一：搜索引擎工作原理介绍

搜索引擎的工作过程大体上分成三个阶段

1.爬行和抓取：搜索引擎蜘蛛通过跟踪链接访问网页，获得页面HTML代码存入。

2.预处理：索引程序对抓取到的页面数据进行文字提取、中文分词、索引处理，以备排名程序调用。

3.排名：用户输入关键词后，排名程序调用索引库中的数据。计算相关性，然后按一定格式生存搜索结果页面。

1.1：爬行和抓取

1.1.1：蜘蛛

搜索引擎用来爬行和访问页面的程序，称之为蜘蛛。

蜘蛛访问任何网站，都会先访问网站根目录下的Robots.txt文件。如果Robots.txt文件禁止蜘蛛抓取某些文件，蜘蛛将遵守协议，不抓取被禁止的网址。

站长可通过日志文件，看到有哪些蜘蛛来爬过。常见的搜索引擎蜘蛛有：百度蜘蛛，雅虎中国蜘蛛，英文雅虎蜘蛛，Google蜘蛛，微软bing蜘蛛，搜狗蜘蛛，搜搜蜘蛛，有道蜘蛛。

1.1.2跟踪链接

整个互联网是由相互链接的网站及页面组成的（外链重要性的体现）。从理论上讲，蜘蛛从任何一个页面出发，顺着链接都可以爬到网上的所有页面。

最简单的爬行遍历策略分为两种，一种是深度优先，一种是广度优先。

深度优先：蜘蛛从A页面的一个链接一直爬，爬到无法前进的时候，才返回A页面，再开始访问第二个链接。直至爬完。

广度优先：蜘蛛先将A页面的所有页面爬完，再继续第二个页面爬行与抓取，直至爬完。

1.1.3如何吸引蜘蛛

实际上蜘蛛不会抓取所有的页面。那么就需要去吸引它来抓取。

网站和页面权重：质量高、资格老的网站被认为是权重比较高的，且爬行的深度也就比较深。

页面更新度：若蜘蛛多次抓取内容相同，蜘蛛就会认为这个网站没有在更新，就会降低抓取频率。

导入链接：指的是进入页面的入口。

距离首页要近:因为首页经常是这个网站的入口，且权重最高。

2.预处理

搜索引擎可以识别Meta标签中的文字、图片替代文字、Flash文件的替代文字、链接锚文字…

3.排名

经过搜索引擎蜘蛛抓取页面，索引程序计算得到倒排索引后，搜索引擎就准备好随时处理用户搜索了。用户在搜索框填入关键词后，排名程序调用索引库数据，计算排名显示给用户，排名过程是与用户直接互动的。