**中文分词技术：**

1. 基于字符串匹配的分词方法

这种方法又叫做机械分词方法，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）。按照扫描方向的不同，串匹配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方法和分词与标注相结合的一体化方法。常用的几种机械分词方法如下：

正向最大匹配法（由左到右的方向）；

逆向最大匹配法（由右到左的方向）；

最少切分（使每一句中切出的词数最小）。

还可以将上述各种方法相互组合，例如，可以将正向最大匹配方法和逆向最大匹配方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，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，还需通过利用各 种其它的语言信息来进一步提高切分的准确率。

一种方法是改进扫描方式，称为特征扫描或标志切分，优先在待分析字符串中识别和切分出一些 带有明显特征的词，以这些词作为断点，可将原字符串分为较小的串再来进机械分词，从而减少匹配的错误率。另一种方法是将分词和词类标注结合起来，利用丰富 的词类信息对分词决策提供帮助，并且在标注过程中又反过来对分词结果进行检验、调整，从而极大地提高切分的准确率。

对于机械分词方法，可以建立一个一般的模型，在这方面有专业的学术论文，这里不做详细论述。

1. 基于理解的分词方法

这种分词方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。

1. 基于统计的分词方法

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。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。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。

1. 基于规则的分词方法

基于规则的分词方法，这种方法又叫做机械分词方法,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) 。常用的方法：最小匹配算法(Minimum Matching)，正向（逆向）最大匹配法(Maximum Matching)，逐字匹配算法,神经网络法、联想一回溯法，基于N-最短路径分词算法,以及可以相互组合，例如,可以将正向最大匹配方法和逆向最大匹 配方法结合起来构成双向匹配法等。目前机械式分词占主流地位的是正向最大匹配法和逆向最大匹配法。

最小匹配算法

在所有的分词算法中，最早研究的是最小匹配算法(Minimum Matching)，该算法从待比较字符串左边开始比较，先取前两个字符组成的字段与词典中的词进行比较，如果词典中有该词，则分出此词，继续从第三个字 符开始取两个字符组成的字段进行比较，如果没有匹配到，则取前3个字符串组成的字段进行比较，依次类推，直到取的字符串的长度等于预先设定的阈值，如果还 没有匹配成功，则从待处理字串的第二个字符开始比较，如此循环。

例如，“如果还没有匹配成功”，取出左边两个字 组成的字段与词典进行比较，分出“如果”；再从“还”开始，取“还没”，字典中没有此词，继续取“还没有”，依次取到字段“还没有匹配”(假设阈值为 5)，然后从“没”开始，取“没有”，如此循环直到字符串末尾为止。这种方法的优点是速度快，但是准确率却不是很高，比如待处理字符串为“中华人民共和 国”，此匹配算法分出的结果为：中华、人民、共和国，因此该方法基本上已经不被采用 。

最大匹配算法

基于字符串的最大匹配，这种方法现在仍比较常用。最大匹配(Maximum Matching)分为正向和逆向两种最大匹配，正向匹配的基本思想是：假设词典中最大词条所含的汉字个数为n个，取待处理字符串的前n个字作为匹配字 段，查找分词词典。若词典中含有该词，则匹配成功，分出该词，然后从被比较字符串的n+1处开始再取n个字组成的字段重新在词典中匹配；如果没有匹配成 功，则将这n个字组成的字段的最后一位剔除，用剩下的n一1个字组成的字段在词典中进行匹配，如此进行下去，直到切分成功为止。

例 如，待处理字符串为“汉字多为表意文字”，取字符串“汉语多为表”(假设比较的步长为5，本文步长step都取5)与词典进行比较，没有与之对应的词，去 除“表”字，用字段“汉语多为”进行匹配，直至匹配到“汉语”为至，再取字符串“多为表意”，循环到切分出“文字”一词。目前，正向最大匹配方法作为一种 基本的方法已被肯定下来，但是由于错误比较大，一般不单独使用。如字符串“处理机器发生的故障”，在正向最大匹配方法中会出现歧义切分，该字符串被分为： 处理机、发生、故障，但是使用逆向匹配就能得到有效的切分。

逆向最大匹配RMM(Reverse Directional Maximum Matching Method)的分词原理和过程与正向最大匹配相似，区别在于前者从文章或者句子(字串)的末尾开始切分，若不成功则减去最前面的一个字。比如对于字符串 “处理机器发生的故障”，第一步，从字串的右边取长度以步长为单位的字段“发生的故障”在词典中进行匹配，匹配不成功，再取字段“生的故障”进行匹配，依 次匹配，直到分出“故障”一词，最终使用RMM方法切分的结果为：故障、发生、机器、处理。该方法要求配备逆序词典。

一般来说根据汉语词汇构成的特点，从理论上说明了逆向匹配的精确度高于正向匹配，汉语语句的特点一般中心语偏后。有研究数据,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。实际应用中可以从下面几方面改进，同时采取几种分词算法，来提高正确率;改进扫描方式，称为特征扫描或标志 切分,优先在待分析字符串中识别和切分出一些带有明显特征的词,以这些词作为断点,可将原字符串分为较小的串再来进机械分词，从而减少匹配的错误率等。

逐字匹配算法

逐字匹配算法，基于TRIE索引树（又称单词索引树 http://baike.baidu.com/view/1436495.htm）的逐字匹配算法,是建立在树型词典机制上，匹配的过程是从索引树的根结点依次同步匹配待查词中的每个字，可以看成是对树 某一分枝的遍历。因此，采用该算法的分词速度较快，但树的构造和维护比较复杂。一种改进的算法是和最大匹配算法相结合，吸取最大匹配算法词典结构简单、 TRIE索引树算法查询速度快的优点。因此词典结构和最大匹配词典构造机制相似，区别在于词典正文前增加了多级索引。匹配过程类似TRIE索引树进行逐字 匹配，在性能上和TRIE索引树相近。

神经网络分词算法

神经网络分词算法，尹峰等提出了以神经网络理论(BP模型)为基础的汉语分词模型,为汉语分词研究开辟了新途径。在实用中,BP算法存在收敛速度慢、易陷 入局部最小等缺点,严重妨碍了分词速度。一种改进算法采用Levenbery2Marquart 算法来加速收敛速度,加快了收敛速度利用神经网络的基本原理进行分词。

联想—回溯法

联想—回溯法(Association－Backtracking Method，简称 AB 法)。这种方法要求建立三个知识库——特征词词库、实词词库和规则库。首先将待切分的汉字字符串序列按特征词词库分割为若干子串，子串可以是词，也可以是 由几个词组合而成的词群；然后，再利用实词词库和规则库将词群再细分为词。切词时，要利用一定的语法知识，建立联想机制和回溯机制。联想机制由联想网络和 联想推理构成，联想网络描述每个虚词的构词能力，联想推理利用相应的联想网络来判定所描述的虚词究竟是单独成词还是作为其他词中的构词成分。回溯机制主要 用于处理歧义句子的切分。联想—回溯法虽然增加了算法的时间复杂度和空间复杂度，但这种方法的切词正确率较高，是一种行之有效的方法。

N-最段路径分词算法

基于N-最短路径分词算法，其基本思想是根据词典，找出字串中所有可能的词，构造词语切分有向无环图。每个词对应图中的一条有向边，并赋给相应的边长(权 值)。然后针对该切分图，在起点到终点的所有路径中，求出长度值按严格升序排列(任何两个不同位置上的值一定不等，下同)依次为第1，第2，…，第 i，…，第N的路径集合作为相应的粗分结果集。如果两条或两条以上路径长度相等，那么他们的长度并列第 i，都要列入粗分结果集，而且不影响其他路径的排列序号，最后的粗分结果集合大小大于或等于N。N一最短路径方法实际上是最短路径方法和全切分的有机结 合。该方法的出发点是尽量减少切分出来的词数，这和最短路径分词方法是完全一致的；同时又要尽可能的包含最终结果，这和全切分的思想是共通的。通过这种综 合，一方面避免了最短路径分词方法大量舍弃正 确结果的可能，另一方面又大大解决了全切分搜索空间过大，运行效率差的弊端。N一最短路径方法相对的不足就是粗分结果不唯一 ，后续过程需要处理多个粗分结果。 但是 ，对于预处理过程来讲，粗分结果的高召回率至关重要。因为低召回率就意味着没有办法 再作后续的补救措施。预处理一旦出错，后续处理只能是一错再错 ，基本上得不到正确的最终 结果。而少量的粗分结果对后续过程的运行效率影响不会太大，后续处理可以进一步优选排 错，如词性标注、句法分析等。

除上面之外，还有基于词频统计的切词法， 基于期望的切词法，有穷多级列举法等。

分词中的难题

有了成熟的分词算法，是否就能容易的解决中文分词的问题呢？事实远非如此。中文是一种十分复杂的语言，让计算机理解中文语言更是困难。在中文分词过程中，有两大难题一直没有完全突破。

1、歧义识别

歧义是指同样的一句话，可能有两种或者更多的切分方法。例如：表面的，因为“表面”和“面的”都是词，那么这个短语就可以分成“表面 的”和“表 面的”。这种称为交叉歧义。像这种交叉歧义十分常见，前面举的“和服”的例子，其实就是因为交叉歧义引起的错误。“化妆和服装”可以分成“化妆 和 服装”或者“化妆 和服 装”。由于没有人的知识去理解，计算机很难知道到底哪个方案正确。

交叉歧义相对组合歧义来说是还算比较容易处 理，组合歧义就必需根据整个句子来判断了。例如，在句子“这个门把手坏了”中，“把手”是个词，但在句子“请把手拿开”中，“把手”就不是一个词；在句子 “将军任命了一名中将”中，“中将”是个词，但在句子“产量三年中将增长两倍”中，“中将”就不再是词。这些词计算机又如何去识别?

如 果交叉歧义和组合歧义计算机都能解决的话，在歧义中还有一个难题，是真歧义。真歧义意思是给出一句话，由人去判断也不知道哪个应该是词，哪个应该不是词。 例如：“乒乓球拍卖完了”，可以切分成“乒乓 球拍 卖 完 了”、也可切分成“乒乓球 拍卖 完 了”，如果没有上下文其他的句子，恐怕谁也不知道“拍卖”在这里算不算一个词。

2、新词识别

新词，专业术语称为未登 录词。也就是那些在字典中都没有收录过，但又确实能称为词的那些词。最典型的是人名，人可以很容易理解句子“王军虎去广州了”中，“王军虎”是个词，因为 是一个人的名字，但要是让计算机去识别就困难了。如果把“王军虎”做为一个词收录到字典中去，全世界有那么多名字，而且每时每刻都有新增的人名，收录这些 人名本身就是一项巨大的工程。即使这项工作可以完成，还是会存在问题，例如：在句子“王军虎头虎脑的”中，“王军虎”还能不能算词？

新词中除了人名以外，还有机构名、地名、产品名、商标名、简称、省略语等都是很难处理的问题，而且这些又正好是人们经常使用的词，因此对于搜索引擎来说，分词系统中的新词识别十分重要。目前新词识别准确率已经成为评价一个分词系统好坏的重要标志之一。

中文分词的应用

目前在自然语言处理技术中，中文处理技术比西文处理技术要落后很大一段距离，许多西文的处理方法中文不能直接采用，就是因为中文必需有分词这道工序。中文分词是其他中文信息处理的基础，搜索引擎只是中文分词的一个应用。其他的比如机器翻译（MT）、语音合成、自动分类、自动摘要、自动校对等等，都需要用到分词。因为中文需要分词，可能会影响一些研究，但同时也为一些企业带来机会，因为国外的计算机处理技术要想进入中国市场，首先也是要解决中文分词问题。在中文研究方面，相比外国人来说，中国人有十分明显的优势。

分词准确性对搜索引擎来说十分重要，但如果分词速度太慢，即使准确性再高，对于搜索引擎来说也是不可用的，因为搜索引擎需要处理数以亿计的网页，如果分 词耗用的时间过长，会严重影响搜索引擎内容更新的速度。因此对于搜索引擎来说，分词的准确性和速度，二者都需要达到很高的要求。目前研究中文分词的大多是科研院校，清华、北大、中科院、北京语言学院、东北大学、IBM研究院、微软中国研究院等都有自己的研究队伍，而真正专业研究中文分词的商业公司除了海量科技以外，几乎没有了。科研院校研究的技术，大部分不能很快产品化，而一个专业公司的力量毕竟有限，看来中文分词技术要想更好的服务于更多的产品，还有很长一段路。