鞍山师范学院学报
Journal of Anshan Normal University
2010-08,12(4):46-48

一种基于近邻匹配的中文分词算法 Jlppeccz

耿新青,陶凤梅,黄宏光
(鞍山师范学院 数学系,辽宁 鞍山 114007)

摘 要:提出一种基于近邻匹配新的分词算法 Jlppeccz,该算法首先把一篇文章以标点符号为界线分成若干个句子,然后用近邻匹配方法把一句话切分成1~4字的词,通过对词库的搜索,对已分的词进行重组,把小词合并成大词,再将处理过的词存储到一个临时的词库里,以备后续的句子查找,并可实现对词库添加词的功能.与经典 MM 算法和词频统计方法相比,本文算法有较大的改进.

关键词:中文分词;近邻匹配;分词系统
中图分类号:TP18　　**文献标识码**:A　　**文章篇号**:1008-2441(2010)04-0046-03

随着信息技术的迅猛发展,人们可以获得的中文电子信息量急剧增长.对自然语言处理的深入研究,将有助于文本挖掘效率的提高,并且推动网页信息提取技术的发展.因此,迫切需要一种有效的分词系统对信息进行管理,以帮助人们从海量的信息中快速准确地获得所需要的信息.通常的处理方法是将文本进行分类,根据文本内容确定类别,从而使分类后的信息具有针对性,避免了无关信息的干扰,提高了文本处理效率.另外,可以从文本中提取主题信息或文摘,对大量信息进行浓缩和提炼,以简洁的方式提高阅读效率.由于信息资源高速膨胀,由人工对文本进行处理已不切合实际,因此需要采用自动化的处理方式.中文文本自动分类、自动文摘、主题自动提取已成为热点.在文本自动处理工作中,主题词提取是基础工作之一,而主题提取又是以分词作为第一步.由此可见,分词是中文信息处理处理的重中之重.中文的自动分词技术虽然已经有一定程度的突破,能满足一般实用性的要求,但是实用的自动分词系统至今尚未针对大规模真实文本展开.这要求有相当可用性的自动分词系统[1].

以往的分词算法中比较经典的是最大匹配算法(MM)[2]和基于词频统计[3]的方法.MM 算法存在预先设定匹配词长初始值和对词典有很强的依赖性的缺陷;基于词频统计方法存在对常用词的识别精度差,时空开销大的缺点.

1 近邻匹配二次重组法和分词系统

MM 分词算法对词典有很强的依赖性,而且分词的精度不高.例如,"当她还是小孩子时"用 MM 法进行分词的结果为"当\她\还是\小孩\子时",成为了歧义字段.采用 Jlppeccz 法,对"当她还是小孩子时"进行切分得到正确的分词结果:"当\她\还是\小孩子\时".该算法克服了局部范围最大匹配的局限,真正体现了"长则优先"的原则.基于词频统计的方法,不需要切分词典,但这种方法也有一定的局限性,会经常抽出一些共现频率高,并不是词的常用字组,例如"这一""之一""有的""我的"等,并且对常用词的识别精度差,时空开销大.建立已分词的临时词库可减少对数据库搜索、检测词的匹配过程,采用相对小很多的临时词库中查找,可缩短搜索时间,提高分词精度.

收稿日期:2009-02-11
基金项目:国家自然科学基金资助项目(60275020).
作者简介:耿新青(1973-),女,江苏江阴人,鞍山师范学院数学系副教授,博士研究生.

中文分词的难点之一是对人名的切分. 专名识别技术[4]是影响中文自动分词精度的的一个重要方面,也是自动分词技术的难点. 难点在于:在真实语料中,专名是一个开放的集合,无法穷举;除了出现频率高,同时比较稳定的专名可以收录进词表外,其它专名只能动态地进行辨识;各种专名之间,专名和普通词之间存在大量的歧义和冲突. 这样由于人名的任意性和无穷尽性,使得要建一个包含所有人名的词库几乎成为不可能.

本文方法是针对最大匹配法(MM)和基于词频的统计方法的改进.

Jlppeccz 算法如下:

(1)以标点符号为标志把一篇文章分成若干个句子.

(2)顺序取文本的一个字,若是标点符号则结束.

(3)和词库,若匹配成功,看待定表中是否有字,若有字把字放到数组中,然后把匹配成功的词放到数组中,转(2). 若不成功,放到待定表,转(2).

(4)看待定表中的字和词库进行匹配,匹配成功放到数组中,转(2).

(5)把一句话分成 1 ~4 字的词.

(6)把两个小词合成一个大词,把小词从数组中删除.

(7)把词存到临时词库.

(8)取下一句话,转(2).

(9)把数组中的词进行输出.

分词系统中向词库中添加词的功能的目的是向词库中添加一些待切分文章中的人名,这样一方面提高了分词的精度,一方面比其他方法要节省时间. 还可以向系统中添加一些英语单词或是一些不是常见的符号,因为文章里不可避免的包含一些英语单词和一些罗马字符号. 如果不能对其正确切分的话,很可能会造成语句的歧义,降低切分的精度.

2　实验验证

程序以标点符号为切分断点,切分成若干个部分. 然后再对每句话进行细致切分,词和词之间用“\”号格开.

系统的运行结果如图 1,词库中不单单有现在已经切分的词,还有“河”,“人”等单字词,分词的时候没有把“河”和“北戴”分开,是因为重组把小词合并成大词. 而且如果词库中没有“北戴河”这样的地名,也可以利用系统中“工具”中的“添加新词”向词库中添加新词.

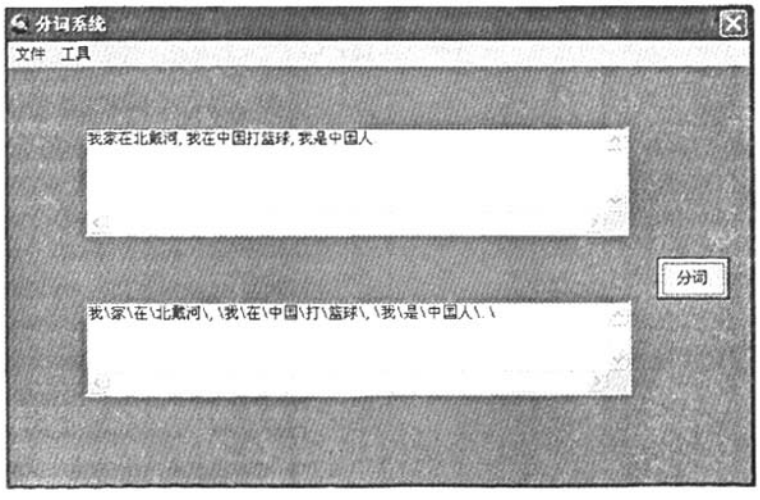

图 1　Jlppeccz 法分词系统对文本的切分

3 实验结论

本文提出一种新的近邻匹配分词算法(Jlppeccz),并且实现了分词系统,解决了对中文词的划分的困难.与其他分词算法相比,Jlppecc 法的准确性和时空开销的综合数据优于其它算法.特别是当文章的长度越大时,优势就越明显.

参考文献:

[1] 刘源,谭强,沈旭昆.信息处理用现代汉语分词规范及自动分词方法[M].北京:清华大学出版社,1992.

[2] 贺艳艳.基于词表结构的中文分词算法研究[D].北京:中国地质大学,2007.

[3] 贾洪晓,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.

[4] 罗智勇,宋柔.一种基于可信度的人名识别方法[J].中文信息学报,2005,19(3):67-72.

Jlppeccz: A New Word Segmentation Algorithm Based on Neiboring Match

GENG Xin-qing, TAO Feng-mei, HUANG Hong-guang

(*Department of Mathematics, Anshan Normal University, Anshan Liaoning* 114007, *China*)

Abstract: This paper presents a new Chinese word segmentation algorithm Jlppeccz based on neighboring match. The traditional MM algorithm which may easily produce ambiguity depends on dictionary strongly. Jlppeccz algorithm divided a article into some sentences with the benchmark of punctuation mark, then one sentence is cut into one word or multiword by neighboring match. The database of the words is searched; the words which have been divided are recombined; the small phrases are combined into the big ones, the words are put into a temporary table to prepare for the following phrases; the words are added into the database of the words. Compared to the classical MM algorithm and the word frequency statistics algorithm, Jlppeccz algorithm has greater improvement. Experiment shows the present algorithm possesses higher precision and efficiency than MM algorithm. The example demonstrates the effectiveness of the present algorithm.

Key words: Chinese word segmentation; Neighboring match; Word segmentation system

(责任编辑:张冬冬)

一种基于近邻匹配的中文分词算法Jlppeccz

作者： 耿新青， 陶凤梅， 黄宏光
作者单位： 鞍山师范学院,数学系,辽宁,鞍山,114007
刊名： 鞍山师范学院学报
英文刊名： JOURNAL OF ANSHAN NORMAL UNIVERSITY
年，卷(期)： 2010，12(4)
被引用次数： 0次

参考文献(4条)

1. 刘源,谭强,沈旭昆.信息处理用现代汉语分词规范及自动分词方法[M].北京:清华大学出版社,1992.

2. 贺艳艳.基于词表结构的中文分词算法研究[D].北京:中国地质大学,2007.

3. 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.

4. 罗智勇,宋柔.一种基于可信度的人名识别方法[J].中文信息学报,2005,19(3):67-72.

相似文献(2条)