中国图书馆学报(双月刊)1997 年第 1 期
ZHONGGUO TUSHUGUANXUEBAO

●吴开华

# 用计算机开发利用图书馆报章信息资源

## ——谈香港报章资料库的建设

**ABSTRACT** The Hong Kong Newspaper and Perical Data Base developed by the Dibrary of Hong Kong Chinese Language University is a national data base and was formally started using in September 1995 which takes the Chinese and English Newspapers published in the Mainland of China, Hong Kong, Taiwan and Macao as the chief information sources. Users may find out electronic materials of the very day already been arranged and processed.

**KEY WORDS** Newspaper and periodical data bases Developments Hong Kong

**CLASS NUMBER** G356.9

### 1 报章数据库建设的意义

报纸刊载的是最新的信息,反映的是最新的动态。报刊资料内容丰富、涉及面广,其汇集则是一部散装的百科全书。报章资料是图书馆文献信息资源的重要组成部分。

图书馆利用现代信息处理技术,对报章文献进行深层次开发,把分散于各报章的有价值的短时信息,进行收集、整理、加工、编排,令其电子化、网络化、系统化,长效化,以达到方便检索、资源共享之目的。报章资料数据库化,具有非常重要的现实意义。笔者工作过的"香港报章资料库",就是这个新技术应用的一个实例。

### 2 香港报章资料库简介

香港报章资料库是由香港政府拨款资助、香港中文大学图书馆开发、1995 年 9 月 15 日正式开启使用的一个图象数据库(Image DataBase)。该资料库收集当天出版的中英文报章,按照一定的制作标准,把报章文献电子化。其内容主要是中国大陆、香港、台湾、澳门的社会、政治及经济新闻。

香港报章资料库通过日渐普及的国际电脑互联网络,为用户提供来自香港报章的最新新闻。作为中国大陆、香港、台湾及澳门的电子储存库,用户可以通过电脑互联网络内的万维网(World- Wide- Web)查阅资料库内容。在一般情况下,香港当日出版的报纸新闻,即日便可提供阅览。

香港报章资料库的基本检索模式是由图书馆工作人员编制关键词词汇索引。关键词包括主题、地名、机构名及人名。用作索引的关键词均有中英文对照,用户既可用中文检索,也可用英文检索。同时,该检索系统还就个别的关键词加上了"用"、"代"、"属"、"分"、"参"等项参照词,以加强词间关系的显示。用户可以根据报章名称、版面、日期及指定关键词进行检索,如果遇到困难,可方便地利用关

键词列表而获得帮助。如有需要, 用户也可以运用布尔逻辑(Boolean Searching)来扩大或缩小检索范围。

截至 1996 年 7 月 23 日, 该资料库已拥有 105590 条记录。其中,《文汇报》24243 条,《大公报》24155 条,《星岛日报》24315 条,《香港联合报》10336 条, 英文《虎报》(Hong Kong Standard) 22541 条。每天都有数百的用户光顾该数据库。目前, 香港报章资料库尚在不断的发展、完善之中。

## 3 香港报章资料库建库技术要点

### 3.1 报章资料的收集

报章资料库的质量首先取决于其所收集的报章文献的质量。茫茫报海, 各种各样的思潮、消息充斥。怎样才能建成一个高水平、高质量的数据库? 这就要求数据库的建设者根据建库的目的和宗旨, 通过比较、分析、鉴别、判断, 从众多的报章中, 选择那些较有代表性、较有影响力、又较具权威性的报纸, 诸如上述香港的几家报纸等作为主要的文献源。当然, 有关报章的版权问题及提供用户使用方式等, 也是数据库的建设者不容忽视的重要问题。

### 3.2 数据采摘的标准

数据采摘标准的制定, 是数据库质量的重要保证。香港报章资料的标准规定:

(1) 新闻方面要选取重要的、影响社会层面较广的问题, 要选取学术性、资料性的专题, 而且同一则新闻的系列报道要收齐。

(2) 社论方面要选取与中国大陆、香港、台湾、澳门的社会、政治、经济密切相关的。关于时事评论, 则要选取具真实姓名者所写的文章。

(3) 财经方面有关中、港、台、澳重要公司的业绩、商业活动、投资发展计划等要重点选取, 并强调每份报纸每天需选一篇对香港股市(恒生指数)总结的文章。

(4) 教育方面有关香港 7 所大专院校的重要消息都选取。

(5) 读者来信包括投诉信, 只选取对整个社区, 甚而是对整个社会都具影响者。

(6) 人物专访及介绍重点选取两岸三地的名人要员, 具新闻价值者。

总之, 报章文献的采摘要坚持三点六性原则。三点, 就是新闻的采摘要选取影响社会层面宽的重点、热点、焦点问题。所谓六性, 就是新闻采摘要注意价值性、学术性、资料性、准确性、实用性以及系统性。只有这样, 才能做到去粗取精, 去伪存真, 全而不滥, 精而不漏。

### 3.3 关键词表的编制

考虑到报章资料读者使用面的广泛性, 决定了标引的专指度不宜过高或过低。标引深度过低, 会埋没有价值的文献; 标引深度过高, 可能会导致增加误检率。只有对报章文献进行全面考察, 科学分析, 权衡利弊, 并在实践中不断总结经验, 才能编制出标引适度、质量较高的词表。

标引的目的在于检索。词表的编制应尽量采用文献中出现的概念明确的自然语言, 尽量尊重读者的检索习惯, 使得作为索引的关键词不需作任何解释就能为读者所接受, 读者根据自己的常识和已往的检索经验就能很容易地确定检索词, 从而达到真正的方便检索。

由于报章文献的特殊性, 决定了人名、地名、机构名称、公司名称等专有名词在检索中的特殊作用, 使得这些专有名词在词表中占有举足轻重的地位。因此, 在选择这些名词作为关键词的时候, 人名应选择社会、政治、经济、文化界的国内外知名人士、商业巨子。也就是说要选择那些具新闻价值者。在选择地名时, 一般应选择省、市级以上的; 在选择公司、企业名称时, 则应选择对国民经济发展有决定性影响, 与国计民生密切相关的公司或企业, 等等。

一个好的词表,还应该能够明确显示相关词的词间关系。香港报章资料库的关键词表可以清楚表达的词间关系有:用词—代用词(use /use for)、上位词—下位词(broader term /narrower term)、前称—后称(earlier term /later term)、相关词(see also)等。

表中的关键词均为中英文对照,而且该词表还是动态的。经常根据新出现的概念增添新词,删除一些已失去检索意义的旧词,以保证词表的旺盛生命力。新增词通常由索引员根据文献标引的实际需要而提出,经主管审批,增入词库,即可使用,关键词的排列,英语关键词依英文字母顺序对比排序,汉语关键词则依汉字的笔划笔顺排。

### 3.4 机助关键词标引

由于事先已将拟定的标引用词全部输入计算机,并建立了索引文档,所以标引员在进行标引的时候,较之于手工标引要快速便捷得多。在一般的情况下,用手工的方法进行标引,一个熟练的标引员,一天只能标引 40～50 篇文献。而使用计算机进行辅助标引,一天则能标引 120～150 篇文献,是手工标引量的 3 倍。而且,机助标引还可以避免手工标引常常出现的各种笔误,防止录入错误,从而提高了标引的质量。计算机辅助标引过程大致如下:

(1) 审读文章标题、文摘、引言及重点与结论部分,分析文章论及的主题,找出文章的检索特征。

(2) 在键盘上输入关键词,系统即以前方一致的方式显示正式的索引词。

(3) 索引员在屏幕上挑选合适的索引词并把它记录在 Selected Keyword (选择关键词)栏。

(4) 校验已选定的索引词,经查证无误,就按一下"OK"(表示认可),该索引词就会自动记录在 Keyword (关键词)栏。若觉有误,就按一下"Details"(详细说明),便可以查询有关该词汇的相关资料,可进一步加以判断。确认之后按一下"OK",若觉不妥,则按"Cancel"(删除),然后重新选择。

试举实例如下:

〈〈公仆资料应包括能力操守 中方力劝英合作确保过渡〉〉

经分析,此文应该用三个关键词进行标引:"公务员"、"主权移交"、"中英关系"。首先输入"公务员",屏幕显示:use (用)"文官制度"。假如您对"文官制度"这个词的意义不太清楚,对标引的正确性没有把握,那么您就按一下"Detail",屏幕就会出现"Keyword Relationship Information"(关键词相关信息);文官制度= Civil Service, UF (代) 公务员= Civil Servants。分析比较结果,这两个词的词义相同,词表规定它们之间的用代关系是正确的,那么就按"OK"。其它词汇的标引,以此类推。

### 3.5 资料处理过程的关键步骤

(1) 在选定的报章中选择当天要闻并剪取;

(2) 将文章以图象扫描方式存入数据库;

(3) 关键词标引;

(4) 数据核对;

(5) 将处理好的文章送上 Internet 网。

### 3.6 系统性能要求

(1) 采用客户/服务器(client/Server)方式,终端用户的界面为WWW Client。

(2) 该系统为多用户系统。在资料处理方面,容许系统管理员及多个资料处理员同时进行资料输入和处理。在终端用户方面,容许多个使用者同时访问并检索。

(3) 可以用中、英文处理资料,检索资料、显示资料。

(4) 资料处理界面是一个把所有功能块都结合起来,构成单一界面的集成系统。

(5) 具有各种统计功能。诸如资料处理员每天完成的工作量以及用户使用量的统计等等。

(下转第 46 页)

## 8 参考咨询服务现代化研究

这类研究主要从以下两个领域展开，且以第二个为重点：其一，对现代化检索技术的研究，涉及 CD—ROM 及某些联机检索系统的检索方法和技术。这类研究在广度和深度上都还很有限，尤其是对联机检索的研究，基本上还局限在对检索某一主题可利用数据库的罗列上；对如何有效利用各种检索指令，如何制定检索式还未作深入系统的研究。其二，对数据库建设的研究。这类研究主要从宏观上探讨建立综合性与专业性两种类型数据库的途径和方法。专业性的，如体育情报检索系统的建设、中国专利信息检索系统、社科文献数据库的建设等。关于建库的途径，其具代表性的观点有：走独立自主、自力更生的道路；政府扶持与市场调节相结合，逐步走商品化、产业化道路；重视数据库法律保护问题的研究。

参考文献

1 方卿 西方参考服务评价的一般方法 山东图书馆季刊，1991，(2)
2 付立宏 近年西方图书馆参考服务评价研究述评 江苏图书馆学报，1994，(5)
3 沈固朝 图书馆咨询环境刍议 图书情报工作，1993，(5)
4 肖红 中西图书馆参考工作之比较 云南图书馆季刊，1993，(4)

詹德优 1963 年武汉大学图书馆学系毕生。现为武大图情学院教授。通讯地址：武汉市，邮编 430072。

赵 媛 武大图情学院硕士研究生。通讯地址同上。

(来稿时间：1996. 7. 22。编发者：李万健)

---

(上接第 40 页)

## 4 几点启示

(1) 文献情报机构必须利用现代信息处理技术对文献信息进行加工、整序、深层次开发，才能方便读者检索，提高文献情报的利用率。

(2) 馆藏文献数据库化，以便于利用的方式将文献保存起来，是解决文献“藏”与“用”矛盾的一个重要途径。

(3) 馆藏资源电子化、数据化的程度，标准化、规范化的水平应是评价一个图书馆先进与否的重要标准。

(4) 馆藏资料电子化应从时效性最强、需求量最大、服务层面最广、最能发挥效益的资料入手。

(5) 将文献以扫描的形式存入，建立图象数据库，提供全文检索，这是建立全文电子资料库的一个简单易行的办法。

(6) 计算机辅助关键词标引大大地提高了文献标引的准确率和标引速度，是文献工作自动化的方向。

(7) 中、英文双语处理问题，是中文文献数据库走出国门，汇入国际信息流大循环的一个关键。

(8) 通过国际互联网络 Internet，是使文献信息得以最迅速地传播，并达到最大范围的资源共享的一个重要手段。

吴开华 现为清华大学图书馆编目部主任，副研究馆员。通讯地址：北京清华大学，邮编 100084。

(来稿时间：1996. 9. 6。编发者：翟凤岐)