搜索引擎工作原理

基本流程

抓取网页：每个独立的搜索引擎都有自己的网页抓取程序爬虫（Spider）。爬虫顺着网页中的超链接，从这个网站爬到另一个网站，通过超链接分析连续访问抓取更多网页。被抓取的网页被称之为网页快照。由于互联网中超链接的应用很普遍，理论上，从一定范围的网页出发，就能搜集到绝大多数的网页。

处理网页：搜索引擎抓到网页后，还要做大量的预处理工作，才能提供检索服务。其中，最重要的就是提取关键词，建立索引库和索引。其他还包括去除重复网页、分词（中文）、判断网页类型、分析超链接、计算网页的重要度/丰富度等。

提供检索服务：用户输入关键词进行检索，搜索引擎从索引数据库中找到匹配该关键词的网页；为了用户便于判断，除了网页标题和 URL 外，还会提供一段来自网页的摘要以及其他信息。

搜索引擎的自动信息搜集功能

提交网站搜索，站长主动向搜索引擎提交网址，它在一定时间内定向向你的网站派出爬虫，扫描你的网站并将有关信息存入数据库，以备用户查询。由于搜索引擎索引规则相对于过去已发生很大变化，主动提交网址并不保证你的网站能进入搜索引擎数据库，因此站长应该在网站内容上多下功夫，并让搜索引擎有更多机会找到你并自动将你的网站收录。

当用户以关键词查找信息时，搜索引擎会在数据库中进行搜寻，如果找到与用户要求内容相

符的网站，便采用特殊的算法——通常根据网页中关键词的匹配程度，出现的位置、频次，链接质量等——计算出各网页的相关度及排名等级，然后根据关联度高低，按顺序将这些网页链接返回给用户。

温馨提示

我们想说的是您应该将您优化的重心和出发点主要放在用户体验上，因为用户才是您网站内容的主要受众，是他们通过搜索引擎找到了您的网站。过度专注于用特定的技巧获取搜索引擎自然搜索结果的排名不一定能够达到您想要的结果。

网站优化基本概念

搜索引擎优化 Search Engine Optimization

定义：是一种利用搜索引擎的搜索规则来提高目的网站在有关搜索引擎内的排名的方式。主要工作原则是，通过了解各类搜索引擎抓取互联网页面、进行索引以及确定其对特定关键词搜索结果排名等技术，来对网页进行相关的优化。

注：请不要针对搜索引擎而采用作弊行为，否则会容易受到处罚。仅仅是模仿甚至抄袭别人的内容，这样对用户来说没有价值的。请牢记：为用户，而不是为搜索引擎提供内容。您网站的设计首先要考虑用户的需求，并同时确保能便于搜索引擎抓取和索引。

站点地图 Sitemap

定义：sitemap 可方便网站管理员通知搜索引擎他们网站上有哪些可供抓取的网页。常见的 sitemap 文件，就是 txt、xml、xml 一级索引这三种格式文件，在其中列出网站中的网址以及

关于每个网址的其他元数据（上次更新的时间、更改的频率以及相对于网站上其他网址的重要程度为何等），以便搜索引擎可以更加智能地抓取网站。

Robots 协议

定义：Robots 协议（也称为爬虫协议、爬虫规则、机器人协议等）也就是 robots.txt，网站通过 robots 协议告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取。Robots 协议是网站国际互联网界通行的道德规范，其目的是保护网站数据和敏感信息、确保用户个人信息和隐私不被侵犯。因其不是命令，故需要搜索引擎自觉遵守。您可以使用 robots.txt 禁止 spider 抓取您不想向用户展现的形式，这也有助于节省您的带宽。

元标签

定义：是使用在网页的 head 标签之间的一种 HTML 标签，主要包括关键词标签和描述标签，现在最常用的也是这两类。与其它的 HTML 标签不同，元标签不会在页面的任何地方显示出来，所以绝大多数的访问者并不会看到它的存在，而且对网站的权威度没有影响，不过仍然是有好处的，特别是在与搜索引擎的 spider 交流的时候。不同的元标签起着不同的作用——但均用来提供关于页面的附加信息。

网页标题 Title

定义：是对一个网页的高度概括，一般来说，网站首页的标题就是网站的正式名称，而网站中文章内容页面的标题就是文章的题目，栏目首页的标题通常是栏目名称。您网站首页的标题可以列出网站或者公司名称和其他一些重要的信息，诸

如您公司的实际地址，一些主要关注的领域或者提供的服务。

助君网络建议站长不要利用网页标题进行恶意作弊，类似于一些网站正文内容与标题不符，或者标题过长、关键词堆砌的网站，搜索引擎不会保证收录，甚至可能处罚。

面包屑导航

定义：是指在网页顶端或者底部放置的一排内部链接，它使用户可以方便地回到上一层结构中的网页或者主页。大多数面包屑导航通常会从最具概括性的页面开始（通常是主页），越往右指向的页面内容越具体。

Alt 属性

定义：是一个用于网页语言 HTML 和 XHTML、为输出纯文字的参数属性。它的作用是当 HTML 元素本身的物件无法被渲染时，就显示 alt（替换）文字作为一种补救措施。当图片因为一些原因不能够显示的时候，alt 属性使您可以指定供替换显示的文字。

为什么使用这个属性呢？如果一个用户在浏览您的网站的时候使用的浏览器不支持图片，或者用户在使用一些类似于屏幕阅读器的设备时，alt 属性的内容就可以提供关于图片的信息。除此之外，使用 alt 属性还有另一个原因。如果您把一张图片作为一个链接，此时这个图片的 alt 属性就能起到与文本链接的锚文本相同的作用。

锚文本 Anchor Text

就是链接文本，是链接的一种形式，即是链接上可以被点击的文字。锚文本可以做为锚文本所在页面内容的评估。正常来讲，页面中增加的链接都会和页面本身的内容有一定的关系。

Heading 标签

定义：Heading 标签也叫做 H 标签，HTML 语言里一共有六种大小的 heading 标签，从最重要的<h1>到<h6>，权重依次降低。是网页 html 中对文本标题所进行的着重强调的一种标签。

Http 状态码

301：（永久移动）请求的网页已永久移动到新位置。服务器返回此响应（对 GET 或 HEAD 请求的响应）时，会自动将请求者转到新位置。

302：代表暂时性转移(Temporarily Moved)。

403：资源不可用。服务器理解客户的请求，但拒绝处理它。通常由于服务器上文件或目录的权限设置导致，比如 IIS 或者 apache 设置了访问权限不当。

404：请求失败，请求所希望得到的资源未被在服务器上发现。404 这个状态码被广泛应用于当服务器不想揭示到底为何请求被拒绝或者没有其他适合的响应可用的情况下。出现这个错误的最有可能的原因是服务器端没有这个页面。