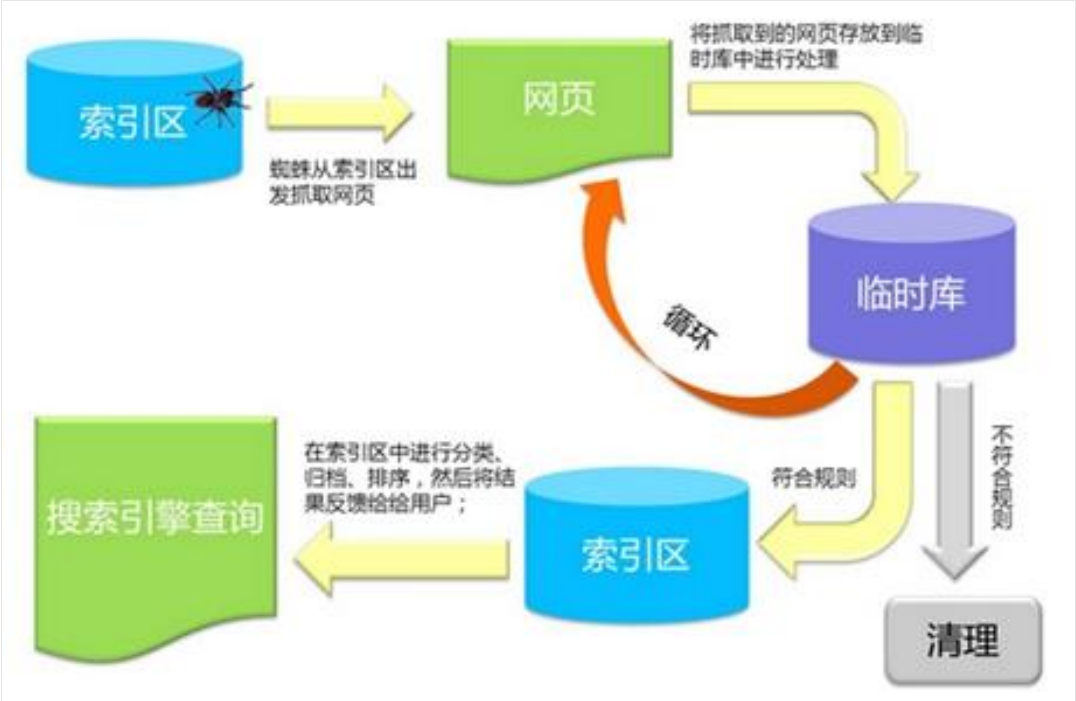
百度搜索引擎工作原理

————————最新更新章节：2014-12-10

关于百度以及其它搜索引擎的工作原理，其实大家已经讨论过很多，但随着科技的进步、互联网业的发展，各家搜索引擎都发生着巨大的变化，并且这些变化都是飞快的。我们设计这个章节的目的，除了从官方的角度发出一些声音、纠正一些之前的误读外，还希望通过不断更新内容，与百度搜索引擎发展保持同步，给各位站长带来最新的、与百度高相关的信息。本章主要内容分为四个章节，分别为：抓取建库；检索排序；外部投票；结果展现。



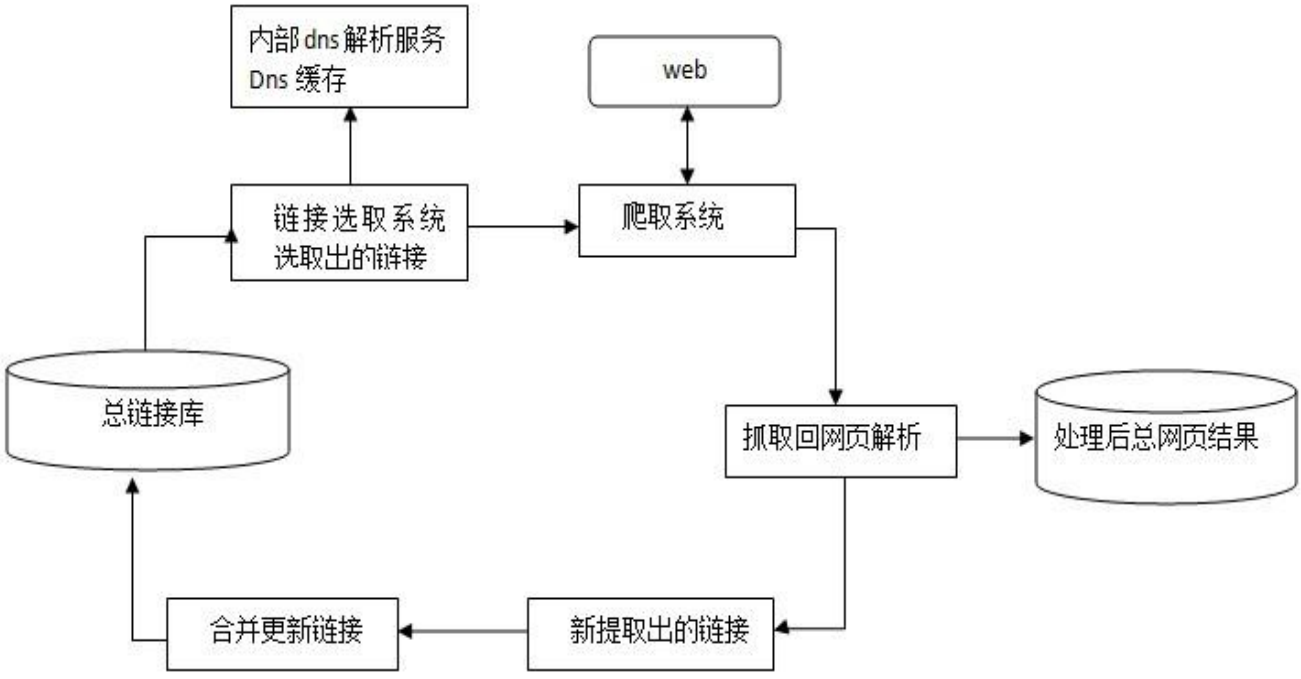
抓取建库

Spider 抓取系统的基本框架

互联网信息爆发式增长，如何有效的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上游，主要负责互联网信息的搜集、保存、更新环节，它像蜘蛛一样在网络间爬来爬去，因此通常会被叫做“spider”。例如我们常用的几家通用搜索引擎蜘蛛被称为：Baiduspdier、Googlebot、Sogou Web Spider 等。

Spider 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 spider 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关系，不断的发现新 URL 并抓取，尽最大可能抓取到更多的有价值网页。对于类似百度这样的大型 spider 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 spider 过去抓取过的页面保持更新，维护一个 URL 库和页面库。

下图为 spider 抓取系统的基本框架图，其中包括链接存储系统、链接选取系统、dns 解析服务系统、抓取调度系统、网页分析系统、链接提取系统、链接分析系统、网页存储系统。Baiduspider 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。



Baiduspider 主要抓取策略类型

上图看似简单，但其实 Baiduspider 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。以下做简单介绍：

1、抓取友好性

互联网资源庞大的数量级，这就要求抓取系统尽可能的高效利用带宽，在有限的硬件和带宽资源下尽可能多的抓取到有价值资源。这就造成了另一个问题，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行为。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的抓取到有价值资源的目的。

通常情况下，最基本的是基于 ip 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（很多大网站）或多个域名对应同一个 ip（小网站共享 ip）的问题。实际中，往往根据 ip 及域名的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 spider 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，例如夜深人静月黑风高时候抓取的可能就会快一些，也视具体站点类型而定，主要思想是错开正常用户访问高峰，不断的调整。对于不同站点，也需要不同的抓取速度。

2、常用抓取返回码示意

简单介绍几种百度支持的返回码：

1）最常见的 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 spider 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，通常网站临时关闭，带宽有限等会产生这种情况。对于网页返回 503 状态码，百度 spider 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，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，spider 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，从库中删除。

4）301 代表是“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

互联网中一部分网页因为各种各样的原因存在 url 重定向状态，为了对这部分资源正常抓取，就要求 spider 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：http 30x 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的巨大以及迅速的变化，对于搜索引擎来说全部抓取到并合理的更新保持一致性几乎是不可能的事情，因此这就要求抓取系统设计一套合理的抓取优先级调配策略。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

spider 在抓取过程中需要判断一个页面是否已经抓取过了，如果还没有抓取再进行抓取网页的行为并放在已抓取网址集合中。判断是否已经抓取其中涉及到最核心的是快速查找并对比，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是存在于网络数据库中，spider 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、孤岛等等问题，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采用数据提交的方式来解决，例如“百度站长平台”“百度开放平台”等等。

7、抓取反作弊

spider 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等。

Baiduspider 抓取过程中涉及的网络协议

刚才提到百度搜索引擎会设计复杂的抓取策略，其实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。spider 抓取系统直接涉及互联网资源提供者的利益，为了使搜素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中我们所说的一些网络协议。

以下简单列举：

http 协议：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，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 httpheader 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 协议：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 协议中的一个属性，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，用以来确定哪些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。

Baiduspider 抓取频次原则及调整方法

Baiduspider 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的呢，主要指标有四个：

1，网站更新频率：更新快多来，更新慢少来，直接影响 Baiduspider 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 Baiduspier 的注意，Baiduspider 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 Baiduspider 判定为低质页面，依然没有意义。

3，连通度：网站应该安全稳定、对 Baiduspider 保持畅通，经常给 Baiduspider 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），是百度内部一个非常机密的数据。站点评级从不独立使用，会配合其它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？百度站长平台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 Baiduspider 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。

造成 Baiduspider 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 Baiduspider 却无法正常访问并抓取，造成搜索结果覆盖率缺失，对百度搜索引擎对站点都是一种损失，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。

下面向站长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，Baiduspider 尝试连接您网站的服务器时出现暂时无法连接的情况；一种是 Baiduspider 一直无法连接上您网站的服务器。

造成服务器连接异常的原因通常是您的网站服务器过大，超负荷运转。也有可能是您的网站运行不正常，请检查网站的 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。您的网站和主机还可能阻止了 Baiduspider 的访问，您需要检查网站和主机的防火墙。

2，网络运营商异常：网络运营商分电信和联通两种，Baiduspider 通过电信或网通无法访问您的网站。如果出现这种情况，您需要与网络服务运营商进行联系，或者购买拥有双线服务的空间或者购买 cdn 服务。

3，DNS 异常：当 Baiduspider 无法解析您网站的 IP 时，会出现 DNS 异常。可能是您的网站 IP 地址错误，或者域名服务商把 Baiduspider 封禁。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，请与域名注册商联系，更新您的 IP 地址。

4，IP 封禁：IP 封禁为：限制网络的出口 IP 地址，禁止该 IP 段的使用者进行内容访问，在这里特指封禁了 BaiduspiderIP。当您的网站不希望 Baiduspider 访问时，才需要该设置，如果您希望 Baiduspider 访问您的网站，请检查相关设置中是否误添加了 BaiduspiderIP。也有可能是您网站所在的空间服务商把百度 IP 进行了封禁，这时您需要联系服务商更改设置。

5，UA 封禁：UA 即为用户代理（User-Agent），服务器通过 UA 识别访问者的身份。当网站针对指定 UA 的访问，返回异常页面（如 403，500）或跳转到其他页面的情况，即

为 UA 封禁。当您的网站不希望 Baiduspider 访问时，才需要该设置，如果您希望 Baiduspider 访问您的网站，useragent 相关的设置中是否有 Baiduspider 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接，包括协议死链和内容死链两种形式:

协议死链：页面的 TCP 协议状态/HTTP 协议状态明确表示的死链，常见的如 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，我们建议站点使用协议死链，并通过百度站长平台--死链工具向百度提交，以便百度更快地发现死链，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几种情况：

1）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，百度建议站长将该无效页面的入口超链接删除掉

2）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：

1）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。

2）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。

3）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页面后发生了跳转的情况。

4）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，请在返回码中返回 503(其含义是“Service Unavailable”)，这样 Baiduspider 会过段时间再来尝试抓取这个链接，如果网站已空闲，则会被成功抓取。

新链接重要程度判断

好啦，上面我们说了影响 Baiduspider 正常抓取的原因，下面就要说说 Baiduspider 的一些判断原则了。在建库环节前，Baiduspider 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&发现新链接的流程。理论上，Baiduspider 会将新页面上所有能“看到”的链接都抓取回来，那么面对众多新链接，Baiduspider 根据什么判断哪个更重要呢？两方面：

第一，对用户的价值：

1，内容独特，百度搜索引擎喜欢 unique 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

4，广告适当

第二，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

百度优先建重要库的原则

Baiduspider 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流量却一直不理想。

那么，哪些网页可以进入优质索引库呢。其实总的原则就是一个：对用户的价值。包括却不仅于：

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，产生了一堆无价值面页，也是百度不愿看到的.

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，比如观点和评论，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。千万不要再问我们伪原创是不是原创。

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。

哪些网页无法建入索引库

上述优质网页进了索引库，那其实互联网上大部分网站根本没有被百度收录。并非是百度没有发现他们，而是在建库前的筛选环节被过滤掉了。那怎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

1）有些内容使用了百度 spider 无法解析的技术，如 JS、AJAX 等，虽然用户访问能看到丰富的内容，依然会被搜索引擎抛弃

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

3）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

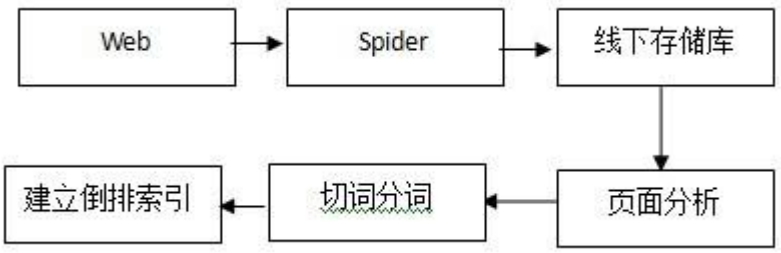
检索排序

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。上一章我们主要介绍了部分抓取存储环节中的内容，此章简要介绍一下索引系统。

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里面捞针，也许一定的时间内可以完成查找，但是用户等不起，从用户体验角度我们必须在毫秒级别给予用户满意的结果，否则用户只能流失。怎样才能达到这种要求呢？

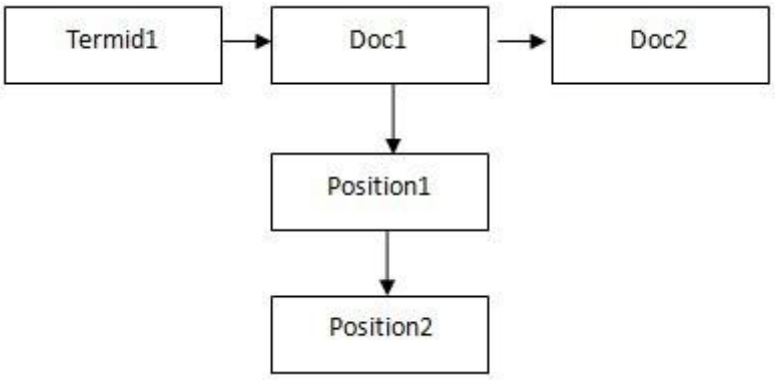
如果能知道用户查找的关键词（query 切词后）都出现在哪些页面中，那么用户检索的处理过程即可以想象为包含了 query 中切词后不同部分的页面集合求交的过程，而检索即变成了页面名称之间的比较、求交。这样，在毫秒内以亿为单位的检索成为了可能。这就是通常所说的倒排索引及求交检索的过程。如下为建立倒排索引的基本过程：



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：title、keywords、content、link、anchor、评论、其他非重要区域等等；

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等，以对某页面 title 分词为例，得到的将是这样的数据：term 文本、termid、词类、词性等等；

3，之前的准备工作完成后，接下来即是建立倒排索引，形成{termàdoc}，下图即是索引系统中的倒排索引过程。

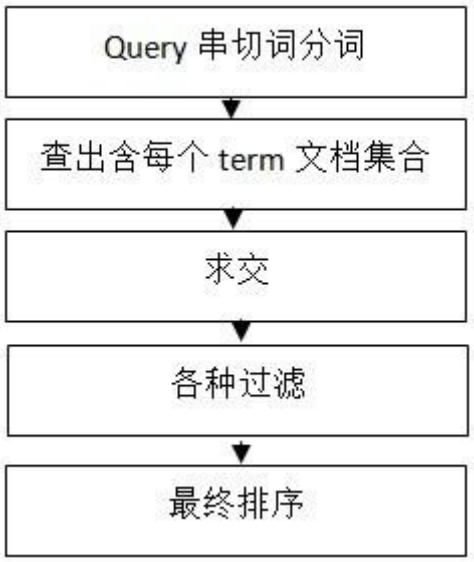


倒排索引是搜索引擎实现毫秒级检索非常重要的一个环节，下面我们要重要介绍一下索引系统建立倒排索引的重要过程——入库写库。

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。在此简要给大家介绍一下索引之后的检索系统。

检索系统主要包含了五个部分，如下图所示：



（1）Query 串切词分词即将用户的查询词进行分词，对之后的查询做准备，以“10 号线地铁故障”为例，可能的分词如下（同义词问题暂时略过）：

10 0x123abc

号 0x13445d

线 0x234d

地铁 0x145cf

故障 0x354df

（2）查出含每个 term 的文档集合，即找出待选集合，如下：

0x123abc 1 2 3 4 7 9…..

0x13445d 2 5 8 9 10 11……

……

……

（3）求交，上述求交，文档 2 和文档 9 可能是我们需要找的，整个求交过程实际上关系着整个系统的性能，这里面包含了使用缓存等等手段进行性能优化；

（4）各种过滤，举例可能包含过滤掉死链、重复数据、色情、垃圾结果以及你懂的；

（5）最终排序，将最能满足用户需求的结果排序在最前，可能包括的有用信息如：网站的整体评价、网页质量、内容质量、资源质量、匹配程度、分散度、时效性等等

影响搜索结果排序的因素

上面的内容好象有些深奥，因为涉及大量技术细节，我们只能说到这儿了。那下面我们说说大家最感兴趣的排序问题吧。用户输入关键词进行检索，百度搜索引擎在排序环节要做两方面的事情，第一是把相关的网页从索引库中提取出来，第二是把提取出来的网页按照不同维度的得分进行综合排序。“不同维度”包括：

1，相关性：网页内容与用户检索需求的匹配程度，比如网页包含的用户检查关键词的个数，以及这些关键词出现的位置；外部网页指向该页面所用的锚文本等

2，权威性：用户喜欢有一定权威性网站提供的内容，相应的，百度搜索引擎也更相信优质权威站点提供的内容。

3，时效性：时效性结果指的是新出现的网页，且网页内承载了新鲜的内容。目前时效性结果在搜索引擎中日趋重要。

4，重要性：网页内容与用户检查需求匹配的重要程度或受欢迎程度

5，丰富度：丰富度看似简单却是一个覆盖范围非常广的命题。可以理解为网页内容丰富，可以完全满足用户需求；不仅可以满足用户单一需求，还可以满足用户的延展需求。

6，受欢迎程度：指该网页是不是受欢迎。

以上便是百度搜索引擎决定搜索结果排序时考虑的六大原则，那么六大原则的侧重点是怎样的呢？哪个原则在实际应用时占比最大呢？其实在这里没有一个确切的答案。在百度搜索引擎早期，这些阈值的确是相对固定的，比如“相关性”在整体排序中的重量可以占到七成。但随着互联网的不断发展，检索技术的进步，网页数量的爆发式增长，相关性已经不是难题。于是百度搜索引擎引入了机器学习机制，让程序自动产出计算公式，推进排序策略更加合理。

低质网页狙击策略——石榴算法

我们理解网站生存发展需要资金支持，从来不反对网站添加各种合法广告，不要再问我们“我们网站加了 XX 联盟的广告会不会被处罚”这类问题。有些站点好不容易在百度有了比较好的排位，却在页面上放置大量有损访问用户体验的广告，已经严重影响到百度搜索引擎用户的使用感受。为此，百度质量团队 2013 年 5 月 17 日发布公告：针对低质量网页推出了石榴算法，旨在打击含有大量妨碍用户正常浏览的恶劣广告的页面，尤其是弹出大量低质

广告、存在混淆页面主体内容的垃圾广告的页面。　　如以下网页截图，用户要花很长时间去寻找真正的下载地址，是百度无法接受容忍的。

百度质量团队希望站长能够多从用户角度出发，朝着长远发展考虑，在不影响用户体验的前提下合理地放置广告，赢得用户的长期青睐才是一个网站发展壮大的基础。

外部投票

外链的作用（2014 版）(在 2015 年 7.1 日百度 htts 化 外链作用几乎为零)

曾经，“内容为王超链为皇”的说法流行了很多年，通过超链计算得分来体现网页的相关性和重要性，的确曾经是搜索引擎用来评估网页的重要参考因素之一，会直接参与搜索结果排序计算。但随着该技术被越来越多的 SEO 人员了解，超链已经逐渐失去作为投票的重要意义，无论是谷歌还是百度，对超链数据的依赖程度都越来越低。那么，在现在，超链在发挥着怎样的作用？

1， 吸引蜘蛛抓取：虽然百度在挖掘新好站点方面下了很大工夫，开放了多个数据提交入口，开避了社会化发现渠道，但超链依然是发现收录链接的最重要入口。

2， 向搜索引擎传递相关性信息：百度除了通过 TITLE、页面关键词、H 标签等对网页内容进行判断外，还会通过锚文本进行铺助判断。使用图片作为点击入口的超链，也可以通过 alt 属性和 title 标签向百度传情达意。

3，提升排名：百度搜索引擎虽然降低了对超链的依赖，但对超链的识别力度从未下降，制定出更加严格的优质链接、正常链接、垃圾链接和作弊链接标准。对于作弊链接，除了对链接进行过滤清理外，也对链接的受益站进行一定程度的惩罚。相应的，对优质链接，百度依然持欢迎的态度。

4， 内容分享，获取口碑：优质内容被广泛传播，网站借此获得的流量可能并不多，但如果内容做得足够，也可以树立自己的品牌效应。

*严格来讲，这并不属于超链的作用。在百度眼里，网站的品牌比超链要重要得多。

切断买卖超链的利刃——绿萝算法 1.0&2.0

百度质量团队 2013 年 2 月 19 日发布公告推出绿萝算法，针对买卖链接行为再次强调：买卖链接行为一方面影响用户体验，干扰搜索引擎算法；另一方面让投机建站者得利、超链中介者得利，真正勤勤恳恳做好站的站长在这种恶劣的互联网超链环境中无法获得应有的回报。因此针对买卖链接行为在清除外链计算的基础上，以下三个类型的网站将会受到不同程度的影响：

1、超链中介：超链本应是互联网上相对优质的推荐，是普通用户及网站之间对页面内容、网站价值的肯定，但是现在种种超链作弊行为使得真实的肯定变成了一些人谋取利益的垫脚石，用户无法根据链接的推荐找到需要的优质资源，并且严重干扰搜索引擎对网站的评价。超链中介便是这畸形的超链市场下形成的恶之花，我们有义务维护超链的纯净维护用户利益，也有责任引导站长朋友们不再支出无谓的花销，所以超链中介将在我们的目标范围内。

2、出卖链接的网站：一个站点有许多种盈利方式，利用优质的原创内容吸引固定用户，引进优质广告资源，甚至举办线下活动，这些盈利方式都是我们乐于见到的，是一个网站的真正价值所在。但是一些网站内容基本采集自网络，以出卖超链位置为生；一些机构类网站或被链接中介所租用进行链接位置出售，使得超链市场泡沫越吹越多。此次的调整对这类站点同样将有所影响。

3、购买链接的网站：一直以来，百度对优质站点都会加以保护和扶植，这是从用户需求以及创业站长的角度出发的必然结果。而部分站长不将精力用在提升网站质量上，而选择钻营取巧，以金钱换取超链，欺骗搜索引擎进而欺骗用户。对于没有太多资源和金钱用于此类开销的创业站长来说，也是一种无形的伤害，如果不进行遏制，劣币驱逐良币，势必导致互联网环境愈加恶劣。此次调整这类站点本身也将受到影响。

以上即百度质量团队首次推出绿萝算法时的具体情况，后来被称为绿萝算法 1.0。事隔 5 个月之后，百度质量团队再次推出绿萝算法 2.0，针对明显的推广性软文进行更大范围更加严格的处理。

惩罚的对象重点是发布软文的新闻站点，同时包括软文交易平台、软文收益站点。惩罚方式包括：

1、针对软文交易平台，将被直接屏蔽；

2、针对软文发布站，将视不同程度而进行处理。例如一个新闻网站，存在发布软文的现象但情节不严重，该网站在搜索系统中将被降低评价；利用子域大量发布软文的，该子域将被直接屏蔽，并且清理出百度新闻源；更有甚者创建大量子域用于发布软文，此种情况整个主域将被屏蔽。

3、针对软文受益站，一个网站的外链中存在少量的软文外链，那么此时该外链将被过滤清除出权重计算体系，该受益站点将被观察一段时间后视情况而进一步处理；一个网站的外链中存在大量的软文外链，那么此时该受益站点将被降低评价或直接屏蔽。

结果展现

结构化数据——助力站点获得更多点击

网页经历了抓取建库，参与了排序计算，最终展现在搜索引擎用户面前。目前在百度搜索左侧结果展现形式很多，如：凤巢、品牌专区、自然结果等，一条自然结果怎样才能获得更多的点击，是站长要考虑的重要一环。

目前自然结果里又分为两类，见下图，第一个，即结构化展现，形式比较多样。目前覆盖 80%的搜索需求，即 80%的关键词下会出现这种复杂展现样式；第二个即一段摘要式展现，最原始的展现方式，只有一个标题、两行摘要、部分链接。

产品原型设计软件(Balsamiq Mockups)下载 v2.0.19(附注册码)_ - ...

下载地址 大小: 21.0M 更新时间: 2012-8-27

简介: Balsamiq Mockups是一种软件工程中快速原型的建立软件，可以做为与用户交互的一个界面草图。Balsamiq Mockups出自加利福...

www.pc6.com/softview/S... 2012-08-27 - 百度快照

推荐两个界面原型设计工具--GUIDesignStudio 和 Mockups For ...

前段时间,有幸参加一次高级软件架构师的培训,授课老师介绍了两个很好玩的界面原型设计工具:GUIDesignStudio 和 Mockups For Desktop,现分享一下,截图说明,洗洗眼球,...

www.cnblogs.com/wuhuac... 2010-01-22 - 百度快照 - 91%好评

很明显，结构化展现能够向用户明确传递信息，直击用户需求痛点，获得更好的点击自然不在话下。目前结构化展现有几个样式：

1， 通用问答：提取答案，方便搜索用户参考，有些结构化数据还提取出了问题

宝宝5个月添加什么辅食？- 已解决 - 搜狗问问

2个回答 - 最新回答: 2008年6月7日

问题描述: 母乳喂养的

最佳答案: 四个月宝宝的辅食添加 添加辅食，步步为营 随着婴儿逐渐长大，4个月后，母乳已经不能完全满足他对营养的需求了。这时，父母就可以考虑给孩子添加辅食了，...

wenwen.soso.com/z/q653... 2008-05-30 - 百度快照 - 评价

您好:宝宝5个半月了,还没有开始添加辅食。请问什么可不..._百度知道

1个回答 - 提问时间: 2013年01月20日

权威专家: 魏健

最佳答案: 你好！ 一般宝宝添加辅食在4个月左右进行，具体要根据宝宝的情况而定，如果宝宝消化能力不是很好的也可以在5个月左右开始...

zhidao.baidu.com/link?... 2013-01-20 - 87%好评

怎么给婴儿拍嗝_育儿问答_宝宝树

11个回答 - 最新回答: 2012年5月23日

最佳答案: 亲，一般宝宝吃完奶以后，就要把宝宝竖着抱起来，让他的头靠在你的肩膀上，然后用手轻轻的拍后背，手是用空心掌最好，一下一下的，然后大概拍十几下...

www.babytree.com/ask/d... 2012-05-23 V2 - 百度快照 - 84%好评

2， 下载:

产品原型设计软件(Balsamiq Mockups)下载 v2.0.19(附注册码)_ - ...

下载地址 大小: 21.0M 更新时间: 2012-8-27

简介: Balsamiq Mockups是一种软件工程中快速原型的建立软件，可以做为与用户交互的一个界面草图。Balsamiq Mockups出自加利福...

www.pc6.com/softview/S... 2012-08-27 - 百度快照

3，时间戳：对于时效性较强的资讯，将时间提取出来，吸引用户点击，还有回复的条目，能够表现这个链接的有效性和热度

2014北京高考论坛_北京高考学习网_北京高考复习资料-e度教育论坛 ...
39条回复 - 发帖时间: 2014年7月14日
[家长交流]【高考汇总贴】2014北京高考重要信息汇总!(已更新分数线) digest 回复:33 查看:29472 最新回复: 都是朋友 5天前 迦南 2014-6-7 18:14 版块...
bbs.eduu.com/forum-96.... 2014-07-14 - 百度快照 - 84%好评

4， 在线文档：出现文档格式示意图

给宝宝拍奶嗝方法大全_百度文库
★★★★★ 评分:4/5 13页
那么为什么要为何要帮宝宝拍嗝?应该怎么帮 宝宝拍嗝?您会正确地拍打嗝吗?希望你看完本期的育儿百宝箱能正确认识宝宝 拍嗝这件事。 为何要帮宝宝拍嗝? 宝宝在...
wenku.baidu.com/link?u... 2012-04-08 - 百度快照 - 88%好评

5，原创标记: 原创标记的使用是最严格的，只有通过人工审核的站点才能拥有原创标志，在抓取和排序上有一定优待，所以审核非常严格，严控质量。

以色列进攻利器包围加沙 200余辆梅卡瓦随时出击_军事_环球网
[原创] 作者：田聿 - 来源：环球时报 - 发表时间：2014年07月14日
以色列国防军发起代号为"护刃行动"的军事行动,重点打击加沙武装人员。... 7月7日开始,以色列国防军发起代号为"护刃行动"的军事行动,重点打击加沙武装...
mil.huanqiu.com/world/... 2014-07-14 - 百度快照 - 74%好评

6，配图：扩大面积，方便用户了解网页内容，吸引点击

绿萝叶子发黄的原因及解决方法-土巴兔装修大学

不少业主在室内绿萝养殖过程中,发现了绿萝叶子发黄的情况,那么,为什么绿萝会出现叶子发黄,绿萝叶子发黄了该怎么办呢?土巴兔小编将在此为大家分析提供最实用的解决方法
www.to8to.com/yezhu/z4... 2012-04-13 - 百度快照

那么站长可以通过什么途径获得结果化展现呢：

1，参与原创星火计划:百度站长平台 VIP 俱乐部提供申请入口，需要经过人工审核后进行数据提交

2，结构化数据提交工具:zhanzhang.baidu.com/wiki/197

3，结构化数据标注工具:http://zhanzhang.baidu.com/itemannotator/index

4，搜索结果配图：具体要求为，在文章主体位置；图片与内容相关；图片上没有文字；图片比例接近 121*91。