham 流体偏心环空螺旋流的流函数-轴向速度方程和 Newton 流体, 0373”中“Bingham 流体”和“Newton 流体”并没有被标引出来,它们不是常用词或外类词,而是流体力学中最新研究中的专业术语。采用“篇名数据法”就可很好地解决这个问题。

从套录出来的 PM.TXT 文件中用程序读出篇名和分类号,建成篇名数据库(PM.DBF),格式如下:

表 2

分类号	篇 名
0346.2	I-II 复合型裂纹亚临界扩展研究
0346.2	II 型加载条件下疲劳裂纹扩展试验研究
0346.3	A1203.SiO ₂ /ZL109 金属基复合材料的强度性能研究
0357.52	Blasius 边界层 C-型失稳的理论分析
0328	摆动对双质体振动筛二次隔振特性的影响
0373	Bingham 流体偏心环空螺旋流的流函数-轴向速度方程和

任意从篇名中选取 10 个语词作试验样本,如:疲劳裂纹、Navier-Stokes 方程、爆炸波、非完整系统等,用这些词分别去匹配 PM 库中的篇名,如果篇名中包含该语词则生成一条分类号-语词的记录,存入临时库 TEMP 中(与 TEMP1 数据格式相同)。如,“疲劳裂纹”在记录“0346.2, II 型加载条件下疲劳裂纹扩展试验研究”匹配成功,则生成“0346.2, 疲劳裂纹”记录。最后共生成 136 条记录,利用条件概率法统计得到这 10 个语词的最佳分类号,见表 3。

除去共现频次小于 3 和 P 值小于阈值(0.23)的 2 条记录,其中 7 个语词的分类号正确,一个错误:爆炸波的分类号应是 0382,正确率达 87.5%。如果有质量较高、规模较大的篇名数据库作为保证,这种方法是可行的。

在这里设置自然语言词汇与分类号共现频次大于 3,不仅仅是为了排除外类词和随机错误,更主

表 3

分类号	WORD	共现频次	词现频次	P 值
0346.1	最小耗能原理	1	3	0.3333
0345	粘弹性	17	36	0.4722
0373	Bingham 流体	2	3	0.6667
0382.2	爆炸波	4	6	0.6667
0346.1	疲劳裂纹	15	21	0.7143
0357.1	Navier-Stokes 方程	14	18	0.7778
0313.3	转动惯量	14	16	0.8750
0316	非完整系统	29	31	0.9355
0346.5	损伤断裂	3	3	1

要的是判断某个自然语言词汇是否已经得到了人们的认同。因为自然语言是最为活跃、变化最快的一种语言,同一概念的表达方式也是千差万别,不是所有专指度高的自然语言词汇都能被人们接受。如果某个概念的语言表现形式没有得到大家的认同,那么也就没必要给它归类了。判断某个自然语言词汇是否已经得到了人们的认同,只需判断它是否被多个著者使用过(至少大于 3 次)。

3 计算机实现

利用程序读出篇名及其对应的分类号,生成由分类号、篇名组成的 PM.DBF 库。输入一个语词,并用该语词依次去匹配 PM.DBF 中的每一条篇名记录,如果该记录包含该语词,则把该篇名对应的分类号添加到一个临时库 TEMP.DBF,直到 PM 库的末尾。

然后,再对 TEMP 库进行统计,自动选出概率最大的记录。具体流程见图 2。

4 实验结果分析

由于篇名是自然语言,包含的词汇很不规范、专指度较强,所以对词汇归类会出现下面几种情况:

4.1 常用词归类出现的情况

常用词包括通用词和较常见的术语。通用词出现的机会比较多,它与某一分类号共现也可达到较高的频次(>3),显然不能给这些通用词归类。分析数据发现,这些通用词虽然出现频率较高,但它们往往会对应不同的分类号,用公式 1 的计算得到的 P 值较低(低于 0.23),所以可设置阈值来筛选掉这些通用词。

较常见的术语专指度较高,一般只对应固定的分类号,统计得到的 P 值较大(介于 0.7~1 之间),与分类号共现频次都大于 3,所以这些专业词汇很容

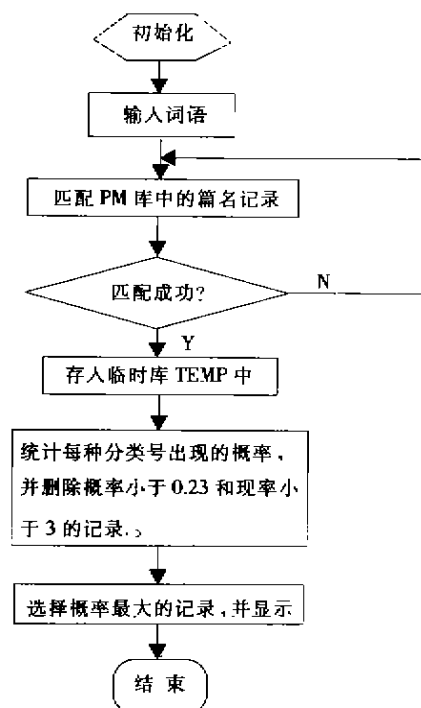


图 2

易被筛选出来,从而取得较满意的结果。

4.2 军用词归类的情况

自然语言是最为活跃、变化最快的一种语言,同一概念的表达方式也是千差万别,同一概念的表达可能分散成多种形式的自然语言词汇。相对来说,某个自然语言词汇使用得比较少,出现的次数就很可能小于 3。譬如词汇“Bingham 流体”,它对应的分类号是“0373”,但它们共现频次为 2,只能被剔除。因此许多自然语言词汇对应的分类号虽然是相符的(计算出的 P 值常为 1!),却因不够“条件”被剔除。一方面,我们可能觉得这些词汇没有被收集是很可惜的,但从另一方面看,这部分词被剔除又是合理的:

在图书馆学中有个著名的“二八律”:20%的文献就可以满足 80%读者的需求。同样道理:20%的后

控词可以满足 80%用户的检索要求,因此没有必要收入那些罕用的、使用效率极低的词汇。当然,这种情况也不是一成不变的,随着这些词汇被使用的增多,它们对应分类号的共现频次会相应提高而符合“条件”被“录入”。

4.3 篇名中出现歧义现象对归类的影响

由于没有经过汉字分词处理,篇名中包含的词汇没有被切分出来,它们隐含在篇名当中,归类处理时采用的是匹配法——用输入词去匹配篇名中的词汇,匹配成功则取出该词,并把篇名对应的分类号赋给该词。这里面就存在一个匹配歧义的问题:篇名中有些词汇与相邻的字、词就可能形成其它意义的词汇,如“0342,《结构力学》教学与辅导”就可切分成四个词汇:“力”、“力学”、“结构”和“结构力学”。对词汇“力学”归类时,在词典收词不全(如未收入“结构力学”)时就会得到“力学-0342”的记录,这显然是错误的。如果这种错误出现概率比较高时,“力学”与“0342”的共现频次、P 值就比较高,从而被筛选出来。这种错误属于随机错误,只要样本库有一定的规模和质量保证,这种错误是可以被屏蔽掉的,即使这种错误仍然存在,它出现的机会也是很小的。

参考文献

- 1 马张华,侯汉清.文献分类法主题法导论.北京图书馆出版社,1999
- 2 张雪英,侯汉清.分类表—叙词表转换系统的设计.全国第三届情报检索语言学术研讨会论文,1999
- 3 侯汉清,李波,戴晶萍.计算机建立分类法和主题词表转换系统的尝试.全国第三届情报检索语言学术研讨会论文,1999
- 4 赖茂生,谭晓冬.基于超文本结构的后控词表管理系统.情报学报,1995(5)
- 5 张琪玉.论后控制词表.图书情报工作,1994(1)

作者单位:张宇萌,浙江宁波大学商学院信息管理系,315211;马张华,北京大学信息管理系,100871

收稿日期:2000年11月6日

首都图书馆新馆“五一”展新颜

[据《北京晨报》报道]今年“五一”,位于朝阳区东三环华威桥东侧的首都图书馆新馆正式开馆。新馆采用计算机综合信息管理系统,从根本上实现馆内和馆外读者查询的一致性。读者除了到图书馆读

书以外,也可以上网浏览馆藏图书资料。据了解,开放的阅览室总阅览座位有 1000 多个,先期可为读者提供的文献达 100 余万册。

该馆全年 365 天都不休息。