中国图书馆学报(双月刊)1995 年第 5 期
ZHONGGUO TUSHUGUANXUEBAO

理论研究·实践研究

陈光祚

我国电子出版物与全文数据库建设

ABSTRACT The manufacturing technology for electronic edition books is mainly the technology of full-text data bases. The structural design of full-text data bases, the combined of automatic and semi-automatic ways of indexing, the design for basic retrieval functions of electronic edition books, the building of post-controlled vocabulary mechanism and the trunk programme of FULLSEA, etc. are introduced.

SUBJECT TERMS Electron edition books—compilations Full—text data bases-Manufactures

CLASS NUMBER G356.1

电子版图书的制作技术，主要是全文数据库的制作技术。主要包括全文数据库的结构设计、标引规则与标引处理软件、全文检索软件及后控词表机制的建立等。

1 电子文本的格式化处理

全文文本不象书目记录那样有规则，它没有一定的格式。有的分章分节，有的却无章节之分；有的文中有小标题，有的却没有；各个自然段的长度彼此差别很大，有的长达数千言，有的仅有一两个字。而且全文文本中的各个段落往往彼此紧密联系相互补充，构成一定的语境，共同组成一个逻辑上的概念单元。这样，给全文数据库的结构设计带来困难。

但是，电子版图书必须进行格式化处理，否则，数以几十万或几百万字的文本难以组成一个有合理结构的数据库，难以进行快速扫描和检索，也难以定义检索结果输出的范围。

全文文本的格式化可有若干选择。一是从文本中的内容含义出发，将一个相对完整的内容单元作为一个记录的单位。这种记录可以包含若干个自然段。例如，对百科全书文本，可以一个条目作为一个记录单位；对一篇科技论著，可以研究问题的提出、研究历史的回顾、研究方法与实验条件、研究结论，以及有关参考文献等各自定义为一个字段。另一种选择是从文献的外表形式来划分，例如以一个自然段作为一条记录。这种方式简单易行，并且可以由计算机程序完成。缺点是不能将一个有始有末的事件、前后彼此关联的假设与论证所包括的若干段落有机地组合在一

* 本文为“中华社会科学基金”研究项目

起,因此,有时会产生割裂现象,使检索所得段落在内容含义上不完整,甚至是断章取义。全文数据库格式化的一个较优的方案是,将上述两种选择结合在一起,既以自然段作为记录单位,同时将内容上密切关联的若干段落组成一个逻辑上相对独立的大单元。这可称为逻辑块。这种单元可作检索结果处理(如显示、打印等)的单位,此外,中文中的目次、脚注、参考文献、目录等均可考虑作为一些特殊的单元进行处理。

全文数据库的格式化,还包括对每个记录和每个逻辑块加上一定的标识符,并进行编号记数,以便作为以后建立的索引之地址。

2 自动标引与半自动标引相结合的标引处理

所谓标引,就是标出全文文献文本中具有情报检索价值和分析价值的知识项。这些标引出来的知识项,可以以其为基础建立各种索引,从而给读者提供检索的入口,即所谓检索点。例如文献文本中的重要人名、地名、年代及关键词就是标引的对象。当它们被标引之后,就成为全文数据库中倒排文件(索引)中的款目词。读者就可以从这些人名、地名、年代、关键词出发,进行单项检索或多项组配检索。标引是揭示全文数据库中情报或知识价值的钥匙。标引的质量在很大程度上决定了全文数据库的质量。

全文数据库的标引,较之书目数据库的标引有自己明显的特色。

全文文本的篇幅巨大。一部著作,少则十数万字,多则数百万字甚至千万字以上。其中的关键词、人名、地名、年代等知识项不可胜数。例如,《国共两党关系通史》150 万字的文本中仅人名就有 4000 多个。而且每个人名,在著作中多则出现千次以上,少则数百、数十次或数次,也有出现 1 次的。地名和关键词的出现个数与出现频率也相当高。例如,我们对《湖北省地方志·大事记》36000 字文本的试标引中,标引项竟达 3000 余个。平均十几个字中就有一个标引项。这么多的标引项,如果采用手工方式标引,其工作量之艰巨是很难想象的。必须借助计算机自动标引或半自动标引,才能达到一定的效果。自动标引的方法,目前以词典法(或词库法)为主,即事先编制人名、地名、年代、关键词的词表,让计算机根据词表扫描全文文本而加以标引。从计算机作业来说,主要是词表中的词与文献全文文本进行字符串的匹配。

全文文本的词是由未经规范化的关键词组成的。这给标引带来许多困难。首先,同义词很多,例如同一概念有许多叫法,如:"鸦片"、"烟土"、"烟毒"、"毒品"等等;官职与人名之间也有实质等价的同义词,如某人有时以其姓名出现(包括他的名、号、字等),有时又以官职或姓加籍贯的称谓出现;实名与代词也有等价的关系,如"他"、"该年"、"该地"、"次年"等等。代词是全文数据库特有的现象,要依赖上下文的语境才能明白其实质内容,如果把代词作为检索词用,便无意义。但也不能把文本中的代词改写成实词,这样就会破坏文献文本的原貌。其次,文本中经常出现缩合词,如"日伪军"、"陈谢兵团"等。第三,简称与全称,如"湘"与"湖南","鄂"与"湖北"。第四,是地名的今昔不同称呼。如"北平"与"北京","蕲水"与"浠水"等等。这类同义词在书目数据库中有,但在全文数据库中较为突出。如果在检索中输出的只是有关的段落、句子,失去了上下文的语言环境,就会给阅读与理解带来问题。对这些问题可作如下处理。

(1) 对人名的各种正式名称、尊号、别名等均取文献中所题字样。对地名的新旧名称,对同一概念的各种同义词,也都以取文献中所提的词汇为准。

(2) 全文数据库的标引仅供计算机系统内部处理用,在显示和打印检索结果时被删除,以恢复文献的原来面貌。唯一例外的是对年代的加注标引不加删除,例如 1854 年 8 月 13 日(**AT**18540813)中的"**AT**18540813"部分

是增加的加注标引。之所以这样处理，是当按段落输出检索结果时，段落中的原文往往不全记年月，如“8 月 13 日”，甚至“13 日”。在输出检索结果时保留对年代的加注标引有助于检索者的识别。

(3) 全文数据库在检索功能上应该做到文中的每一个字都是可供检索的。全文数据库的标引政策是深度较大的标引，但不是完全彻底的标引。同一人名、地名或名词(概念的关键词)在同一句子中出现两处以上的，只标引一次即可；专指性不很强的关键词(如描述时间的“春”、“夏”、“秋”、“冬”，描述职务的“书记”，“长官”，描述过程的“进攻”、“退却”等)可不标引。

(4) 文本中的人名、地名、文献名、关键词及年代，均冠以下列不同的标识符，以便让计算机识别、抽取和聚类①人名冠以 **N**，并用尖括号括起，如〈**N** 林则徐〉等。一人在文中不同地方分别以名、字、号等出现时，均按原文所题标引，另设后控词表来解决这类问题，保证查全率。②地名冠以 **P**，并用尖括号括起，如〈**P** 武昌〉等。文中同一地名有新旧名之别时，均以原文所题为准。例如〈**P** 北平〉等。③文献名冠以 **D**，并用尖括号括起，如〈**D**《筹议严禁鸦片章程折》〉等。④关键词冠以 **K**，用尖括号括起，如〈**K** 起义〉⑤年代冠以 **AT**，并将年月日改注××××××××8 位阿拉伯数字，再用尖括号括起，如 1854 年 8 月 13 日，标引成：1854 年 8 月 13 日〈**AT**18540813〉。这样便于计算机对年月日进行大于等于或小于等于的运算，以检索一定年限范围的资料。

(5) 文本中某一人名、地名、文献名、年代等，在其后的行文中往往出现“他”、“其”、“该地”、“同年”、“次年”等代词。这种情况需进行加注标引。例如：他〈**AN** 林则徐〉，其〈**AN** 钟人杰〉，该地〈**AP** 黄冈〉，同年〈**AT**18540000〉，次年〈**AT**18550000〉。加注标引的目的是使不定代词明确化，增加检索入口。

(6) 标引词的选取。简称缩合性关键词，如“日、伪军”(即“日军”。与“伪军”)、“两江”(即“浙江”和“江西”)、“一、四纵队”(即第一纵队和第四纵队)等等。为了提高检索功能，除按“**K** 日、伪军”、“**K** 两江”等进行标引外，还可标引成“**AK** 日军”、“**AK** 伪军”、“**AP** 浙江”、“**AP** 江西”、“**AK** 第一纵队”、“**AK** 第四纵队”等。凡有“**A**”开头的标识符，皆为加注性标识符。人名、地名、年号同其他词汇构成的复合关键词，如“陈得才部”、“刘邓大军”、“孟良崮役”、“咸丰帝”等，对这类复合关键词，原则上按关键词处理，标识符是“**K**”。但对“将领姓名＋部队”的关键词，如“陈得才部”，可标引成“**N** 陈得才”，适应检索者的习惯。对“官衔＋姓名”(如湖广总督官文)的称呼分别标引。例如标引成“**K** 湖广总督”和“**N** 官文”，这种做法是把复合概念拆成单元词，既可增加检索点，又便于灵活组配。这种后组式的标引更适合计算机检索的特点。同样，“短程航线”标引成“**K** 短程”和“**K** 航线”，而不标引成“**K** 短程航线”。

由于中文书写或排印中，将词与词、关键词与非关键词都紧密相连成句的特点，如果采用计算机高速自动标引就会出现错误标引。这种错误，有下列五种类型。

(1) 字面嵌套引起的错误标引。例如，人名中的“陈云涛”与“陈云”，就可能标引上两个“陈云”。地名中的“檀香山”与“香山”，也可能标上两个“香山”。这种情况，在关键词中尤其多见，如“反社会主义”与“社会主义”，“西山会议派”与“西山会议”，都可能出现标引上的错误，难以区别。在不同属性的词之间，也可能产生另一类型的混淆，例如人名“范长江”与作为关键词的“长江”，人名“谭平山”与作为地名的“平山”县。特别是以地名和关键词组成的名词，如“大理石”、“黄岩蜜桔”、“美洲虎”等，也会造成地名与关键词的混淆。当然，可采用自动标引中的“最长匹配”原则来处理，但增加了自动标引程序的复杂性。

（2）词的同形异义造成的错误标引。例如，“日照时间长”中的关键词“日照”，与山东省的地名“日照”，计算机就难区别。同样，地名“平定”与关键词“平定”也是这样。

（3）字的假组合而引起的错误标引。例如，“泰山西部地区”，可能导致标引出“山西”这一地名。“他上海校学习”可能标引出“上海”。这类问题在全文文本中出现的概率较高，也是计算机导致“杂凑”产生的原因。

（4）中文文本中出现的转义字句也会导致错误标引，如“马虎”和成语“一丘之貉”，并无探讨动物的意思。同样，“福如东海”，决不实指“东海”。这些都给自动标引设置了障碍。

（5）由于计算机中文处理的局限，难于标引一字词（即由一个字构成的词）。在标引一字词时，可能导致前一汉字的第二个字节与第二个词的第一字节错误地拼成另一汉字。然而在全文文本中存在不少一字词，例如，地名的简称“沪”、“京”、“鄂”、“湘”等又例如朝代的名称“唐”、“汉”、“陈”、“宋”等。一些稍带文言的文本，更会出现如“降”、“俘”、“银”、“渔”等一字词，汉字编码上的问题，容易造成一字词自动标引的“禁区”。这也可以通过在匹配扫描时右移一个整字（一个汉字或一个 **ASCII** 字符）来解决，但这增加了匹配扫描的时间。

因此，对于全文数据库来说，完全自动的标引是不完全适合的，还需要在此基础上进行有人工干预的半自动标引。半自动标引，可利用文字处理软件中的“查找并取代”的现成软件模块，选用其中的“**G**”操作键，回答“**Y**”或“**N**”作判别，以此进行标引。这样，上述五种不易自动标引的问题就可得到解决。这就是自动标引和半自动标引相结合的标引模式。

实行两者相结合的标引模式，就要将标引用的词表分成两个部分。第一部分是既有独立性的检索价值，又不会引起错误标引的词，如“李大钊”、“周恩来”等人名、“武汉”、“杭州”等地名，以及专指性高的关键词，如“中共中央”、“台儿庄战役”等等。选取这些词时，可以将所有的标引词按区位码排序，从相邻排列的词表中，有助于识别如“陈云”与“陈云涛”这样可能引起标引问题的词。当然，如果字面嵌套出现在词的中间时，不易觉察，如“檀香山”与“香山”这两个词按区位码排列，并非相邻。因此，有些错误标引是始料不及的。若发现此类问题（如“檀香山”标引成“香山”）也仍然有补救的办法。就是仍然用文字处理软件的“查找并取代”模块，查找“檀〈**P** 香山〉”改为〈**P** 檀香山〉”即可。这种半自动的标引又可称为二次标引。因为它一般是在完全自动的标引之后进行的。

全文数据库的标引最好是进行词的属性标引。所谓属性标引，就是在标识标引词时，同时指出该词的属性。词的属性有：人名、地名、关键词、年代、文献名等。可分别用一个字母来代表各种属性。由于加上属性指标符，就能区分同形异义词。例如，检索作为地名的“**P** 黄岗”，就不会误检出作为人名的“**N** 黄岗”，以降低误检率。其次，更为重要的是，便于对检索结果按词的属性进行聚类输出。例如，检索“**N** 陈毅”之后，可以对命中记录中的所有人名进行抽取，并在统计频率的基础上按频率高低排成一个人名表。这个人名表就是与陈毅有关的人氏的名单，频率高的人名，显然与陈毅的关系较为密切。频率较低的人名，也表示多少同陈毅有一定的关系。同样，如果将与陈毅有关段落文字中的地名进行聚类输出，也可列出与陈毅有关的地名表，从中大致可看出他南征北战的地方。其中频率高的地名，是他主要活动地点。如果按关键词进行聚类输出，则输出的关键词表大致可以说明哪些事件或事物与陈毅有关。如果按文献名称进行聚类输出，所获得的聚类表则会表示哪些历史性文献与陈毅有关。同样，如果检索某一关键词（例如淮海战役），则也可按人名、地名等聚类，排出与此战役有关的人名、

地名和关键词。这种聚类对于图书内容的分析和情报计量的研究都有重要意义。如果全文数据库不进行词的属性标引，就不可能实现这种聚类，全文数据库的价值就会降低。

在全文数据库的标引中，有时还需进行加注标引。可以充分提供检索入口点，发掘文献的情报信息含量，节约检索者的智力劳动。

全文数据库标引时所用的符号，为〈，〉，**N**，**P**，**K**，**AN**，**AK**，**AT** 等等，在检索结果显示或打印时应该删除，以恢复文献文本的原貌。因此，这些符号和字符应该采用区别于汉字码的 **ASCII** 码。加注标引与一般标引不同的是，一般标引仅删除这些符号与字符本身，而汉字仍然保留。然而加注标引则不同，不仅要删除所加的符号与字符，而且连同加注的汉字一并删除，以恢复文本的原貌。为了达到这一目的，在采用词属性指示符时应与一般标引所用的词属性指示符有所区别，如“**AN**”区别于“**N**”；“**AP**”区别于“**P**”等等。当然，这些指示符以及各种标引符号可以改变，只要便于记忆和程序执行即可。

全文数据库的标引应该适度，凡作为标引的词，应该有一定的检索价值。对于没有被标引的人名、地名、关键词，虽然不能直接作为检索用词进行检索，但仍然有一定的方法补救。这就是通过全文检索软件中的“二次检索”功能又称为“文中扫描检索”来弥补。即在第一次检索的初步结果中，进行顺序扫描，而将未被标引的人名、地名、关键词查找出来。有关“二次检索”的功能，将在全文检索软件设计部分详细讨论。

把握适度标引的原则，其依据主要是文献保障原则和用户保障原则。所谓文献保障，就是文献中出现频率极高和极低的词可以不必作为标引词。用户保障，是指用户在检索时，有可能作为检索出发点的词才应作为标引词进行标引。如果有些词，用户从来不会作为检索词进行检索的，就不必作为标引词。

设立方面词对于全文数据库标引来说也是有必要的。所谓“方面词”，是指有情报检索价值的，但并非在概念上是独立完整的词。这些词仅仅作为事物的一个“方面”。例如“决议”、“文告”、“惨案”、“谈判”、“战役”等等。方面词可以在不影响关键词完整标引的情况下，通过加注标引而进行标引。例如，“沙面惨案”可以标引成〈**K** 沙面惨案〉和〈**AK** 惨案〉两个词。这样，检索用户可以分别从这两个词进行检索。通过方面词检索，可以允许用户进行从同一方面进行情报的检索和统计。如有哪些惨案发生过，举行过哪些起义等。事实上这也是从另一个角度进行的聚类。方面词的设立，可以增加全文数据库的情报价值和分析功能。

3 电子版图书基本检索功能的设计

检索功能的设计与实现是电子版图书制作的核心问题。电子版与印刷版比较，主要的优势是电子版能借助计算机对著作中各个知识项和每个文字句段进行高速、准确、全面的检索。这种检索是印刷版图书的手工翻阅远远不能比拟的。电子版图书的格式化处理、自动和半自动标引处理等环节，目的也在于为用户提供检索功能。作为具有较高学术价值和资料价值的大部头著作，读者的检索要求有如下五种类型。

(1) 查找某一知识项在著作中的出处。例如，书中在什么地方提到了某人、某地、某事？这些知识项所在的段落文字或上下文是什么？见之于印刷版图书中的哪些页码？这是用户的起码要求，是单项检索的要求。

(2) 查找两个或两个以上知识项相结合的文字出处。例如，某人在某一年代的活动，或某人在某年某地的活动，或某一事件中某人的言行，或某两个或某三个人的关系，或某人除参与某个事件之外的所有记载等等。这种检索需要涉及的项目较多，并且这些知识项是相互联系在一起的。这是多项组配检索。

(3) 片言只字的检索。这种检索，所查找的某句话，例如某句名言、成语、典故等。一般

来说，它是不作为知识项标引的，因而实现这种检索功能，需要让计算机合理地扫描电子文本。这是“文中扫描”或“二次检索”。

（4）查找各种人物或事物之间的联系。例如，某一事件中涉及到的地方；某人与其他人的关系，该人与哪些人的关系密切程度如何；某地发生过哪些事件，其中重大的事件是哪些；某人主要从事什么工作或参与什么活动；等等。这种检索要求并不一定需要找到书中的段落文字，而只要求计算机告诉这些问题的答案本身。显然，在印刷版图书的条件下，读者要获得这些答案是相当困难的。但在电子版图书的条件下，计算机可以轻而易举地让用户获得答案。这种检索，就是所谓“知识项综合”或“聚类”，或称为情报的“缔合”。

（5）情报计量学的统计排序要求。这是读者用户的高层次的情报检索需求。例如，想要了解著作中先后出现过多少人名地名，多少个关键词，以及每个人名、地名、关键词的各自出现次数。计算机应极其精确地统计出来，并且按各种次序（如频率的高低，或人名、地名、关键词的字顺）进行排序列表。读者用户根据这些排序表，可以了解著作内容所涉及的广度，帮助估计一定人物的历史作用与影响程度。这种检索就是“频率排序”。

针对读者的上述情报检索要求，作为电子版图书的全文数据库系统的功能设计是：

第一，设置一级检索机制。第一级是作为标引词检索。要为标引词构造倒排文件，使之能直接查找。查找的模式为布尔逻辑检索，包括“与”、“或”、“非”，以及括号嵌套、截词检索（前方一致）等。布尔逻辑的三个运算符以及括号的连用，可以允许检索者表达检索意图，便于调整检索的专指度水平，任意进行扩检或缩检。截词（词的右截断）可以使用户进行非精确一致的检索，允许从词的片段进行“模糊检索”。由于建立了倒排文件，第一级检索的速度很快。第二级检索在第一次检索基础上进行，又称为二次检索。二次检索的对象可以是未经标引的词或字符串，甚至是一个汉字。二次检索无倒排文档可用，因而采用右序扫描的方法进行检索目标与文本的匹配。因此，二次检索又称为“文中查找”，或具体意义上的“全文检索”。在第二次检索中，应实现字符串的位置检索功能，即按指定的字符间的相邻度进行匹配，第二次检索进行的速度较慢。

第二，建立法定数检索的机制。所谓法定数，是用户在检索时指定的检索结果的期望值。例如，他要求检索到有关某一事物的三条段落的资料。当输入该检索词时，可能有10条段落命中，这时系统就要求用户再输入1个检索词，以便在系统内部作逻辑乘的运算，以缩小检索的命题范围。若第二次检索结果仍大于用户的期望值，系统会再要求用户输入一个检索词，再作逻辑乘的运算，进一步缩小检索的范围。反之，如果检索结果小于期望值，系统要求用户再输入一个检索词，在内部作逻辑加的运算，以扩大检索的命题范围，直至接近用户指定的期望值为止。法定数检索是一种逐步逼近的算法。这种检索能改变布尔逻辑检索的盲目性与机械性，能使检索结果具有不同的贴近性，同时也避免用户的检索结果为零。法定数检索是符合人们搜集情报的要求和行为的。

第三，为了实现人名、地名、关键词的聚类。需要在软件中配置对检索结果的再次数据处理功能。它包括自动抽词、频率统计及排序。

第四，全文数据库系统软件还应包括全书规模的处理功能，如全书标引项统计、排序、编制书本式索引的功能。

第五，应实现检索结果输出的多样化，在软件上应有一定的自动编辑功能，并能提供显示、打印、套录的多种输出选择。在结果显示中，对导致检索结果的检索词应有反相显示，以便于用户检查检索结果的准确性。

第六，电子版图书的制作还需一系列辅

助处理软件。例如消除不必要的控制符号、超长记录的监测与处理、排除汉字错位等软件。同时须掌握发现不正确标引及进行纠正的技巧。

电子版图书检索功能的设计应从用户的需要出发，分析其获取情报信息的目的、途径和方法，注意其心理和行为的表现。而检索功能的实现，即软件的编制，应力求选择良好的编程语言和工具，采用软件工程学的方法。

用户界面是整个系统的窗口。它直接面向用户，应该有菜单驱动和命令驱动相结合的驱动方式，应有明确提示。需要进行长时间作业时，应该让用户事先知道，使之有等待的思想准备，并随时了解系统内部处理的进程。对用户的误操作，应提供有礼貌的说明提示。这些都是"用户友好"原则的体现。

4 设置后控词表的机制

全文数据库的标引，是自然语言的标引，很难做到控制词汇标引。这是全文数据库与书目数据库的主要区别之一。电子版图书的标引用词是随著作内容而转移的。即使是同类著作，不同的著作行文用词也会有所不同。由于电子版图书的篇幅较大，不可能对其中的各种用词做规范化工作，因此，电子版图书的标引必须以文中用词为依据，即以自然语言为基础，在标引时不加规范控制。这样，可以节约标引的时间和费用，减少自动标引的复杂性。

当然，应当承认，没有控制词汇的标引，在查全率和查准率方面会产生不良影响。同义词的存在，会使读者用户对某一问题的检索导致漏检。例如，"起义"这一概念，用词很多，如"起义"、"起事"、"举义"、"暴动"、"兵变"、"哗变"、"谋反"等。在没有规范化的情况下，读者用户想要把电子版图书中所记述的各次起义的段落文字全部检索出来，就必须要把各种可能设想到的同义词和近义词统统列举出来，并用逻辑"或"连接成一个检索式。同样，著作中对一个人的称呼，往往也有不同的称谓。例如"孙中山"就会有以"孙总理"、"大总统"、"孙文"、"孙逸仙"、"孙先生"等不同的称呼。事实上，这也可以看作是同义词。

相关词的存在，往往无上位词来统率，因而也会导致检全率的下降。例如，各种自然灾害有地震、洪水、干旱、冰雹、风灾、虫灾、瘟疫等。在上述每种灾害之下，又有不少相关词。要把各种自然灾害的段落文字都检索齐全，则在构造检索式时需要花费很大的智力。而且，在构造检索策略时，用户付出的智力劳动，系统不能记载下来，加以积累继承。举个例子，如果有一个高水平的用户拟订了一个很好的检索式，取得了良好的检索结果。当他完成检索后，第二个用户再来检索同一课题，还必须从头考虑检索式的构造。这就是说，前人的成功检索经验，不能成为后人可以享用或参考的财富，这是以自然语言为基础的全文数据库标引带来的主要问题。

要解决这个问题，一个可取的办法是在电子版图书的检索系统中建立后控词表的机制。所谓后控词表，不是供标引用的，只是辅助检索的一种手段。这是后控词表与一般词表主要的区别。它的构成是：用户认为是等价的同义词并用逻辑"或"相连接的各词，被纳入同一个词组号，构成后控词表的一个片断。例如，"孙中山"、"孙文"、"孙逸仙"、"孙总理"等等，构成同一组的检索词。当用户的检索式中只检索其中的一个词，而需求助于后控词表机制时，后控词表的机制就把这一词组内的所有词均以逻辑"或"相连纳入用户的检索式，使用户的检索式变成"孙中山＋孙逸仙＋孙总理"。这就无疑使检索达到较高的查全率。后控词表可以由专家事先编制，但是更重要的是随着用户在检索中使用逻辑"或"的式子，系统自动捕获逻辑"或"前后连接的各词，而将其纳入后控词表，或者是由系统管理人员根据用户的成功经验，有选择地将相关的一批词纳入后控词表。这种纳入是随检索的不断进行而进行的，是不断增长的。例如，第

一个用户检索"洪水"的检索式只是"洪水＋水灾"，第二个用户检索时想得更全面，检索式是"洪水＋水灾＋大水"，这样，"大水"就纳入原来"洪水"与"水灾"两词的后控词组，使之变成第三个成员。第三个检索者如果又想到一个新词，那么这个新词就可以成为第四个成员。这种后控词表机制，是在自然语言标引的情况下，帮助用户找到同义词和相关词以进行合理检索的工具和手段。借助于这种机制，前人在检索上所花的智力劳动及其成果，可以变成后来检索者能够分享的财富。后控词表可以帮助用户尽可能找全文献中各种可能的标引词，是弥补自然语言标引的局限性的一种办法。作为全文数据库的电子版图书，应当具备这种建立和使用后控词表的机制。

5 FULLSEA 主干程序

Fullsea 并非一个单纯的单汉字系统，也不仅仅为开发全文系统所用。在我们的研制过程中，自始至终贯穿着这样一种思想，即 **Fullsea** 应当是一个融合多种标引方案、书目全文均通用的集成化的情报管理系统。

为了验证 **Fullsea** 的各种性能指标，并检测系统运行的稳定性、安全性和实用性，我们在实际工作中利用 **Fullsea** 作了多次的开发应用。

比如，武汉大学图书情报学院校友通讯录数据库系统。这是一个事实型数据库。我们首先利用 **CDS/ISIS** 建成了此库，然后通过 **ISO**2709 格式将数据交换到 **Fullsea** 系统下。该库容量为 9000 多条记录，每个记录包括姓名、性别、专业、类型、工作地址、在校年月 6 个字段。系统提供 5 种检索途径，姓名、专业、类型均为全字段索引方式，工作地址为单汉字索引。根据检索结果，获得了各种统计数据，例如历届校友男女性别比例、各专业学生变化情况、各地生源人数等，这些结论为我们作进一步的定量分析提供了依据，系统运行良好，其速度明显快于 **CDS/ISIS**。

又比如，《中国名胜诗词大辞典》电子出版物。它是在本人的主持下，由武汉大学图书情报研究所和武汉大学出版社共同研制的全文检索系统。该系统以 **Fullsea** 为原型，经过适当修改以适应诗词全文检索的需要，系统容量达到 120 多万字。

我们针对名胜诗词的特点，拟定了诗名、作者、景点、名胜介绍、诗词正文、注释 6 个字段。提供多种索引方式，保证用户从不同的途径均可找到所需的诗词，其中诗名、作者、景点、名胜介绍为全字段方式，诗词正文做单汉字索引。在将北大方正电子排版中间文本转化为自由文本的过程中，我们保留了各个字段的标识符，如诗名、作者等，初步满足了聚类的要求；同时对各检索点在菜单中均提供对应的入口，减少用户不必要的输入。整个操作轻松方便，随时提供辅助说明及在线帮助。

Fullsea 实际上是这样一个综合体：以 **Dbase** 为代表的微机 **DBMS**，以 **CDS-ISIS** 为代表的典型 **IMS**、以 **Dialtwig**（**Diaiog** 微机模拟系统）为代表的典型 **IRS** 三种系统工具和功能在一定程度上的集成。尽管它在全文本的系统开发上取得了一定的成效，但远未达到成熟的实用阶段，还有待于进一步开发。

Fullsea 在保持记录的最大长度不变的情况下，采用动态策略分配字段空间所获得的最大长度明显优于其他系统，因此它适合于字段少、字段长度较大的全文本数据的处理。这是本系统设计的出发点。**Fullsea** 较大的限制是字段数，但字段数的增加并不困难，只需在内存中再划出一小片空间即能实现。

从数据库的设计、标引、索引到检索，**Fullsea** 系统具有了大多数检索系统的基本功能，特别在软件设计过程中，针对全文检索的要求，**Fullsea** 对一些原有的功能作了扩充，使之适合于全文本检索。某种程序上，**Fullsea** 是一个开发全文系统的工具。

Fullsea 比较注重数据存 （下转封三）

（上接第 33 页）

贮的方便、索引更新的简单及对大数据量处理的要求，采取了 B+树索引技术和索引地址参照链接文件的方式，避免了 B+层数的限制，而且索引更新也极为容易。但是这种链接方式造成了存取数据效率低下。在正常情况下，检索一个索引项地址的读盘次数是 3 次，这段时间用户几乎觉察不到，但读取记录集合号费时颇多，一个记录号就要定位一，检索的绝大部分时间都浪费在磁盘的读状态中。造成效率低下的另一个原因是位置运算。Fullsea 没有记录词的位置信息，它必须在检索结果的基础上重新搜索记录的字段内容才能判断位置是否紧邻。对于大的全文本这种操作极不合算。对效率的改进涉及到许多模块的修改，这是十分棘手的事情。

Fullsea 作为一个开发全文检索的工具，有一定的实用价值，但尚需进一步的完善与优化。这里略举一二。

（1）索引参照文件结构的重新设计。系统目前采用单链文件结构，主要为了实现索引更新功能，但它以牺牲时间为代价，成为系统效率的一个掣肘因素。迄今为止，笔者尚未找到一种合理的文件结构同时满足索引更新和检索速度的要求。

（2）附加索引词的位置信息。这是系统改进的另一重要点。采用这种方式势必增加内存开销，而且必须在检索过程中实施内存与外设相结合的技术，否则难以达到实用水平。这就为开发者们提出了另一个课题，即怎样在软件中使用扩展内存，扩充内存及临时交换文件技术。

（3）与自动标引结合。自动标引对改善全文检索系统的性能有重要影响。Fullsea 今后准备将自动标引作为它的一个功能模块。

（4）后控词表。后控词表对于采用自然语言标引的系统是一种好的、值得采用的方法。CDS/ISIS、IMS-tool 具有简单的后控词表功能。这一点值得 Fullsea 系统借鉴。

6 电子出版物制作中尚待解决的一些技术问题

我国电子出版物的出版发行目前处于起步阶段。从技术上说，还存在一些尚待解决的问题。例如汉字库字量不足，各种电子排版系统不得不造字拼字，而彼此之间缺乏标准化，如果离开电子排版系统的支持，那么电子排版文本中的某些字就无法复原；电子出版物的制作系统与电子排版系统二者不一定相同，但存在二者之间的沟通问题；由于电子排版文本在制作中需要进行全文数据库所需格式的转换，汉字易产生错位；电子出版物的标准化的起步阶段需及时提上日程；电子出版物的定价依据也需进行科学的分析。

陈光祚 1957 年北京大学图书馆学系毕业。现为武汉大学图书情报学院教授。长期从事图书情报学科研教学。发文近百篇，出版专著 3 种，译著 3 种。主要成果有《科技文献检索》、《计算机情报检索系统导论》、《情报检索系统：特性、试验与评价》和《新一代的情报检索系统》等。通讯地址：武汉，邮码 430072。

（来稿时间：1995-01-16，编发者：徐苇。）