

基于倒排表的中文全文检索研究

杨安生

(惠州学院数学系 广东 516015)

摘要 对全文检索倒排表技术作了较细致的研究,对全文检索的基本原理和技术进行了深入的探讨。对中文分词方法作了重点研究和总结,并对词典分词法中的最大匹配法加以改进,使用C++设计了一个程序,真正实现了最大匹配法。

关键词 全文检索 倒排表 中文分词 词典分词法 N元素引法

中图分类号:G354.45

文献标识码:A

文章编号:1005-8095(2009)07-0077-04

1 中文倒排表的建立

倒排表索引法是受书目索引启发而诞生的,是一种将文本中出现的各个索引项和索引项出现的位置信息存储在称为表结构的索引中,检索时,仅查找倒排表来检索查询词的方法。在查询的时候由于可以一次得到关键字所对应的所有文档,所以效率很高。在全文检索中,检索的快速响应是最为关键的,而索引建立是在后台进行,效率会相对低一些,所以目前在全文检索中大多数是采用倒排索引法。

倒排表结构描述如下:

假设索引项集合 $W=\{w_1, w_2, \dots, w_m\}$, 文本项集合 $D=\{d_1, d_2, \dots, d_n\}$

如果索引项 w_j 在 i 个文本中出现,则 w_j 按以下形式组织:

$w_j: d_1, d_2, \dots, d_i$ (w_j 的值按升序排列)

一般索引项在文本中出现的位置也存放在倒排文件中,则按以下形式组织:

$w_j: (d_1: p_{11}, \dots, p_{1k_1}), (d_2: p_{21}, \dots, p_{2k_2}), \dots, (d_i: p_{i1}, \dots, p_{ik_i})$

其中 p_{ik} 表示 w_j 第 k 次出现在 d_i 文本中的位置。

中文倒排表的建立与检索在原理上与英文相同。以汉语为查询对象的倒排表的建立方法分为两大类:一类是基于字符的索引,不进行词法分析,直接从文本建立倒排表的方法;另一类是基于词的索

期调查和统计读者的满意程度,不断调整采购方针,以最大程度满足读者需求。总之,提升人文社科资源的质量是一项长期的持续不断的系统工程,必须严格遵循文献资源建设的规律,循序渐进、可持续发展,容不得急功近利。

2.2 开发人文信息资源

在网络技术飞速发展的今天,利用现代信息技术,开发人文信息资源,对于丰富理工科院校图书馆人文信息资源馆藏,满足读者对人文信息资源的需求,开展人文教育服务具有重要意义。为此,一方面图书馆工作人员应主动了解本校人文学科发展概况,围绕用户需求,结合国内外人文科学发展状况,将各种人文信息资源进行整合加工,获得具有二、三次文献特征的知识产品;另一方面,要按用户特定需要,利用图书馆文献资源的优势,对人文社科馆藏文献进行有效组织,并建立各种目录、索引、文摘及具有良好检索功能的专题数据库。最后,图书馆还应利用丰富的网络资源,建立人文信息资源的专题网址库,作为本馆馆藏的补充,为读者提供更多的获取人文信息的渠道。

2.3 实现资源共建共享

由于办学资本有限,任何一个图书馆的文献信息资源都无法满足所有读者对信息的需求。因此,理工科院校利用自己理工科文献资源优势与其它人文

基础坚实的综合性图书馆进行广泛协作,通过馆际互借或全文传递等方式,实现人文信息资源的共建共享,使读者能轻松自由地获得其它图书馆的人文信息资源,弥补本馆人文资源的不足,缓解由经费紧张人文资源建设发展缓慢所带来的问题,也是目前许多高校促进文献资源发展的有效途径之一。

2.4 提高图书馆工作人员的业务水平和综合素质

为保证人文社科馆藏质量及信息资源的开发利用,提高采访工作人员和人文信息资源开发人员的业务水平和专业素质显得尤为必要。因此,为了提高他们的综合素质,图书馆应积极制定人才培养计划,鼓励他们通过各种途径提高知识层次和业务水平,以适应网络环境下文献信息资源建设的需要。

参考文献

- 1 王雅琴. 网络环境下高校图书馆文献资源建设的思考. 现代情报, 2007(12)
- 2 瞿学惠. 图书馆与人文资源开发研究. 河南图书馆学刊, 2006(1)
- 3 刘华等. 图书馆信息资源与服务现状调查分析. 图书馆情报工作, 2007(12)
- 4 董正宇. 论地方合并综合性大学人文学科发展策略. 当代教育论坛, 2004(1)
- 5 万利. 论 21 世纪图书馆员素质的培养和提高. 情报探索, 2005(3)

(责任编辑:黄建)

引,需要事先分割单词进行词法分析,建立倒排文件的方法(称为“词典法”)。前者是从字这一观点出发确定索引项,避免了复杂且昂贵的语义索引过程,是将文本内含有的所有的字符作为索引项建立倒排文件。这种基于字的索引由处理 N 个字单位的索引,称为 N 元索引,特别是由一个字($N=1$)的索引项组成的索引称为单字索引, N 元索引的主要特点是不需要词表维护成本,实现简单,缺点是索引效率低。

2 中文倒排文件的检索

中文倒排文件的检索根据倒排表的内容,可以分成两种方法,一种是基于字的 N 元字索引检索,另一种是基于词的索引检索。

2.1 采用 N 元字索引建立的汉语倒排文件检索

采用 N 元字索引的检索方法分两种情况来探讨,一种是单字索引,另一种是 N 元字索引。

2.1.1 单字索引的倒排表的检索

由于中文文献中词与词之间没有任何区分标志,因此,中文文献的自动标引,首先要解决中文词的自动切分,但汉语自动分词存在着许多困难。由于英文检索的最小单位是英文单词,词与词之间有分隔号,一些学者从英文检索中得到启示,从而提出了单汉字索引,单汉字是构成文献的最小单位,每个汉字占2个字节,有固定的长度,计算机处理方便。

采用单字索引的倒排表,具有能查询任意的字符串的优点。在单汉字索引中,将检索对象文本分割成每一个汉字为单位的索引项。例:查询对象文本“全文检索技术”,按单汉字索引需要分割为“全”、“文”、“检”、“索”、“技”、“术”6个索引项,将查询词字符串分解成单个汉字,在索引表中进行位置信息匹配,从而获得检索结果。为了减少汉字的匹配次数,可以将索引项按位置信息量的个数从小到大的顺序排列。

单字索引具有以下特点:①不需要词法分析,便于计算机自动抽取,节省了人工标引的大量劳动,而且标引客观一致;②由于单字组配功能灵活,有利于“字面成族”;③系统维护简单;④字与字之间的匹配运算量大,检索效率低;⑤无检索意义和分辨率低的字占有很大比例。

随着查询表达式的长度增长,位置信息的匹配次数会大幅度增长,所以,存在检索时间效率下降的问题。为了解决这个问题,就出现以2个汉字为单位形成二元字索引和以3个汉字为单位的三元字索引。

2.1.2 N 元字索引倒排表的检索

在二元字($N=2$)索引中,将检索对象文本分割成每2个汉字为单位的索引项。例:查询对象文本“全文检索技术”,采用二元字索引,以字符串的开始每2个字符为分割单位,每次向右移动一个字符,形成了“全文”、“文检”、“检索”、“索技”、“技术”和“术”

几个索引项,这里每次向右移动一个字符,生成的索引项,是为了能够检索任意部分的字符串,在字符串的末尾特意给出一个字符“术”的索引项,是为了用“术”检索时也能匹配成功,这种用小于 N 的字符串形成的索引项叫末尾处理。在三元字($N=3$)索引时,索引项可以分割成“全文检”、“文检索”、“检索技”、“索技术”、“技术”和“术”。下面讨论二元字索引的情况。

对于“与检索技术”,二元字索引的倒排文件如下:

序号	索引项	位置信息
1	与检	18
2	检索	10,20
3	索技	22
4	技术	24
5	术	26

采用二元字索引法,根据查询表达式字符的个数,可以采用奇数和偶数个数处理方法。

当查询表达式的字符个数为偶数时,从表达式的开始位置,每2个字符为分割单位,形成查询词,但每个查询词字符是不重复的。例如“检索技术”在双元组分割为“检索”和“技术”

如果查询字符串的字符个数为奇数时,也是以2个字符为单位分割,但总会在某个位置上会出现字符重复。一般,重复字符的位置,按位置信息匹配次数最小来确定。例如:查询“与检索技术”时,查询单位的组合可以是以下2种:

模式A:“与检”,“检索”,“技术”

模式B:“与检”,“索技”,“技术”

与他们的位置信息比较,模式A有4个,模式B有3个,所以采用模式B。

从上述情况分析不难得出:采用二元索引和采用单元组索引比较而言,二元汉字索引减少了查询单词数,不仅仅从倒排文件中读入的位置信息减少了,与位置信息的匹配次数也减少了,所以在检索时间效率上得到了提高。

2.1.3 N 元字索引倒排表的检索分析

二元字索引,是将 N 固定为2的 N 元字索引,那么最佳的 N 值如何来确定,以下讨论 N 值如何影响空间效率和时间效率。

首先,从空间效率来看,如果 N 值越大,位置信息越分散,各个位置索引项的信息量越少,但由于各个索引项的模式增加,结果索引的量将增加,即从空间效率来说, N 值过大并不好。

其次,从时间效率来看, N 值小,查询词会增加,位置信息的匹配次数也增加,因此影响了时间效率;反之, N 值越大,查询词越少,越能够缩短检索时间。但是, N 值越大,就要频繁地查找比 N 小的字符串

(设 M 个字符)的查询表达式,这样,对于 M 比 N 小的查询表达式,需要从倒排文件中查找所有的与查询表达式前缀 M 个字符一致的索引项,检索时间就会增长许多。再来分析索引的建立时间,由于 N 值越大,索引量就越大,建立时间也必然增加,即从检索效率看, N 值过小或过大都不好。

N 值一般取 2 或 3 比较好。在实际运用中,往往采用几种索引单位。

2.2 词典分词法

以词为索引项的技术重点是词的切分问题。根据一定的原则和方法对文章进行自动“切词”,然后按词建库,对用户的检索结果按词汇匹配来进行查询,这种处理方法具有较高的查询命中率,但对“切词”技术的要求很高,要求配备词典库。目前采用的分词方法主要有基于神经网络和专家系统的算法,正向、逆向最大匹配法,逐词遍历法,最佳匹配法,词频统计法,此外还有穷多层次列举法、二次扫描法、基于期望的分词法、双向扫描法、邻接约束法、邻接知识约束法、最少分词词频选择法等。但归纳起来不外乎两类:第一类是在生成关键词时将语法、句法、语义结合起来,模仿人类的阅读过程,但有时语法、句法、语义连开发人员都不是很清楚,故一般情况下不采用。第二类由词典匹配法和基于频度方法组成,这些方法比起上一种更具体、更实用。目前常用的是最大匹配法(又称 MM 法)。

MM 法是一种得到广泛应用的分词方法,它在分词过程中除了依靠一个分词词典以外不再拥有其他词法、句法和语义知识。MM 法的基本思想如下:假设词典中最长的词由 Max 个字组成,则每次从句子开始位置截取一个长度为 Max 的字符串,然后与词典中的词依次匹配,如果匹配成功,就把这个字串作为一个词从句子切分出去,然后再从句子余下部分的起始位置截取另一个 Max 个字符串,重复上述过程,直至句子被切分完为止。如果在词典中找不到与之匹配的词,就从该字符串尾部删去一个字(一个汉字占两位),用 $Max-2$ 字长的字符串到词典中去查找。若匹配成功,则同样把该字串作为一个词从句子中切分出去;若匹配失败,则从该字符串尾部再删去一个字,再用 $Max-4$ 的字串去词典中匹配,直至匹配成功。

但 MM 方法有 2 个不足:

一是将词典的最大词长 Max 作为每次取出待匹配字符串的长度(Max 是一个固定值),在分词匹配过程中,如果 Max 较大,会造成不必要的比较匹配,影响分词效率。例如许多分词系统习惯将 Max 的值取为 8 个汉字,则对于句子“我们是中华人民共和国的公民”,第一次匹配取出的待匹配字符串即为“我们是中华人民共和国”,很明显,该字符串需要匹配到“我

们”一词才可以完成此次任务。如果以后的每次匹配都取出长度为 8 的待匹配字符串,会产生不必要的匹配,造成时间的浪费,影响整个分词系统的分词速度。但是,如果 Max 取值太小,则无法识别出更长的词语,会影响分词的精度。例如,如果将 Max 设为 4,则对于句子中的“中华人民共和国”一词将无法正确识别。二是有时并没有体现最大匹配特点。仔细分析其原因,MM 方法都是从句子左边开始的 Max (词典的最长词)个字符范围内找最大的词,而且最大词必须包含最左边的字,而不是在整个句子中找最大词。为了真正体现“长词优先”原则,笔者以为应该每次尽可能找出句子中的最大词。

为了弥补这 2 个不足,我们可以做以下改进:

①在词索引表中增加一个临时数据项,代表每个首字的所有词条中长度最大的词条所含字符数 Len 。这样,在每次匹配查找时,首先可以根据首字信息,找到词索引表中最大词条所含字符数 Len ,将其作为此次最大匹配词长,然后再根据最大匹配法进行分词处理。

②改进算法,每次找出查询词的最大词。具体算法如下:假设字典中的最长词为 Max ,输入查询表达式的汉字个数为 N 。

从句子第 1 个字开始取长度为 MAX 的字串,在字典中寻求匹配;如果匹配不成功,就从句子第 2 个字开始截取一个长度为 MAX 的字串重复以上过程。如果还找不到,则依次从第 3,4,... $N-MAX$ 个字开始截取长度为 MAX 的字串进行匹配。如果在某一次匹配中匹配成功,就把这个字串作为一个词从句子中切分出去,把原句中位于这个字串左右两边的部分看作为两个新的句子,递归调用这一过程。如果所有的匹配都不成功,说明句子中没有长度为 MAX 的词,则开始寻找长度 $MAX-1$ 的词。重复这个过程直到整个句子被切分。

根据这种思路,笔者用 C++ 设计了一个程序,实现了这种设想。以下以“当五星红旗升起的时候”为例,介绍程序的应用。

```
#include<iostream.h>
#include<string.h>
char *cd[ ]={"五星红旗","升起","时候"};//词典数组
char s[ ]="当五星红旗升起的时候";//检索词数组
const int MAX=4; //词典中的最大汉字长度
char *strcut( char *s,int m ,int n ) //取子串函数
{
    static char substr[20];
    int i;
```

```

for (i=0;i<n;i=i+1)
    substr[i]=s[m+i];
substr[i]='\0';
return substr;
}
void search(char s[ ],int m,int n)
{
for(int j=MAX*2;j>=2;j=j-2) //一个汉字占两个
字节
{ int y=0;
for( int i=m;i<n;i=i+2)
{
for(int k=0;k<=2;k++)
if((! strcmp(strcut(s,i,j),cd[k])&&i+j<=n+1)||
j==2)
{ cout<<strcut(s,i,j)<<" "<<i<<endl;y++;
if(i>m) //如果匹配成功,截取余下字串,递归
调用,继续分词
search(s,m,i-1);
if(i+j<=n+1)
search(s,i+j,n);
break; }
if(y>0) break;}
if(y>0) break;}
}
void main()
{
search(s,0,19);
}

```

运行后,得到分词结果为“当/五星红旗/升起/的/时候”。

2.3 词典法和N元字索引法的比较

我们将词典法和N元字索引法作以下比较:

①检索速度:词典法不仅索引项数少,由于一个索引项具有的位置信息也少,位置信息的匹配次数得到限制;而N元字索引法由于一个查询词需要多个索引项,所以,和词典法相比,检索速度低。

②索引量:词典法因为经过词法分析后,不必去除关键词之外的认为无用词性位置信息,索引项数少,索引容量也少;N元字索引法基本上需要将文本内的所有字符建立索引,加上随着N值的增大,各个索引项的字符间出现重叠,索引量将更大。

③检索完备性:词典法对于文本进行词法分析时,若达不到100%的精确度,就有可能遗漏需要的位置信息,特别是,对于查询词进行词的切分如果失败,就不能正常检索;N元字索引法由于用文本内的所有的字符建立索引,索引项是一个字符一个字符

移动取N个字符,所以可以检索文本内任意的字符串。

④建立检索的时间:词典法为进行词法分析,需要有另外的分析词典,存储过程中因为要进行词法分析,建立索引的时间花费就更多;N元字索引法建立索引时,由于可以不进行词法分析的预处理,索引建立的时间短,不过随着N值的增大,索引项数的增多,随之带来索引建立的时间也增多。

⑤同义词、相关词的处理:词典法可通过词典中的参照关系来标出,由于N元字索引法没有办法处理文本中的隐含概念,若文中未出现则无法标出。

⑥新词的处理:词典法需要对词典经常更新,将新出现的词加入词典中,否则不能处理;N元字索引法则没有新词的概念,都是通过字与字的组配关系得到。

⑦适用范围:词典法如果要将所有学科领域的词包含在词典中,词典会很大,难以维护;N元字索引法可以适用各个学科领域。

⑧误检情况:词典法虽然有词典保证,但是无论采用哪种分词方法都很容易出现歧义现象,从而导致出现漏检或错检的情况;N元字索引法由于单个汉字经常不具有独立的含义,存在错误匹配现象,误检率相对较高,如检索“华人”时,会将包括“中华人民共和国”的文档检索出来。

参考文献

- 靖培栋,宋雯斐. 基于混合索引的中文全文检索系统研究. 中国图书馆学报,2008(1)
- 何莘,王琬芩. 自然语言检索中的中文分词技术研究进展及应用. 情报科学,2008(5)
- 吴凡. 信息检索中的中文分词问题研究. 情报杂志,2008(7)
- 钱爱兵. 全文检索算法设计及全文检索系统的优化. 现代图书情报技术,2003(2)
- 熊回香. 全文检索中的汉语自动分词及其歧义处理. 中国图书馆学报,2005(2)
- 祈延莉,赵丹群. 信息检索概论. 北京:北京大学出版社,2006
- 郭琦娟,陈通照. 一种动态更新索引结构的设计与实现. 计算机系统应用,2006(12)
- 曹元大,贺海军等. 全文检索字索引技术的研究与实现. 计算机工程,2002(6)
- 靖培栋,宋雯斐. 中文全文检索系统截词检索的实现研究. 情报科学,2006(6)
- 陈玮,陈玉鹏,石晶等. 一种高效的全文检索索引技术. 计算机应用研究,2004(7)

(责任编辑:黄建)

收稿日期:2009-01-12

作者简介:杨安生(1964—),男,大学本科,讲师,研究方向:信息管理、算法设计,已发表论文13篇。