汉字全文数据库管理系统及其应用

沈 艺 (南京师范大学)

在过去一段时间内,国内外已经建立了许多实用的信息管理系统,在这些系统中存储的信息大多是结构化的,数据之间的关系根据所采用数据库管理系统的不同有关系统、网状型或层次型,它们是经用户加工过的二次信息,因此用户在使用时一般只能获得一些指示性数据。对于这些信息管理系统,要想获得实质性数据则只有依靠这些指示性的结果再进一步从原始资料中获得。全文数据库管理系统(Full-text Database Management System)作为一种新的信息管理方法,从八十年代以来在国外发展十分迅速。以世界上最大的情报检索系统DIALOG系统为例,1983年在228个数据库中全文数据库为7个,占数据库总量的3%,到目前为止DIALOG系统数据库总量为345个,其中全文数据库达86个,占25%。近年来,随着大容量光盘、高速CPU、高速数字信号处理器、宽带网络等硬件技术以及多媒体技术的发展,为全文数据库的开发与应用提供了良好的基础。

目前,我国使用的汉字全文数据库管理系统,有汉化西文的全文数据库管理系统,并在其中增加了处理汉字全文数据的各种功能;也有自行研制汉字全文数据库管理系统,这些系统规模较小,大多数是在微机上运行。汉字全文数据库管理系统根据其索引方式可分成以词索引和以字索引两类。汉字全文数据库管理系统可提供以下几个方面的功能:

1.能建立全文数据库,并能对全文数据库进行词(字)、句、段一级的加工检索与编辑。

2.能对经过标引的文献资料的全部内容进行检索。

3.能提供文中查找功能的系统。

4.可对文献中任何有意义的词进行检索操作,它除能表示"与"、"或"、"非"逻辑外,还能表示检索词之间的复杂位置关系。

5.允许用户从一个字段、子字段、一句、一词组、一个词或一个词的一部分(或者是单个汉字)去查找,允许用户在上述不同级别上的任意逻辑组合的系统。

6.使用自然语言的正文查找.即把文献以机器可读的形式逐字存储于电子计算机内,然后用自然语言表达检索课题,借助截词、邻接等匹配方法,直接对文献正文进行查找以检索出所需的文献。

总而言之,全文数据库管理系统除具有全文数据库和布尔逻辑组合检索功能外,还具有文本检索功能,即位置逻辑、字符串检索、截词检索等,允许用户以自然语言检索,直接获得原文中有关章节、段、句、词、字等内容。显然全文检索系统对于帮助人们迅速准确地从浩如烟海的文献中猎取的有关记载或论述的文字,具有重大意义。因为用户总希望以最少的努力获得实质性的数据。

通过全文数据库管理系统,可以找到一切输入的文本,用途至为广大,如重要会议记录咨询;图书馆书目、期刊篇目、论文篇目检索;法院中的法规条文、案例检索;政府机关中的公文查阅;研究机关中的学术论文、研究报告阅览;新闻单位中的联机新闻稿件披阅;公司中信件、商务资料检索;医院的病历、治疗方案检索;情报检索;百科全书和辞典查询;电子邮件选阅等。可见,全文数据库内容通常是经典著作、法律条文及案例、重要科技期刊、新闻报道以及百科全书、手册、年鉴、公文等其全部文字或虽非全部文字但包含原著的全部内容。全文数据库包含的信息量是非常大的。

近年来我国已建立了一些汉字全文数据库,如上海交通大学的"法律条目全文数据库",武汉大学的"湖北省地方志全文数据库",陕西省中医院完成了<素问>、<灵枢>、<甲乙>、<难经>等多部中医经典古籍的全文数据库,江苏省中医研究所对<伤寒论>、<金匮>、<脾胃论>等20余本中医古籍原著建立的全文数据库,深圳大学建立了古典名著《红楼梦》的全文数据库等,这些全文数据库都对用户提供了有效的检索服务。

一、全文数据库的数据组织

按文件组织形式,数据存储与检索技术的发展大致经历了三个阶段。第一阶段为顺序检索方法,文件组织只有一个主文件和一个查询文件,这是一种典型的批处理方式。主文件的每一个记录(文献)与查询文件的每个记录(提问式)逐个进行比较,然后成批输出结果。由于检索速度慢,且不能随时改变检索策略,这种检索方法已经被淘汰。第二阶段为顺序检索与倒排文件组成,处理方式也从批处理方式发展到联机形式,该方法要求检索者分别提出两个提问式,第一个提问式必须由具有倒排文件的检索点组成;第二个提问式是其他非倒排文件的检索点组成(有时也可以没有)。这种检索方法的缺点是,快速检索点有限,没有检索命令语言,且如果第一个检索命中的文献集较大,则第二次检索要花较多时间。到 70 年代末期,西文检索技术已发展到成熟阶段,即第三阶段,这一阶段文件的组织特点是,文献记录的全部字段都可以倒排,主文件的记录采用可变长存储,在传统倒排文件的基础上增加了效率更高的索引文件(如 ISAM,VSAM,B 树等)用户可以对任何字段,子字段进行快速查找,并使用丰富的检索命令语言随时修改检索策略。

随着计算机存储设备价格的降低及检索技术的发展,逐步形成全文数据库管理系统。全文数据库的建立以及检索功能的实现是全文检索的两大技术支持。全文数据库一般由一个变长的主文件和一个在索引文件控制下的倒排文件组成,索引文件倒排文件在物理上是分开的。检索时,由索引文件指向倒排文件,倒排文件指向主文件。主文件中一般定义以下几种数据类型的字段:

1.文本型字段(text),适于由若干段落和句子组成的文本,如普通书信、论文、文摘、产品说明等。

2.短语型字段(phrase),适于较短的文本,如论文标题、书名、人名、地址、产品名等。

3.数字型字段(number),适于数值信息(整数的实数),每个数字可分配一个子字段。

4.日期型字段(date)。

5.时间型字段(time)。

全文索引与检索是指上述前两类字段而言,后三类字段则按整个字段或子字段被索引。西文以词做索引;汉字既可以以字做索引,也可以以词做索引。就目前来说,由于汉语计算机分词的困难目前尚未解决,全文检索系统自动生成词索引难度较大,而以单个汉字做索引则较为简单,单个汉字根据各种规则可方便地组成词。

1.以字索引的数据组织

下面用一封短信为例,说明单汉字全文检索系统中全文数据库的数据组织。限于篇幅,例子中的主文件记录只列出记录号、收信人、收信地址和书信正文这四个字段,并用 F 和 P 分别标识"字段"和"段落(子字段)"。具体实现时主文件可按连续可变长存储:

^ 1F 1588

^ 2F^1P 隆。史密斯先生

^ 3F^1P 斯巴克林大街 16 号^ 2P 斯巴克莱顿^ 3P 美国

^ 6F

史密斯先生:

您的电传已收到。我们将于 11 月在纽约展示这个系统,那时我们送给您一些有关那套系统的新资料。

希望您再坚持几天,我记着您的事。

您的忠实的

全文数据库管理系统扫描这个汉字记录,生成全部汉字和阿拉伯数字的倒排文件,并按数字与汉字内码排序后生成关键字(词)索引文件,下面选择 8 个汉字(或数字)予以说明,其余字(词)类同。

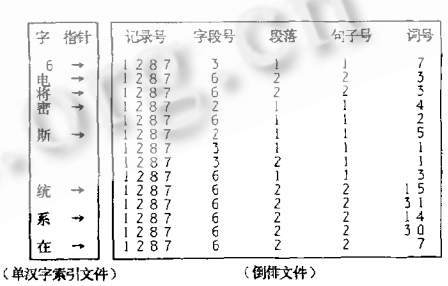

图 1

由上图可以看出,这是一个不使用"非用词典"的单个汉字的全文索引。其左边的"字"排序后按叶→结点→根的生成次序生成一个 B 树结构的索引文件,并指向左边的倒排文件。由于索引文件中包含了非汉字串(如数字 16)。非汉字串的扫描原则应遵循西文件倒排的原则。这样才有可能保证中西文兼容。

2.以词索引的数据组织

对于以词索引的全文数据库,词作为检索的基本单元,标引与检索的着眼点是体现相对独立完整概念的词,比较符合人们的思维和表达习惯。借助于词表,较易实现对同义词、多义词、近义词、反义词、相关词等的的规范和控制,从而保证检索的标引和检索在一定的检索语言基础上进行,显然有利于查全率和查准率的提高。其主库和索引倒排文件通常采用以下结构:

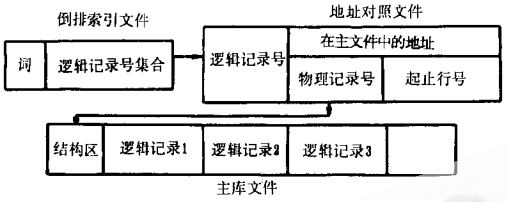

图 2

但是,计算机自动分词是一个很大的困难,而手工分词对于一个稍大一点的全文库来说工作量极大。

二、汉字全文数据库的检索

全文数据库检索系统除具备一般的逻辑检索功能外,还应具备邻近(位置逻辑)检索、字符串检索、截词检索、同义词控制、后控词表对检索策略的自动调整等功能。

邻近检索是指在一个允许用户使用文本中的词构成提问的检索系统中常常需要对两个或多个提问词指定相对位置,这主要有以下几种:直接相邻;在二词(字)间有一定数量的有意义的间隔或符号;二个词(字)间间隔的意义的字符数量有一个上限;二个(词)间可间隔任意数量的词(字)等。实现这种邻近操作所采用的方法主要有二种,一种是先利用基本倒排文档获得符合二个词(字)"与"逻辑的条目,例如要求"情报"与"检索"直接相邻,则通过"情报"AND"检索"获得文献的条目号,然后从主文件取出这些条目(称一次检索结果),对这些条目中有关字段用字符串匹配的方法逐篇顺序检索,把符合要求相邻的条目作为最终结果,也就是通常所说的二次检索结果。另一具实现相邻操作的方法是利用改进的倒排文档,由于在倒排档中每个词条中加入了其在正文中的位置信息,这些信息包括条目编号、段号、句号以及在句中的词(字)号,如图 1 所示。通过二个词(字)的位置信息,可以进一步确定检出条目是否符合用户指定的邻近检索要求。因此查找"系统"与查找"系统"要求这两个汉字必须处于同一个记录、同一字段、同一句子且位置相邻;而查找"系"AND"统"则不论它们是否处于同一字段、同一句子中,只要在同一记录中就算命中。

在自然语言检索中,字符屏蔽和字的截断操作是非常需要的。截断是字符屏蔽的特殊形式。屏蔽通常包括三种:屏蔽指定的字符;屏蔽小于指定数的字符;屏蔽字符数不受限制。它们的实现主要依赖数据库的组织。如在倒排档的词典采用字顺序排列的话,检索时就很容易实现后屏蔽(后截断)。将字典中的每个词的字符倒置后按字顺排列,用以实现前屏蔽。若词典利用两维散列技术组织可以实现前、后任意截断。在顺序检索中,对不同屏蔽要求的提问词前后加上不同的标识,然后构筑成有限状态自动机,以此有限状态自动机对文本中的字等逐字扫描,同样可以完成截断操作。

在全文检索系统中采用文本中的自然语言进行检它,由于异形同义词存在,将影响检索效率,人们为提高检索效率进行了许多研究。如建立同义词典,使全文数据库中的异异同义词得以联系,这样使用户、文献所用的词汇通过词典得以统一,从而提高查全率。

此外,全文检索系统还可以运用模糊集合理论及人工智能技术,通过字、词、句的分析,获得一些具有记忆与联想功能的语义信号,运用模糊规则建立不确定性规则库的推理策略,从而保证理解和检索在正确、可能的轨道上起到积极作用,对标题、篇章进行上下文统一理解,减少单句分析中遇到的不确定因素。同时从全文理解吸取的新知识或数据,加到相关库中,作为历史记录以便以后查询和进一步检索全文的需要。

三、汉字全文数据库管理系统存在的问题

1.汉语的切分

基于词的汉字全文检索的一个重要工作是检索词的选取。选词有手工选取,如针炙古籍全文检索需要从一篇文章中提出病症、俞穴、作者、针炙方法等。又如湖北省地方志中提取的主题词、人名、地名、年代等。也有计算机自动标引,自动标引实现方法主要有三种:词典切分法、单汉字标引法和逻辑推理法。手工标引对于较小的全文数据库尚能实现,但也费时费力,容易出错。计算机自动标引的词典切分法则会遇到判断哪几个字在一起可能构词困难,如"热能发电"是切分成"热能"、"发电",还是切分成"热"、"能"、"发电",计算机难以作出。

总得来说,基于词在的主要问题是:(1)词表的维护是一件需要付出相当代价的,并且永无尽头。各门学科和社会生活的发展,使新的概念的新的词汇层出不穷,而词表的更新总有一定的"时滞"性。(2)词汇是人类的知识产物,知识体系的不断丰富往往导致原有的一些词汇的合并、分离,更改含义范围,使词与词之间出现新的关系。(3)词表所收的词,不可能达到彻底的专指。因为词表必须控制自己的规模。而单汉字标引法正好弥补词典切分法的不足。当然,单汉字模式并不是十全十美,从下面的几点讨论,可以看出其不足的方面。

2.汉字分布不均匀

对于以字索引的汉字全文数据库,在做单汉字标引时,由于计算机中标准化的汉字不足 7000 个,因此汉字关键词索引文件(不考虑汉字文本中的非汉字串)比西文关键词索引文件小得多,但汉字倒排文件比西文倒排文件要大得多("指针"个数多),且不同的汉字在倒排文件中的分布不均匀。在计算机中可处理的标准汉字共 6763 个,其中一级汉字 3755 个(常用字),二级汉字 3008 个(次常用字),而一级汉字的使用频度在 90%以上,即使在一级汉字中,每个汉字的使用频度也相对很大。另外,汉字的分布情况还与记录的具体内容有关,每个汉字在自然科学与社会科学不同领域所出现的频度也不一样。这样对一二个较大型汉字全文数据库来说,全文索引后的不同的汉字在 4000 个左右,而高频汉字的"指针"数甚至超过 4000 个。如果用传统的顺序链表结构来组织倒排文件,则检索速度非常慢。解决的办法是,对每个汉字的"指针"链表采用多级索引,这样可以大大加快查找链表的速度。下图是采用三级顺序索引方法来组织。

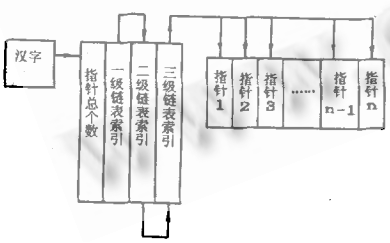

图　3

如果每级索引包含 10 个索引道值,则三级索引可控制 10000 个指针。对 4000 个不同的汉字在倒排文件中对应 4000 个这样的链表,这个庞大的指针链表群有时要求在物理上必须是一个文件。

3.汉字机内码问题

在不同的计算机系统中,汉字机内码表示也不同。就目前我国汉字处理而言,最常用的机内码有两类:与 ASCⅡ代码集适配的异形国标码和 EBCDIC 代码集适配的带标识的国标码。前者广泛用于微小型机上,后者用于大中型上。异形国标码用两个高位置"1"的字节表示一个汉字,故与 ASCⅡ字符互不混淆,这种代码体系中任意两个高位为"1"的字节都可以组成一个汉字的代码,因此在选取两个字节作索引时,可能会错位,造成索引指针在一个不正确的位置上。对于带标识的国标码,虽然也用两个字节表示一个汉字,但在中西文混合字符串中,必须用"SO"(西文转入汉字)和"SI"(汉字转入西文)这两个控制码作标识,否则 EBCDIC 字符将与汉字发生混淆,当处理到控制码"SO"时,说明后面将转入汉字,做索引倒排时可依次取两个字节,然后在这两个字节前面加上一个"SO"控制码字节,把这三个字节代表的一个汉字存入到关键词索引文件中,但是这种代码体系检索处理很复杂,实现困难较大。实施 ISOIEC 10646"通用编码字符集"后,这些问题可望解决。

4.汉字的输入

一个汉字全文数据库往往有数百万、数千万字甚至更多,汉字的手工输入不仅速度慢,成本也高。为提高建库速度,降低成本,可采用计算机汉字扫描识别(OCR)技术。利用它还可以提高录入的准确性,减少录入中的丢段漏字,降低人工校对修改和费用。OCR 可将印刷的和手写的资料录入到计算机中,对于印刷的资料,可保证每天十万字的入库数据,每万字的录入费用不高于 3 元人民币。手写体的资料录入效果略差于印刷体的资料。

四、结束语

在计算机自动分词问题解决前,实现基于字的全文检索系统较为符合我国的国情,具有一定的优势。对单汉字索引其他一些不完善的地方,可以用下述措施进行优化:

1.建立停用字表,以便在建立索引时滤掉虚词和不必要的常用词,压缩索引的篇幅。

2.在单汉字索引的全文检索系统中适当增加控制词汇的因素,增强主题词与自由词检索功能。

3.在单汉字系统中增设后控制表,以加强检索功有。

4.建立与单汉字检索系统并进行的完整控制词表文档,把它作为检索用户的辅助手段。