

# 一种改进的最大匹配中文分词算法

闻玉彪, 贾时银, 邓世昆, 李远方

(云南大学 信息学院, 云南 昆明 650091)

**摘 要:**最大匹配算法包括正向最大匹配和逆向最大匹配两种算法,是中文分词领域的基础性算法,目前被广泛应用于众多领域。文中在详细分析了最大匹配算法的优缺点的基础上,提出了一种改进的最大匹配分词算法。改进算法在分词前先对词库进行了规范化预处理,分词时由汉字检索到该字开头的词组,再按词组长度由长到短的顺序使用传统最大匹配算法检索词库。目的是解决传统方法匹配效率低下和不能切分长词的问题。经算法分析结果表明,改进的算法较传统的最大匹配算法高效,分词能力更强。

**关键词:**最大匹配;索引;词库;分词

**中图分类号:**TP391.1

**文献标识码:**A

**文章编号:**1673-629X(2011)10-0092-03

## An Improved Algorithm for Maximum Matching of Chinese Word Segmentation

WEN Yu-biao, JIA Shi-yin, DENG Shi-kun, LI Yuan-fang

(College of Information, Yunnan University, Kunming 650091, China)

**Abstract:**Maximum matching algorithm includes two, they are forward and reverse maximum matching algorithm. It is a fundamental algorithm in the field of Chinese word, is widely used in many fields. In this paper, get a detailed analysis of advantages and disadvantages of the maximum matching algorithm, based on it, proposed an improved algorithm for maximum matching of word segmentation. In improved algorithm deal thesaurus with some rules at first, when doing Chinese word segmentation it searches the relative Chinese phrases with the beginning character of the word, then matches word with the traditional maximum matching algorithm from long to short of the order of phrases' length. The aim is to reduce the inefficiencies of traditional methods and solve the problem that the long word can not split well. The algorithm results show that the improved algorithm is better than the traditional maximum matching algorithm in efficiency, and more powerful.

**Key words:**maximum matching; index; thesaurus; segmentation

### 0 引言

随着社会、经济的飞速发展,特别是计算机和计算机网络的出现和应用普及,人类从此步入了信息化时代。伴随信息化不断推进,Internet上中文网页急剧增加,中文电子出版物、中文数字图书馆迅速普及<sup>[1]</sup>。词是语言的最小单位,所以汉语的文本分词是解决该问题的第一步<sup>[2]</sup>。同时,中文自动分词又是信息提取、信息检索、机器翻译、文本分类、自动文摘、语音识别、文本语音转换、自然语言理解等中文信息处理领域的基础研究课题,也是中文信息处理的瓶颈问题<sup>[3-6]</sup>。可

见,中文分词技术的地位和重要性日益突现。众所周知,英文以空格作为自然的分隔符,而中文由于继承自古代汉语的传统,词语之间没有分隔。中文只是字、句和段能通过明显的分界符来划界,唯独词没有形式化的分界符,虽然英文也同样存在短语的划分问题,不过在词这一层上,中文的独特构成决定了中文远比英文要复杂、困难。比如:歧义句、文言文,普通人去理解就有极大的困难,更不用谈让计算机去处理所面临的困难了。

中文分词,即把一个汉字序列按照某种规则切分成一个个单词序列的过程<sup>[7]</sup>。例如:“我是一个学生”,经中文分词程序切分为“我/是/一个/学生”。

中文分词的困难主要体现在两方面<sup>[3]</sup>:语言学和计算机科学。

#### ●语言学方面困难:

(1)词没有统一的定义形式。目前,语言学界还没有对词语给出一个为大家所共同认可和具有严格而

收稿日期:2011-03-08;修回日期:2011-06-21

基金项目:云南省自然科学基金(2007F174M);云南大学研究生科研课题资助项目(200928)

作者简介:闻玉彪(1984-),男,云南人,硕士生,研究方向为Web信息挖掘与提取、中文信息处理;邓世昆,教授,研究方向为计算机网络、智能建筑。





能汇聚多个链路的带宽。

### 3 实验结果及分析

本实验使用了虚拟机 VM 6.0.2 来搭建网络环境。采用 ROS 2.9.27 作为 PPPoE 服务器;将 PPPoE 服务器的上下行流量限制为 1 Mbps/s,然后启动 NAT 进行登陆,运行虚拟机的 Windows2003,将网关的地址设计为 INIC 的 IP,并开启迅雷下载。测试结果如表 1 所示。

表 1 不同帐号个数下的下载速率

帐号个数	帐号带宽(kbps)	下载速率(kbps)
1	1000	987.6
2	1000	1924.4
3	1000	3915.2

从测试结果来看,在帐号带宽同为 1Mbps 的情况下,下载速率随帐号个数的增加而呈线性增长,达到了系统设计的目的,实现了多帐号带宽汇聚的功能。

### 4 结束语

文中在 PPPoE 连接获取公网 IP 后,设计并实现了一种可以汇聚多帐号带宽的 NAT,并最终测试成功。由于时间、水平的限制,还有些需要改善的地方,整个设计可以在 NDIS 内核驱动中实现,而文中是使用 WinPcap 在应用层上实现的,所以效率不及前者,此外,由于一般情况下 IP 头选项几乎没用,ICMP 协议大都对网络进行差错报告,所以文中所设计的系统暂时不支

持这两样功能。

#### 参考文献:

- [1] 王艳平. Windows 网络与通信程序设计[M]. 第 2 版. 北京:人民邮电出版社,2009.
- [2] 赵雪峰. 基于 PPPoE/PPP 协议的带宽接入客户端拨号软件的实现[D]. 北京:中国地质大学,2005.
- [3] RFC2516. A Method for Transmitting PPP Over Ethernet (PPPoE)[S]. [s. l.]:Network Working Group,1999.
- [4] 邹航,杨元晔,苟光磊. NAT 网络地址转换技术分析[J]. 重庆工学院学报,2007,21(7):89-91.
- [5] 肖辽亮. NAT-PT 簇负载均衡的设计与实现[J]. 计算机技术与发展,2006,16(3):80-82.
- [6] 王南,孙保锁,王月平. P2PSIP 系统中 NAT 穿越方案的研究与设计[J]. 计算机技术与发展,2009,19(10):66-69.
- [7] RFC3022. Traditional IP Network Address Translator[S]. [s. l.]:Jasmine Networks,2001.
- [8] 郭士秋. IP 协议体系[M]. 北京:电子工业出版社,2002.
- [9] Stevens W R. TCP/IP 详解 卷 1:协议[M]. 范建华,张涛,译. 北京:机械工业出版社,2007:33-34.
- [10] WinPcap Team. WinPcap Documentation[EB/OL]. 2002. [http://www.winpcap.org/docs/docs\\_411/html/main.html](http://www.winpcap.org/docs/docs_411/html/main.html).
- [11] 罗军舟,黎波涛,杨明,等. TCP/IP 协议及网络编程技术[M]. 北京:清华大学出版社,2004:22-26.
- [12] Comer D E. 用 TCP/IP 进行网际互联(第一卷:原理、协议与结构)[M]. 林瑶,蒋慧,杜蔚轩,等译. 第 4 版. 北京:电子工业出版社,2001.

(上接第 94 页)

高,达到了预期效果;对长词的切分能力更强。

同时,算法也还存在不足之处。中文是一种较为复杂的语言,其结构复杂多样,用法灵活多变,应用也无处不在,这就决定了其词库非常庞大,而且要求高效;另外,随着社会、经济的飞速发展,大量的新词语往往随机涌现,这些新词语必然在词库中尚未收录,这给词库的更新及维护带来了新的挑战。因此,词库的建立是个巨大而艰难的工程,对词库的维护、更新及新词语的识别还有待进一步研究。

#### 参考文献:

- [1] 孙茂松,邹嘉彦. 汉语自动分词研究评述[J]. 当代语言学,2001(1):22-32.
- [2] 李淑英. 中文分词技术[J]. 商丘科技职业学院学报,2007(36):95-95.
- [3] 张春霞,郝天永. 汉语自动分词的研究现状与困难[J]. 系统仿真学报,2005,17(1):138-147.
- [4] 金在全,赵照,杜秀全. 一种改进的增字最大匹配算法[J]. 科学技术与工程,2007,18(7):4161-4164.

- [5] Li Haizhou, Yuan Baosheng. Chinese Word Segmentation [C]//Language, Information and Computation (PACLIC 12). [s. l.]:[s. n.],1998:212-217.
- [6] Xue Nianwen. Chinese Word Segmentation as Character Tagging [C]//Computational Linguistics and Chinese Language Processing. [s. l.]:[s. n.],2003:29-48.
- [7] 徐飞,孙劲光. 中文分词切分技术研究[J]. 计算机工程与科学,2008,30(5):126-128.
- [8] 文庭孝,邱均平,侯经川. 汉语自动分词研究展望[J]. 现代图书情报技术,2004(7):6-10.
- [9] 冯书晓,徐新,杨春梅. 国内中文分词技术研究新进展[J]. 情报检索,2002(11):29-30.
- [10] 曹卫峰. 中文分词关键技术研究[D]. 南京:南京理工大学,2009.
- [11] 宋国柱,陈俊杰. 基于双字词的动态最大匹配分词算法的研究[J]. 太原科技大学学报,2009,30(3):199-202.
- [12] 龙树全,赵正文,唐华. 中文分词算法概述[J]. 电脑知识与技术,2009,10(5):192-193.
- [13] 尹锋. 汉语自动分词研究的现状与新思维[J]. 现代图书情报技术,1998(4):22-26.