电子商务师考试辅导之网络数据的挖掘电子商务师考试 PDF 转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/631/2021_2022__E7_94_B5_E5_AD_90_E5_95_86_E5_c40_631219.htm 搜索引擎（SearchEngine）是随着WEB信息的迅速增加，从1995年开始逐渐发展起来的技术。用户在海量的数据里查找所需要的信息，自然会有“ 迷航 ”问题，搜索引擎正是为了解决这个问题而出现的技术。搜索引擎以一定的策略在互联网中搜集、发现信息，对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而起到信息导航的目的。 1、网络信息采集 考虑到效率问题，搜索引擎不可能在用户搜索时才去实时的检查每个网页，因而需要把网页先抓取下来，按照关键词建立好索引，考|试/大每次搜索的结果都会直接从搜索引擎建立好索引的数据库中查找，然后把结果返回给访问者。而抓取信息的任务就由网络爬行器或网络蜘蛛来完成。 网络蜘蛛是通过网页的链接地址来寻找网页，从某一个页面开始，读取网页的内容，找到在网页中的其它链接地址，然后通过这些链接地址寻找下一个网页，这样一直循环下去。如果把整个互联网当成一个内部相互关联的网站，那么网络蜘蛛就可以用这个原理把互联网上所有的网页都抓取下来。在爬行的下一步选择上，网络蜘蛛一般有两种策略：广度优先和深度优先。 作为网络蜘蛛，在爬行中除了获得网页的数据外，还有一项重要的工作：从网页的数据中提取出新的链接，以维持爬行的继续进行。一般的网页内容都是HTML类的描述性语言，具有自身的语法。显然，这项工作应由一个语法分析器来完成。 搜索引擎建立网页索引，处理的对象是文本文件

。对于网络蜘蛛来说，抓取下来的数据存在各种格式，考|试/大包括html、图片、doc、pdf、多媒体、动态网页及其它格式等。这些文件抓取下来后，还需要把这些文件中的文本信息提取出来。准确提取这些文档的信息，一方面对搜索引擎的搜索准确性具有重要作用，另一方面对于网络蜘蛛正确跟踪其它链接有一定影响。 2、数据存储与搜索 获得数据后，如何将这些数据进行存储以及如何为用户提供服务，本文将较详细地介绍关于用户在进行数据查询及搜索时所涉及到的核心问题：网页在搜索结果中的排名和中文分词技术。用户在提交被查询的关键字后，希望所得到的结果集是自己需要的内容，至少在得到的结果集中，开头几条或十几条所包含的内容基本是用户所期望和关心的。考|试/大实际上，考|试/大网页本身已经包含了很多关于网页内容及网页重要性的描述信息，如：网页的标题，HTML报文头的META描述，以及前面提到的加粗表示的重点关键字等等。充分利用这些信息，是可以给出一种排序的算法对用户搜索的结果集进行排名，以满足要求。以下给出在网页的权值运算中典型的算法。（1）Google和PageRank算法搜索引擎Google最初是斯坦福大学的博士研究生SergeyBrin和LawrencePage实现的一个原型系统，现在已经发展成为WWW上最好的搜索引擎之一。它与传统的搜索引擎最大的不同在于对网页进行了基于权威值的排序处理，使最重要的网页出现在结果的最前面。Google通过PageRank元算法计算出网页的PageRank值，从而决定网页在结果集中的出现位置：PageRank值越高的网页，在结果中出现的位置越靠前。 （2）PageRank算法PageRank算法基于下面2个前提： 前提1：一个网页被多次引用，则它可能是很重

要的；一个网页虽然没有被多次引用，但是被重要的网页引用，则它也可能是很重要的；考|试/大一个网页的重要性被平均的传递到它所引用的网页。这种重要的网页称为权威网页。 前提2：假定用户一开始随机的访问网页集合中的一个网页，以后跟随网页的向外链接向前浏览网页，考|试/大不回退浏览，浏览下一个网页的概率就是被浏览网页的PageRank值。 简单PageRank算法描述如下：u是一个网页，F(u)是u指向的网页集合，B(u)是指向u的网页集合，N(u)是u指向外的链接数，显然N(u)=|F(u)|，c是一个用于规范化的因子（Google通常取0.85），另外还有一些特殊的链接，指向的网页没有向外的链接。PageRank计算时，把这种链接首先除去，等计算完以后再加入，这对原来计算出的网页的rank值影响是很小的。Pagerank算法除了对搜索结果进行排序外，还可以应用到其它方面，如估算网络流量，向后链接的预测器，为用户导航等。 3、中文分词技术 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Iamastudent，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。考|试/大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。 中文分词技术属于自然语言处理技术范畴，对于一句话，人可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。现有分词算法大致可分为三类： （1）基于字符串匹配的分词方法：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）。（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。（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。考|试/大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。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，都需要综合不同算法。 4、结论 要实现一个真正的引擎所涉及的问题还包括：被搜网站的诊断，流量分析，引擎的优化（SEO）等等。另外，针对服务的时效及质量，还应考虑数据的更新周期，多语种的兼容及转换等。 随着搜索引擎的发展，各种新型的技术已经被使用。今天的搜索引擎所包含的数据更复杂，提供的服务也更广泛，从图片到音视频数据都成为了用户的搜索对象，准确性与可信度也越来越高。随着搜索市场价值的不断增加，越来越多的公司继已开发出了自己的搜索引擎。实际上，搜索引擎经济的崛起，又一次向人们证明网络所蕴藏的巨大商机。编者推荐：电子商务师考试－电子商务员辅导电子商务师考试－综合辅导 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