电子商务员辅导之搜索引擎之中文分词技术电子商务考试
 中文分词是将一句话或一个短语按照日常阅读习惯进行机械分解。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，我很喜欢搜索引擎，分词的结果是：我|很喜欢|搜索引擎。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。 中文每个字都可以直接作为一个词来使用，没有断词，正因为此它才多变。虽然多变，但是在表达上灵活。但是对于搜索引擎来说这是非常难以解决的问题。在中文分词当中，有三种难分类型。 1、交集型歧义 假设“ ABC ”是一个由A、B、C三个汉字构成的字串，如果“ AB ”、“ BC ”都是词，那么计算机在切分时可以把“ ABC ”切分为“ AB/C ”，也可以切分为“ A/BC ”。这种切分歧义称为交集型歧义。 2、组合型歧义 如果“ AB ”是词、“ ABC ”也是词，那么产生的切分歧义称为组合型歧义。 3、混和型歧义 混和型歧义是包含交集型歧义和组合型歧义的切分歧义。 目前解决这些问题主要通过字典和统计学的方法。 首先我们先说说字典分词法。考|试/大字典一般采用前缀树和后缀树的数据存储结构。什么是前缀树呢？其实就是我们把一个句子从左向右扫描一遍，遇到字典里有的词就标识出来，遇到复合词就找最长的词匹配，遇到不认识的字串就分割成单字词，于是简单的分词就完成了。后缀树就是从右向左扫描一遍。 统计学的方法，虽然字典分词已经解决了很多分词上出现的问题。但是面对

很多新出的词汇，分词也面临着挑战。统计学的分词方式是基于概念和信息学方面的知识进行处理。基本原理就是寻找那些经常一同出现的字，总是相互的字很有可能构成一个词。为此需要分析大量内容。考|试/大即使到现在中文分词还在不断发展，还没有一个分词方法能彻底解决一切问题。F8F8"