兰 州 商 学 院

本科生毕业论文（设计）

论文（设计）题目：汉语分词技术初探

学　　院、　　系：信息工程学院

计算机科学与技术系

专　　业　（方　向）：计算机科学与技术

年　　级、　　班：

学　生　姓　名：

指　导　教　师：

2011 年 5 月 18 日

声　　明

本人郑重声明：所呈交的毕业论文（设计）是本人在导师的指导下取得的成果。对本论文（设计）的研究做出重要贡献的个人和集体,均已在文中以明确方式标明。因本毕业论文（设计）引起的法律结果完全由本人承担。

本毕业论文（设计)成果归兰州商学院所有。

特此声明

毕业论文（设计）作者签名:

年　　月　　日

汉语分词技术初探

摘　要

所谓汉语分词，就是将中文语句中的词汇切分出来的过程。由于汉语的书写习惯，汉语语句中词与词之间的标志是隐含的，英文的单词与单词之间有空格，所以不存在分词问题.而中文的每一句中，词与词之问是没有空格的，因而必须采用某种技术将其分开。

分词技术作为自然语言处理的基础环节，同时也是关键环节之一，它的质量好坏直接影响到后续处理步骤的效果。汉语分词工作看似细微，但作为计算机自然语言处理的第一步，它的关键作用是不容忽视的。如今汉语分词已成为自然语言处理的研究热点与难点。

本文讨论了中文分词的概念、目标及其所面临的一些基本问题，详细介绍了三种基本中文分词算法，并对中文分词词典的索引及常用词典结构进行了介绍，最后说了正向最大算法的实现及测试结果。

[关键词] 中文分词 最大匹配 分词词典 自然语言处理

ABSTRACT

Chinese word segmentation ， is to cut the sentence in the Vocabulary sub —out process ． Since the writing habits of Chinese ， Chinese sentence symbol between words is implied ． the English words have the spaces between the words ， So there is easy to separate ． The Chinese word for each sentence ，there is no space between words ，and therefore must be some kind of technology to separate sentence ． Chinese sentence segmentation algorithm from the 20th century ，since the 80’S has been a research focus ， due to the complexity of the Chinese language has been in a stage of development ．

Segmentation of natural language processing technology as the basic link ， but also one of the key links，and its direct impact on the quality of the subsequent processing steps results ． Chinese word segmentation the first step in natural language processing,and its importance can not be ignored．

［Key Words] Chinese Word Segmentation ， Maximum match ， Segmentation Dictionary, Chinese Information Processing

目　录

汉语分词技术初探

一、引言

在自然语言处理中，词是最小的能够独立活动的有意义的语言成分.我们知道，在英文文本中，单词之间是以空格作为自然分界符的.中文和英文比起来，有其自身的特点，就是中文以字为基本书写单位,句子和段落通过分界符来划界，但是词语之间没有一个形式上分界符。也就是说，从形式上看，中文没有“词”这个单位。

因此，进行中文的自然语言处理通常都是先将中文文本中的字序列切分为合理的词序列，然后再在此基础上进行其它分析处理.将中文连续的字序列按照一定的规则重新组合成词序列的过程,就叫做中文分词。

作为中文信息处理基础的中文分词技术，已经被广泛应用于中文信息领域的信息检索、自动摘要、中文校对、汉字的智能输入、汉字简繁体转换、机器翻译、语音合成等技术中.自20世纪80年代初，中文信息处理领域提出自动分词以来,相关方面的众多专家学者、科研院所、业机构为之付出了不懈的努力，取得了一些重要的进展和一些实用性的成果，提出了许多中文分词方法，有些成熟的技术已经应用于产品当中。但这些分词方法或多或少都存在着不足之处，比如对于检索系统，由于近年来信息的多元化、复杂化，对信息处理技术的研究、发展、应用提出了新的挑战，跨越了单纯文本的检索，例如问答系统必须对关键字进行语义分析与处理.这就要求信息处理技术必须跟上信息发展的速度，在速度与性能方面具备更高的指标。

要让计算机能够自动地处理信息就必须借助分词技术让计算机理解自

然语言。分词技术作为自然语言处理的基础环节，同时也是关键环节之一，它的质量好坏直接影响到后续处理步骤的效果。其中，中文分词由于中文结构的特点,与西方国家文字相比更难于处理.汉语的词汇与词汇之间没有显式的边界，汉语的分词需要通过计算机对文字内容的分析，来达到切分词汇的目的。中文分词已成为计算机处理汉语面临的首要基础性工作。只有提高中文分词系统[2]的准确率和工作效率,才能使自然语言处理系统稳定高效地工作。

本文的主要目标是通过对目前中文分词关键技术的研究，设计并实现最大正向匹配算法。本文的主要研究内容为：

(1）分析了当前中文分词的研究背景,介绍了中文分词的研究意义.

(2)对中文分词技术进行了简介,简单介绍了中文分词、中文分词的应用、中文分词系统的目标、中文分词的基本问题。

(3)研究了三类基本的分词算法:基于字符串匹配的分词方法，基于统计的分词方法，基于知识理解的分词方法。举例说明了实际应用中如何使用合适的分词算法.

（4） 介绍了目前中文分词技术中常用的索引方法和词典机制。

（5）对正向最大匹配算法进行了实现与测试.

相对于研究内容,本文的结构安排为：

第一章 引言。本章首先介绍了本文的研究背景和研究意义,然后介绍了本文的主要工作和论文的结构安排。

第二章 中文分词简介.本章首先介绍了中文分词的基本概念、中文分词的应用,接着介绍了中文分词系统的目标、中文分词的基本问题等.

第三章 基本中文分词算法。本章首先介绍了常用的中文分词算法,包括基于字符串匹配的分词方法，基于统计的分词方法,基于知识理解的分词方法；然后举例说明了实际应用中如何使用合适的分词算法。

第四章 中文分词词典。本章首先介绍了中文分词技术中的索引方法，然后又介绍了目前中文分词技术中常用的词典机制。

第五章 正向最大匹配算法的实现。本章根据正向最大匹配算法的流程，编写了采用正向最大匹配算法的分词程序，并对程序进行介绍和测试。

第六章 结论。此部分总结论文的所有工作，分析论文中存在的不足和一些未解决的问题。

二、中文分词简介

中文分词是中文信息处理技术中最基础、最关键的一个环节.所谓分词，就是把一个句子中的词汇按照使用时的意义切分出来。

(一)中文分词的概念

将连续的字序列按照一定的规范重新组合成词序列的过程被称为分词；中文分词就是把中文的汉字序列分成有意义的词.分词只是中文信息处理的一部分，分词本身并不是目的,而是后续处理过程的必要阶段，是中文信息处理的基础技术。

1、什么是中文分词

分词就是利用计算机识别出文本中词汇的过程。在英文中，词汇之间一般会有空格等比较明显的分隔符.而中文中，是以字为基本书写单位,只有段与段、句语句之间有分隔符，而词汇之间没有分隔符.所以,虽然在英

语中也存在分词问题，但远没有中文分词那么复杂苦难。

比如:The table tennis bat is sold out。中文意思就是乒乓球拍 卖完了.对于通过空格和标点来切分的英语例句,一般不会产生歧义.但是在中文中,“乒乓球拍卖完了”则因为乒乓球和乒乓球拍、卖与拍卖都是词语而又没有明显分隔而产生了:“乒乓球/拍卖/完了 和“乒乓球拍/卖/完了”两种完全不同的意义。所以，要让计算机完成上述过程,相对于英语,难度有质的不同。

中文分词的过程,就是要把一句话中有意义的词汇都切分出来，并给出所有正确结果。由于中文词汇中间是有分隔符的，所以对中文词组的正确识别就显得很重要。词是汉语中最小的有意义的独立单位，但是这最小的单位却是没有显式分割的。若要使计算机与人类达到自由无障碍的语言交互，就必须让计算机能理解自然语言。只有当汉字串组成的句子被准确地转化为词之后，才能继续进一步工作.比如一个中英翻译系统，如果连词汇都不能正确切分，翻译得到的英文是不可能符合原义的。

2、中文分词的应用

互联网绝大部分需要分词，其中典型的实例有：

（1）汉字处理.这方面主要包括拼音输入法、手写识别、简繁转换等。

(2)信息检索。如Google、Baidu、Yahoo等检索工具。

雅虎中文网页搜索技术部总监张勤认为：中文分词是中文搜索技术的基础，只有做好了分词，才能有好的搜索。可见，掌握了优秀的分词技术就可以在中文搜索中占有一席之地。

众多优秀的搜索引擎都有自己的分词技术。如雅虎、百度，都是使用

.谷歌也是采用的美国 Basis technology 提供的中文分词技术。

(3)内容分析。这方面主要包括了机器翻译、广告推荐、内容监控等。

现在的翻译技术，无论是在线的还是单机的，在翻译句子或段落的时候总会让我们不知所措，语法错误明显，词不达意等等。究其根本原因就是因为中文分词技术的滞后和一些多义词汇选义的把握。

中文分词的其中一个重要功能就是为词语的计量分析,词频的统计提供可靠的依据，比如汉语中最常用的词是哪个。这使我们可以做一些广告推荐(哪些广告被更多次的提起）等。

（4）语音处理。 语音识别、语音合成等。

由于汉语中的多音字、一些发音习惯使得语音识别和合成过程中需要一个可靠地中文分词作为基础。比如:

发音的不同——如：的(d ē)，目的（d ì ）

变声—-如：好酒(h ǎojiu —-〉h áojiu)

轻声——如:桌子(zi ）

要处理这些中文特有的由于发音习惯而引起的分词问题，一个好的分词技术显然是必不可少的。

二）中文分词的目标

中文分词系统的目标为达到信息处理的需求，达到所要求的相应水平，具体来说，主要是准确、高效、通用及适用四个方面。

准确率是分词系统性能中最重要的核心指标。现有的分词系统中，有些准确率已达到98%一99%,光从数据上看似乎已经相当高了，其实不然。这样的分词系统如果被用来支持中外文翻译系统，现在假设平均每句语句有 10 个汉语单词,那么以直前的概率来计算，10 句语句中就会切分错 1-2 个词,含有错误分词的 1-2 句就不可能被正确翻译.于是仅仅因为分词系统的准确率欠佳，中外文翻译系统的翻译准确率就降低了 10—20 个百分点。进一步分析，对自动分词来说,其更大的作用是对大规模语料库进行加工，从而为上层应用系统提供统计数据和各种知识.如果分词产生错误则会在最后的统计结果中累积起不可忽视的“垃圾”，从而给上层的应用系统带来相当严重的影响.由此可见，分词系统的准确率应当达到 99．9%以上，这样才能基本满足上层的使用要求，换句话说,即使提高千分之一的准确度，对实际应用都是非常有意义的。

2、运行效率

分词是各种汉语处理应用系统中共同的、基础性的工作，这步工作消耗的时间应尽量少，应只占上层处理所需时间的一小部分，并应使用户没有等待的感觉，在普遍使用的平台上大约每秒钟处理 1 万字或 5 千词以上为宜。

3、通用性

随着互联网的普及应用，中文平台的处理能力不能仅限于我国，仅限于字处理，仅限于日常应用领域.作为各种高层次中文处理的共同基础，自

.

(1）中文自动分词系统应支持不同地区（包括中国香港、澳门、台湾，以及新加坡和澳洲、欧洲、美洲的华语社区）的中文字符处理；

（2）中文自动分词系统应能适应不同地区的不同用字、用词，不同的语言风格，不同的专名构成方式（如港澳台地区一些妇女名前冠夫姓,外国人名地名的汉译方式与我国人名地名很不一样)等；

(3）中文自动分词系统应能支持不同的应用目标，包括各种输入方式、简繁转换、语音合成、校对、翻译、检索、文摘等等:支持不同领域的应用，包括社会科学、自然科学和技术,以及日常交际、新闻、办公等等；

（4）中文自动分词系统应当同现在的键盘输入系统一样成为中文平台的组成部分.为了做到足够通用又不过分庞大,必须做到在词表和处理功能、处理方式上能灵活组合装卸,有充分可靠和方便的维护能力，有标准的开发接口.同时,系统还应该具有良好的可移植性，能够方便地从一个系统平台移植到另一个系统平台上而无需很多的修改。当然,就当今的现状来说,完全的通用性很难达到。

、适用性

中文自动分词只是手段而不是最终目的，任何分词系统产生的结果都是为某个具体的应用服务的。好的分词系统具有良好的适用性，可以方便地集成在各种各样的汉语信息处理系统中.

（三)中文分词的基本问题

我们可以看出，中文分词技术必然是以后计算机发展必不可少的一项

30 年的研究，仍存在诸多不足.而这主要是由于中文分词有很多难点难以兼顾，总体归纳起来，中文分词主要有三大困难:分词规范、歧义识别以及未登录词。

、分词规范

(1)“词”是否有清晰的定义？ 在每本汉语语法教科书中,我们都可以找到对“词”的这样一条定义：语言中有意义的能单说或用来造句的最小单位。这个定义相当抽象，从计算的层面上讲,这种模棱两可的定义是不可计算的,即不可操作的。而产生如此定义涉及多个方面[3]：

①核心词典问题:在进行分词时需要有一个核心（通用的、与领域无关的)词典，即普通词典，凡在该词典中存在的词，在分词时就应该切分出来。但是应该将哪些词组收入到核心词典中去，虽然已经提出各种收词的条件，但是对每个词组按照这些条件的进行判断却难以操作，因此目前还没有合理的可操作的理论和标准。

②词的变形结构问题：汉语中的动词和形容词有些可以产生变形结构,例如“打牌”、“开心”、“看见 、“相信”可能变形为“打打牌 、“开开心”、“看没看见”、“相不相信”等。在对变形结构进行切分时,如果切分出“打打＼牌”、“开开＼心”就不怎么合理,“看＼没＼看见”还说得过去,但“相＼不＼相信”就说不过去了。在进行中文分词时,对这些变形结构的切分缺少可操作的、合理的规范.

③词缀的问题:例如语素“者 在现代汉语中单独使用是没有意义的,因此“作者 、“成功者”、“开发者”内部不能切开。依据这个标准,“开发中国第一个操作系统软件者”、“做出了巨大个人财产和精神牺牲者”、“克

,这样复杂的结构在本质上就与词的定义相矛盾。又如职务名称“外交部长”，语义上理解为“外交部之长”，切成“外交＼部长”、“外交部＼长”、“外交＼部＼长”或不予切分，都会有人提出异议.

④非词语素问题：现代的书面汉语并非纯粹的“现代汉语”，其中夹杂着不少文言成分，如“为民除害”、“以逸待劳”、“帮困济穷”等等。探寻白话文中夹杂文言成分的规律，是中文信息处理需要解决的一大问题。

(2）词频对领域有一定的敏感性.即使一些统计信息是从精心挑选的“平衡语料库”中计算而来，将之应用于不同领域也会产生偏移，从而导致切分过程中切分的精度下降。而且不同目标的应用对词的切分规范的要求又有所不同,理论上讲汉语自动分词规范，作为规范，那么必须支持各种不同目标的应用，但不同目标的应用对词的要求是不同的，甚至是有矛盾的.

①以词为单位的键盘输入系统，为了提高输入速度,一些互现频率高的相互邻接的几个字也常常作为输入的单位,比如:“每一”、“再不”、“这就”、“也就”等。

②检索系统，检索系统的词典注重术语和专名，并且一些检索系统倾向于分词单位较小化.比如,在构造倒排文档及创建索引时把“分布式计算”切成“分布式＼计算”，使得无论用“分布式计算”还是用“分布式”检索,都能查到。

上述的两个实例，前者把不是词的几个字放在了一起组成了“词 ，而后者把是词的却切分开了。事实上，许多中文信息处理系统，都是根据自

性普遍不足，其分词结果很难采用统一的通用的分词标准来评价。

、歧义识别

歧义是汉语中普遍存在的问题，因此切分歧义词也是汉语分词中的一大难题。形式上相同的一段文字,在不同的场景或语境中，可以切分出不同的结果，有不同的含义。

（1）交集型歧义

对于汉字串 AJB, AJ， JB 同时成词。

例:他说的/确实/在理. 他说/的确/实在/理。

(2）组合型歧义

对于汉字串 AB， A， B, AB 皆可独立成词。

例：门/把手/坏/了，请/把/手/拿/开.

将来， 学生会

(3）混合型歧义

同时包含交集型和组合型歧义.

这些歧义有的会产生不同的分词结果，这些结果有时都有含义，这种情况就是真歧义；有时，只有一种结果是在所有真实语境中是有实在意义的。这种情况叫作伪歧义。

(4) 真歧义

歧义字段在不同的语境中确实有多种分隔形式

例：地面积

这块/地/面积/还真不小。 地面/积/了厚厚的雪.

（5）伪歧义

歧义字段单独拿出来看有歧义,但在所有真实语境中，仅有一种分隔形式可接受。

例：　　挨批评

挨/批评（√）　　挨批/评(X)

对于交集型歧义字段，真实文本中伪歧义现象远多于真歧义现象。

3、未登录词

在文本处理过程中，会遇到很多词典中未囊括的词语.如：人名等.这些不断增加的词汇没有可能和必要都加入到词典中。所以，分词中遇到未登录词汇是不能避免的.

例如：

实体名词和专有名词

—人名：张三、李四

-地名：三义庙、白洋淀

-机构名:方正、联想

专业术语和新词语

—专业术语:万维网、主机板

—缩略词：三个代表、扫黄打非

未登录词和歧义现象是影响中文分词准确率的两大因素,两者之中，未登录词造成的影响更为严重。在真实的文档和语料库中，专有名词和术语占了很大比例，词典在多数情况下很难包括这些词。分词算法能否对新词进行有效识别对应用来说十分重要，目前新词识别的准确率已经成为一个

评价分词系统好坏的重要指标。

三、基本中文分词算法

自从 1983 年,背景航空航天大学实现了我国第一个实用性的自动分词系统到现在，国内外的研究者在中文分词领域进行了广泛的研究,提出了许多有效的算法。

（一）中文分词算法介绍

现在最常用的中文分词系统主要采用以下 3 种算法：

1、基于字符串匹配的分词算法

这种方法又叫做机械分词算法，机械分词法按照一定策略将待切分字符串与机器里预先准备的词条进行匹配，然后找出一个最长的结果.按照扫描方向的不同,串匹配分词算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）匹配和最小(最短）匹配；按照是否与词性标注过程相结合，又可以分为单纯分词算法和分词与标注相结合的一体化算法。常用的几种机械分词算法如下:

(1）正向最大匹配法(由左到右的方向);

正向最大匹配分词是基于词典的分词系统。所谓最大匹配,就是要求每一句的分词结果中的词汇总量要最少。正向最大匹配分词又分为增字和减字匹配法[4]。增字匹配法需要一种特殊的词典结构支持，能够达到较高的分词效率。

减字法的流程为：首先读入一句句子，取出标点符号，这样句子就被分成相应的若干段，然后对每一段进行词典的匹配，如果没有匹配成功就

从段末尾减去一个字,再进行匹配,重复上述过程,直到匹配上某一个单词。整句句子重复这些流程，直到句子全部分解成词汇为止。如果事先知道词典中最长词的长度,那么在一开始的匹配中，不用将分割出来的整段语句与词典匹配，只需要以最长词的长度为最大切分单位进行切分就可以了。

(2)逆向最大匹配法(由右到左的方向)；

逆向最大匹配分词与正向最大匹配分词相反，从句子结尾开始进行分词。

（3）最少切分(使每一句中切出的词数最小).

这种算法使每一句中切出的词数最小.如果将上述各种方法相互组合，例如，可以将正向最大匹配算法和逆向最大匹配算法相结合来构成双向匹配法。由于汉语单字成词的特点，正向最小匹配和逆向最小匹配一般很少使用。可以把机械分词作为初步的处理手段，然后再通过进一步工作提高结果的正确率.

实际使用中还可以将上述各种算法相互组合，例如，可以将正向最大匹配算法和逆向最大匹配算法结合起来构成双向匹配法。由于汉语单字成词的特点，正向最小匹配和逆向最小匹配一般很少使用。一般说来，逆向匹配的切分精度略高于正向匹配,遇到的歧义现象也较少.统计结果表明[5]，单纯使用正向最大匹配的错误率为 1/169,单纯使用逆向最大匹配的错误率为 1/245。但这种精度还远远不能满足实际的需要。实际使用的分词系统,都是把机械分词作为一种初分手段，然后通过利用各种其它的语言信息来进一步提高切分的准确率.

2、基于理解的分词算法

这种分词算法是通过让计算机模拟人对句子的理解,达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式,因此目前基于理解的分词系统还处在试验阶段。

3、基于统计的分词算法

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的反映成词的可信度[4]，可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息.定义两个字的互现信息,计算两个汉字X、Y的相邻共现概率。互现信息体现了汉字之间结合关系的紧密程度.当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词。这种方法只需对语料中的字组频度进行统计，不需要切分词典，因而又叫做无词典分词法或统计取词方法。但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组,例如“这一”、“之一 、“有的”、“我的”、“许多的”等,并且对常用词的识别精度差，时空开销大。它的优点在于可以发现所有的切分歧义，但是统计语言模型的精度和决策算法在很大程度上决定了解决歧义的方法,需要大量的标注语料，并且分词

速度也因搜索空间的增大而有所缓慢.实际应用的统计分词系统都要使用一部基本的分词词典（常用词词典)进行串匹配分词，同时使用统计方法识别一些新的词，即将串频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点.

（二）根据具体应用使用合适的分词算法

在实际应用中，对于某一具体的应用系统，并不是单纯使用某种分词算法就能解决问题,我们可以根据具体应用的所需满足条件使用不同的方法.在此以中文信息检索中所用到的分词算法为例进行说明。

1、混合分词

对于实际应用中的中文信息检索系统来说，当弄不清楚使用哪种分词算法更好的话，可以试着合并使用多种方法，混合分词就是一种简单且容易实现的方法，也是大型检索系统中常用的一种方法，使用混合分词方法能够涵盖更多的词汇。

混合分词的原理就是“先用专业词典进行一遍分词，再用普通词典进行一遍分词”,我们用一个实例对为何要进行两次分词进行说明。例如，对“搜索引擎知识”这句话进行分词，如果我们的词典中含有“搜索引擎”这个词,那么这句话的切分结果就是“搜索引擎＼知识”。如果词典中没有“搜索引擎”这个词，而只含有“搜索 ，“引擎”,“知识”这三个词，那么这句话的切分结果就是“搜索＼引擎＼知识”.因此我们可以得到这样一个结论，对同一文本进行切分,如果使用的词典不同，会导致不同的分词结

果。显然,如果用第一种方法分词，当一个用户想要查找包含“搜索”这个关键字的相关资源时，他就不会搜索到结果。同理，假设检索系统不对用户输入的词进行切分,如果用第二种方法分词，当一个用户想要查找包含“搜索引擎”这个关键字的相关资源时，同样也找不到结果。所以，只进行一遍分词必然有一定得局限性,如果采用两遍、甚至多遍分词,便会解决上述问题.对于上面这个例子,我们采取组织两个词典的措施：一个为专业词典，一个为普通词典。其中，专业词典放置一些比较专业的词组,比如名人人名、专有名词、地点名、机构名等,普通词典就是我们常用的词组.那么我们可以将“搜索引擎 放入专业词典，将“搜索”、“引擎”放入普通词典。先用专业词典进行一遍分词,再用普通词典进行一遍分词，最后将结果合并到一起,那么结果如“搜索引擎＼搜索＼引擎＼知识”。这样既满足了查询“搜索引擎”的要求，又满足了查询“搜索”的要求.

据了解[6]，百度的分词采取了至少两个词典，一个是普通词典，一个是专用词典。而且是专用词典先切分，然后将剩余的片断交由普通词典来切分。一般专业的搜索引擎对分词速度要求要达到 1M/ s 以上,因此为了提高处理速度，百度的普通词典切分采用双向最大匹配算法，这种分词算法舍弃了一定得精度来达到极快的切分速度.因为对于搜索引擎来说，在查询切分和文档切分时采用相同的分词算法，如果有一些文档切分是分词是错误，在查询切分时也产生相同的切分错误。那么即使两次切分阶段错误,但最后相同错误却使匹配成功，使得仍然可以正确检索到结果。

2、基于字的切分法

现实中，无论一个词典所包含的词组有多么齐全，其还是包含不了一