

●宋明亮

我国报纸信息数据库开发的现状与对策

ABSTRACT Since 1990 when the "Economy Daily" set up the first newspaper information data base of the country, 19 of the kind have been completed. After analyzing their basic circumstances and problems existed, the paper suggests to adopt 3 countermeasures, e. g. the indexing method of "free indexing + post-controlled vocabulary", etc. 2 tables. 4 refs.

SUBJECT TERMS Newspaper information data base — China Full-text retrieval—Post-controlled system

CLASS NUMBER G254.36

报纸因其详实、丰富和新颖的内容,而成为人们最喜闻乐见的出版物,是当今社会主要的信息传播媒介之一。在社会的信息化趋势日趋明显的今天,沿袭了几十年的《缩印本+分类索引》已无法适应时代的需求。1987年,《经济日报》首先采用了激光照排技术,从此我国有了电脑存贮的报纸信息原始数据。这一质的飞跃,为报纸信息计算机检索打下了基础。1990年底,又是《经济日报》,建立了我国第1个报纸信息数据库。截止1994年2月,已经建成并投入使用的还有《人民日报》、《文汇报》、新华社以及中信所重庆分所的医学、化工等专题新闻题录数据库。《科技日报》和《解放日报》正在进行数据处理和运行调试工作。《解放军报》、《广州日报》和《光明日报》等也在筹划之中。

1 基本情况

(一)从收录范围看,在建立本报社信息数据库的同时,根据各自特点,选择一些具有通用推广价值的热门专题,建立专题报纸信息数据库。从收录的时间跨度看,除《人民日报》的40年累积题录库和人物索引库、《经济日报》的国家领导机构及领导人库以外,其它都只收录建库以后的资料(见表1)。

(二)数据库结构一般由外表特征、内容特征、出处和原文4部分组成,是典型的全文型数据库。用户可在自己的终端编辑稿件,同时检索和调用库内文献。但因各个报社的传统习惯和数据库内容范围的不同,而导致具体字段的设置也略有不同。这种状况表明,各种报纸信息数据库将提供的是一种不同的检索入口(见表2)。

表1 我国已建成的报纸信息数据库

单 位		数据库名称	收录时间	简 介
《人民日报》社	1	人民日报全文库	91~93	以年为单位,每篇文章为一条记录。存放每日出版的《人民日报》版面全文信息,每日入库约120条。已制成光盘发行。
	2	人民日报索引库(题录库)	54~	收录《人民日报》建报40年来的索引数据,提供19个检索入口。
	3	人民日报评论库	91~	存放《人民日报》的社论、评论及短篇文章的全文信息。
	4	市场信息库	91~	收录市场动态、海外市场、新产品介绍等经济方面全文信息。
	5	人物索引库	54~	从国内多种报刊上整理出来的各行业知名人物资料,约12万条数据。
	6	国际动态库	91~	该报社所编《世界各国国情概况》等全部文章的全文信息。
	7	新闻通稿库	91~	内容为新华社每天向人民日报传送的电讯稿。每日入库125条记录,现约有7万余条。
《经济日报》社	8	经济日报库	90.10~	存放该报全文信息,日均增加80条记录,已存近8万条。
	9	综合经济信息库	91~	从全国200多种报纸中精选出的国内国际经济新闻资料,日均增加40多条记录,现存4万多条。
	10	中华人民共和国领导机构及领导人库	49~	存储建国以来各党政机构领导人变更情况,已存800多条。
	11	新华社电讯稿库	90.9~	存储新华社电传该报的通讯稿,日增约150条,现存16万多条。
新 华 社	12	中文新闻稿库	92~	存放该社每日国内播发的新闻稿,日增约300条,现有16万多条。
	13	英文新闻稿库	92~	存放该社向世界播发的英文新闻稿,日增约250条,现有12万条。
	14	参考消息库	93.3~	收录《参考消息》的全文信息。
	15	科技信息库	92~	收录国内外报纸刊登的最新科技动态及发展趋势的全文信息。
	16	国内国际人物库	92~	从100多种报纸搜寻来的副部级以上党政军领导及各界知名人士资料,并据每日报纸所载消息进行跟踪修改。
	17	全国选报库	92~	摘选认为有保存价值的各省市重点报纸的全文信息。
	18	国际商业信息库	92~	收录国内外报纸刊登的招标、求购、展览会、海关税率方面的全文信息。
《文汇报》社	19	文汇报全文库	92.11~	唯一没有采用小型机建库,而用微机光盘联网形式运行的系统。日增约120条记录,现约有4万条。

表 2 可检字段一览

单 位 名 称	外表特征字段	内容特征字段
《人民日报》社	日期、星期、版次、版名、连载号、栏目名、来源(发稿部门)、作者、作者单位、作者职务、发电地点、标题 1(主)、标题 2(副)	分类号、类目名称、主题词、关键词
《经济日报》社	日期、版面、栏名、标题 1、标题 2、作者、连载号、国家与地区、来源	类号、类名、标引词
新华社	稿号、发稿日期、入库时间、作者、信息来源、标题	分类号
《文汇报》社	标题、日期、版次、作者、来源	分类号、主题词、自由词

(三)从我国已建成的 19 个报纸信息数据库来看,具有 3 个相同的检索性能:

1、都能比较快速简便地从作者、标题、日期、版次、人名和栏目等外表特征查找资料。其中《人民日报》和《经济日报》将传统的计算机检索方式——菜单选择改为表格式查询,把所有可检字段在屏幕上以表格形式显示,用户只需移动光标来选择表中的一个或多个字段作为检索入口,并填入提问内容,系统就会自动默认字段之间为逻辑与的关系。

2、由于用户对分类体系和分类方法不熟,使花费较多时间的分类工作没有起到应有的作用。

3、不论是题录库还是全文库,其中的全文检索使用率都最高,几乎成为唯一的内容特征查找途径。

(四)各系统都有专人负责每日报纸的标引工作,标引的内容一是分类,二是从文章中提出作者、标题、栏目、人名、地名和机构名等外表特征。主题标引只有《文汇报》在做,但数量极为有限。

2 存在的问题

(一)用户友好信息太少,整个系统的易用性较差,表现在两个方面:

1、要求用户输入的检索命令太多。如《文汇报》查作者要输入“ZZ×××”,查日期要输入“RQ×××××”,查两个词出现在同一句内要输入“×××AND/SEG×××”等,共有 11 种特定检索表达式命令。《人民日报》13 种。《经济日报》和新华社采用的引进版 TRIP 检索软件也存在同样问题。一般用户往往只采用常见的几种,影响了数据的使用效果。

2、用户要想查全资料,必须想出尽可能多的同义词、近义词和不同词序的词。另外,要扩大或缩小检索范围,只能边检边试,缺乏科学的依赖工具。

(二)全文检索虽有加工简单、标引深度高、便于查找文章细节等优点,但在不进行控制的情况下,有 4 个因素制约着全文报纸检索效率,而使其不尽如人意:

1、由于对全文进行字面的扫描匹配,容易出现假义词,如“中华人民共和国”中的“华人”、“发展中国家”中的“中国”。特别是以地名和关键词组成的名词,更易造成混淆,如“大理石”并非报道大理地区的新闻。

2、由于文章中的每一个字、词均可成为检索点,使每次检索命中的文献偏多,筛选量大。因为在许多新闻中虽然出现了某检索提

问词,但并没有真正叙述该主题。这样的新闻对该提问来说是很次要的,甚至无用的,造成了检索噪音^[1]。

3、同义词大量存在,造成文献分散。

4、同形异义词和相关词得不到反映,也影响了检索效果。

目前,几个系统在频繁使用全文检索的过程中,已意识到了以上问题,尤其是其中的第2和第3个问题的严重存在。为了减少误检和漏检,已采取以下措施:

1、允许进行位置限定和相邻检索。如只出现在标题中或两个提问词出现在同一句、同一段中才算命中以及限定两个提问词之间的距离不超过N个字等。

2、用“or”、“and”、“not”等逻辑表达式和用表格查询进行多字段的逻辑组配检索。

3、北京信息工程学院在为《人民日报》设计TRS1.0版的基础上,最近推出了TRS2.0版新闻检索系统软件,其特定之一是试图一改过去按“字面词”匹配为按“实义词”匹配。其主要思想是借助一个8000多词的分词词典,对库中文献进行分词处理。分词时采用“最长匹配原则”,即当“中华”、“华人”、“人民”、“共和国”、“中华人民”、“人民共和国”、和“中华人民共和国”均为合法词时,取最长的“中华人民共和国”。这样就不会将“华人”抽出来了。如果分词词典中没有的仍按全文字面匹配或单汉字组配查找。同时将用户确认的提问词增加到分词词典中去。

4、通过分类途径提高系统的查全和查准性能。

5、根据提问词在命中文献中的出现频率和出现位置(标题、段首和段尾等)对检索结果进行排序,将“最相关”的文献先提供给用户,方便甄别,这也是TRS2.0的新功能之一。

6、进行主题标引。即从新闻信息中分析提炼出若干主题词作为检索点,以减少噪音。《文汇报》社是在运行一段时间后,发现检索

噪音问题比较严重而采用“《新闻叙词表》规范词+自由词”的方式进行主题标引的,但其进展较为缓慢。《人民日报》社、《经济日报》社和新华社虽未做这一工作,但都备设了主题词或关键词字段以及词表管理子系统,试图今后加以改进。

综观以上措施,笔者以为有以下几点不足:

1、位置限定和相邻检索只是在某些特定情况下才有用的一种措施。这是因为:首先,限定检索词在标题中出现,其前提是标题能反映文章的主题,而据笔者的抽样统计,只占70%左右的报纸标题能够反映主题,大大低于期刊论文的同类比例。其次,两个词出现在同一句、同一段或限定间隔多少个字也不一定能使检索结果更为准确。事实上,相邻检索只适用于中间截词的模糊检索,其前提是已知某词组的中心词素。显然,局限性较大。

2、这些办法都没有解决检索者负担过重这一根本问题,还是依赖人的“想象”来选择提问词和构造检索策略。用逻辑表达式进行检索时,要求用户能把握哪些词要累加到一起而进行“逻辑与”运算,哪些词要排除掉而进行“逻辑非”运算,这实际也是很困难的事情。

3、分类方法虽可按范畴聚集同类主题的文獻,但报纸信息恰恰是学科系统性不强,绝大部分新闻都不可能象专业论文那样归入某一学科。毫无疑问,报纸中大量有关某人、某地、某时、某事的特定信息,最好方法是按事物聚类的主题查找,而不是按学科聚类的分类查找。遗憾的是,这些系统由于过份依赖全文检索而使主题检索显得异常薄弱。归结起来在报纸信息数据库中,主题检索功能未能很好实现的原因有:

(1)总想依赖一部现成的高质量规范词表进行受控标引,但又无法如愿。原因有3:一是国内新闻行业仅有的《新闻叙词表》(新华社1988年编),不具备通用于各类报纸的

性能;二是报纸信息变化快,新概念、新术语不断涌现,词表中相对固定的词汇无法适应变化万千的需要;三是受控标引要求高,需进行查词、选词,工作量大,面对每天一期和几十年积累的报纸,令人望而生畏。正因如此,新华社虽然编制了《新闻叙词表》,但也没有使用。

(2)自动抽词标引虽省时省力,但受计算机语义分析能力的限制,加上汉语分词的复杂性,目前还达不到实用程度。

(3)检索结果按提问词在文献中的位置和出现频率等加权因素进行排序均有局限性,并不一定能决定文献与提问的相关程度。人为地确定标准,又得考虑这些标准是否符合用户检索需要。否则,只会适得其反,将用户认为不怎么相关的文献排在了前面。

(三)标准化程度不高。编程语言兼容性差,常用的实用程序难于互相移植。同时,用户界面、检索命令和数据格式也是十人十指,互不相同。

3 今后对策

随着电脑排版技术在报纸行业的普及,必将有越来越多的报社会建立自己的数据库。因此,总结经验,吸取教训,扬长避短,找到切合实际的方案,乃当务之急。

(一)采用后控方式。要解决影响全文检索效率的4个因素,传统方法采用的是编制词表,进行受控标引。实践证明,它不完全适用于报纸信息数据库的建立。笔者以为,最好的方法是在标引阶段使用自然语言,而在检索阶段则通过后控词表来实现词汇控制,即采取“自由标引+后控词表(由轮排索引和分类主题表两部分组成)”^[2,3]的方式,其根本出发点是减轻用户检索过程中的“想词”负担,增强系统易用性,提高查全率和查准率。

1、在自由标引中,其主题词可直接从标题和正文中抽取,也可由标引员自己措词,允许用词不统一,对标引员要求不高,上马较为

容易。由于免去了词表翻查和词汇转换过程,处理速度虽比不上自动标引,但比受控标引要快得多,且质量优于自动标引,加上后控词表辅助,效果更佳^[4]。

2、后控词表可灵活应变,具有多种用途。它与全文检索配合使用,利用词表的词间关系显示,可自动找出提问词的所有同义词后进行全文查找;这样,可以提高查全率。也可先用后控词表进行粗泛查找,再用专指提问词对全文进行二次查找;这样,可以提高查准率。同时,后控词表也可适应未来的自动标引的需要。因为,只有人工抽词标引和措词标引经验积累到一定程度并总结出规律后,才能由计算机代替人工,找到适合自己的自动标引方案,而作为标引积累的后控词表是可靠的第一手材料。同时若采用“规范词典+抽词词典”的规范化自动标引方法,后控词表可作为编制这些词典的直接依据。

3、从工作量来看,进行自由标引和编制后控词表要花费一定的人力、物力,但由于后控词表的分类主题表部分已具备分类检索功能,可免去对报纸文献的分类工作。这样,总的工作量和这些系统现有的工作量差不多,但效率较高,功能更强。只要确定标引规则,便可开始处理数据。标引文献达到一定数量后,又可进行后控词表的编制工作,使词表编制和系统建立同步进行,不影响系统的时效性。

4、方案适用于整理电脑排版前的非机读资料。象《解放军报》,现有1956年创办以来的近50万篇文献,约4.4亿字。如此庞大数量的资料,全文输入不太可能,受控标引更不现实,因此,最好的办法是象《人民日报》那样只做题录库。但自由标引弥补了新闻标题情报性差的不足,加上后控词表的辅助,准确、全面检出所需资料是可能的。

(二)加强协词,减少重复浪费。从表1可以看出,几个系统在数据库选题方面有些雷同。比如,新华社已有了专门的中文新闻库,

《经济日报》和《人民日报》又建新华社通讯稿库和新闻通稿库。再如,《经济日报》的中国领导人库、新华社的副部级以上国内国际人物库和《人民日报》的人名索引库,也有些交叉重复。其实,应进行一定的分工协调,如新华社可专建国际人物库、《经济日报》社专建经济人物库、《人民日报》社专建国内人物库、《解放军报》社专建军事人物库。这样,不仅可使各单位重点突出,而且通过交换,可拥有更多更全的信息资源。要做到资源共享,应做好以下工作:

1、应有专门机构进行组织协调。中国新闻资料学会在这方面可发挥应有作用。

2、加强数据库结构的标准化。要象机读目录那样,制订相应的数据库标准格式及著录规则。只有在统一的基础上,才可能实现数据的交换。

3、应加强用户友好界面建设,利用弹出型菜单、下拉式菜单、图形窗口和热键等新技术,使系统能被不同人员所掌握。同时,对实现相同功能的工作,应形成规律性的操作步骤。

(三)发挥图书情报部门优势,积极参与开发工作。从各个报社现在的情况来看,新闻资料工作长期以剪贴报纸为主,尚无独立的数据库建设能力,现有几个系统都是在和有关科研单位合作开发的情况下完成的。图书情报部门应抓住良机,发挥自己在信息处理方面的特长,既可象美国 MEAD 数据中心那

样,建立《纽约时报》、《新闻周报》、《商业周报》和《华盛顿邮报》等多种报纸的综合新闻数据库系统;也可象中信所重庆分所那样,利用馆藏报纸建立专题新闻数据库。目前,在各类报纸中经常出现的旅游信息、影视娱乐、报刊文学和轶闻趣事等专栏文献,很受大众欢迎。它们为建立专题信息库,开辟信息市场提供了物质前提。当然还可象复旦大学图书馆编制《新民晚报索引》那样,主动和有关报社寻求合作。只要我们努力去做,这的确是图书情报部门投身信息产业,提高经济效益和社会效益的一条好路子。

参考文献

- 1 Long, A. Full text newspaper retrieval is hard to manage: fact of fiction. New Jersey: processings of the 9th national online meeting, 1988:213~216
- 2 张琪玉. 论后控制表. 图书情报工作, 1994, (1): 1~4
- 3 宋明亮. 论对自然语言的控制. 情报理论与实践, 1994, (2): 30~34
- 4 张琪玉. 汉语关键词法研究. 图书馆论坛, 1993, (1): 10~14

宋明亮 1990年毕业于空军政治学院图档系,现为该院情报检索语言在读研究生,发表本专业论文5篇。通讯地址:上海空军政治学院图档系,邮编200433。

(来稿时间:1994-04-04。编发者:丘峰。)

