●靖培栋　宋雯斐

基于混合索引的中文全文检索系统研究

摘　要　在中文全文检索系统中引入了混合索引,建立了混合索引之 Hash 索引,给出了 Hash 索引在内存中的存储结构,并给出了这种索引下的检索算法。这种索引既能保证索引的全面性,又能提高系统检索效率。通过实际构建系统,探讨了基于混合索引的中文全文检索系统的实现。图 3。表 4。参考文献 5。

关键词　全文检索　单汉字索引　混合索引　Hash 索引

分类号　G354

ABSTRACT　The authors introduce mixed indexing into Chinese full-text retrieval system, establish Hash indexing in mixed indexing, provide a storage structure of Hash indexing in memory, and provide a retrieval algorithm in such kind of indexing. This indexing can not only ensure the comprehensiveness of the indexes, but also enhance the efficiency of system retrieval. By actual examples, the authors discuss the realization of a Chinese full-text retrieval system based on mixed indexing. 3 figs. 4 tabs. 5 refs.

KEY WORDS　Full-text retrieval.　Single-character indexing.　Mixed indexing.　Hash indexing.

CLASS NUMBER　G354

本文对中文全文检索系统引入了混合索引,并且建立了混合索引之 Hash 索引。这种索引既能保证索引的全面性,又能提高系统检索的效率,这些技术与方法在中文全文检索系统中有通用性。笔者开发了《邓小平文选》全文检索系统进行试验,开发工具是 VC + + 及 SQL Server 2000。

1　系统设计原则

系统的设计遵照人性化原则:在确保用户检索高自由度的前提下,减少用户的智力负担,同时也考虑系统的反馈速度和空间开销,在这些要素间作一个折中,达到以上各方面的相对平衡。

索引全面性原则:本着要全面、深度地开发利用隐含在正文内部全部的信息的目的,为正文中的每个字或词建立索引。

索引细粒度原则:索引粒度是指索引能够定位到检索词在文档中出现位置的准确度,粒度的粗细直接影响到索引所占的系统空间,索引粒度越细,占用的空间也越大。一般的索引只返回检索词出现的文档,而本系统则要求定位到原文的准确位置。

限度系统开销原则:本系统的细粒度原则决定了本系统索引空间会相当大,并且单纯的字索引系统的查询速度也有所欠缺。在索引中引进单元词与关键词,以压缩索引、改变查询策略来提高系统查询速度。

系统总体包括:全文数据库模块,混合索引模块,Hash 索引模块和检索模块。

本系统以《邓小平文选》作为试验性全文数据库。因为本系统是一个细粒度索引,索引需要定位到文章段落的句子中,所以全文库中每一段落作为一个记录,数据库字段如表 1。

表 1　全文数据库结构

字段名	记录号	文献号	题名	段落号	正文段落
字段标识	ID	ArticleID	Title	ParaID	Content
字段类型	int(4)	int(4)	varchar(50)	int(4)	varchar(2000)

其中,记录号 ID 用来标识记录;段落号 ParaID,指本记录的正文在所在文章的第几段落。

2　混合索引

一个比较完备的中文全文检索系统需要高速的索引、健全的检索途径等多项技术。其中最核心的就是索引机制和检索算法,它们直接影响到系统的响应速度和用户查询的结果。标引的语言单位可以是单字、词或短语。

目前,中文全文检索系统是以单汉字或者词作为索引单元。单汉字索引可由系统自动完成,词索引实现的关键技术是对原文档进行分词处理,将文档分解为若干词的集合,用这些词建立索引。本系统所用的索引是由单汉字索引、单元词索引及关键词索引共同构成的一个索引,本文称之为混合索引。它既能保证系统的查全率,又能克服单汉字索引查找效率较低的问题。图 1 是混合索引倒排文件建立的数据流图。

全文索引要求对全文数据库中的每个字词都提供检索入口,哪怕是被认定无检索意义的虚词。因为既为全文检索,就不排除用户的检索提问中引用某篇章的句子片段中包含虚词。再者,哪些词是无检索意义的虚词也不易定论。比

如:"的"可能被认为是当然的无检索意义的虚词,而被作为停用词,如果这样,像"的确良"这样的词将会被漏检。当然这种准确无误的检索效果会以系统的空间和速度作为代价,至于做怎样的折中处理,要看系统的目标如何定位。本系统为了保证查全率,未使用停用词表,因为本系统采用混合索引,绝大多数情况下不会影响系统的速度。

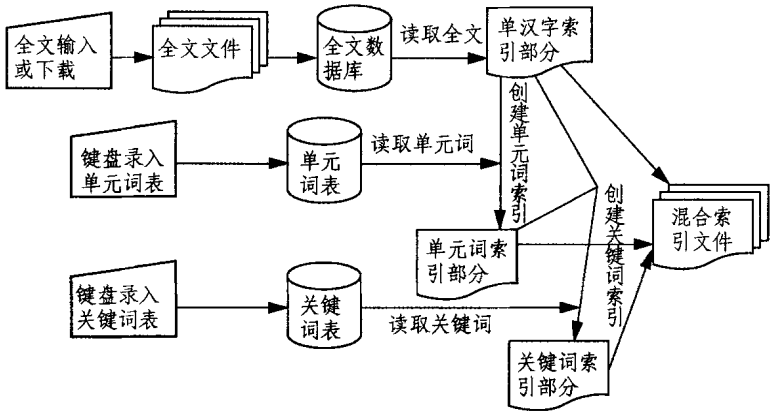

图1　混合索引建立流程

2.1　混合索引结构

本系统的混合索引采用倒排文档结构:

<字或词　记录数　记录号1　记录1中此字或词的个数　位置1　位置2　…

记录号2　记录2中此字或词的个数　位置1　位置2　…

…

记录号n　记录n中此字或词的个数　位置1　位置2　…>

结构中,记录数、记录中字或词的个数、位置都为整型int.占2个字节。

在混合索引中,每一字或词有一行索引信息,即一个字或词的索引信息写入文件后,写入回车换行符。

2.2　混合索引的创建

混合索引由3部分构成:单汉字索引部分、单元词索引部分、关键词索引部分。分别给出各部分的创建。

(1)所谓单汉字索引法,就是将文本正文中的每个汉字均作为索引词,不加选择地进行索引。这种索引可以由计算机自动完成。本系统混合索引的单汉字部分,首先由系统自动完成。

单汉字索引法避开了语词切分的问题,利用计算机自动索引,大大加快了索引速度,也增强了索引的客观性和一致性,再就是单纯的字索引适用的学科领域比较宽广。由于不需要建立词典,打破了中文分词词典的学科领域限制,用一个单汉字索引文件就可以快速完成全文检索。而且索引处理新词的能力较强,因为它打破了词的限制,把词看成是若干索引单元的集合,任何新词都是单字组配的结果,这也是以词典为后盾的分词索引系统所不及的。

本系统录入《邓小平文选》第一卷共41篇文章,每篇文章分段入库,共946个记录。经统计,全文库容量约20M,共167948个汉字,并且文档中的字剔除重复汉字共2432个,其中一级汉字2174个,二级汉字258个且字频都没有超过10。假设平均每个字在m个记录中出现,根据索引倒排档结构,可以计算出单汉字索引倒排档部分的空间大小:40M+2432×2(记录记录数的空间)+2432×m×2<45M。表2清楚地列出了上面各项数据。

表2　字频率分析

名　称	资源数
记录数	946
全部文档大小	40MB
需作索引汉字个数	167948
去同后汉字个数	2432(其中一级汉字2174,二级汉字258个)
索引倒排文件大小	<45MB

在本系统的全文数据库中,汉字的频率相对集中。频率500以上常用的近70个汉字,其占用了71364个汉字空间,占了需作索引汉字的42.5%,也就是说这些汉字的索引在单汉字索引部分中占了约42.5%的空间。

单汉字索引系统的检索方案有两个,以一个有n个汉字的提问式为例。一种检索方案是从单汉字索引倒排文件中找出这n个汉字的位置信息,然后进行n-1次的汉字索引集合的逻辑乘运算,找出符合提问式的全文记录;另一种则是对上一种方案的改进,采用"首字定位,全词匹配"的检索算法[1],利用检索词的首字查找单汉字索引,获取记录,然后直接用检索词与该记录进行固定位置与长度的比较。

(2)单元词索引。一般用户都是以词作为完整的概念词进行检索提问的,词一般都由多个汉字组成,检索时都要经过二次、三次乃至更多次逻辑乘运算,耗时较多,降低了系统检索速度[2~3]。单元词是最小和最基本的词汇,使用单元词进行索引可以提高系统的查询速度,比如一个由n个汉字组成的检索语句,如果用单汉字索引进行检索,就得经过n-1次的集合乘积操作,而采用单元词索引可以把检索语句分解成若干个单元词、字的组合,可以大大减少集合乘积操作的次数,弥补单汉字索引的查询速度不快的缺陷。

本系统采用《现代汉语频率词典》(北京语言学院语言教学研究所编,北京语言学院出版社出版)的报刊政论(包括经济、政治、哲学、法律、历史、地理等)类词,使用了出现

频度最高的2000个字词,并去掉对相应字的索引。《现代汉语频率词典》中报刊政论类词统计来源于1951～1981年人民日报、光明日报、国家领导人的报告及法律政治历史地理等方面34种语料,这与《邓小平文选》的主题基本一致,都是政论性的文章,措辞用语比较相似,有较强的参考价值。但是由于该词频词典出版年份较早,依据的文章都集中在中华人民共和国成立之初到改革开放之前的这段时间,对解放前及改革开放后的文章未做统计,并且这个时期中国的状况有不小差别,欠缺全面性,但这并不妨碍本系统的试验性运行。录入《现代汉语频率用语》的报刊政论类词统计中的词(除去单字),本系统录入了1405个单元词(它们在《现代汉语频率用语》中出现的频率在前2000个字词中,除去单字)作为试验性单元词表进行索引。

对组成单元词的各字,从单汉字索引倒排文档中取出它们的位置信息,如果这些位置信息在原文中的逻辑位置符合单元词中各字的顺序逻辑关系,就可以记录下单元词中首字的这个位置作为这个单元词在单元词索引中的位置信息。并且删除单汉字索引部分中关于此单元词的重复信息,这样可以减小索引倒排文件。假设一个单元词包含m个字($m \geq 2$),此单元词在原文中出现n次,这样,用单元词索引在索引中需要n个位置信息,而用单字就需要$m \times n$个位置信息。用这种方法,不需要再对原文数据库进行扫描、分词就可以编制单元词索引部分。

对单元词索引部分进行细致的统计分析,可得单元词索引部分的基本状况,如表3。

表3　单元词统计分析

名　称	资源数
单元词数	1405
单元词在文中出现的频率	46362(其中双音节45969,三音节378,四音节15)
频率为0的单元词个数	197
实际被作单元词索引的数目(去重后)	1208
少作索引汉字个数	46770
节省索引空间	$46770 \times 2B = 93540B$

(3)关键词是直接从文献的题目、正文或文摘中抽出的具有实际意义的词语。引入关键词索引的原因是:

首先,关键词索引可以弥补单元词表主题分散的缺陷。单元词一般通过组配才能表达具有一定专指度的概念,增加系统负担。

其次,单元词表中的单元词录自《现代汉语频率用语》的报刊政论类词统计中词频最高的词(除去单字)。统计标引法实验发现,频率太高的词(高频词)分辨率很低,有些甚至接近于零。因为它们一般都是一些只起语法作用而无实际内容的功能词,或是一些很泛指的词。单元词表中的词多为泛指的词,它们专指度低,不能起到区分不同文献的作用,因而用单元词表的索引效果质量不高,不能满足用户的检索期望。而关键词索引则具有完整的检索概念,有更高的专指度和检索效率、检索质量。

关键词从《<邓小平文选>索引》[4]中提取。该索引共24个专题,各专题下一般有四级标题,并且标题都尽可能引用邓小平同志的原话。笔者根据这些标题原话,提取《邓小平文选》第一卷相关的关键词作为本系统的关键词表。

我们已经建立了单汉字索引、单元词索引,全文库中各字、单元词表中的各单元词在全文中出现的位置都已经标明,可以根据这两个索引的字、词位置关系明确某关键词在原文中的逻辑位置。和单元词索引倒排文件的建立一样,关键词索引倒排文件也不需要对原文数据库进行扫描、分词,从关键词的组成字、词的单汉字索引、单元词索引的位置信息就可以确定关键词在原文中的位置。并且删除单汉字索引部分与单元词部分中关于此关键词的重复信息。同单元词索引一样,使用关键词索引也可以减小索引倒排文件。

3　混合索引之Hash索引

如果需要装载混合索引倒排文件入内存,很可能导致系统内存空间的溢出。为此需建立混合索引倒排文件之Hash(哈希)索引。它是索引的索引,它的使用也可以提高检索效率。考虑到字或词的首字内码作为Hash索引的键值很直观,因为不同的词的首字可以是同一个字,内码相同,采用链地址法解决首字相同的词的存储问题。

表4　混合索引的Hash索引结构

键值(int)	同首词数(short)	字或词数、字或词、Hash地址(行号)(int, string, int)
key1	$wordscount_1$	(n_{11}, $word_{11}$, $line_{11}$)(n_{12}, $word_{12}$, $line_{12}$)…
key2	$wordscount_2$	(n_{21}, $word_{21}$, $line_{21}$)(n_{22}, $word_{22}$, $line_{22}$)…
…	……	……
keyh	$wordscount_h$	(n_{h1}, $word_{h1}$, $line_{h1}$)(n_{h2}, $word_{h2}$, $line_{h2}$)…

Hash索引是常住内存的,它在内存中的结构如图2。

图2中左侧是一个结构数组,结构中的第一个成员是字的内码,第二个成员是同首字词数,第三个成员是指向单链表的指针,单链表中存储的是同首字、词在混合索引中的行号。结构数组中的信息按内码从小到大排序。

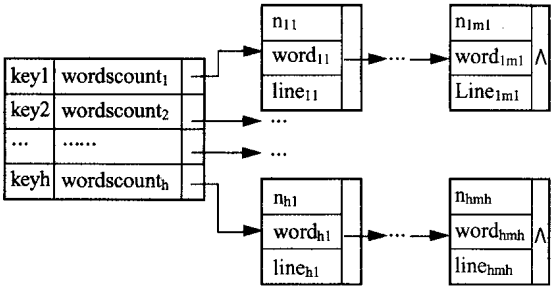

图2　Hash索引在内存中的结构

(下转第98页)

7 周文骏. 文献交流引论. 北京:书目文献出版社,1986
8 宓浩,黄纯元. 知识交流与交流的科学——关于图书馆学基础理论的建设. 图书馆研究与工作,1985(3)
9 南开大学图书馆学系等. 理论图书馆学教程. 天津:南开大学出版社,1986
10 宓浩. 图书馆学原理. 上海:华东师范大学出版社,1988:299
11 程焕文. 北刘南杜 世纪大师——论刘国钧先生在20世纪中国图书馆学术史上的历史地位. 见:北京大学信息管理系等. 一代宗师——纪念刘国钧先生百年诞辰学术论文集. 北京:北京图书馆出版社,1999: 144
12 吴慰慈. 图书馆学书目举要. 北京:北京图书馆出版社,2004:3
13 于良芝. 图书馆学导论. 北京:科学出版社,2003:174
14 范并思等. 20世纪西方与中国的图书馆学——基于德尔斐法测评的理论史纲. 北京: 北京图书馆出版社,2004
15 刘宏波. 第三代图书馆学家与图书馆学多元化. 图书馆,1988 (4)
16 韩继章. 新一代图书馆学家及其时代使命. 图书馆论坛,2005(6)
17 王余光. 图书馆学前辈学术著作的传与读. 图书情报工作,2005(1)

陈源蒸 中宣部出版局离休干部。通讯地址:北京东厂北巷4号楼553室。邮编100086。

(来稿时间:2007-03-13)

(上接第87页)

当检索需要读取索引信息时,根据首字应用折半查找算法[5],在结构数组中找到相应的键值,根据此键值定位到相应的链表,在链表中匹配相应的字或词,取出它在倒排索引文件中的行号,便可以找到相应的索引信息入内存,然后进行逻辑乘运算,便可以得到查询结果。

4 系统检索

在系统检索界面输入检索字符串,如果有多个检索词,词间以空格作为分界,这些检索词被处理为"逻辑与"的关系。根据词间空格进行检索串的分割,得到检索词个数。

我们采用的检索算法如下:

(1)取出一个未处理的检索词,建立一个bool型变量cut,初始值为false。

(2)在Hash索引的结构数组中,运用折半查找算法确定其中是否存在此词的首字,若不存在,则查询最终结果集为空集,转向步骤7;若存在,则在相应的单链表中查找出最长的与检索词前方一致的词,通过混合索引得到此前方一致的最长词的查询结果,若此最长词的长度等于检索词的长度,转向步骤4。

(3)若此最长词的长度小于检索词的长度,则将检索词截去此最长词,将剩下的部分作为一个新的检索词,cut = true,转向步骤2。

(4)若cut = false,则可从混合索引直接得到关于此检索词的结果;若cut = true,要将结果进行逻辑乘运算,以确定原文中包含此检索词的信息。

(5)对于其余检索词重复步骤1~4。

(6)得到每一个检索词对应的检索结果集后,对它们进行集合求交的运算,最终得到包含所有检索词的段落集合,即为最终结果集。

(7)检索算法结束。

从上面的检索算法明显看出,使用混合索引,在多数情况下可以减少逻辑乘运算的次数。前面已经说过,使用混合索引可以减小索引倒排文件。

若步骤6中得到的最终结果集非空,则将相关段落显示出来,用黑体显示段落中出现的检索词;若最终结果集有多个段落,按它们在原文中的次序显示;若最终结果集为空集,提示用户未找到相关信息。

系统检索流程如图3。

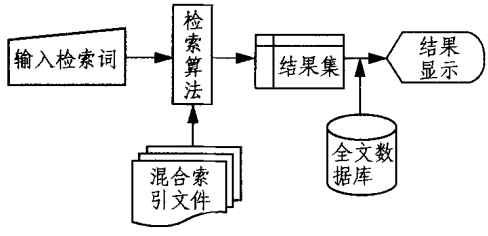

图3 系统检索流程

5 结束语

引入了混合索引的全文系统既节省了索引空间,也提高了系统检索的速度,并且不影响全文检索系统的索引全面性。本文虽然采用了《邓小平文选》为全文库,词频统计词典中时政类词为单元词表的取词依据,《<邓小平文选>索引》为关键词表的取词依据,具有特殊性。但是,混合索引适用于任何中文全文检索系统。

参考文献

1 王森. 单汉字标引技术的改进研究. 现代图书情报技术,1997(2)
2 余海燕,张仲义. 基于单汉字索引的全文检索系统的优化研究. 中文信息学报,2001(15)
3 钱爱兵. 全文检索算法设计及全文检索系统概述. 现代图书情报技术,2003(2)
4 中央社会主义学院编写组.《邓小平文选》索引. 北京:华文出版社,1996
5 严蔚敏,吴伟民. 数据结构(C语言版). 北京:清华大学出版社,1997

靖培栋 北京师范大学管理学院教授,博士生导师。通讯地址:北京。邮编100875。

宋雯斐 北京师范大学管理学院情报学硕士生。通讯地址同上。

(来稿时间:2007-03-02)