第 7 章

百 度 篇

百度是国内访问量最大的搜索引擎。百度优化是 SEO 从业者的重点。

7.1 百度搜索引擎的工作原理

在国内要想做好 SEO，百度搜索引擎是必须要重点研究的对象。百度公司通过百度搜索资源平台对百度搜索算法进行了比较系统的解读。本章内容基于百度搜索资源平台，对该平台的内容进行了整理，以帮助读者更好地理解该平台的特点和优化逻辑。

有关百度搜索引擎的权威资料，可以通过百度搜索资源平台 https://ziyuan.baidu.com/ 找到。

7.1.1 抓取建库

1. 百度蜘蛛抓取系统的基本框架

现在互联网上的信息呈爆发式增长，如何有效地获取并利用这些信息，是搜索引擎工作中的首要环节。数据抓取系统作为整个搜索系统中的第一个环节，主要负责互联网信息的搜集、保存和更新，它像蜘蛛一样在网页中爬来爬去获取信息，因此被叫作蜘蛛或者网络爬虫。

蜘蛛抓取系统是搜索引擎数据来源的重要保证，如果把 Web 理解为一个有向图，那么蜘蛛的工作过程可以认为是对这个有向图的遍历。从一些重要的种子 URL 开始，通过页面上的超链接，不断地发现新的 URL 并抓取，尽最大可能抓取到更多有价值的网页。对于类似百度这样的大型蜘蛛系统，因为每时每刻都存在网页被修改、删除或出现新的超链接的可能，因此，还要对蜘蛛过去抓取过的页面保持更新，维护一个 URL 库和页面库。

图 7-1 所示为百度蜘蛛抓取系统的基本框架图，包括总链接库、链接选取系统、内部 DNS 解析服务、抓取系统、抓取回网页解析等。百度蜘蛛即是通过这种系统的通力合作完成对互联网页面的抓取工作。

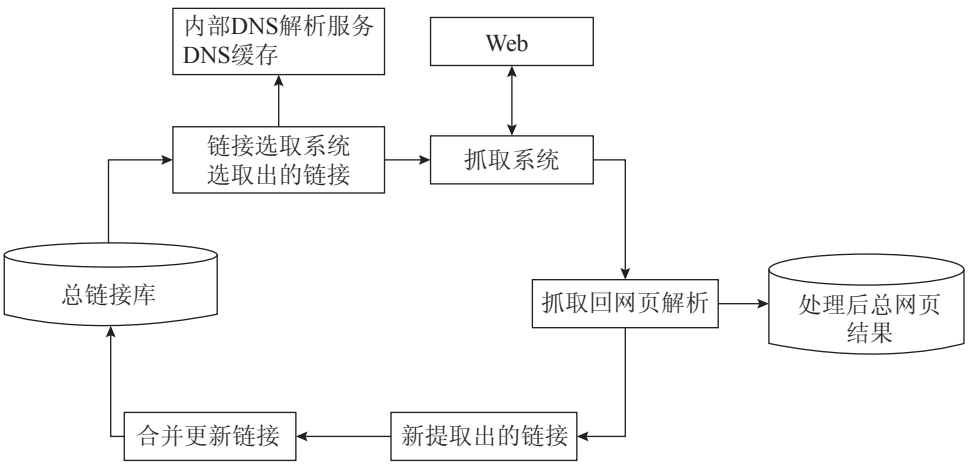

图 7-1　百度蜘蛛抓取系统框架图

2. 百度蜘蛛主要抓取策略

图 7-1 看似简单，但其实百度蜘蛛在抓取过程中面对的是一个超级复杂的网络环境，为了使系统可以抓取到尽可能多的有价值资源并保持系统及实际环境中页面的一致性，同时不给网站体验造成压力，会设计多种复杂的抓取策略。

1）抓取的友好性

互联网资源庞大的数量，要求抓取系统尽可能地高效利用带宽，在有限的硬件和带宽资源下尽可能多地抓取到有价值资源。这就产生了另一个问题——耗费被抓网站的带宽造成访问压力，如果程度过大将直接影响被抓网站的正常用户访问。因此，在抓取过程中就要进行一定的抓取压力控制，达到既不影响网站的正常用户访问又能尽量多地抓取到有价值资源的目的。

通常情况下，最基本的是基于 IP 地址的压力控制。这是因为如果基于域名，可能存在一个域名对多个 IP 地址（很多大网站）或多个域名对应同一个 IP 地址（小网站共享 IP 地址）的问题。实际工作中，往往是根据 IP 地址及域名多种条件进行压力调配控制。同时，站长平台也推出了压力反馈工具，站长可以人工调配对自己网站的抓取压力，这时百度蜘蛛将优先按照站长的要求进行抓取压力控制。

对同一个站点的抓取速度控制一般分为两类：其一，一段时间内的抓取频率；其二，一段时间内的抓取流量。同一站点不同的时间抓取速度也会不同，例如夜间抓取可能就会快一些，视具体站点类型而定，主要思想是错开正常用户访问高峰，不断调整。对于不同站点，也需要采用不同的抓取速度。

2）常用抓取返回码

下面简单介绍几种百度支持的返回码。

（1）404 代表 Not Found，认为网页已经失效，通常将在库中删除，同时短期内如果百度蜘蛛再次发现这条 URL 也不会抓取。

（2）503 代表 Service Unavailable，认为网页临时不可访问，通常在网站临时关闭，带宽有限时会产生这种情况。对于网页返回 503 状态码，百度蜘蛛不会把这条 URL 直接删除，同时短期内将会反复访问几次，如果网页已恢复，则正常抓取；如果继续返回 503，那么这条 URL 仍会被认为是失效链接，将之从库中删除。

（3）403 代表 Forbidden，认为网页目前禁止访问。如果是新 URL，百度蜘

蛛暂时不抓取，短期内同样会反复访问几次；如果是已收录URL，不会直接删除，短期内同样反复访问几次。如果网页正常访问，则正常抓取；如果仍然禁止访问，那么这条URL也会被认为是失效链接，将之从库中删除。

（4）301代表Moved Permanently，认为网页重定向至新URL。当遇到站点迁移、域名更换、站点改版的情况时，我们推荐使用301返回码，同时使用站长平台网站改版工具，以减少改版对网站流量造成的损失。

3）多种URL重定向的识别

互联网中有一部分网页因各种各样的原因存在URL重定向情况，为了对这部分资源正常抓取，就要求百度蜘蛛对URL重定向进行识别判断，同时防止作弊行为。重定向可分为三类：HTTP30X重定向、Meta Refresh重定向和JS重定向。另外，百度也支持canonical标签，在效果上可以认为也是一种间接的重定向。

4）抓取优先级调配

由于互联网资源规模巨大且变化迅速，对于搜索引擎来说全部抓取并合理更新保持一致性几乎是不可能的事情，因此就要求抓取系统有一套合理的抓取优先级调配策略。该策略主要包括：深度优先遍历策略、宽度优先遍历策略、PR优先策略、反链策略、社会化分享指导策略等。每个策略各有优劣，在实际情况中往往是多种策略结合使用以达到最优的抓取效果。

5）重复URL的过滤

百度蜘蛛在抓取过程中需要判断一个页面是否已经抓取过，如果还没有抓取则进行抓取网页的行为并记录在已抓取网址集合中。判断网页是否已经抓取其中涉及到最核心的功能是快速查找并对比，同时涉及到URL归一化识别，例如一个URL中包含大量无效参数而实际是同一个页面，这将视为同一个URL来对待。

6）暗网数据的获取

互联网中存在着大量的搜索引擎暂时无法抓取到的数据，被称为暗网数据。一方面，很多网站的大量数据存在于网络数据库中，搜索引擎难以采用抓取网页的方式获得完整内容；另一方面，由于网络环境、网站本身不符合规范以及孤岛等问题，也会造成搜索引擎无法抓取。目前来说，对于暗网数据的获取主要思路仍然是通过开放平台提交数据的方式来解决，例如“百度站长平台”“百度开放平台”等。

7）抓取反作弊

搜索引擎在抓取过程中往往会遇到所谓抓取黑洞或者面临大量低质量页面的困扰，这就要求抓取系统中同样需要设计一套完善的抓取反作弊系统。例如分析 URL 特征、分析页面大小及内容、分析站点规模对应抓取规模等。

3. 百度蜘蛛抓取过程中涉及的网络协议

搜索引擎与资源提供者之间存在相互依赖的关系，其中搜索引擎需要站长为其提供资源，否则搜索引擎就无法满足用户检索需求；而站长需要通过搜索引擎将自己的内容推广出去获取更多的受众。百度蜘蛛抓取系统直接涉及互联网资源提供者的利益，为了使搜索引擎与站长能够达到双赢，在抓取过程中双方必须遵守一定的规范，以便于双方的数据处理及对接。这种过程中遵守的规范也就是日常我们所说的网络协议。下面介绍一些常用的协议。

HTTP：超文本传输协议，是互联网上应用最为广泛的一种网络协议，是客户端和服务器端请求和应答的标准。客户端一般情况是指终端用户；服务器端即指网站。终端用户通过浏览器、蜘蛛等向服务器指定端口发送 HTTP 请求。发送 HTTP 请求会返回对应的 HTTP Header 信息，可以看到包括是否成功、服务器类型、网页最近更新时间等内容。

HTTPS：实际是加密版 HTTP，一种更加安全的数据传输协议。

UA 属性：UA 即 User Agent，是 HTTP 中的一个属性，代表了终端的身份，向服务器端表明“我是谁，来干嘛”，进而服务器端可以根据不同的身份来做出不同的反馈结果。

robots 协议：robots.txt 是搜索引擎访问一个网站时要访问的第一个文件，用以确定哪些网页是被允许抓取的，哪些是被禁止抓取的。robots.txt 必须放在网站根目录下，且文件名要小写。详细的 robots.txt 写法可参考 http://www.robotstxt.org。百度搜索严格按照 robots 协议执行。另外，同样支持网页内容中添加的名为 robots 的 meta 标签，index、follow、nofollow 等指令。

4. 百度蜘蛛抓取频次原则及调整方法

百度蜘蛛根据上述网站设置的协议对站点页面进行抓取，但是不可能做到对所有站点一视同仁，会综合考虑站点实际情况确定一个抓取配额，每天定量

抓取站点内容，即我们常说的抓取频次。那么百度搜索引擎是根据什么指标来确定对一个网站的抓取频次呢？主要有以下四个指标。

（1）网站更新频率。更新频率快的网站多抓取，更新慢的网站少抓取。网站更新频率直接影响百度蜘蛛的来访频率。

（2）网站更新质量。更新频率提高了，仅仅是吸引了百度蜘蛛的注意，百度蜘蛛对质量是有严格要求的，如果网站每天更新出的大量内容都被百度蜘蛛判定为低质页面，则依然没有意义。

（3） 连通度。网站应该安全稳定，对百度蜘蛛保持畅通，经常给百度蜘蛛吃闭门羹可不是好事情。

（4）站点评价。百度搜索引擎对每个站点都会有一个评价，且这个评价会根据站点情况不断变化，是百度搜索引擎对站点的一个基础打分（绝非外界所说的百度权重），是百度内部非常机密的数据。站点评级从不独立使用，而是配合其他因子和阈值一起共同影响对网站的抓取和排序。

抓取频次间接决定着网站有多少页面有可能被建库收录，如此重要的数值如果不符合站长预期该如何调整呢？

百度提供了抓取频次工具（http://zhanzhang.baidu.com/pressure/index），并已完成多次升级。该工具除了提供抓取统计数据外，还提供“频次调整”功能，站长根据实际情况向百度站长平台提出希望百度蜘蛛增加来访或减少来访的请求，工具会根据站长的意愿和实际情况进行调整。

5. 造成百度蜘蛛抓取异常的原因

有一些网页，内容优质，用户也可以正常访问，但是百度蜘蛛却无法正常访问并抓取，造成搜索结果覆盖率缺失，对百度搜索引擎和站点都是一种损失，百度把这种情况叫“抓取异常”。对于大量内容无法正常抓取的网站，百度搜索引擎会认为网站存在用户体验上的缺陷，并降低对网站的评价，在抓取、索引、排序上都会受到一定程度的负面影响，最终影响到网站从百度获取的流量。

下面介绍一些常见的抓取异常原因。

（1）服务器连接异常。服务器连接异常会有两种情况：一种是站点不稳定，百度蜘蛛尝试连接网站的服务器时出现暂时无法连接的情况；另一种是百度蜘蛛一直无法连接上网站的服务器。

造成服务器连接异常的原因通常是网站服务器的访问量过大，超负荷运转，也有可能是网站运行不正常，需检查自己网站的 Web 服务器（如 Apache、IIS）是否安装且正常运行，并使用浏览器检查主要页面能否正常访问。网站和主机还可能阻止了百度蜘蛛的访问，需要检查网站和主机的防火墙。

（2）网络运营商异常。网络运营商有电信和联通，百度蜘蛛通过电信或网通无法访问网站。如果出现这种情况，需要与网络服务运营商联系，购买拥有双线服务的空间或者购买 CDN 服务。

（3）DNS 异常。当百度蜘蛛无法解析网站的 IP 时，会出现 DNS 异常。这可能是网站 IP 地址错误，或者域名服务商把百度蜘蛛封禁了。请使用 Whois 或者 host 查询自己网站 IP 地址是否正确且可解析，如果不正确或无法解析，需与域名注册商联系，更新 IP 地址。

（4）IP 封禁。IP 封禁是指限制网络的出口 IP 地址，禁止该 IP 段的用户进行内容访问，在这里特指封禁了百度蜘蛛 IP 地址。当自己的网站不希望百度蜘蛛访问时，才需要使用该设置，如果希望百度蜘蛛访问自己的网站，需要检查相关设置中是否误添加了百度蜘蛛 IP 地址。另一种可能是网站所在的空间服务商把百度 IP 地址进行了封禁，这时就需要联系服务商更改设置。

（5）UA 封禁。服务器通过 UA 识别访客身份。当网站针对指定 UA 的访问返回异常页面（如 403、500）或跳转到其他页面，即为 UA 封禁。当网站不希望百度蜘蛛访问时，才需要该设置，如果希望百度蜘蛛访问网站，需要确认 UA 相关的设置中是否封禁了百度蜘蛛 UA，并及时修改。

（6）死链。页面已经无效，无法对用户提供任何有价值信息的页面就是死链，包括协议死链和内容死链两种形式。

协议死链：页面的 TCP 状态或者 HTTP 状态明确表示的死链，常见的有 404、403、503 状态等。

内容死链：服务器返回状态是正常的，但内容已经变更为不存在、已删除或需要权限等与原内容无关的信息页面。

对于死链，笔者建议站点使用协议死链，并通过百度站长平台的死链工具向百度提交，以便百度更快发现死链，减少死链对用户以及搜索引擎造成的负面影响。

（7）异常跳转。将网络请求重新指向其他位置即为跳转。异常跳转指的是

以下情况：

当前该页面为无效页面（内容已删除、死链等），直接跳转到前一目录或者首页，对这种情况下百度建议站长将该无效页面的入口超链接删除；跳转到出错或者无效页面。

注意：对于长时间跳转到其他域名的情况，如网站更换域名，百度建议使用 301 跳转协议进行设置。

（8）其他异常。

针对百度 refer 的异常：网页针对来自百度的 refer 返回不同于正常内容的行为。

针对百度 UA 的异常：网页对百度 UA 返回不同于页面原内容的行为。

JS 跳转异常：网页加载了百度无法识别的 JS 跳转代码，使得用户通过搜索结果进入的页面发生了跳转的情况。

压力过大引起的偶然封禁：百度会根据站点的规模、访问量等信息，自动设定一个合理的抓取压力。但是在异常情况下，如压力控制失常时，服务器会根据自身负荷进行保护性的偶然封禁。这种情况下，可在返回码中返回 503（其含义是 Service Unavailable，服务不可用）。这样，百度蜘蛛会过一段时间再来尝试抓取这个链接，如果网站已空闲，则会成功抓取。

6. 新链接重要程度判断

上面我们说了影响百度蜘蛛正常抓取的原因，下面就要说说百度蜘蛛的判断原则了。在建库环节前，百度蜘蛛会对页面进行初步内容分析和链接分析，通过内容分析决定该网页是否需要建索引库，通过链接分析发现更多网页，再对更多网页进行“抓取—分析—是否建库与发现新链接”的流程。理论上，百度蜘蛛会将新页面上所有能“看到”的链接都抓取回来。百度蜘蛛判断新链接重要性的依据有以下两个方面。

1）对用户的价值

（1）内容独特。百度搜索引擎喜欢原创的内容。

（2）主体突出。切不要出现网页主体内容不突出而被搜索引擎误判为空短页面不抓取。

（3）内容丰富。可以解决用户查询的问题。

（4）广告适当。网站的广告不可以影响用户访问页面主体内容。

2）链接重要程度

（1）目录层级——浅层优先。

（2）链接在站内的受欢迎程度。

7. 百度优先建重要库的原则

百度蜘蛛抓了多少页面并不是最重要的，重要的是有多少页面被建索引库，即我们常说的“建库”。众所周知，搜索引擎的索引库是分层级的，优质的网页会被分配到重要索引库，普通网页会待在普通库，再差一些的网页会被分配到低级库去当补充材料。目前 60% 的检索需求只调用重要索引库即可满足，这也就解释了为什么有些网站的收录量超高而流量却一直不理想。

那么，哪些网页可以进入重要索引库呢？其实总的原则就一个：对用户的价值。重要索引库中的网页包括却不限于：

（1）有时效性且有价值的页面。在这里，时效性和价值是并列关系，缺一不可。有些站点为了产生时效性内容页面做了大量采集工作，产生了一堆无价值页面，是百度不愿看到的。

（2）内容优质的专题页面。专题页面的内容不一定完全是原创的，即可以很好地把各方内容整合在一起，或者增加一些新鲜的内容，如观点和评论，给用户更丰富、全面的内容。

（3）高价值原创内容页面。百度把原创定义为花费一定成本、大量经验积累提取后形成的文章。伪原创不是原创。

（4）重要个人页面。这里仅举一个例子，科比在新浪微博开户了，即使他不经常更新，但对于百度来说，它仍然是一个极重要的页面。

8. 哪些网页无法进入索引库

上述优质网页进入了索引库，其实互联网上大部分网站根本没有被百度收录。但这并非是百度搜索没有发现它们，而是它们在建库前的筛选环节已被过滤掉了。什么样的网页在最初环节就被过滤掉了呢？

（1）重复内容的网页：互联网上已有的内容，百度必然没有必要再收录。

（2）主体内容空短的网页。

有些内容使用了百度蜘蛛无法解析的技术，如 JS、Ajax 等，虽然用户访问时能看到丰富的内容，但依然会被搜索引擎抛弃。

加载速度过慢的网页，也有可能被当作空短页面处理。注意，广告加载时间也算在网页整体加载时间内。

很多主体不突出的网页即使被抓取回来也会在这个环节被抛弃。

（3）部分作弊网页。

7.1.2 检索排序

1. 搜索引擎索引系统概述

众所周知，搜索引擎的主要工作过程包括抓取、存储、页面分析、索引、检索等几个主要过程。7.1.1 节主要介绍了部分抓取存储环节中的内容，下面简要介绍索引系统。

在以亿为单位的网页库中查找特定的某些关键词犹如大海捞针，也许给予一定时间可以完成查找，但是用户等不起。从用户体验角度搜索引擎必须在毫秒级别给予用户满意的结果，否则用户只能流失。怎样才能达到这样的要求呢？

如果能知道用户查找的关键词（关键词分词）都出现在哪些页面中，那么用户检索的处理过程即可以想象为包含了关键词中分词后不同部分的页面集合求交的过程，而检索即变成了页面名称之间的比较、求交。这样，在毫秒内以亿为单位的检索成为了可能。也就是通常说的倒排索引及求交检索的过程。图 7-2 所示为建立倒排索引的基本过程。

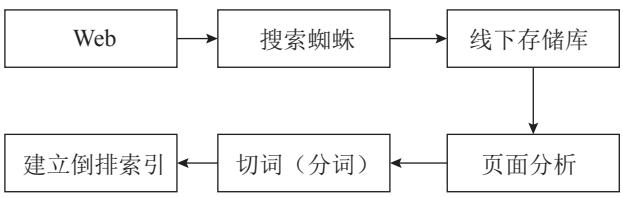

图 7-2　倒排索引流程图

（1）页面分析的过程实际上是将原始页面的不同部分进行识别并标记，例如网页标题、关键词、网页描述、链接以及其他非重要区域等。

（2）分词的过程实际上包括了切词分词、同义词转换、同义词替换等。以对某页面标题分词为例，得到的将是这样的数据：搜索词、搜索词 ID、词类、词性等。

（3）前面的准备工作完成后，接下来是建立倒排索引，图 7-3 所示即是索引系统中的倒排索引过程。

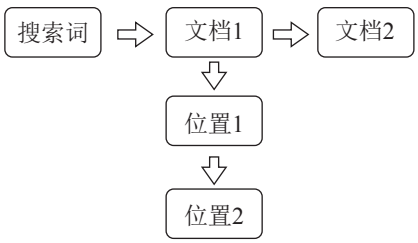

图 7-3　百度倒排索引过程示意图

倒排索引是搜索引擎实现毫秒级检索非常重要的一个环节。

2. 倒排索引的重要过程——入库写库

索引系统在建立倒排索引的最后还需要有一个入库写库的过程，而为了提高效率这个过程还需要将全部 term 以及偏移量保存在文件头部，并且对数据进行压缩，这涉及到的过于技术化在此就不多提了。在此简要介绍一下索引之后的检索系统。

检索系统主要包含五部分，如图 7-4 所示。

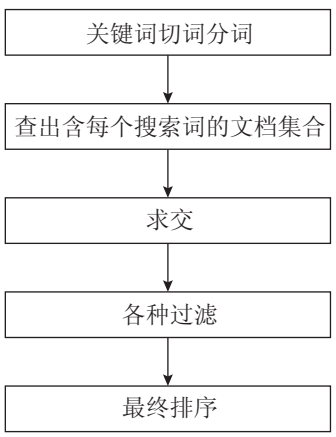

图 7-4　检索系统