【科普】搜索引擎的工作原理

原创

caoz 于 2016-12-12 21:05:43 发布

昨天的文章 全球化的误区，本地化的机会 ，评论里，有人说，搜索引擎技术似乎不需要本地化，这一看就是彻底不懂这个领域的人讲的。当然，实话说，如果有人说，google在中文本地化方面做得非常好，我是可以部分同意的，同意的比例可能会比google工程师少一些。但我相信google工程师也会告诉你，搜索引擎是需要本地化的。

今天写篇科普文，讲讲搜索引擎的技术机理和市场竞争的一些特点。当然，作为从事或有兴趣从事流量运营的朋友，是可以用另一个角度去理解本文。

搜索引擎的核心技术架构，大体包括以下三块，第一，是蜘蛛/爬虫技术；第二，是索引技术；第三是查询展现的技术； 当然，我不是搜索引擎的架构师，我只能用比较粗浅的方式来做一个结构的切分。

1、 蜘蛛，也叫爬虫，是将互联网的信息，抓取并存储的一种技术实现。

搜索引擎的信息收录，很多不明所以的人会有很多误解，以为是付费收录，或者有什么其他特殊的提交技巧，其实并不是，搜索引擎通过互联网一些公开知名的网站，抓取内容，并分析其中的链接，然后有选择的抓取链接里的内容，然后再分析其中的链接，以此类推，通过有限的入口，基于彼此链接，形成强大的信息抓取能力。

有些搜索引擎本身也有链接提交入口，但基本上，不是主要的收录入口，不过作为创业者，建议了解一下相关信息，百度，google都有站长平台和管理后台，这里很多内容是需要非常非常认真的对待的。

反过来说，在这样的原理下，一个网站，只有被其他网站所链接，才有机会被搜索引擎抓取。如果这个网站没有外部链接，或者外部链接在搜索引擎中被认为是垃圾或无效链接，那么搜索引擎可能就不抓取他的页面。

分析和判断搜索引擎是否抓取了你的页面，或者什么时候抓取你的页面，只能通过服务器上的访问日志来查询，如果是cdn就比较麻烦。 而基于网站嵌入代码的方式，不论是cnzz，百度统计，还是google analytics，都无法获得蜘蛛抓取的信息，因为这些信息不会触发这些代码的执行。

一个比较推荐的日志分析软件是awstats。

在十多年前，分析百度蜘蛛抓取轨迹和更新策略，是很多草根站长每日必做的功课，比如现在身价几十亿的知名80后上市公司董事长，当年在某站长论坛就是以此准确的分析判断而封神，很年轻的时候就已经是站长圈的一代偶像。

但关于蜘蛛的话题，并不只基于链接抓取这么简单，延伸来说

第一，网站拥有者可以选择是否允许蜘蛛抓取，有一个robots.txt的文件是来控制这个的。

一个经典案例是 https://www.taobao.com/robots.txt

你会看到，淘宝至今仍有关键目录不对百度蜘蛛开放，但对google开放。

另一个经典案例是 http://www.baidu.com/robots.txt

你看出什么了？你可能什么都没看出来，我提醒一句，百度实质上全面禁止了360的蜘蛛抓取。

但这个协议只是约定俗成，实际上并没有强制约束力，所以，你们猜猜，360遵守了百度的蜘蛛抓取禁止么？

第二，最早抓取是基于网站彼此的链接为入口，但实际上，并不能肯定的说，有可能存在其他抓取入口，比如说，

客户端插件或浏览器， 免费网站统计系统的嵌入式代码。

会不会成为蜘蛛抓取的入口，我只能说，有这个可能。

所以我跟很多创业者说，中国做网站，放百度统计，海外做网站，放google analytics，是否会增加搜索引擎对你网站的收录？我只能说猜测，有这个可能。

第三，无法被抓取的信息

有些网站的内容链接，用一些javascript特殊效果完成，比如浮动的菜单等等，这种连接，有可能搜索引擎的蜘蛛程序不识别，当然，我只是说有可能，现在搜索引擎比以前聪明，十多年前很多特效链接是不识别的，现在会好一些。

需要登录，需要注册才能访问的页面，蜘蛛是无法进入的，也就是无法收录。

有些网站会给搜索特殊页面，就是蜘蛛来能看到内容（蜘蛛访问会有特殊的客户端标记，服务端识别和处理并不复杂），人来了要登录才能看，但这样做其实是违反了收录协议（需要人和蜘蛛看到的同样的内容，这是绝大部分搜索引擎的收录协议），有可能遭到搜索引擎处罚。

所以一个社区要想通过搜索引擎带来免费用户，必须让访客能看到内容，哪怕是部分内容。

带很多复杂参数的内容链接url，有可能被蜘蛛当作重复页面，拒绝收录。

很多动态页面是一个脚本程序带参数体现的，但蜘蛛发现同一个脚本有大量参数的网页，有时候会给该网页的价值评估带来困扰，蜘蛛可能会认为这个网页是重复页面，而拒绝收录。还是那句话，随着技术的发展，蜘蛛对动态脚本的参数识别度有了很大进步，现在基本上可以不用考虑这个问题。

但这个催生了一个技术，叫做伪静态化，通过对web服务端做配置，让用户访问的页面，url格式看上去是一个静态页，其实后面是一个正则匹配，实际执行的是一个动态脚本。

很多社区论坛为了追求免费搜索来路，做了伪静态化处理，在十多年前，几乎是草根站长必备技能之一。

爬虫技术暂时说到这里，但是这里强调一下，有外链，不代表搜索蜘蛛会来爬取，搜索蜘蛛爬取了，不代表搜索引擎会收录；搜索引擎收录了，不代表用户可以搜索的到；

site语法是检查一个网站收录数的最基本搜索语法，我开始以为是abc的常识，直到在新加坡做一些创业培训后交流才发现，大部分刚进入这个行业的人，或者有兴趣进入这个行业的人，对此并不了解。

一个范例，百度搜索一下 site:4399.com

2、索引系统

蜘蛛抓取的是网页的内容，那么要想让用户快速的通过关键词搜索到这个网页，就必须对网页做关键词的索引，从而提升查询效率，简单说就是，把网页的每个关键词提取出来，并针对这些关键词在网页中的出现频率，位置，特殊标记等诸多因素，给予不同的权值标定，然后，存储到索引库中。

那么问题来了，什么是关键词。

英文来说，比如 this is a book，中文，这是一本书。

英文很自然是四个单词，空格是天然的分词符，中文呢？你不能把一句话当作关键词吧（如果把一句话当作关键词，那么你搜索其中部分信息的时候，是无法索引命中的，比如搜索一本书，就搜索不出来了，而这显然是不符合搜索引擎诉求的）。所以要分词。

最开始，最简单的思路是，每个字都切开，这个以前叫字索引，每个字建立索引，并标注位置，如果用户搜索一个关键词，也是把关键词拆成字来搜索再组合结果，但这样问题就来了。

比如搜索关键词“海鲜”的时候，会出现结果，上海鲜花，这显然不是应该的搜索结果。

比如搜索关键词“和服”的时候，会出现结果，交换机和服务器。

这些都是蛮荒期的google也不能幸免的问题。

到后来有个梗，别笑，这些都是血泪梗，半夜电话过来，说网监通过搜索发现你社区有淫秽内容要求必须删除，否则就关闭你的网站，夜半惊醒认真排查，百思不得其解，苦苦哀求提供信息线索，最后发现，有人发了一条小广告，“求购二十四口交换机”。 还有，涉嫌政治敏感，查到最后“提供三台独立服务器”， 看出其中敏感词了没？你说冤不冤。 这两个故事可能并不是真的，因为都是网上看到的，但是我想说，类似这样的事情真的有，并非都是空穴来风。

所以，分词，是亚洲很多语言需要额外处理的事情，而西方语言不存在的问题。

但分词不是说说那么简单，比如几点，1：如何识别人名？2、互联网新词如何识别？比如“不明觉厉”。3、中英混排的坑，比如QQ表情。

做一个分词系统，说到底也不难，但是要做一个自动学习，与时俱进，又能高效率灵活的分词引擎，还是很有技术难度的。 当然，这方面我不是专家，不敢妄言了。

现在机器学习技术发达了，特别是google在深度学习领域拥有领先优势，以前很多通过人工做标定，做分类的工作可以交给算法完成，从某种意义来说，本地化的工作可以让机器学习去完成；未来，也许深度学习技术可以自己学习掌握本地化的技巧。 但我想说两点，第一，从搜索引擎发展历史看，在深度学习技术还没成熟的情况下，本地化的工作是非常重要的，也是很重要的决定竞争成败的要素；第二，即便现在深度学习已经很强大，基于当地语言的人工参与，标定，测试，反馈，一些本地化的工作依然对深度学习的效率和效果拥有不可替代的作用。

索引系统除了分词之外，还有一些要点，比如实时索引，因为一次索引库的更新是个大动静，一般网站运营者知道，自己网站内容更新后，需要等索引库下一次更新才能看到效果，而且索引库针对不同权重的网站内容，更新的频次也不太一样。 但诸如一些高优先的资讯网站，以及新闻搜索，索引库是可以做到近似实时索引的，所以我们在新闻搜索里，几分钟前的信息就已经可以搜索到了。

我以前经常吐槽一个事情，我在百度空间发表的文章，每次都是google率先索引收录，当时他们的解释是，猜测是因为很多人通过google阅读器订阅我的博客，而google阅读器很可能是google快速索引的入口。（然并卵，百度空间已经没有了，google阅读器也没有了。）

索引系统的权值体系，是所有SEOER们最关心的问题，他们经常通过不同方式组合策略，观察搜索引擎的收录，排名，来路情况，然后通过对比分析整理出相关的策略，这玩意说出来可以开很长一篇了，但今天就不提了。

但我说一个事实，很多外面的公司，做SEO的，会误认为百度里面的人熟悉这里的门道和规律，很多人高价去挖百度的搜索产品经理和技术工程师去做SEO，结果，呵呵，呵呵。 而外面那些草根创业者，有些善于此道的，真的比百度的人还清楚，搜索权值的影响关系，和更新频次等等，比如前面说到的，身价几十亿的那个80后创业者。

基于结果反推策略，比身在其中却不识全局的参与者，更能找到系统的关键点，有意思不。

3、查询展现

用户在浏览器或者在手机客户端输入一个关键词，或者几个关键词，甚至一句话，这个在服务端，应答程序获取后处理步骤如下

第一步，会检查最近时间有没有人搜索过同样的关键词，如果存在这样的缓存，最快的处理是将这块缓存提供给你，这样查询效率最高，对后端负载压力最低。

第二步，发现这个输入查询最近没有搜索，或者有其他条件的原因必须更新结果，那么会将这个用户输入的词，进行分词，没错，如果不止一个关键词，或者是一句话的情况下，应答程序会又一次分词，将搜索的查询拆成几个不同的关键词。

第三步，将切分后的关键词分发到查询系统中，查询系统会去索引库查询，索引库是个庞大的分布式系统，先分析这个关键词属于哪一块哪一台服务器，索引是一种有序的数据组合，我们用可以用近似二分法的方式思考，不管数据规模多大，你用二分法去查找一个结果，查询频次是log2(N)，这个就保证了海量数据下，查询一个关键词是非常快非常快的。 当然，实际情况会比二分法复杂很多，这样说比较容易理解而已，再复杂些不是我不告诉大家，是我自己都不是很清楚呢。

第四步，不同关键词的查询结果（只是按权值排序的部分顶部结果，绝对不是全部结果），基于权值倒序，会再汇总在一起，然后把共同命中的部分反馈回来，并做最后的权值排序。

记住，搜索引擎绝对不会返回所有结果，这个开销谁都受不了，百度也不行，google也不行，翻页都是有限制的。

再记住，如果你多个关键词里有多个不同品类冷门词，搜索引擎有可能会舍弃其中一个冷门词，因为汇总数据很可能不包含共同结果。搜索技术不要神话，这样的范例偶尔会出现。

这是三大部分，多说一点，其实还有第四部分。

用户点击行为采集和反馈部分

基于用户的翻页，点击分布，对搜索结果的优劣做判定，并对权值做调整，但这个早期搜索引擎是没有的，后面才有，所以暂时不列为必备的三大块。

此外，一些对搜索优化的机器学习策略，对易混词识别，同音词识别等等，相当部分也都基于用户行为反馈进行，这是后话，这里不展开。

关于第四部分，我以前说过一个词，点击提权，我说这个词价值千金，我猜很多人并没理解。没理解就好，要不我要被一些同行骂死了。

以上是单指搜索引擎的工作原理，和一些技术逻辑，当然，只是入门级的解读，毕竟再深入就不是我能讲解的了。

但搜索引擎的本地化，并不局限于搜索技术的本地化。

百度的强大，不只是搜索技术，当然有些人会说百度没有搜索技术，这种言论我就不争论了，我不试图改变任何人的观点，我只列一些事实而已。

百度的强大还来自于两大块，第一是内容护城河，第二是入口把控。

前者是百度贴吧，百度mp3，百度知道，百度百科，百度文库

后者是hao123和百度联盟。

这两块都是本地化，google进中国的时候，在这两块都有动作

投资天涯，收购265，以及大力发展google联盟，这些都是本地化。

此外，重申一下，百度全家桶的出现以及，百度全家桶和hao123的捆绑，是360崛起之后的事情，hao123从百度收购到360崛起之前，一直风平浪静的没做任何推广和捆绑，从历史事实而言，请勿将本地化等同于流氓化。