中文信息检索中词典机制分词算法的研究

宗 中
(江苏省邮电规划设计院有限公司,江苏 南京 210006)

摘 要:中文自动分词是实现搜索引擎信息检索的基础,分词词典是汉语自动分词系统的一个重要组成部分,词典的加载和查询速度直接影响到分词系统的速度。文中在研究传统词典机制的基础上,分析了基于双字哈希词典机制对词条除首次字外剩余词的不足,给出了一种改进的双字哈希的词典机制。最后,文中对改进算法从准确率、分全率和分词速度等方面进行了测试,结果表明,改进后的分词算法在不提升已有典型词典机制维护复杂度的情况下,提高了词条匹配的查询速度和效率。

关键词:信息检索;中文分词;数据结构;哈希
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)04-0118-04
doi:10.3969/j.issn.1673-629X.2014.04.030

Study of Segmentation Algorithm of Dictionary Mechanism Orienting Chinese Information Retrieval

ZONG Zhong
(Jiangsu Posts & Telecommunications Planning and Designing Institute Co. ,Ltd,
Nanjing 210006,China)

Abstract:Chinese automatic segmentation is the base of the information retrieval search engine. Word dictionary is an important part of Chinese word segmentation system. The loading and querying efficiency is a key impact fact of the word segmentation system. Based on the study of the traditional dictionary mechanism, analyze the weak point of the double word hash dictionary, and propose a modified double hash dictionary. At last test the method from the accurate, full-rate, word speed, etc. With the result of the test, this improved hash mechanism enhances the entry speed and efficiency of matching queries, without completing the maintenance complexity of the traditional dictionary.

Key words:information retrieval;Chinese word segmentation;data structures;hash

0 引 言

信息检索是将信息按一定的方式组织和存储起来,并根据用户的信息需求查找所需信息的过程和技术。对中文文本信息检索来说,由于中文文本是按句连写的,每个句子中的词没有空格,需要用分词来处理[1]。因而在中文文本信息检索处理中,对歧义切分字段的处理能力,严重影响到中文自动分词系统的精度[2],词的正确切分是进行中文文本信息检索处理的必要条件。因而,分词能有效地提高文本检索的效率[3]。

基于词典的分词算法作为当前分词技术的主流,由于分词系统所需要的各类信息都要从词典中获取,所以其精确度依赖于词典的完全性和歧义的有效消除[4],速度则取决于所设计的加载词典的数据结构和相应的切分算法[5]。因而,分词词典是基于词典机制的汉语自动分词系统的重要组成部分,其性能的优劣直接影响到分词系统的速度和效率,建立高效而快速的分词词典机制势在必行。

1 传统的词典机制

分词词典是汉语自动分词系统的一个重要组成部分。词典的加载和查询速度直接影响到分词系统的速度[6],对于基于词典的分词算法,影响其精度的因素有[7]:分词词典中词库的选择和词条的数量;机器可读

收稿日期:2013-06-03 **修回日期**:2013-09-15 **网络出版时间**:2014-01-28
基金项目:江苏省自然科学基金项目(BK2009425)
作者简介:宗 中(1984-),男,江苏南京人,硕士,法国工程师,研究方向为计算机应用、信息化咨询、项目管理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140128.1144.030.html

词典与待切分文本中词汇的匹配关系;未登录词;分词方法等。设计的词典机制和词条的完备率对分词结果的准确性将产生重要影响,词典的性能在一定程度上决定着整个系统的性能。

根据构造词典的数据结构的不同,可将分词词典机制分为四种[8]:整词二分的词典机制、TRIE 索引树的词典机制、基于逐字二分法的词典机制和双字哈希的词典机制。

整词二分的词典机制是一种广为使用的机制。该机制的词典结构分为首字哈希表、词索引表、词典正文等三级。整词二分的词典机制的分词算法的优点是数据结构简单、占用空间小,构建及维护也简单易行,但由于采用全词匹配的查询过程,效率较低。

TRIE 索引树的词典机制是一种以树的多重链表形式表示的键树[9]。汉字接近 7 000 个,面向中文的 TRIE 索引树的节点应允许指针个数变化。TRIE 索引树的优点是在对被切分语句的一次扫描过程中,不需预知待查询词的长度,沿着树链逐字匹配即可[10],避免了整词二分的词典机制中不必要的多次试探性查询;缺点是它的构造和维护比较复杂,是单词树枝(一条树枝仅代表一个词),浪费一定的空间。

基于逐字二分法的词典机制是对前两种词典机制的改进方案。逐字二分与整词二分的词典结构完全一样,只是查询过程有所区别,但由于采用的仍是整词二分的词典结构,效率的提高受到一定的限制。

双字哈希的词典机制主要结合了词典中的多字词条(3 字词以上)数量少[11-12],使用频度低的特点,对基于 TRIE 索引树的词典机制做出了改进,把 TRIE 索引树的深度限制为 2,词的剩余字串则按序组成类似“整词二分”的词典正文。其结构分别由首字哈希索引表,次字哈希索引表,剩余字串组三层组成。

基于双字哈希的查询机制相当于使 2 字词以下的短词用 TRIE 索引树机制实现,3 字词以上的长词的剩余部分用线性表组织,从而避免了深度搜索,一定程度上提高了查询性能,提高了分词的速度,对三字词以上的长词的剩余字串组,用的是逐词匹配,仍有改进的空间。

2　改进的双字哈希的词典机制——mDHash

为了最大限度地提高匹配的时间效率并兼顾空间利用率,文中提出了 mDHash—改进的双字哈希的词典机制。改进的词典机制充分吸纳了“整词二分”和“TRIE 索引树”二者的优点,仅对词条的前两个字顺次建立 Hash 索引,构成深度仅为 2 的 TRIE 子树,实现对 2 字词以下的短词的快速查找,在次字哈希表中记录下每个词条首次字在词典中开始和结束的位置,这样可以在这个空间中“二分查找”三字词以上的长词的剩余字符串。这种改进比原词典机制中“顺序查找”的匹配效率、速度优势是非常明显的。

2.1　词典的结构组织

对于基于词典的分词算法而言,文本的切词效果主要依赖于系统的基本词典,当基本词典的结构组织不合理或词典中存储的词条信息不满足信息查询的需要时,将影响到对信息的查全、查准率。

在文中根据 ICTCLAS 词典的特点,对词典结构进行了重新设计,将以往单一的分词词典分解成通用词典和特征词典两个部分组成的词典集。通用词典由 ICTCLAS 词典的 coreDict(核心词典)组成,里面收录有平时使用的所有词汇。特征词典由 ICTCLAS 词典的 nr、ns 等表示人名、地名的专有词典组成,主要用于处理未登录词时识别常见的中文人名和地名。核心词典的逻辑结构如图 1 所示。

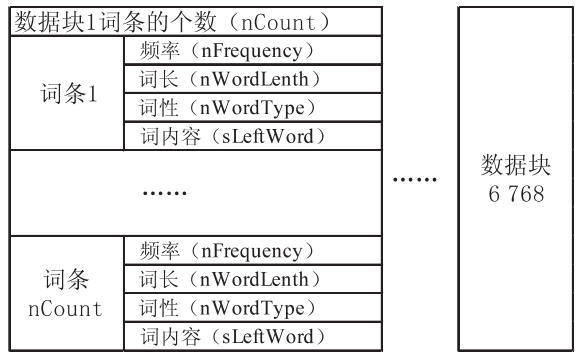

图 1　核心词典的逻辑结构

在图 1 中含有 6 768 个数据块,分别由 GB2312 编码标准收录的常用汉字开始的词条有序排列组成。每个数据块的词条都是同一汉字开始的,并且依次按照后面字的内码顺序排列。如在匹配以字“中”开始的词组时,只需要在词典中所有以字“中”开始结束的数据中查询,无需遍历整个词典,这样可以尽量缩小每一次的查找范围,提高查询速度。

从图 1 可见,每个数据块由该区域词条的个数和每个词条的相关属性组成,每个词条的属性有:词长、频率、词性、词内容(除首字),见图 2。

词长 (nWordLenth)	频率 (nFrequency)	词性 (nWordType)	词内容 (sLeftWord)

图 2　词条的字段属性

从图 1 逻辑结构中可以看出,分词系统采用中科院 ICTCLAS 词典的优势有:$O(1)$时间内得到每个字开始词的个数,读取词条时无须统计计数;方便取得每个字开始最长词的最长长度,适于用正向减字相对最大匹配法作分词算法;每个字开始的词都临近在一起,快速地节省了词典的加载时间和切分算法中二分查找的时间。

2.2 词典的结构设计

在基于词典的词条模式匹配分词方法中,词条的组织和词典结构的设计是否合理是影响文本分词效率和信息查全、查准率的一个重要因素。其次词典应考虑到具体分词算法的需要,因此,设计一种高效的词典内存数据结构,为分词提供准确有效的切分保证。

根据权威统计,双字词语所占比例很大,使用频率也很高,多(四字以上)字词语占比例很小,使用频率较低。汉语各字词条的词频统计如表 1 所示。从表 1 中可以看到,两字词的词频占了一半以上,而五字以上的词的出现频率就非常小,使用频率低。

表 1 各类词条出现的概率统计

词条中的字数/字	1	2	3	4	5	6	7
词频/%	6.98	50.034	20.010	16.477	3.891	1.857	0.744

考虑到中文字的编码体系和中文词的分布以及哈希索引结构的词典查询速度快的优点,文中设计了一个基于双字哈希的三级索引词典结构。对词的前两个字依次建立哈希索引,构成双层哈希索引结构,词的剩余字串按内码大小的顺序组成“词典正文”(剩余字串组),并且在次字哈希表中记录下每个词条首次字在词典中开始和结束的位置,这样就可以对 2 字词以下的短词直接哈希查找,对三字词以上的长词的剩余字符串运用“二分查找”, 最大限度地提高匹配的时间效率。

文中改进的词典的数据结构由首字哈希表、次字哈希表、词索引表、词典正文四部分组成。

(1)首字哈希表。

首字哈希表由四部分组成,如图 3 所示:nCount 表示以该首字开始的词条的个数,isSingleWord 表示该字是否是单字词,nMaxLength 表示以该字开始的最长词条的长度,secPointer 表示指向次字哈希表的指针。

nCount	isSingleWord	nMaxLength	secPointer

图 3 首字哈希表的字段属性

汉字在计算机内部是以内码的形式存放的,国内的常用汉字编码标准是 GB2312 编码。GB2312 标准共收录 6 768 个常用汉字,采用区位码的形式来处理。对 6 768 个常用汉字进行数值化(函数映射),映射到首字哈希表中去,映射关系(首字哈希函数)为:

$$\text{offset} = (c_1 - 176) * 94 + (c_2 - 161) \quad (1)$$

式中,offset 为首字在首字哈希表中的索引值;c_1 和 c_2 分别代表首字的区码和位码。

(2)次字哈希表。

文中采用除留余数法的散列函数处理次字哈希索引。

次字哈希函数为:

$$\text{offset} = (c_1 - 176) * 94 + (c_2 - 161) \bmod \text{Length} \quad (2)$$

式中,offset 为次字的索引值; c_1 和 c_2 分别代表次字的区码和位码;Length 是求模系统,文中对其长度进行统一分配。

文中采用链地址来解决地址冲突问题,对相同的关键字放到同一个线性链表中去。

线性链表的节点由两部分组成:char 表示该汉字,nStart 表示次字开头的词条在词典开始出现的位置,nEnd 表示次字开头的词条在词典最后出现的位置,见图 4。指针域 next 指向产生冲突的词次字结构组成的下一节点,如图 5 所示。

char	nStart	nEnd	next

图 4 线性链表节点结构

first	last	count

图 5 线性链表结构

(3)词索引表。

词的索引表由四部分组成,如图 6 所示,lefPointer 指向剩余字,nWordLength 表示剩余字的长度,nWordType 表示该词的类型,nFrequency 表示该词使用的频率。

lefPointer	nWordLength	nWordType	nFrequency

图 6 词索引表的字段属性

(4)词典正文。

词典正文是由词中除首次字外剩余部分组成的按内码大小排序的有序表。

程序设计中主要用到的数据结构如下:

```
public class FirstHashTable {      // 记录首字哈希表的结构
    public int      nCount;   //以首字开始词条的个数
    public SecondHashTable[ ]  secondHashTables;  //指向次字哈希表
    public WordLinkList[ ]   list;  //储存次字哈希的链表
    public boolis SingleWord; //表示该字是否是单字词
    public static int nMaxLength;  //该首字开始的最长词条的长度
}
public class SecondHashTable {   //记录次字哈希表的结构
    public string   sLeftWord   //储存除首字外剩余字
    public int nWordLenth;   //除首字外词条剩余字的字节数
    public int nWordType;  //词条的词性
    public int nFrequency;  //词条的频率
}
```

3 测试与性能分析

文中的测试语料采用比较常用的北京大学计算语

言研究所加工的 1998 年 1 月份《人民日报》,该语料是纯文本文件,共计 2 305 896 字,内容涵盖政治、经济、体育、娱乐等方面的题材。文中从语料中选取了 25 篇文本进行综合测试,按国际、社会、政治、财经、体育分为五类,每类 5 篇文本。将所选文本在同一个分词系统分别采用不同的词典机制进行测试,测试结果数据(取其 5 篇的平均结果)如表 2 所示。

表 2　不同类型文本的测试结果

文本类别		国际	社会	政治	财经	体育
DHash	准确率/%	98.21	98.40	98.39	98.01	98.25
	分全率/%	98.44	98.51	98.58	98.42	98.22
	分词速度	413	424	420	419	425
mDHash	准确率/%	98.22	98.38	98.47	98.03	98.32
	分全率/%	98.49	98.54	98.63	98.48	98.14
	分词速度	549	538	642	551	543

从表中数据可看出采用 mDHash 词典机制,其分词速度有着明显的提高。

4　结束语

文中在分析基于双字哈希词典机制对词条除首次字外剩余词的不足的基础上,给出了一种改进的分词词典机制。将词典中的每个词条加载到基于 Hash 的三级索引数据结构。在加载的过程中记录每个首次字开始词条在词典中开始和结束的位置,在查找词条过程中采用“首次字双层哈希索引+剩余词二分查找”的算法,在不提升已有典型词典机制维护复杂度的情况下,提高了中文分词的速度。

参考文献:

[1] 郑晓刚,韩立新,白书奎,等. 一种组合型中文分词方法[J]. 计算机应用与软件,2012,29(7):26-28.

[2] Liang Xiongyou, Xue Yongsheng. Algorithm of solving interlink overlapping ambiguity and combinatorial ambiguity and compound ambiguity in Chinese word segmentation[J]. Journal of computational information systems, 2007, 3(3): 1189-1200.

[3] Li Sheng, Zhao Tiejun. Chinese information processing and its prospects[J]. Journal of computer science and technology, 2006, 21(5): 838-846.

[4] Tsai R T H, Dai Hongjie, Hung Hsieh-Chuan, et al. Chinese word segmentation with minimal linguistic knowledge[C]// Proceedings of the 2006 IEEE international conference on information reuse and integration. [s. l.]: [s. n.], 2006: 274-279.

[5] Zhang Ruiqiang, Keij Y, Eiichiro S. Chinese word segmentation and statistical machine translation[J]. ACM transactions on speech and language processing, 2008, 5(2): 4-9.

[6] Jiang Bin, Yang Chao, Zhao Huan. A kind of dictionary mechanism based on the two-word-bitmap for Chinese word segmentation[J]. Journal of Hunan University (Natural Sciences), 2006, 33(1): 121-123.

[7] Zhang Liyi, Li Yazi, Meng Jian. Design of Chinese word segmentation system based on improved Chinese converse dictionary and reverse maximum matching algorithm[C]//Proceedings of 2006 international workshops on web information systems engineering. [s. l.]: [s. n.], 2006: 171-181.

[8] 曹勇刚,曹羽中,金茂忠,等. 面向信息检索的自适应中文分词系统[J]. 软件学报,2006,17(3):356-363.

[9] 姜　维,王晓龙,关　毅,等. 基于多知识源的中文词法分析系统[J]. 计算机学报,2007,30(1):137-145.

[10] Qiao Wei, Sun Maosong, Menzel W. Statistical properties of overlapping ambiguities in Chinese word segmentation and a strategy for their disambiguation[C]//Proceedings of the 11th international conference on text, speech and dialogue. [s. l.]: [s. n.], 2008: 177-186.

[11] 李庆虎,陈玉健,孙家广. 一种中文分词词典新机制—双字哈希机制[J]. 中文信息学报,2003,17(4):13-18.

[12] Qin Ying, Zhang Suxiang, Wang Xiaojie. Combining multi-knowledge for Chinese word segmentation disambiguation[C]//Proceedings of the sixth international conference on intelligent systems design and applications. [s. l.]: [s. n.], 2006: 551-556.

(上接第 117 页)

林大学出版社,2011.

[9] 王利香. 高等学校毕业生质量的粗软集评价方法[J]. 潍坊学院学报,2012,12(2):40-41.

[10] Meng Dan, Qin Keyun. Soft rough fuzzy sets and soft fuzzy rough sets[J]. Computers & mathematics with applications, 2011, 62(12): 4635-4645.

[11] Leoreanu-Fotea V, Jun Y B. Soft sets and soft rough sets[J]. Information sciences, 2011, 181(6): 1125-1137.

[12] Johnston J, Eloff J H P, Labuschagne L. Security and human computer interfaces[J]. Computers & security, 2003, 22(8): 675-684.

[13] Special column of Shanwenbang. Talk about software human-computer interface[EB/OL]. 2012-12-15. http://blog.csdn.net/shanwenbang/article/details/7109026.

中文信息检索中词典机制分词算法的研究

作者： 宗中， ZONG Zhong
作者单位： 江苏省邮电规划设计院有限公司,江苏 南京,210006
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(4)