**中文分词技术总结**

一、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。

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。

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。  
（以上内容摘录自附录1）

二、计算所汉语词法分析系统 ICTCLAS

中国科学院计算技术研究所在多年研究基础上，耗时一年研制出了基于多层隐马模型的汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)，该系统的功能有：中文分词；词性标注；未登录词识别。分词正确率高达97.58%(最近的973专家组评测结果)，基于角色标注的未登录词识别能取得高于90%召回率，其中中国人名的识别召回率接近98%，分词和词性标注处理速度为31.5KB/s。ICTCLAS 和计算所其他14项免费发布的成果被中外媒体广泛地报道，国内很多免费的中文分词模块都或多或少的参考过ICTCLAS的代码。

下载页面： http://www.nlp.org.cn/project/project.php?proj\_id=6

由于 ICTCLAS 是由 C 语言写成的，现在主流的开发工具用起来不太方便，于是有一些热心的程序员把 ICTCLAS 改为 Java 和 C# 等其他语言。

（1）fenci，Java 的 ICTCLAS，下载页面： http://www.xml.org.cn/printpage.asp?BoardID=2&id=11502

（2）AutoSplit，另一个 Java 的 ICTCLAS，已经找不到下载页面，点击本地下载

（3）小叮咚中文分词，曾经有下载页面，现在找不到了。据作者介绍，从 ICTCLAS 中改进，有 Java，C# 和 C++ 三个版本，介绍页面： http://www.donews.net/accesine

三、海量智能分词研究版

海量智能计算技术研究中心为了使中文信息处理领域的研究者们能够共同分享海量智能中心的研究成果，共同提高中文信息处理水平，特此发布《海量智能分词研究版》，供专家、学者和爱好者进行研究。

下载页面： http://www.hylanda.com/cgi-bin/download/download.asp?id=8

四、其他

（1）CSW中文智能分词组件

运行环境：Windows NT、2000、XP 或更高，可以在 ASP，VB 等微软的开发语言中调用。

简介: CSW中文智能分词DLL组件，可将一段文本自动的按常规汉语词组进行拆分,并以指定方式进行分隔，且可对其拆分后的词组进行语义、词频标注。其广范应用于各行各业的信息资料检索、分析。

下载页面： http://www.vgoogle.net/

（2） C# 写的中文分词组件

据作者介绍，一个 DLL 文件，可以做中英文分词组件。完全C#托管代码编写，独立开发。

下载页面： http://www.rainsts.net/article.asp?id=48

附录：  
1. Winter；中文搜索引擎技术揭密：中文分词；http://www.e800.com.cn/articles/98/1091788186451.html；2004-04-19.