使用二级索引的中文分词词典

张庆扬,柴 胜
ZHANG Qing-yang,CHAI Sheng

吉林大学 计算机科学与技术系,长春 130062
Department of Computer Science and Technology,Jilin University,Changchun 130062,China
E-mail:zhqyzhqychn@sina.com

ZHANG Qing-yang,CHAI Sheng.Chinese word segmentation dictionary using two-level index.Computer Engineering and Applications,2009,45(19):139-141.

Abstract: As the basis of Chinese information processing,Chinese word segmentation plays a very important role in the fields of searching engine,automatic and so on.Chinese word dictionary is the basis of mechanic segmentation algorithm,it tells the algorithm what is a Chinese word.Because the algorithm needs the content of dictionary in order to match the string in the text,the storage structure of the dictionary will decide the method of the algorithm and its performance.Through making research into the existed theory and refinement,this paper adds multi-level index for the dictionary,and based on this formulates a new mechanism of Chinese word segmentation dictionary—dictionary based on two-level index.On the basis of this new theory,this paper also improves the positive matching method,reduces the complexity of matching process,moreover,elevates the speed of the segmentation.
Key words: Chinese word segmentation;two-level index;positive maximum matching

摘 要:中文分词是中文信息处理的基础,在诸如搜索引擎,自动翻译等多个领域都有着非常重要的地位。中文分词词典是中文机械式分词算法的基础,它将告诉算法什么是词,由于在算法执行过程中需要反复利用分词词典的内容进行字符串匹配,所以中文分词词典的存储结构从很大程度上决定将采用什么匹配算法以及匹配算法的好坏。在研究现存分词词典及匹配算法的基础上,吸取前人的经验经过改进,为词典加上了多级索引,并由此提出了一种新的中文分词词典存储机制——基于二级索引的中文分词词典,并在该词典的基础上提出了基于正向匹配的改进型匹配算法,大大降低了匹配过程的时间复杂度。从而提高了整个中文分词算法的分词速度。
关键词:中文分词;二级索引;正向最大匹配
DOI:10.3778/j.issn.1002-8331.2009.19.043 **文章编号:**1002-8331(2009)19-0139-03 **文献标识码:**A **中图分类号:**TP391.1

中文文本在书面表达或计算机内部表示时,字与字、词与词之间没有明显的切分标志。对于一句话,人可以通过自己的知识来得到哪些是词,但如何让计算机也能做到这一点,就是中文分词算法的目的。中文分词是中文信息处理的基础,在诸多领域起着至关重要的作用,例如,机器翻译、文本检索、搜索引擎、自动文摘等很多应用都需要以中文分词结果作为基础。一个中文分词算法的好坏,会对其后续的应用产生极大的影响。

现有的分词算法,基本上可以分为三种:机械式分词、理解式分词、统计式分词。在这三种算法中,机械式分词的准确率最高,对于机械式分词,需要一部词典作为后台依据,简单的说是通过将文档中的汉字串与字典中的词相匹配来完成词的切分。 这是一种基于字符串匹配的分词方法, 它是按照一定的策略将待分析的汉字串与一个(充分大的)机器词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功(识别出一个词)。按照对文本串的扫描方向的不同,机械分词方法可以分为正向匹配和逆向匹配;按照不同长度优先匹配的情况,可以分为最大(最长)匹配(Maximum Matching Method,简称 MM法)和最小(最短)匹配。

正向最大匹配算法的原则是切分出来的词越长越好。因为词越长,包含的信息量就越大。机械式分词算法之所以成为“机械”,是因为这种算法没有考虑任何语法及语义问题,而是单纯的,机械式的将文本与词典中的词相匹配。由此可以看出词典所含词汇量的大小将决定分词精度的好坏。由于很多机构已经构造出了极为丰富的词典所以在这点上基本上已没有改进的余地。由于机械式分词是基于字符串匹配的。所以词典的存储数据结构和匹配算法的好坏将会直接影响该算法的时间和空间复杂度。

1 传统词典机制

想要提高机械式分词算法的执行效率,就要降低匹配算法的时间复杂度,而一个匹配算法的时间复杂度,从很大程度上取决于匹配数据的存储结构[1]。在机械式分词算法中,最重要的就是分词词典的存储结构。下面介绍几种传统的分词词典机制[2]:

(1)基于整词二分的分词词典机制

作者简介:张庆扬(1986-),男,主要研究搜索引擎及中文信息处理;柴胜(1976-),男,博士生,主要研究方向为软件工程。
收稿日期:2008-04-15 **修回日期:**2008-07-23

该词典的存储机制把词典分为词典正文、词索引表、首字Hash表等三级。词典正文是以词为单位的线性表,词索引表是指向词典正文中每个词的指针表。通过首字Hash表的散列函数和词索引表很容易确定指定词在词典正文中的可能位置范围,进而在词典正文中通过整词二分进行定位。

(2)基于TRIE索引树的分词词典机制

TRIE索引树是一种以多重链表形式表示的键树。基于TRIE索引树的分词词典机制有首字散列表和TRIE索引树节点两部分组成。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;缺点是它的构造和维护比较复杂,而且都是单词树枝(一条树枝仅代表一个词),浪费了一定的空间。

(3)基于逐字二分的分词词典机制

这种词典机制是前两种机制的一种改进方案。逐字二分与整词二分的词典机构完全一样,只是查询过程有所区别:逐字二分吸收了TRIE索引树的查询优势,即采用的是"逐字匹配",而不是整词二分的"全词匹配",这就一定程度地提高了匹配的效率。但由于采用的仍是整词二分的词典结构,使效率的提高面临很大的局限。

2 一种新的词典机制

在上一章介绍了几种传统的分词词典机制,通过介绍,可以了解到已有的词典机制都存在很明显的优缺点。对于基于整词二分的词典机制,词典结构简单,维护容易,但由于采用的是全词匹配,所以速度较慢。而对于TRIE树,采用逐字匹配,速度较快,但词典的存储数据结构复杂,较难维护。而对于第三种相当于两种算法的折中,对两种算法的优点都不能全部发挥。

为了最大程度的提高算法的匹配效率和词典的可维护性。设计了一种新的分词词典存储结构—基于二级索引的分词词典。

在检测以某一个汉字开始的汉字串是否是一个词的时候,算法的工作是检测在词典中是否有以该汉字开始的词条与此汉字串相同,如果有则就是词。算法的关键在与怎么才能迅速实现判定。为了迅速定位以某一个汉字开始的所有词条首先为首汉字建立索引[3]。这里利用汉字的机内码和区位码相互关系,根据转换公式:

内码高位=区码+A0H(H表示十六进制)

内码低位=位码+A0H

很容易将一个汉字的机内码转换为区位码,而在将区位码转换为十进制表示。例如"中"字的最终转换结果为十进制"5448",由于汉字的区位码最大为"8794",所以可以用一个固定的结构体数组来存储这第一级索引。例如该数组的第5 448项(第0项不用),里面存储的就是所有以"中"开头的词条的存储位置的头指针,当然如果某一个汉字不能组词,则该指针为空。这样已知一个汉字串的首字就可以在复杂度O(1)的时间里定位到以该汉字开始的词条。这是词典的第一级索引。

经过一级索引可以将时间复杂度大大地降低,但是文献[4]对汉语中的词条分布进行了统计,具体情况见表1。

表1 词条字数分布统计

词条字数	2	3	4	5	6	7
词条个数	33 527	3 693	3 622	83	26	3

由统计结果可以看出汉语中双字词较多,而三字、四字次之,其他词较少。经过为词典建立的一级索引后,由于二字词较多,例如以"中"字开始的二字词就有100多个,这样如果顺序匹配的话,仍然会花费很多的时间,所以在一级索引的基础上,在为分词词典建立一个二级索引。

由于词的第二个汉字分布的不确定性,利用第一级索引的方法是不现实的,所以建立二级索引是本文采用除留余数的Hash方法来记录不同的词条存储位置。Hash函数为:

$Offset=((ch1-0xB0)\times 94+(ch2-0xA1))\%length$

其中$ch1$,$ch2$为汉字机内码的高、低字节。$length$代表以某一首汉字开始的词条的个数。

既然是散列函数,就有可能出现冲突,在解决冲突方面采用了链地址法,即如果有两个以上的词有相同的Hash值,则分别将其链在一起。

当然,利用上述建立索引的方法可以为词典建立三级[5],甚至四级索引,但是考虑到汉语中三字、四字等词比较少,而且以相同的两个汉字开始的多字(三字及以上)词较少[6],所以没有必要建立更多级的索引。

经过二级索引,很容易就能够识别二字词,由于在汉语中二字词较多,所以这大大提高了匹配速度。但在匹配过程中不可避免会遇到三字词或更长的词,而且也说过没有必要再建立更多的索引,为了进一步提高匹配的速度,对除去前两个汉字的剩下串按升序排列,这种存储结构的优势在匹配时将会见到。

为了节省存储空间经过两次索引之后,在词的存储过程中,只存储剩下的词,例如在下面给出的例子中"中心汇率"这个词时,其实在正文词典里只存储了"汇率"这个词,而其首字和次字可以在匹配过程中自动得到,无需存储,这从很大程度上节省了内存空间(为了提高匹配速度,词典常驻内存)。

介绍完词典的存储机制,下面简要介绍一下基于该分词词典的正向最大匹配算法。例如匹配"我不知道中心词是什么",现在从"中"字开始进行正向最大切分,

(1)根据首汉字"中"的区位码"5448"作为首字索引表的下表得到其索引结构体。

(2)由于结构体中的次字索引表基指针base不为空,说明可以继续匹配汉字串,则根据base找到以"中"开始的次字索引表。

(3)通过Hash("心"),求出次字为"心"的所有词的索引项在次字Hash表的偏移量offset,根据base+offset求出以"中心"开始的所有词的存储位置的头指针,当然在这一过程中,需要解决Hash值的冲突问题。在本例中不存在这一问题。

(4)在剩下的串中采取临近匹配算法[7],这就是要把剩余汉字按升序排列的原因。理由是,在找到某个字符串后,在其后增加一个字得到一新字串,如果新字串在词典中出现,那么新词一定在原字串的后面,且相隔不会太远。最终匹配出"中心词"。词典的存储机制及示例的匹配过程如图1所示。

3 实验结果与分析

根据以上叙述的理论,用VC++6.0实现一个中文分词系统。并在AMD2600+,512 M内存的情况下对本系统和没有改进前的系统做了比较,输入文本统一采用1 M网页,用正向最大匹配算法切分。实验数据如表2所示。

从表2可以看出,分词词典索引对于分词算法的好坏有着很大的影响。如果不加索引则每次都要匹配词典里所有的词,切分速度非常慢。本文所设计的词典为词典加了两级索引,并

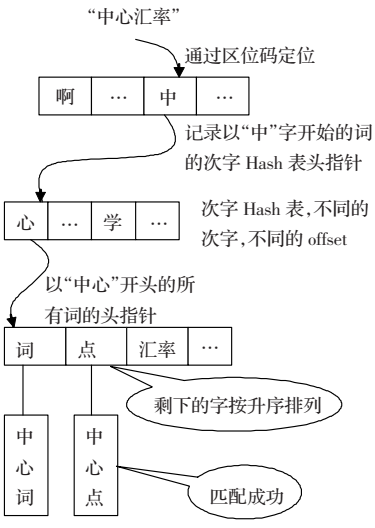

图 1　分词词典机制

表 2　系统测试结果

词典机制	整词二分	TRIE 树	二级索引
时间/ms	1 016	865	500

且为两级索引以后的汉字采用了临近匹配算法，最大程度上避免了全词匹配[8]。由于在汉语中二字词占大多数，而在本文所设计的词典中所有的二字词都采用了逐字匹配，这即节省了时间，也符合正向最大匹配算法的要求。

4　结束语

在已有分词词典结构的基础上经过改进，为词典加上多级索引，并相应地改进了匹配算法，大大降低了分词算法的时间复杂度。目前，有很多对于分词词典的改进方法，到底哪一种更有效并无定论，当然一个好的中文分词系统也不能只使用一种算法，而是要结合各种算法的优点才能处理不同的输入问题。

参考文献：

[1] Sproat R，Gale W，Shih C.A stochastic finite-state word-segmentation algorithm for Chinese[J].Computational Linguistics，1996，22(3)：377-404.

[2] 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.

[3] 翟凤文，赫枫龄，左万利.字典与统计相结合的中文分词算法[J].小型微型计算机系统，2006，27(9)：13-18.

[4] 李振星，徐泽平，唐卫清，等.全二分最大匹配快速分词算法[J].计算机工程与应用，2002，38(11)：106-109.

[5] 张科.多次 Hash 快速分词算法[J].计算机工程与设计，2007，28(7).

[6] 肖红，许少华，李欣.具有三级索引词库结构的中文分词方法研究[J].计算机引用研究，2006(8)：49-51.

[7] 陈桂林，王永成，韩客松，等.一种改进的快速分词算法[J].计算机研究与发展，2000，37(4).

[8] Teahan W J，Mcnab R，Wen Ying-ying，et al.A compression-based algorithm for Chinese word segmentation[J].Computational Linguistics，2000，26(3)：375-393.

(上接 138 页)

对实验结果采用客观和主管两个评价标准。增强后的信噪比(帧间平均 SNR)提升见表 1 所示，二维维纳滤波再高斯白噪声环境下对语音客观质量有巨大改善，信噪比在一维维纳滤波的基础上又有大幅提升。二者的 MOS 得分见表 2，分别由 13 位试听者按 5 分制打分后平均所得，5 分最高，0 分最低。从主观评价上可知，二维维纳滤波有较小的失真，听起来更易接受，MOS 评价结果显示，二维维纳滤波得分比一维维纳滤波平均提高 13.8%。

表 1　客观评价

客观评价(SNR/dB)	0	5	10	20
1D Wiener	21.4	25.7	28.8	33.6
2D Wiener	39.8	42.9	44.8	45.8

表 2　主观评价

主观评价(MOS)	0	5	10	20
1D Wiener	2.7	3.0	3.2	3.3
2D Wiener	3.0	3.4	3.6	3.9

二维维纳滤波更能提高客观语音质量，原因在于目标谱估计使用了前后帧间的关联信息(语音在短时间内是平稳变化的，噪声不是)，使得估计值更准确，因此有更好的语音质量。而之所以没有过强的乐性噪声，是因为噪声估计时采用块平均，而且每个采样点在 3×3 范围内也进行了平滑，使得不稳定的噪声毛刺较少，这就从根本上抓住了乐性噪声产生的机理并进行了有效抑制。

5　小结

在充分考虑语音帧之间的关联信息的基础上，提出并实现了二维维纳滤波算法，在多帧组合构成的块结构中，用二维窗函数和二维维纳滤波器进行目标信号估计，并采用在 3×3 个采样点范围内平滑的策略，消除过多的乐性噪声。实验采用对比的形式，用文献[4]的一维维纳滤波方法和本文的二维维纳滤波方法比较，结果表明，无论客观质量还是主观评测，本文的二维方法都较一维方法有更好的评价，不但信噪比大幅提升，语音可理解度也得到保持。

参考文献：

[1] Hu Y，Loizou P C.Subjective comparison and evaluation of speech enhancement algorithms[J].Speech Communication，2007，49(7/8)：588-601.

[2] You C H，Koh S N，Rahardja S.An invertible frequency eigendomain transformation for masking-based subspace speech enhancement[J].IEEE Signal Processing Letters，2005，12(6)：461-464.

[3] 欧世峰，赵晓晖，顾海军.基于 DCT 与维纳滤波的单通道语音增强算法[J].通信学报，2006(10)：86-93.

[4] Quatieri T F.离散事件语音信号处理——原理与应用[M].北京：电子工业出版社，2004.

[5] Soon I Y，Koh S N.Speech enhancement using 2-D fourier transform[J].IEEE Transactions on Speech and Audio Processing，2003，11(6)：717-724.

[6] 程正，赵鹤鸣.基于多频带谱减法的语音增强算法的研究[J].计算机工程与应用，2007，43(36)：40-42.