一种消除中文分词中交集型歧义的方法

魏博诚,王爱平,沙先军,王 永
(安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:切分速度和精度是中文分词系统的两个主要性能指标。针对传统的中文分词中出现的分词速度慢和分词精度不高的问题,采用了双层 hash 结构的词典机制来提升分词的速度,对于匹配结果中出现的交集型歧义字段,通过互信息的方法来消除,以提高分词精度。并对该分词系统进行了实现。通过与传统的中文分词系统的分词速度以及分词效果的对比,发现该系统在分词速度和精度上都有所进步,从而取得较好的分词效果。

关键词:中文分词;互信息;交集型歧义
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2011)05-0060-04

A Method about Removing Overlapping Ambiguity Producing in Chinese Matching

WEI Bo-cheng, WANG Ai-ping, SHA Xian-jun, WANG Yong
(Ministry of Education Key Lab . of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: Segmentation accuracy and speed are the two main performance indexes of the Chinese word segmentation system . According to the question of slow speed and precision of the word in the traditional Chinese word segmentation, it uses the structure dictionary of double-decked hash mechanism to promote the speed of word segmentation. To improve the segmentation accuracy, use the method of the mutual information to eliminate the overlapping ambiguity string which appeared in the matching results, the Chinese word segmentation system is achieved. The system is improved in the speed and accuracy compared with the traditional Chinese word segmentational system. The experiment results make the good participle progress.

Key words: Chinese word segmentation; mutual information; overlapping ambiguity

0 引 言

中文信息处理是文本挖掘重要的研究内容之一,而中文自动分词又是中文信息处理中的关键技术之一,尤其是在海量信息处理方面有很大的作用。分词的速度极为重要,对整个系统的效率有较大的影响[1]。中文词语的划分往往存在着歧义性,分词的难点在于切分歧义。如何最大限度地消除歧义,提高分词的准确性,对整个文本挖掘后面的工作都会产生十分重要的影响,因此,对于这个问题的研究有很大的现实意义。

文中从提高分词的速度和精度角度出发,实现了一个基于词典的中文分词系统。

收稿日期:2010-10-19;**修回日期**:2011-01-22
基金项目:安徽省自然科学基金项目(090412054)
作者简介:魏博诚(1985-),男,硕士研究生,研究方向为数据挖掘;王爱平,教授,研究方向为数据挖掘、人工智能、编译技术、计算机仿真以及滤波算法收敛性等领域。

1 中文词语切分的方法

随着信息时代的来临,中文分词显得越发的重要。众多专家学者通过对中文分词技术进行深入的研究,提出了许多种分词算法。主要包括以下几类:词典分词算法、统计分词算法和规则分词算法。这几种方法都有自己的优缺点。文中采用的是基于词典的分词方法。

1.1 分词词典机制的选择

由于不同的分词词典的机制因素,会影响词典分词系统的分词速度快慢。因此对中文分词的词典机制的研究就显得十分的必要。研究好中文分词的词典机制对提高分词算法的效率有着直接的关系。通过研究发现简单的顺序排列是最开始的分词词典机制,然而通过对分词词典机制的进一步深入探讨,学者们一致发现原先的结构并不能取到优良的分词效果。可喜的是,通过大量专家学者的不屑努力,多种具有代表性的分词词典机制被相继提了出来,有基于 Trie 树的词典分词机制[2-4]、双层 hash 词典分词机制[5-7]等等许多

种不同的词典机制。

由于对分词速度的考虑,选用了双层 hash 词典机制。通俗的说,双层 hash 词典分词机制就是一种多次 hash 结构的循环形式,是对单层 hash 的一种扩充。它不仅是对短语的第一个字进行一次 hash 查找,并且对于这个短语的第二个字还用 hash 查找的办法来确定。该词典由以下几个部分组成:短语第一个字 hash 表、短语第二个字 hash 表、短语余下的字索引表、短语余下的字词典正文。相对于单层 hash 词典机制,这种词典机制能较大提高系统的分词速度。

1.2 分词方法描述

在采用双层 hash 词典机制的基础上,该方法通过词典切分方法对中文文本进行切分,找出文本中的词语,作为理解中文的前提。具体如下:使用了一个具有 275786 条词汇的语料库,这是对大量的文档统计后得到的结果。对于需要分解成一个个独立的词的一段文章来说,按下面的方法进行:首先从语料库的第一个词的位置开始,对目标字段进行从左向右方向扫描,看目标字段是否包含这个词。如果目标字段中包含有这个词,则把这个词记录下来,并把这个词在整个目标字段中出现的总的次数(n_{wi})记录下来,一直到使用完整个语料库中的全部词汇。这样进行下去:给定的一段字段 $C_1C_2\cdots C_iC_{i+1}\cdots C_n$,按照从左向右的方向逐个扫描 $C_1C_2\cdots C_iC_{i+1}\cdots C_n$。假如 $C_i\cdots C_j$是一个词,那么从 C_{i+1}处开始接着按从左到右的顺序扫描字段,直到扫描完整个字段为止。例如:"工人们按时下班"这个目标字段则可以分解出以下五个词:"工人"、"人们"、"按时"、"时下"、"下班"。很显然,在分词的初步结果中,存在着歧义。下面介绍怎么样消除交集型歧义。

2 交集型歧义消除方法

2.1 为什么会产生歧义

在分词过程中下面3个原因会产生歧义:

1)自然语言二义性的原因而必然引起的一些歧义。可以举出许多例子。如:"棒球拍卖完了"既能切分成"棒球/拍卖/完了"又能切分成"棒球拍/卖/完了"。从表面上看这两种切分方式不管在语法上还是在语义上都是准确无误的,这里就留下这样一个比较难以解决的问题:在实际运用中到底怎么切分呢?恐怕只有结合上下文语境才能得出正确的形式了。

2)因机器自动分词所产生的一些特有歧义。如:"他只会诊断一般的疾病",用计算机切分,可以切分成"他/只会/诊断/一般/的/疾病",也可以切分成"他/只/会诊/断/一般/的/疾病",可以看出,只有前者切分方式是正确无误的,符合人们的思维习惯。而人工分词是不会出现这样的歧义的。

3)分词词典容量的大小所引起的一些数量众多的歧义。如:"汪大民是一位学生",用计算机切分被分为"汪/大/民/是/一位/学生","汪大民"在这个语境中是一个名字,在汉语中本应该是一个独立的词,因此这个划分显而易见是不正确的。

2.2 歧义字段的分类

专家学者们通过大量的研究表明:通常歧义字段可分成以下三大类:交集型歧义字段、组合型歧义字段和真歧义。

2.2.1 交集型歧义

在以 ABC 构成的字段中,如果 AB 是词表中的一个独立的词汇,而且 BC 也是词表中的一个独立的词(其中 A、B、C 为字串),于是可以把形如 ABC 这样的一类称为交集型歧义字段。比如交集型歧义字段"白天鹅"可切分成"白天/鹅"和"白/天鹅"这两种不同的切分形式,在这里 A="白",B="天",C="鹅"。再如交集型歧义字段"从小学"可切分成"从小/学"和"从/小学"这两种不同的切分结果,其中 A="从",B="小",C="学"。交集型歧义在实际的生活中经常遇见并且数量庞大。

2.2.2 组合型歧义

在以 AB 构成的字段中,如果 AB 是词表中的一个独立的词汇,而且 A 也是词表中的一个独立的词,并且 B 也是词表中的一个独立的词(其中 A、B 为字串),于是可以把形如 AB 这样的一类称为组合型歧义字段[8,9]。下面举几个例子来说明:组合型字段"将来"可切分成(1)"魏教授/将/来/当涂/讲学"和(2)"当涂/将来/文化/更/繁荣"两种切分方式。组合型字段"十分"可切分成(1)"现在/差/十/分/就/十一点/了"和(2)"她/十分/欣慰"两种切分方式。

2.2.3 真歧义

存在这样的一些字段:如果仅给字段本身,而不结合另外的信息,人就不能对这个字段的正确的切分方式判断出来,那么称这样的字段为真歧义。比如:"棒球拍卖完了"这个语句既可分成(1)"棒/球拍/卖/完/了"又可以分成(2)"棒球/拍卖/完/了"。这里如果仅给出语句本身而缺少上下文的相关联的有用的信息,就不会知道"棒球拍卖完了"这个字段到底是什么意思,因此更不会给出这个字段的正确划分了。

2.3 歧义的发现

文中利用的双向最大匹配法[10]是在分词算法中运用最广泛也是最基础的算法。它通过对正向最大匹配算法与反向最大匹配算法的分别运用,对目标字符串分别进行分词,从而得到两种不一致的分词结果。接着对上述的结果进行全面的比较:两次切分一致的

分词就被认为是非歧义字段，而不一致分词结果就被认定为歧义字段。

下面举例来说明。

待分的字符串为“按时下的风气”，则：

(1)用正向最大匹配切分，得到：“按时/下/的/风气”；

(2)用逆向最大匹配切分，得到：“按/时下/的/风气”；

(3)将上述的两个算法切分的结果进行比较，可以发现“按时下”是两次不同的切分；

(4)认定“按时下”是歧义字段，发现歧义字段后，接下来的是如何进行处理。

双向最大匹配分词方法在使用的过程中，要且仅只需要对目标句子进行一遍扫描，就能发现是否有交集型歧义字段在句子中出现，方法如下所述：对于用双向最大匹配分词方法进行分词而得出的结果，在里面任取两个 W_i 和 W_j，假如 W_i 与 W_j 的字串由下面的形式构成：AJ 和 JB，而且在语句中必须同时出现 AJB，那么目标字段中一定有交集型歧义的存在。

2.4 歧义的消除

从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好地反应成词的可信度。文中采用互信息[11,12]的方法来对歧义进行消除。

互信息算法是指对于两个字符 A 和 B，通过互信息公式来计算出 A 和 B 的互信息值 $M(A,B)$ 的大小，从而对字符 A 和 B 的关联程度进行判断。

$$M(A,B)=\log_2\frac{P(A,B)}{P(A)P(B)}$$

$$P(A,B)=\frac{n_{A,B}}{N}\quad P(A)=\frac{n_A}{N}\quad P(B)=\frac{n_B}{N}$$

其中，$n_{A,B}$ 是在中文子句中的同现次数，n_A 和 n_B 分别是它们各自在中文子句中出现的次数，N 为整个文本里中文子句的个数。

通过变通思维就不难发现，对于形如 AJB 的这样的交集型歧义字段，如果能够判断出 J 到底是与 A 组成一个独立的词还是与 B 组成一个独立的词，那么就能消除交集性歧义的问题了。而通过互信息算法的研究学习知道判断的依据可以是 AJ 和 JB 谁的成词几率大。可以通过计算 AJ 和 JB 的互信息值的大小来消除交集型歧义，谁的互信息值小就作为歧义消除。

3 中文分词系统的设计与实现

在上述工作的铺垫之下，笔者设计了一个简单的中文分词系统。

正是因为双层 hash 结构的词典机制所具备的能可观的提高分词速度和易于实现的特点，因此文中采用双层 hash 结构来作为词典机制。利用 C++语言来设计并实现了双层 hash 结构的词典，最终对给定的文本进行了分词。

歧义字段的发现和消解是歧义处理工作中最主要的两个部分。在文中设计并实现了双向最大匹配检索法来实现对文本中歧义字段的识别；同时文中采用互信息的方法实现了对歧义字段的歧义消解。

这个系统主要有预处理、分词、歧义发现和歧义消解四个模块组成，如图 1 所示。

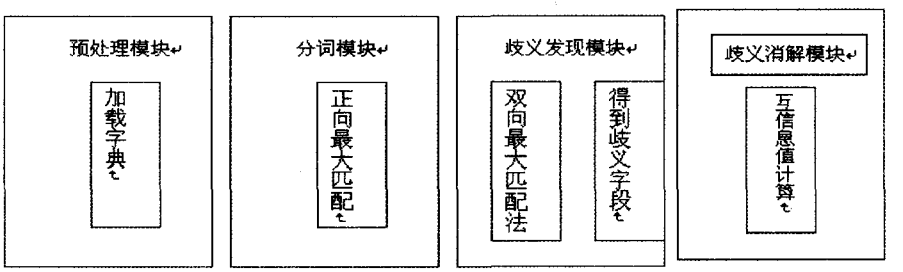

图 1 系统的结构示意图

4 实验结果及分析

(1)分词精度比较。

文中的实验数据来源是随机提取了 6 篇语料，全部来自语料库中。

表 1 所示的是正向最大匹配算法的切分的精度。

表 2 所示的是歧义消解算法的切分的精度。

图 2 所示的是两种算法的切分精度的比较结果。

表 1 正向最大匹配算法切分精度结果

	1	2	3	4	5	6
正确切分词数	55	61	64	72	76	87
切分总词数	60	66	70	78	82	95
切分精度	0.917	0.924	0.914	0.923	0.927	0.916

表 2 分词歧义消解算法切分精度结果

	1	2	3	4	5	6
正确切分词数	57	63	67	75	79	90
切分总词数	60	66	70	78	82	95
切分精度	0.95	0.955	0.957	0.962	0.963	0.947

图 2 中的数据表明，文中所采用的歧义消解算法在分词精度上比传统的正向最大匹配算法要精准。它很好地说明了采取这种歧义消解算法，对提高系统的分词准确率和提高分词精度有很大的帮助。

通过对实验结果的分析,切分错误原因主要有以下几个方面:

a. 存在组合型歧义字段和真歧义字段;

b. 未登录到字典中的词;

c. 含有错别字的字段。

因此,要提高分词的精度就必须要考虑到这几方面的问题,要使语料库更加丰富。

(2)分词速度比较。

通过采用 Trie 索引树词典机制的分词系统分词结果和文中所采用的双层 hash 结构词典机制结果来比较各自系统的分词速度。文中的实验数据是三个较小型语料和三个较大型语料,并在上述两种不同的词典机制上分别进行测试,图 3 所示的是两种不同的结果。

从图 3 可以看出,本分词系统所采用的双层 hash 结构在分词速度方面要强于 Trie 索引树,因此对于分词系统的效率和分词速度来说是明显提高的。

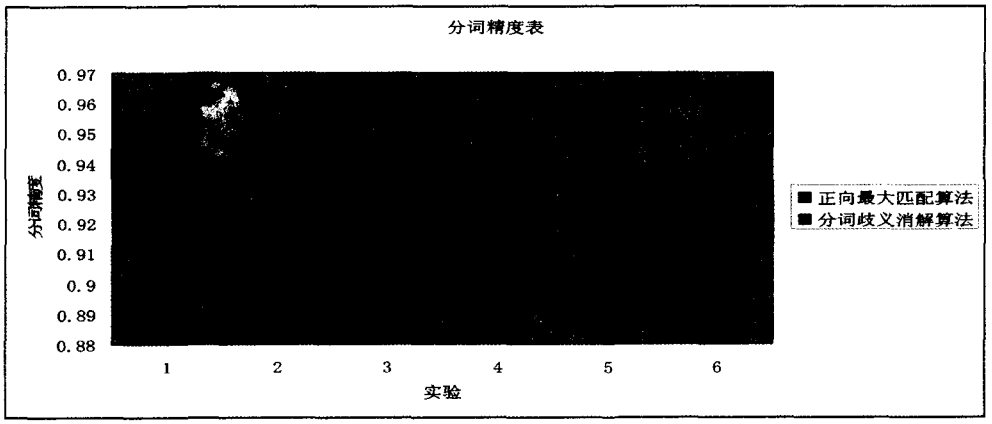

图 2　分词精度比较表

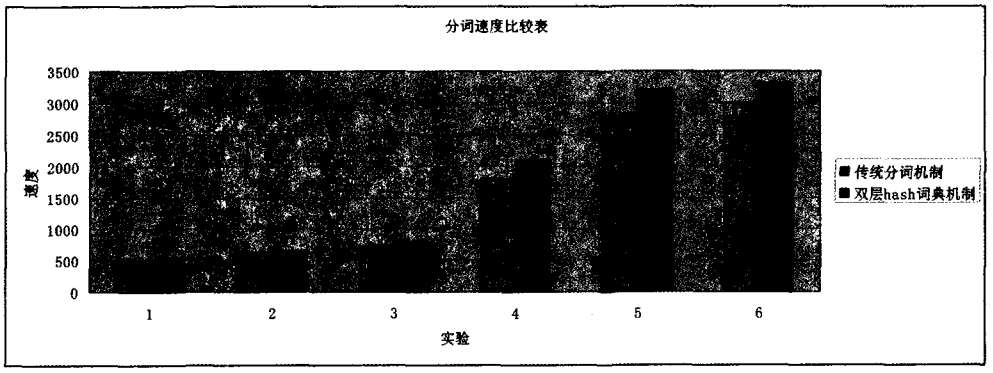

图 3　两个系统的分词速度比较

5　结束语

文中所设计的自动分词系统以及歧义消除方法,是出于对分词速度和精度的考虑,通过采用了双层 hash 结构的词典机制并对常见的交集型歧义字段性质的研究,运用互信息的方法来消除歧义,从而提高分词精度。

通过实验的验证证明该系统取得较好的分词效果。

参考文献:

[1] Allen J. 自然语言理解[M]. 第 2 版. 刘　群,译. 北京:电子工业出版社,2005:12-20.

[2] 王思力,张华平,王　斌. 双数组 Trie 树算法优化及其应用研究[J]. 中文信息学报,2006,20(5):24-30.

[3] XU Kc. A Non-collision Hash Trie-tree Based Fast IP Classification Algorithm[J]. J. Comput. Sci. & Techol. ,2002,17(2):219-226.

[4] Jiang Y,Shang F. Research on Multibit-Trie Tree IP Classification Algorithm[C]//International Conference on Communications. [s. l.]:[s. n.],2006.

[5] 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.

[6] 张　科. 多次 Hash 快速分词算法[J]. 计算机工程与设计,2007,28(7):1716-1718.

[7] 张培颖,李村舍. 一种中文分词词典新机制——四字哈希机制[J]. 微型电脑应用, 2006, 22(10):35-37.

[8] 冯素琴,陈惠明. 利用上下文信息解决汉语组合型歧义[J]. 电脑开发与应用,2007,20(1):23-25.

[9] 肖　云,孙茂松,邹嘉彦. 利用上下文信息解决汉语自动分词中的组合型歧义[J]. 计算机工程与应用, 2001, 37(19):81-87.

[10] Luo Zhi yong, Song Rou. Disambiguation in a Modern Chinese General - Purpose Word Segmentation System[J]. Journal of Computer Research and Development,2006,43(6):1122-1128.

[11] 费洪晓,康松林,朱小娟,等. 基于词频统计的中文分词的研究[J]. 计算机工程与应用,2005(7):67-68.

[12] 朱小娟,陈特放. 词频统计中文分词技术的研究[J]. EIC,2007,14(3):78-79.