简谈搜索引擎工作流程 PDF转换可能丢失图片或格式，建议阅读原文

https://www.100test.com/kao_ti2020/62/2021_2022__E7_AE_80_E8_B0_88_E6_90_9C_E7_c40_62562.htm 互联网是一个宝库，搜索引擎是打开宝库的一把钥匙。然而，绝大多数网民在搜索引擎的相关知识及使用技巧上能力不足。国外的一次调查结果显示，约有71%的人对搜索的结果感到不同程度的失望。作为互联网的第二大服务，这种状况应该改变。互联网的迅速发展，导致了网上信息的爆炸性增长。全球目前的网页超过20亿，每天新增加730万网页。要在如此浩瀚的信息海洋里寻找信息，就像“ 大海捞针 ”一样困难。搜索引擎正是为了解决这个“ 迷航 ”问题而出现的技术。搜索引擎的工作包括如下三个过程：1.在互联中发现、搜集网页信息；2.对信息进行提取和组织建立索引库；3.再由检索器根据用户输入的查询关字，在索引库中快速检出文档，进行文档与查询的相关度评价，对将要输出的结果进行排序，并将查询结果返回给用户。发现、搜集网页信息 需要有高性能的“ 网络蜘蛛 ”程序(Spider)去自动地在互联网中搜索信息。一个典型的网络蜘蛛工作的方式，是查看一个页面，并从中找到相关信息, 然后它再从该页面的所有链接中出发，继续寻找相关的信息，以此类推，直至穷尽。网络蜘蛛要求能够快速、全面。网络蜘蛛为实现其快速地浏览整个互联网，通常在技术上采用抢先式多线程技术实现在网上聚集信息。通过抢先式多线程的使用，你能索引一个基于URL链接的Web页面，启动一个新的线程跟随每个新的URL链接，索引一个新的URL起点。当然在服务器上所开的线程也不能无限膨胀，需要在服务器的正

常运转和快速收集网页之间找一个平衡点。在算法上各个搜索引擎技术公司可能不尽相同，但目的都是快速浏览Web页和后续过程相配合。目前国内的搜索引擎技术公司中，比如百度公司的网络蜘蛛采用了可定制、高扩展性的调度算法使得搜索器能在极短的时间内收集到最大数量的互联网信息，并把所 获得的信息保存下来以备建立索引库和用户检索。索引库的建立 关系到用户能否最迅速地找到最准确、最广泛的信息，同时索引库的建立也必须迅速，对网络蜘蛛抓来的网页信息极快地建立索引，保证信息的及时性。对网页采用基于网页内容分析和基于超链分析相结合的方法进行相关度评价，能够客观地对网页进行排序，从而极大限度地保证搜索出的结果与用户的查询串相一致。新浪搜索引擎对网站数据建立索引的过程中采取了按照关键词在网站标题、网站描述、网站URL等不同位置的出现或网站的质量等级等建立索引库，从而保证搜索出的结果与用户的查询串相一致。新浪搜索引擎在索引库建立的过程中，对所有数据采用多进程并行的方式，对新的信息采取增量式的方法建立索引库，从而保证能够迅速建立索引，使数据能够得到及时的更新。新浪搜索引擎在建立索引库的过程中还对用户搜索的查询串进行跟踪，并对查询频率高的查询串建立Cache页。用户检索的过程这是对前两个过程的检验，检验该搜索引擎能否给出最准确、最广泛的信息，检验该搜索引擎能否迅速地给出用户最想得到的信息。对于网站数据的检索，新浪搜索引擎采用Client/Server结构、多进程的方式在索引库中检索，大大减少了用户的等待时间，并且在用户查询高峰时服务器的负担不会过高（平均的检索时间在0.3秒左右）。对于网页信息的

检索，作为国内众多门户网站的网页检索技术提供商的百度公司其搜索引擎运用了先进的多线程技术，采用高效的搜索算法和稳定的UNIX平台，因此可大大缩短对用户搜索请求的响应时间。作为慧聪I系列应用软件产品之一的I-Search2000采用的超大规模动态缓存技术，使一级响应的覆盖率达到75%以上，独有的自学习能力可自动将二级响应的覆盖率扩充到20%以上。 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