指导:中文搜索引擎技术揭密：中文分词 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/64/2021_2022__E6_8C_87_E5_AF_BC__E4_B8_AD_c40_64660.htm 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、中国搜索等大型搜索引擎一直是人们讨论的话题。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（http://www.baidu.com/）、中搜（http://www.zhongsou.com/）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。 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。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，在现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 ” 在中搜上输入 “ 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。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。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 100Test 下载频道开通，各类考试题目直接下载。详细请访问 www.100test.com