**中文搜索引擎技术揭密：中文分词**

作者[Winter](http://stlchina.org/twiki/bin/view.pl/Main/WinterWen)  
首发于:[e800.com.cn](http://ditital.e800.com.cn/articles/98/1091788186451.html)

# 1前言

信息的飞速增长，使搜索引擎成为人们查找信息的首选工具，Google、百度、yisou、中搜等大型搜索引擎一直人们讨论的话题。随着搜索市场价值的不断增加，越来越多的公司开发出自己的搜索引擎，阿里巴巴的商机搜索、8848的购物搜索等也陆续面世，自然，搜索引擎技术也成为技术人员关注的热点。

搜索引擎技术的研究，国外比中国要早近十年，从最早的Archie，到后来的Excite，以及altvista、overture、google等搜索引擎面世，搜索引擎发展至今，已经有十几年的历史，而国内开始研究搜索引擎是在上个世纪末本世纪初。在许多领域，都是国外的产品和技术一统天下，特别是当某种技术在国外研究多年而国内才开始的情况下。例如操作系统、字处理软件、浏览器等等，但搜索引擎却是个例外。虽然在国外搜索引擎技术早就开始研究，但在国内还是陆续涌现出优秀的搜索引擎，像百度、中搜等。目前在中文搜索引擎领域，国内的搜索引擎已经和国外的搜索引擎效果上相差不远。之所以能形成这样的局面，有一个重要的原因就在于中文和英文两种语言自身的书写方式不同，这其中对于计算机涉及的技术就是中文分词。

# 2什么是中文分词

众所周知，英文是以词为单位的，词和词之间是靠空格隔开，而中文是以字为单位，句子中所有的字连起来才能描述一个意思。例如，英文句子I am a student，用中文则为：“我是一个学生”。计算机可以很简单通过空格知道student是一个单词，但是不能很容易明白「学」、「生」两个字合起来才表示一个词。把中文的汉字序列切分成有意义的词，就是中文分词，有些人也称为切词。我是一个学生，分词的结果是：我 是 一个 学生。

# 3中文分词和搜索引擎

中文分词到底对搜索引擎有多大影响？对于搜索引擎来说，最重要的并不是找到所有结果，因为在上百亿的网页中找到所有结果没有太多的意义，没有人能看得完，最重要的是把最相关的结果排在最前面，这也称为相关度排序。中文分词的准确与否，常常直接影响到对搜索结果的相关度排序。笔者最近替朋友找一些关于日本和服的资料，在搜索引擎上输入「和服」，得到的结果就发现了很多问题。下面就以这个例子来说明分词对搜索结果的影响，在目前最有三个中文搜索引擎上做测试。测试方法是直接在[Google](http://www.google.com)、[百度](http://www.baidu.com)、[中搜](http://www.zhongsou.com)上以「和服」为关键词进行搜索：

在Google上输入「和服」搜索所有中文简体网页，总共结果507,000条，前20条结果中有14条与和服一点关系都没有。在第一页就有以下错误：

“通信信息报：瑞星以技术和服务开拓网络安全市场”

“使用纯HTML的通用数据管理和服务- 开发者- ZDNet ...”

“陈慧琳《心口不一》 化妆和服装自己包办”

“::外交部：中国境外领事保护和服务指南(2003年版) ...”

“产品和服务”

等等。第一页只有三篇是真正在讲「和服」的结果。

在百度上输入「和服」搜索网页，总共结果为287,000条，前20条结果中有6条与和服一点关系都没有。在第一页有以下错误：

“福建省晋江市恒和服装有限公司系独资企业”

“关于商品和服务实行明码标价的规定”

“青岛东和服装设备”

在中搜山输入「和服」搜索网页，总共结果为26,917条，前20条结果都是与和服相关的网页。

这次搜索引擎结果中的错误，就是由于分词的不准确所造成的。通过笔者的了解，Google的中文分词技术采用的是美国一家名叫Basis Technology（http://www.basistech.com）的公司提供的中文分词技术，百度使用的是自己公司开发的分词技术，中搜使用的是国内海量科技（http://www.hylanda.com）提供的分词技术。业界评论海量科技的分词技术目前被认为是国内最好的中文分词技术，其分词准确度超过99%，由此也使得中搜在搜索结果中搜索结果的错误率很低。由此可见，中文分词的准确度，对搜索引擎结果相关性和准确性有相当大的关系。

# 4中文分词技术

中文分词技术属于自然语言处理技术范畴，对于一句话，人可以通过自己的知识来明白哪些是词，哪些不是词，但如何让计算机也能理解？其处理过程就是分词算法。

现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

#### 4.1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：

1. 正向最大匹配法（由左到右的方向）；
2. 逆向最大匹配法（由右到左的方向）；
3. 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。

#### 4.2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。

#### 4.3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。

到底哪种分词算法的准确度更高，目前并无定论。对于任何一个成熟的分词系统来说，不可能单独依靠某一种算法来实现，都需要综合不同的算法。笔者了解，海量科技的分词算法就采用「复方分词法」，所谓复方，相当于用中药中的复方概念，即用不同的药才综合起来去医治疾病，同样，对于中文词的识别，需要多种算法来处理不同的问题。

# 5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。

#### 5.1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。

#### 5.2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。

# 6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。

<http://stlchina.org/twiki/bin/view.pl/Main/SESegment>

**参考**   
1. [The Anatomy of a Large-Scale Hypertextual Web Search Engine](http://www-db.stanford.edu/~backrub/google.html)   
2. 召回率: recall. 即得到的正确结果占所有应该得到的正确结果的比例。如：包含'雅虎'的正确的网页应该有500个，但搜索得到了600个结果，其中有400个是正确的，还有200个是错误的。那么准确度是:400/600=66.67%, 召回率是：400/500=80%.