SEM 一家之言合集

SEO策略

Google工具栏收集用户的行为应用到算法里

N 年前我在阿里的第一个主管 BEN 面试中问我有没有自己独特的想法时，我说：不知道你是不是觉得我胡思乱想，我觉得 google 利用工具栏在收集用户行为，并且已经应用到算法中。我当时就看到 BEN 两眼放光了。后来，我才知道这句话在我这个面试当中的含义。有人告诉我，其实在那个时候 BEN 也一直在和别人说 google 在利用工具栏收集用户行为，但是那时候周围的人，要么不懂这个对 SEO 有什么意义，要么就是没有什么人相信他的话。

我在 04 年是用自己的实验证明了这一点的。04 年的时候，我同时有 4 台电脑可以用。受到 alexa 工具条收集用户行为来统计网站流量的启发，我觉得既然 google 工具栏普及率比 alexa 工具条大得多，会不会有可能也在收集用户行为来应用到算法里面。当时就新做了一个网站，完全让爬虫没有任何机会访问这个网站。然后给 4 台电脑都装上 google 工具栏，启用工具栏里的“显示 PR 值”（google 那时的规则就是只要你启用显示 PR 值，就会向 google 服务器发送数据），把此网站设为 IE 首页，这样每天可以访问很多次网站。 不记得是过了多久，google 就把这个新网站收录了。 后来又继续试验，主页上设立一个主关键词，为这个关键词做好内部优化，每天继续用这 4 台电脑访问网站。和猜想的一样，发现这个关键词的排名上升了，而且越排越好。

到现在，上面提到的两点：google 利用工具栏收集新网站和利用工具栏收集到的网站流量来决定网站的排序，是被证实成立的。

其实证据非常简单，在 google 工具栏的隐私条款里面明明白白的写着 google 如何利用工具栏的数据。

地址
是： http://www.google.com/support/toolbar/bin/static.py?page=privacy.html&hl=zh_CN&v=

这份更新于 07 年的隐私条款里面可以看到这样的叙述：

只要您使用 Google 工具栏与 Google 联系，例如向 Google 发送搜索查询，工具栏就会发送事先限定的标准信息，包括您的计算机的 IP 地址 以及一个或多

个 Cookie。 这些数据保存在 Google 服务器日志中， 并且受到我们通用隐私政策的保护。

这是 07 年 12 月份更新的版本，在更早以前的版本中，google 说明了除了收集 IP 地址和用户 COOKIE，还收集用户搜索的关键词等 。

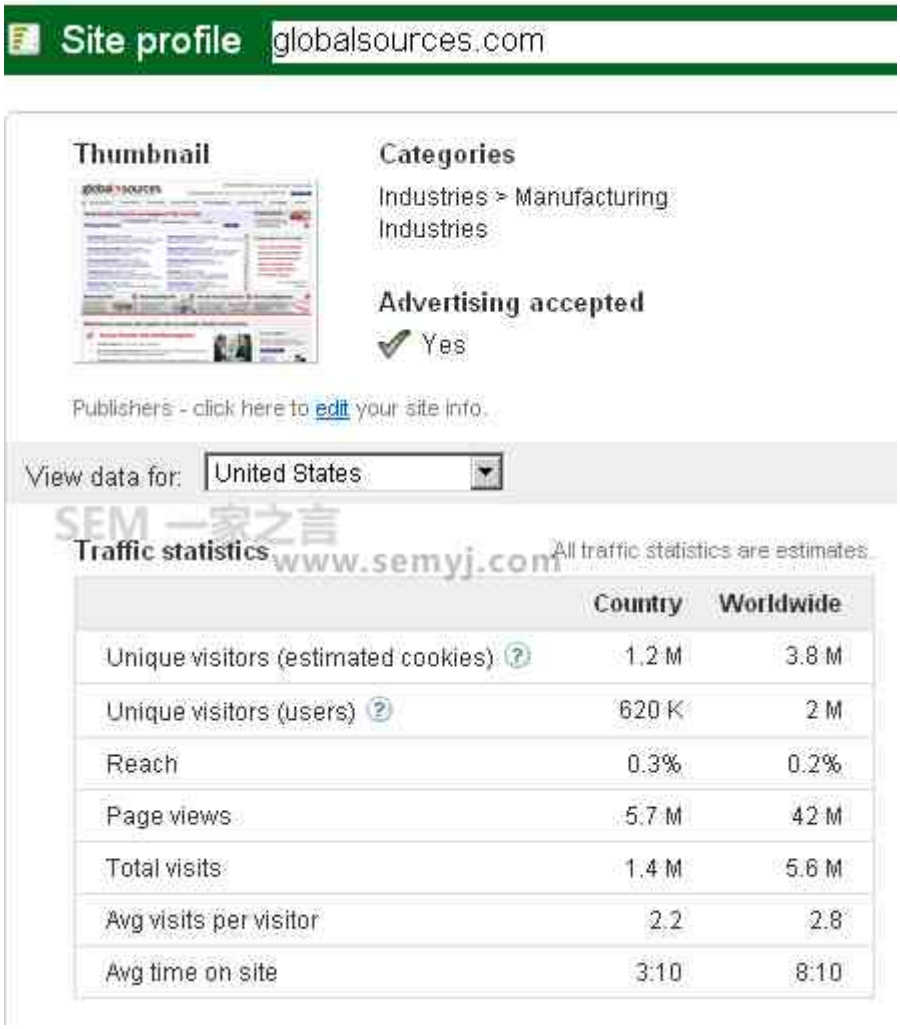
隐私条款里还有更多的信息说明 google 收集了很多用户的数据。

而在搜索结果页面，又有另外的跟踪方法，大家可以随意搜索一个词语，然后再“查看源代码”，就会发现这样的代码：<a href=”http://www.alibaba.com/” target=_blank onmousedown=”return clk(0,”,”,’res’,’2′,”)”> ，这是用来跟踪哪个搜索结果被点击的。

工具栏配合固有的跟踪方法，就可以的知道哪些网站好，哪些网站差了。

还有工具栏应该会跟踪用户在一个网站上的停留时间，粘度等等。

google trends for website 和 google adplanner 可以看到所有那些流量大的网站的流量。如下图：



以上是 adplanner 里对 globalsources.com 的流量描述。 停留时间和粘度都写得清清楚楚。globalsources.com 没有安装 Google Analytics 来统计流量，从 adplanner 的帮助文档里可以猜想到是 google 工具栏收集的了。

这种利用 google 工具栏改进算法是比较合理。

google 工具栏比 alexa 工具条装机量要大得多，它应该是现在装机量最大的工具条了。既然 alexa 现在还能在统计各个网站的世界排名，那 google 统计到的数据显然比 alexa 准确很多。最低限度，google 工具栏统计到的信息不一定知道哪个网站好，但是一定知道有哪些网站差。

Alexa 为了防止作弊，每个工具条都有一个全球唯一的编号的。这点 google 工具栏也有，在隐私条款里这样写着：

几乎所有的 Google 工具栏版本 （低于 4.0 的 Internet Explorer 版 Google 工具栏除外） 都包含一个或多个唯一应用程序编号。 这些唯一应用程序编号在运行 Google 工具栏时 必不可少，因此不能禁用。 当您安装或卸载 Google 工具栏时，这些编号和表示运行是否成功的消息 会发回 Google。此外，Google 工具栏会定期 与我们的服务器联系，以自动请求最新版本， 此请求还将发送唯一

应用程序编号以及 可选 工具栏使用 和配置统计信息。 Google 不会将这些唯一应用程序编号与您或您的 Google 帐户相关联。

Alexa 工具条统计网站流量的方法已经被一些人拿来作弊。在 05 年，也有国外的牛人曾经用 google 工具栏作弊，不过很快被封杀了。 以 google 的技术和快速反应，相信 google 工具栏统计到的数据能保证权威性。

而 google 的这个做法，给如何做好 SEO 也是很有启发的。

SEO作弊与反作弊

network

先讲一个作弊方法。以下的一个作弊方法，至今还能行得通的。

代码如下：

```
<TABLE>
<TR>
<TD HEIGHT=”1000″ BGCOLOR=”#000000″ BACKGROUND=”White.jpg”>
<FONT COLOR=”#FFFFFF”>隐藏文字 隐藏文字</FONT>
</TD>
</TR>
</TABLE>
```

这段代码，搜索引擎看到的是一个黑色背景下有一些白色的文字，这是不算作弊的。但是用户看到的就是一片白色，不会看到里面的文字。原因就是用一张白色的图片作为了背景。在以 table 布局的网页里，如果同时定义了一个 table 的背景颜色和背景图片，它是优先显示图片颜色的。这样，用户看到的是一片白色背景下的白色文字，当然就看不到这些文字了。这种作弊方法利用了一点：就是搜索引擎至今不能识别一张图片的颜色。

当然搜索引擎还有很多其他弱点。迈克·摩尔曾经说过：确实有办法愚弄搜索引擎，但是只有少数人能真正办到。其实他就是其中的一个，因为他自己做了 20

年的搜索引擎技术研究，在搜索引擎领域有很多专利，能从头到尾建立一个搜索引擎。

但是他那样的专家，是不会用一些作弊的方法来做 SEO 的，原因就是这样做太蠢了。

这要从搜索引擎反作弊策略说起。一个搜索引擎成功的反作弊策略一定是这样的：

1，允许算法被探测出来，而且即使算法被公布，搜索结果的公正性都不会受太大影响。要这么做的原因就是不希望和作弊的人陷入到一种猫捉老鼠的死循环当中。如果老是以堵漏洞的做法来修正算法，那永远都没有尽头。出于这样的考虑，搜索引擎会把那些无法被作弊的的因素在排序算法里放到比较重要的程度。

2，尽可能用一切技术手段自动检测，当技术手段不能解决问题，就用人工来解决。然后把人工发现的问题又反馈给自动检测机制，使自动检测越来越完善。

现在的 google 基本上就是这样来做的。在现有的排序规则中，那些无法作弊的和能精确反应内容的因素，都是很重要的排序因素。

当然 Google 也不排斥频繁的调整算法，这也有出于给用户一个最好的搜索体验考虑的。

至于技术检测和人工审查，google 也一直在做。

google 很早就有匿名蜘蛛来检测一个网站是不是在作弊的。如果去分析网站的服务器 LOG 日志，就会发现它们。

你会发现，有的爬虫，通过 IP 查询是来自 google，但是它没有自己的声明（user-agent），这就是 google 的匿名爬虫。它会判断你有没有对 google 爬虫特别对待，做一些隐藏页面，还会解析 Javascript 文件和 CSS 文件等等。有人用 CCS 文件来隐藏内容，这种事情现在是不用去做的，google 都能查出来。

Google也有人工审核机制，从webmaster tool 里提交的问题，都是有人工跟进审核的。以下就是号称google内部流传出来审核规则，可以点此下载。

既然 google 反作弊那么优秀，那文章一开始提到的那个作弊方法怎么解决呢？

那个方法 google 确实检测不出来，但是用这个方法的人，到最后还是会被 google 发现作弊。

google 的反作弊是“善意原则”优先，是假设你这个网站是没有作弊的，但是用其他所有作弊的特征来检查。 用了我提到的这个方法，在用颜色隐藏内容这一块是没事了，但是会在堆砌关键词，反向链接，以及其他很多方面路出马脚来。google 就是相信，一个在页面上隐藏内容的人，也一定会去做垃圾链接群发等

等其他作弊的事情。就像现实生活中一个吸毒的人，当然也是爱打架的，或者爱偷东西的，总有一件事情让你进局子里。

而你假设其他什么都不做，就是用那个方法隐藏一点内容，其实你也不能得到什么。因为你仅仅是隐藏内容的话也不会有排名的。

google 就是这样捍卫了自己排名的公正性。

对这些了解得越多，就越发现作弊实在是费力不讨好了。（作弊源于不了解，通过正常途径提升 SEO 流量的方法有的是，为什么放弃那么多好的方法而选择差的方法呢？在现在的 SEO 界，你会发现一个现象，越是 SEO 刚入门的人越喜欢搞一些作弊的事情，而 SEO 从业越久的人，就越不会参与这些。）

想做一个优秀 SEOer 的人，对所有这些因素都要有一定程度的了解的。这样做即可以避开无意中犯下的错，又可以避免不必要的恐慌。

比如沙盒效应，很多人总觉得很神秘，其实从搜索引擎的角度出发没什么好神秘的。你要是站在搜索引擎的角度考虑问题，就觉得这是一个很有必要的措施了。你也会知道如何发展自己的外部链接。避免 google 的反作弊手段落到你网站上。

还有，关于重复内容，google 一定是“善意原则”优先的，它甚至会帮你处理掉因为网站大量采用模板带来的重复问题。

要做到了解这些，就是不断的实践，学习和实验。

最近的美剧《Lie to me》非常好看，有一个印象我很深刻，就是他们会定期做一些实验，来了解人类各种复杂微妙的表情后面隐藏着怎样的心理活动，会定期形成报告。这是一种非常好的研究程序。

面对 google,我们就像那些心理学家面对人类的心理一样，很多东西是你不了解的。你去测试，就能得到独家的资料和信息。这也是我博客很多东西的来源。

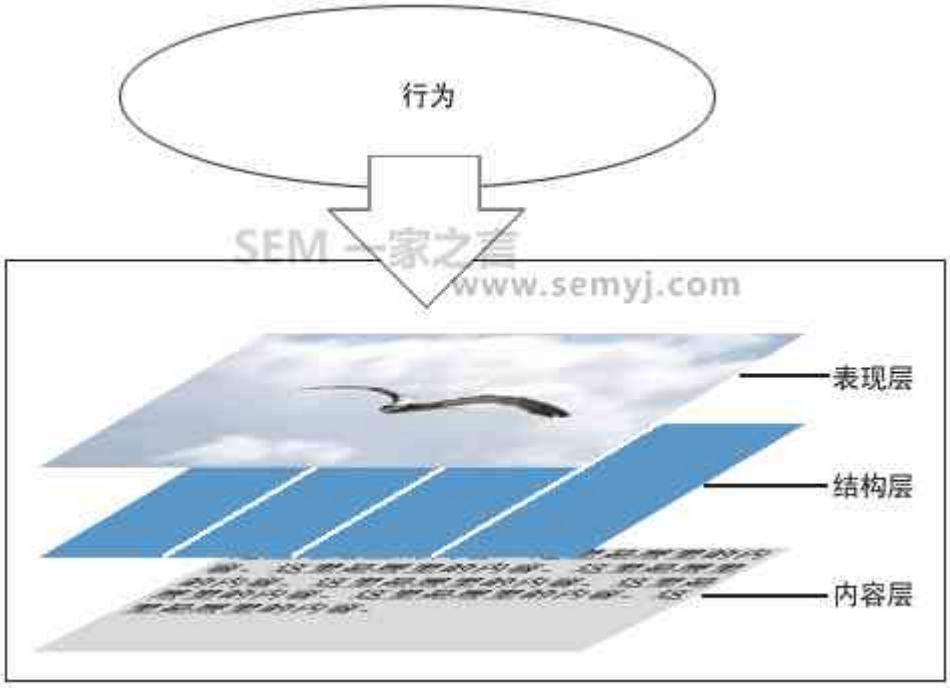
上篇博客中提到的那个GSA，就是一个绝佳的测试工具，具体的做法还是以后再讲。

把Web标准化进行得更彻底一点

说起 Web 标准化，大家马上就想到 DIV+CSS 网页布局。用 DIV+CSS 做网页布局的优点至少包含以下这些：对开发人员可以减少开发和维护成本，对网站可以减少服务器带宽消耗，对用户可以提高加载速度，对搜索引擎可以有利于内容索引。

DIV+CSS 是 Web 标准化的一种体现。但是不要认为只要把网站做成 DIV+CSS 的表现形式就可以了。关于 Web 标准化在 SEO 上的应用，还可以走得更远的。

一个网页应用了 DIV+CSS 排版，就有了以下的一个网页模型。



这个网页模型体现在网页代码里面是这样的：

内容层，就是一些纯文字信息，还有一些非背景的图片。

结构层，就是一些 html 标签，如 、 <h1> 、 <p>等。

内容层和结构层的内容是搜索引擎能看到的。

表现层，就是 CSS 文件。为了让搜索引擎看不到这些代码，一般是作为一个 CSS 文件外调的，如：<link rel=”stylesheet” href=”
http://www.****.com/style.css” type=”text/css” />

这个 CSS 文件里的代码差不多是以下这样的，也可以有图片，做背景用：

```
html, body {
padding:0px;
background-image:url(../images/bg.gif);
}

#logo {
width:258px;
float:left;
}
```

在这三层之上的，就是一些负责交互的 Javascript 文件。也为了让搜索引擎看不到那些代码，是外调的。

转换一下角色，站在搜索引擎的角度想象一下，当搜索引擎分析一个网页的时候，怎么来理解这个网页讲了什么内容呢？

有些所谓的搜索引擎爬虫模拟程序（http://tool.chinaz.com/Tools/Robot.aspx），会把一个网页上所有的文字抓下来，然后看看爬虫看到了什么。其实这是很不科学的。因为光看这些文字，搜索引擎怎么来理解上下文的关系，还有哪些内容是重点？

要做到理解上下文还有理解重点，就必须借助结构层里的 html 标签。因为这些标签表达了信息的层次。比如<h1>~<h6>表达了是一个标题。<a href=” *” title=> 中的 title 表达了一个链接的替代描述文字。<b>和表达了内容中应该重视的部分。

table 布局的网页也有这些标签，但在表达信息的时候，它的嵌套结构，还有大量对搜索引擎无效的代码防止了信息层次的表达。

即使 DIV+CSS 排版的页面，有时候一段内容后的一个 没有关闭，搜索引擎都不能很好的理解网页内容。div+css 排版的页面采用嵌套结构也不友好。

还有，很多人可能一味的追求关键字密度而不太重视这些标签。其实与关键字密度相比，搜索引擎更关注关键字所处的位置。搜索引擎的对于站内因素的重要性排列充分模仿了报纸的排版。所以，你加了大量的关键词发现排名上升了，可能并不是关键词密度上升的缘故，而是因为碰巧把关键词加在了比较重要的位置。

还有一些人这几个标签应该是注意到了的，但是因为没有从根本的角度去考虑过问题，所以还有很多其他事情没有去做。比如：

Web 标准里化有很多可以利用的东西，文字标签就有<abbr>、<acronym>、<cite>、<dfn>等，这些标签分别表达的意思是搜索引擎完全能识别的；还有，title 这个属性不光是可以加到<a>标签里的，还可以加到<abbr>、<acronym>等标签里的，你要是弄个关键字密度什么的，这些标签里完全可以利用起来；加粗文字还可以用标签……甚至<table>标签都还有一个 summary 属性可以加内容的。

还有很多SEO人一直没有关注的“微格式”。（ “微格式”维基百科的解释。 ）

通俗的讲，微格式就是一种类似 XML 的结构化的语义标记。它试图让机器更加容易的理解网页上的内容。现在这种微格式搜索引擎是可以读取的。我这个博客用的是 wordpress 程序，你现在查看源代码的话就可以看到微格式的一个具体应用。如：

<a href=”http://www.semyj.com/archives/74″ rel=”bookmark” title=”Permanent Link to 几句开场白”>

（这是 wordpress 系统自带就有的。）

你只要遵照这些标准，开始去应用，慢慢进化，就会发掘很多技巧类的东西。有些技巧我或许以后会在各个章节中体现出来。

把 Web 标准化进行得更彻底一点，一定让你受益匪浅。

“丰富网页摘要”，让你的网站与众不同。

在 6 月 15 的《谷歌中文网站管理员博客》中，出现了一篇介绍“丰富网页摘要”的文章。

看到这篇博客，我就知道 3 年前对 google 的预测，现在终于变成现实了。不过没想到在众多网站还没普及这些新技术的情况下，google 这么快就开始应用了。

关于“丰富网页摘要”的详细介绍，大家可以点上面的链接去了解。“丰富网页摘要”，用一句话说明一下就是：你给你的网站的某些内容，用特定的格式标注一下，就可以让你的网页在 google 的搜索结果中，显示更多的结构化的信息。比如那个博客上的截图：

丰富网页摘要

当用户搜索“drooling dog”，用户能在搜索结果页直接看到网页上的产品有多少人评论以及价格范围。

这些信息是在原网页中本来就有的：http://www.yelp.com/biz/drooling-dog-bar-b-q-colfax

如果你的网站也能这样呈现搜索结果给用户，有什么好处自然是不用说的。接下来要讲的是 google 为什么会采用“丰富网页摘要”。下面给大家呈现一下以前我的分析过程。这样你也能预测搜索引擎以后会做些什么了。

我们如果站在搜索引擎的角度去看，就发现作为一个通用搜索引擎，其实是非常不容易的。搜索引擎面对的是上百亿的网页，先不说分词、索引、以及抓取和排

序等等的技术。先来看搜索引擎如何判断网页上有什么内容，就发现是件很复杂的事情。

互联网上可以说什么样的网站都有，体现在网页的代码里，什么样的 HTML 写法都有的。 在具体的网页设计上，有些网站用模版做网站，可能仅仅只有一个区域内的内容是有效的；而有些不用模版，网页上从头到尾都是有价值的内容；每个网站的代码都是如此不相同而且混乱，但是搜索引擎还要通过这些 HTML 代码来判断你的重点内容。搜素引擎要从这么多繁杂的网页里提炼有价值的内容给用户，那个过程非常的痛苦。

我曾经和前 yahoo 中国的工程师一起做过一些事情，发现这样的互联网现状太考验一个公司的技术水平了。到如今，像 google 这样的公司，在处理网页噪音的时候都还遇到很多困难的。 所以，如果有一种统一的格式和标准，让大家来遵守，大家把网站里的内容都用这个标准把信息结构化的话，那对搜索引擎来说是一件非常幸福的事情。

现在博客搜索里，google 已经开始应用一些现成的标准了。如：

在google的博客搜索里搜索“SEM一家之言”，出现这样的搜索结果：

google 博客搜索

大家看这个搜索结果，已经把博客发布的时间和作者给列出来了的。但是可以看到，这两个页面的排版和页面代码都是不一样的。那搜索引擎是怎么准确的知道这两个信息的呢？ 特别是第二个结果，google 列出的那个时间的格式和我博客上的格式是不一样的。

原因倒非常简单。因为这两个博客都提供了 RSS 供稿，在 RSS 文件里，都用一个标准的格式写明了时间和作者这些信息的。

RRS 文件

google 通过读取这个标准化的 RSS 文件准确的抓取到了这些信息。

除了应用 RSS 这个通用的标准，google 还试图创立一个自己的标准来规范化很多信息。像 google base 就是其中的一个实验的项目。google base 这个项目以后会有专门的介绍。它试图把很多的信息都规范化，如你卖的东西的价格和产地，是否提供运输等等；你的房屋租售价格和位置；甚至一个学校的课程表，都可以用特定的格式标准化。

google base 还在不断的发展和完善中，在目前的应用中，凡是 google base 里的信息，都有可能在相关的搜索结果中排在靠前的位置。

比起自己去创立和推广一个标准来，应用现成的标准无疑是最省事的。“微格式”和 “RDFa”就是一个这样的现成的标准。具体的应用在《谷歌中文网站管理员博客》的那篇文章中已经说明了。

这些标准的应用起来效果是非常好的。

还是同一个词语“drooling dog”， 用美国 IP，在英文版的 google 上搜索，在第 2 个搜索结果中，还是出现那那家“Bar B Q”，而且有个地图标明了地址，地图旁边还有地址和电话。

单独的地图和地址电话信息

这个搜索结果不是那个“本地商家”的搜索结果。因为这个结果只显示这一家店的地址，而且特别标注了这家店的地址和电话。

这幅地图和地址信息在那个网页上本来就有的，只是这个网站把它们用特定的标准标注了才有了这种效果。

我的这个分析方式，就是站在搜索引擎的角度，来考虑如何提供更好的搜索结果给用户。这是一种很好的 SEO 方法。以后大家也可以从这个角度来考虑 SEO。

顺便说一下的是，上面那个博客搜索里的时间都比 RSS 文件里的时间慢一天的，这是因为 google 服务器所在的时区比中国时间慢的缘故。当然谷歌是不会处理这些小细节的，包括最近谷歌的首页上找不到登陆的地方也是。

一个关于SEO的故事

94 年，杨致远创立了 yahoo 搜索引擎。95 年就出现了第一批人数比较多的 SEOer。不过当时没有 SEO 这个概念，那些人也不觉得自己是 SEOer。很少有人是想获得商业上的利益的，大部分人都是为了好玩或者获得一点小小的虚荣心。

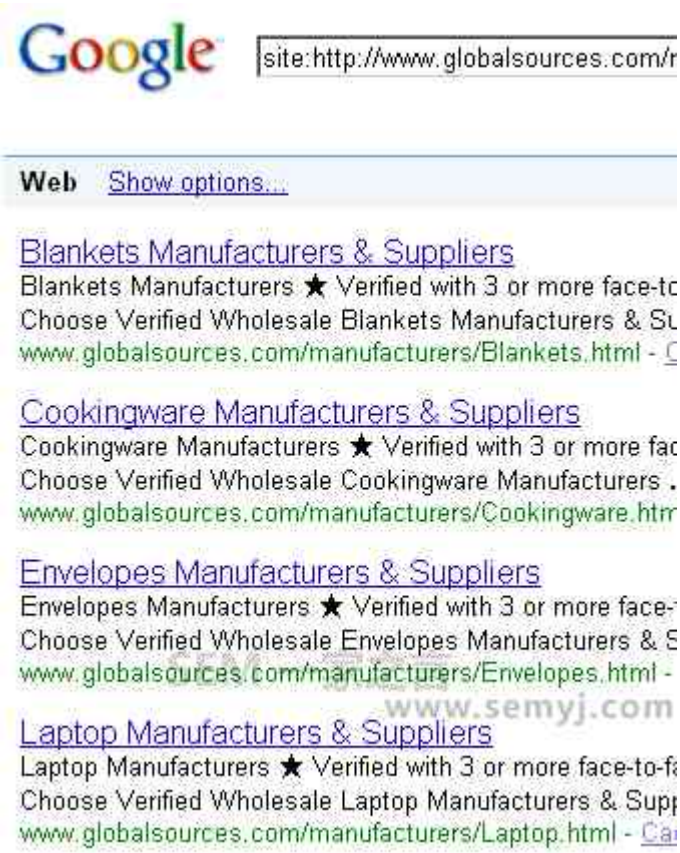
当时 yahoo 这个搜索引擎还没有爬虫，所有的网站都是靠人工编辑审核的。而在 yahoo 网站，除了用搜索框，访问者主要是依靠一个树形的分类目录来找网站。在某个具体的类目里，说起来大家可能不信，里面的排序竟然主要是按网页的第一个英文字母从大到小排序的。如：

排序

这里的排序规则很简单，就是新加入的网站置顶一段时间，排在最前面。后面的就按网页标题中第一个英文字母从大到小排序。网站开始增多以后，很多人的网站被挤下去了。有些人就想了一些“花招”来把自己的网站提上去。

不是按字母排序吗，有些人就觉得数字肯定比字母排得好，结果还确实是这样的。他们就硬生生的把标题第一个字符改为数字。后来发现字符要比数字排在前面一些，那就又在数字前加一个符号，如键盘上能直接打出来的字符：@,#,$,%,& 等等。所以，在 95 年的 yahoo 目录上，出现了很多乱七八糟的字符。但是这么多字符，本身也是有一个排序先后的，更多的字符还有“↑⊙★∵■♀∠……”等等。最后，经过少部分“高人”的摸索，发现“★”这个字符是能排在最前面的字符。后来，很多网站主在写标题的时候，都有意无意的在标题前加一个或者多个“★”。

若干年后，我在分析 globalsources 的 description 的时候，看到了下面这样的东西，真的是会心一笑。



很多地方都用星号

http://www.google.com/search?hl=en&q=site%3Ahttp%3A%2F%2Fwww.globalsources.com%2Fmanufacturers%2F*.html&btnG=Google+Search&aq=f&oq=&aqi=

给 globalsources 做 SEO 的这个人叫 Stephen，是从 97 年就开始做 SEO 的。后来我和他有机会一起工作，我一直没和他谈起这段历史，倒是他跟我说了他要把这个“★”放在 description 里的原因。他现在把“★”放在 description 里，是出于提高转化率的角度考虑的，关于这点以后再说。

yahoo 在 96 年就清理掉了那些带字符的标题，不过还是保留了字母排序这个规则直到 04 年。不过，很多网站主都保留了在标题中加符号的爱好。他们的想法是：虽然 yahoo 取消了符号，但是还有其他一些搜索引擎或目录还沿袭着 yahoo 的做法。这个理由倒还勉强说得通。但是还有一些人，是信息不灵通，还在看以前很老的一些人的文章，按他们传授的方法在做。

这种情况，其实到今天了还是一摸一样。我今天看到很多的 SEO 人，还在用 4，5 年前的方法在做 SEO，很多还是一些作弊的方法。这个是值得我们深思的。我们一定要弄清楚很多 SEO 做法的来源，然后再在这个基础上改进。

比较早上网的人，应该还记得以前的网站标题中，有很多是加了符号的。但是那个时候已经基本上只有装饰作用了。

下一篇博客分析 globalsources 的一个 SEO 做法，看这个网站是怎么应用前面谈到的几个 SEO 因素的。

SEO关键词的选择

这篇博文属于老调重弹。下面要叙述的内容，有些国外 SEO 人也曾经说过，但是我还是从我的角度说一说吧。

SEO 关键词的选择是一个策略方面的问题。它一开始就决定了你是不是走在正确的路上。

要做好 SEO 关键词的选择就要从研究用户的搜索习惯出发。现在把你自己当做一个搜索者，你大概会在搜索引擎上搜索一些什么关键词呢？

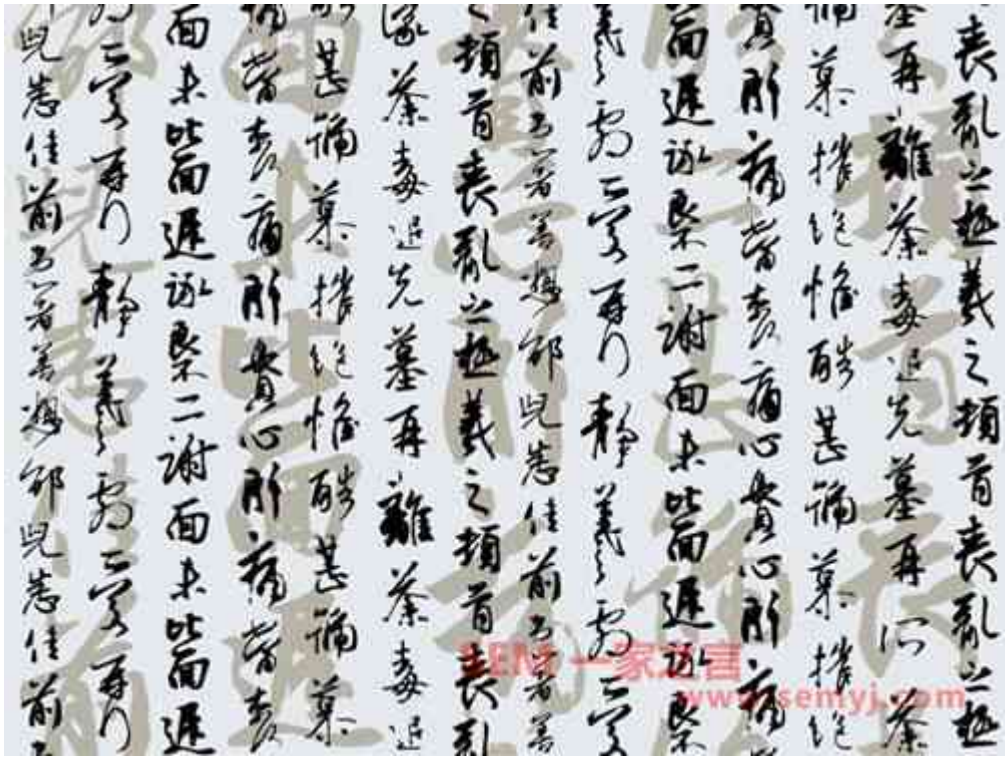
一般的搜索者的搜索行为可以分为以下三种：

1，导航型搜索。就是用户清楚的知道某个网站（或信息）是一定存在的，只需要搜索引擎导一下航，找到这个网站即可。如搜索“搜狐”、“youtube”、“163 邮箱”等等。

2，问答型搜索。就是用户搜索一些问题，希望得到这个问题的解决办法。如搜索“怎么做宫保鸡丁”、“如何做 SEO”、“怎样买一个好的 MP3”等等

3，产品型搜索。就是搜索具体的某个产品或事物。如搜索“爱国者 MP3”、“免费邮箱”、“IBM 笔记本 T61”等等。

大家先不要看下面的答案，就上面提出的三种搜索猜想一下：哪种搜索类型的搜索量是最大的？哪种小一点？哪种最小？



关键字选择

整体来看，真实的数据就是：

第 2 种“问答型搜索”的搜索量是第 3 种“产品型搜索”的 1.2 倍。

第 1 种“导航型搜索”至少是第 2 种“问答型搜索”的 4 至 5 倍。

这个问题，我问过很多人，但是至今还没有一个人的答案是上面我提的那样。上面 1、2、3，三种搜索，按搜索量大小排序，大部分人排的次序是：3>2>1，而实际情况是：1>2>3。

在大家的印象里，“产品型搜索”应该是最大的。但是你要是用 google 关键词工具和百度指数去查一些信息，就明白真实的状况了。

以下是用百度指数搜索“免费邮箱”和“163 邮箱”得到的数据：

两个关键词的百度指数对比，不一定反映搜索量

从 google 关键词工具得到的数据：

关键字	广告客户竞争程度	本地搜索量：6月	▼全球每月搜索量	匹配类型：精确
与所输入字词相关的关键字 - 按相关性排序				
[163邮箱]		数据不足	246,000	添加完全匹配
[免费邮箱]		数据不足	49,500	添加完全匹配

精确匹配得到的两个词语的搜索量

整体上，“导航型搜索”是占绝对的优势的。至于“问答型搜索”，因为很多问题提问的方式是有各种变形的，所以不能直接和其他两种搜索类型比较搜索量的大小。

做 SEO，看数据是必不可少的，而且要站在一个整体角度考虑问题。在上面的列子中，其实除了我们这些沉浸在互联网行业的人，对于普通大众，很多人都不知道 163 邮箱的地址其实就是 mail.163.com 的。所以“导航型搜索”才大行其道。

这个现状在国外也是一样的。我因为一些特殊渠道，看到了“美国 yahoo”08 年全年下来用户搜索的一些数据。前 150 个搜索量最大的词语，无一例外，都是那种“导航型搜索”。还有，让我自己也很吃惊的是，“google”这个词语竟然排在了前 10。他们竟然不愿意或者不知道在浏览器地址栏里直接输入 google.com 来访问。

知道了上面这些，你就不需要仅仅在“产品型搜索”里和别人争得头破血流了。可以看到，现在的 SEO，无论国内国外，还有好大一片蓝海。

“问答型搜索”非常适合和软文一起做，有时候转化率做到 10%以上也不是难事。

百度是一直知道哪些搜索量大，哪些搜索量小的。“百度知道”应该就由此而来。在具体的“应用”上，百度也知道怎么去做。看看下图就明白了。

利用“导航型搜索”

有时候三种搜索的关键词可以一起配合着做。三种搜索关键词，分别代表处于不同阶段的客户需求。

所有具体的做法都留给大家去探索。不要 SPAM,不去 SPAM 反而可以想出很多效果更好的方法来。

“锚文本”在SEO方面的重要性

“锚文本”，google 的中文称呼叫“定位文字”。要清楚“锚文本”在 SEO 方面的重要性，就要先看一个案例。

在 google.com 上搜索三个中文字“国际站”，排在第一的是 alibaba.com。但是我们来查看这个页面，这是一个英文页面，网页中从头到尾都没有出现“国际站”三个中文字。其实，页面代码中一个中文字都没有，看到的仅有的几个中文字其实是图片。

产生这种现象的原因，就是在阿里其他的子公司的网站页脚下，在介绍 alibaba.com 这个网站的时候，用的锚文本就是“国际站”三个字。如：

淘 宝

雅 虎 口 碑

中 文 站

支 付 宝

各个网站页脚的锚文本

仅仅是因为有这么多的锚文本，就可以让一个英文网站在搜索一个中文字的时候排名第一。这个案例应该能给人启发的。

如果大家碰到一些 SEO 新手，以为 SEO 就是修改 title,内容里加一加关键词等等，觉得 SEO 很简单的人，就可以拿这个案例给他们看。这个案例，站内关键词因素是零，但是依然可以排第一。

对于热衷于 SEO 理论的，觉得 PR 值决定排名的人。就可以知道 PR 值的重要性是多少了。这个网页的 PR 值是 7，在一定程度上确实帮助了“国际站”这个词语排在第一。但是，如果你看看网页中其他好几个密度很高的词语，排名是不怎么样的。这个时候高 PR 值不能让那几个站内因素也很好的关键词排名上去。

————————————————-

在这里，我稍微说一下上次和ZAC的nofollow的争论。我对ZAC本人是没有意见的，我见过ZAC，觉得他是个不错的人。以前我在ZAC博文里也看到过一些不同意的观点也没有反驳。这次之所以要这么“高调”的说明一下，最直接的想法是不想一些新手陷入误区，也不想一些观点以讹传讹。到时候真有那么多人弃用nofollow，对很多SEO人是件受损失的事情。还有一点可能大家不知道的是，有时候我们其他部门的人会用ZAC的观点来质疑我正在做的事情，因为他们认为ZAC是专家。

所以我才开始针对 ZAC 的说法写一篇博文，最终目的是不希望大家盲目迷信专家。其实 ZAC 本人也没有说自己是专家。他在博客上写得很明白了：“介绍和研究世界最先进搜索引擎优化 SEO 技术。我的目标是每天都总结国际上搜索引擎排名研究的最新动态。 - Zac” 。把国外的 SEO 理念和方法传播到国内，我也觉得意义蛮大的。

但是最终，就像我在 “SEO作弊与反作弊”中说的那样：SEO还是靠自己的实践，学习和实验。我开这个博客的目的，也是希望树立起“一种凡事从实际出发，看数据，讲事实的SEO方法。”

回到 PR 值的问题，这可能是一些人最关心的一个问题，国外也有很多 SEO 人喜欢讨论这个。那些 PR 值计算公式，衰减因子这些当然是要了解的。但是要务实一点的话，在实战中就忘记这些东西好了。PR 值和排名是有关系的，但是还有其他众多因素和 PR 值一样重要，甚至可能更重要。所以单纯说 PR 值和排名的关系，是不大的。

站在局部的角度，是需要对所有的细节了解清楚的，我上篇文章就建议大家去了解搜索引擎爬虫。但是站在整体的角度，很多细节是不需要那么去死抠的。那些“用nofollow控制PR的流动”的人，还在斤斤计较有PR“损失”掉了没被计算进去，那就是玩理论玩过头了。这种“损失”就让它损失好了，不妨碍你拿到好的SEO排名。

而且“用 nofollow 控制 PR 的流动”也是和 google 对着干，所以 google 现在才会改 PR 算法。

用了 nofollow,网站权重是更好的集中了的。至于 PR 值，不存在浪费一说。

SEO 做得好，有两个标志：1，和 goolge 是双赢关系。2，一切因素都是可控的。关于这个以后再说。

————————————————-

关于“锚文本”，google 至少从三个地方“告诉”了我们它的重要性。

1，《google 网站质量指南》，在关于 nofollow 的描述中，这样写到：“Google 不会传送这些链接中的 PageRank 或定位文字。”

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=96569

“定位文字”就是“锚文本”。大家可以想一想为什么一个链接被 nofollow 以后，锚文本不会被传递呢？

在《google 网站质量指南》的另一篇文章：

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=94034

这样写到：“在评估网页的价值和相关性时，Google 不仅会考虑网页本身的内容，还会考虑指向此网页的链接的定位文字。”

2，在 google webmaster tool 中，google 已经给你收集了所有指向你网站的外部链接中的前 200 个“锚文本”。

定位文字

这个“定位文字”的统计，很多人可能没怎么关注。你要是用这里的前几个关键字去搜索，很有可能就发现你的网站在这些词语上排名很好。

3，google 官方发布的 SEO 入门指南当中也说明了锚文本的重要性。

http://googlewebmastercentral.blogspot.com/2008/11/googles-seo-starter-guide.html

（这个链接需要代理服务器或者 VPN 才能访问）

大家可以看到，Google 已经明确的说明了如何做好 SEO 了。

但是基于快速排名的想法，很多人不会看这些基本的说明，希望从一些专家那里得到排名的秘笈。

这也是我希望很多人摈弃的想法。所谓的秘笈，都包含在我们平常所熟悉的东西中。即使是国外的有些 SEO 专家，也是会有臆测和道听途说的观点的。

经常参加搜索引擎营销大会，有一个很大的感触。就是每次 google 的代表人员都在极力强调《google 网站质量指南》以及 google webmaster tool 等等的重要性。但是下面很多提问的人，就是视而不见的。我参加过这么多次搜索引擎营销大会，还很少听到过在 《google 网站质量指南》不能找到答案的问题。下次

如果 google 的代表还来搜索引擎营销大会，建议他们这样说：“我们把所有的排名秘笈都写在了《google 网站质量指南》当中了。”

最后再重复一下的是，“锚文本”这个细节你是需要了解的，但是在全局角度，就要合理的应用了。我不提倡用“锚文本”作弊的方法，到时候损失了流量不要说是从这个博客上得到的观点，详情我不便于说。你就用“锚文本”得到你该得到的流量即可。

我也谈一下nofollow

写这个文章是因为看到 ZAC 的两篇博文。觉得很多人都可能受到误导，所以特意说明一下。

一篇是ZAC一年前写的nofollow控制站内权重，一篇是最近写的nofollow会浪费PR和权重。

我要说的是，这两篇文章里表达的观点都错了。如果 ZAC 有在大型网站做 SEO 的经验，那他一定会知道他错误的理解了他引用的文章。

nofollow 是 05 年 google 推出的一个属性，理论上加了 nofollow 属性的链接爬虫都不抓取。当初推出这个属性主要是为了应对日益泛滥的群发作弊。后来小部分 SEO 人认识到了 nofollow 链接对 SEO 的好处，有了一些应用 nofollow 的技巧。其实，直到现在，nofollow 都还是一个有利的 SEO 手段。

我们先来看看 Matt Cutts（google 反作弊组的老大）的博客里关于 nofollow 的描述。（这也是 ZAC 引用的链接）

http://www.mattcutts.com/blog/pagerank-sculpting/

这里面从来没有说明 nofollow 会浪费 PR 和权重，这里面只有一个观点，那就是你即使加了 nofollow，也不会使你的 PR 值增高。原文中有个例子：

“So what happens when you have a page with “ten PageRank points” and ten outgoing links, and five of those links are nofollowed? Let’s leave aside the decay factor to focus on the core part of the question. Originally, the five links without nofollow would have flowed two points of PageRank each (in essence, the nofollowed links didn’t count toward the denominator when dividing PageRank by the outdegree of the page). More than a year ago, Google changed how the PageRank flows so that the five links without nofollow would flow one point of PageRank each.”

意思就是说：你原来有一个页面 PR 值有 10 点，这个页面中有 10 个链接。nofollow 之前每个链接分到 1 点的 PR 值。如果你 nofollow 掉其中 5 个链接，你以为剩下的 5 个链接每个链接能分到 2 点的 PR 值，但是实际上，每个链接还是只能分到 1 点 PR 值。

也就是说，在单个链接的 PR 值的计算上，根本不会听从 nofollow 这个属性。博客中的意思是，那些被 nofollow 的链接的 PR 值和锚文本不会被传递。没被传递，不是意味着 PR 值就节省下来被传递到其他链接了，计算单个链接的 PR 值的时候，nofollow 还是不能影响到。google 这样做，是不想网站所有者为了控制 PR 值，把一些好的内容给 nofollow 了。但是，也就只有这样而已，并不代表 nofollow

就没有用了。更不会浪费 PR 和权重了。原文的中 Matt Cutts 的一个回答说的很明白了，就是你要想一想没有 nofollow 之前是什么状况。

要清楚 nofollow 的作用以及为什么说 ZAC 错了，要从头说起。

一个网站，只要页面稍微一多（比如只要有个几百页以上），就遇到一个问题，就是搜索引擎在短短的几天内，没办法把你的所有网页都抓取一遍。几百个网页都这样，那一些 B2B、B2C、招聘网站、分类网站等等稍微大一点的网站这个问题就更加严重，如果你有 google webmaster tool ，去“抓取统计信息”里看看就明白了，爬虫一天访问的页面量可能不到你页面总量的 1%。页面没有被爬虫抓取，就意味着这些页面要被收录是不可能的。一旦收录情况不理想，整个网站要获取 SEO 的机会也是少了很多。

当 nofollow 属性出来，一些做 SEO 的人合理应用了 nofollow 属性以后，发现爬虫每天的抓取量就应声上去了，接着网站整体的收录量上去了，整体的 SEO 流量也上去了。（这种方已经是一少部分 SEO 人屡试不爽的技巧）为什么呢？

因为只要你合理的应用 nofollow 属性，就会帮助爬虫节省很多时间，还可以让爬虫更多的抓取那些有收录价值的页面。比如：你网站上有一些链接是“注册”、“发送反馈”、或者“添加到购物车”等等的链接，这些链接是没有收录价值也不会有排名的。这种“垃圾页面”放上 nofollow 属性以后，爬虫就不爬了，就会去爬别的没有放上 nofollow 属性的链接。这样，你节省了爬虫时间（在一定的的期间，爬虫呆在你网站上的总时间是相对固定的）。然后，也可以“控制”爬虫抓取重要的页面，让那些还没被爬虫抓取的好页面有被收录的机会。虽然爬虫可能还是没有百分百抓取完你的全部网页，但是已经改善太多了。

然后，一般大型网站都是用模版的，理论上，你在一个页面上 nofollow 掉 10 个链接，如果这样的页面有 100 万个，那你就节省了 1 千万个爬虫抓取“垃圾页面”的机会。而我的实际工作中，有时候一个页面上可以 nofollow 掉 50 个链接，以及涉及到上千万页面。

nofollow 正确的做法就是这样的，nofollow 要控制站内权重，也是通过这种做法实现的。所以 ZAC 的两篇博文都没有理解那些在一线 SEO 人员的做法。nofollow 又怎么浪费权重了呢？

至于 PR 值，真正务实的 SEO 人从来不在意这个，因为事实上 PR 值和排名关系不大。我相信以前擅用 nofollow 的人也没有想过用它来控制 PR 值。

前段时间，有个朋友问我快速提高SEO流量的做法，现在这就是一个。如果不是为了让大家不受到误导，我可能不会公开讲出来。正是因为以前有人误导 nofollow可以控制PR值，才有人去把自己有价值的页面也nofollow掉了。google 这么做非常及时，免得大家误入歧途。那些企图用nofollow控制PR值的人也真的是自食其果。我在另一篇博文里也说了，SEO其他提高流量的方法有的是，作弊是最蠢的。

做整站优化，如果你老板给你一个网站要你马上提高 SEO 流量，那你就合理的应用 nofollow，我敢说，在 2 个月内涨 30%以上的流量是完全可能的。这是白帽 SEO,国外很多顶尖的网站也都是这么做的。发展到后面，nofollow 还有很多非作弊的技巧。

如果有人还不信我说的，其实，在《google 网站质量指南》里，明确的说明了 nofollow 的应用。

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=96569

这份说明更新于 2009 年 5 月 12 日。正是 Matt Cutts 宣布 nofollow 不能塑造 PR 值的时候。里面说的意思就是，被 nofollow 的链接，PR 值或锚文本不会被传递。但没有说这些没被传递的 PR 值就被节省下来了。

另外，Matt Cutts 在博客中也说，google 明明就是从 webmaster tool 后台通知网站主了，或者在《google 网站质量指南》里写明白了，很多人就是不会看的。

《google 网站质量指南》是最权威的 SEO 教材，这个我要到下篇博文里讲为什么。我在阿里内部的每次培训都要推荐大家去看《google 网站质量指南》。另外我的博客很多读者是上过我培训课的阿里人，可以看看我发的 PPT，上面说了，nofollow 是提高爬虫效率的方法。

怎么样去学**SEO**（一）

我前面写的这些东西，很多看起来是很偏门的。不过我并不是故意挑这样的内容去写。这些东西，其实正是我想要很多 SEOer 去探索的。也都是我平常工作中会用到的知识和工具。

本来这篇文章我打算讲一下分词和索引库，但是写到一半，发现如果我不把怎么学 SEO 讲明白的话，有些人可能又觉得我写了一个偏门的东西了。 还有，上次网站备案的时候很多人想让我写一下怎么学 SEO。所以这篇文章就讲一下怎么来学 SEO。我希望这篇文章能广为传播一下，让很多 SEOer 意识到这些。

先定义一下，这里所说的 SEO，是纯粹的指从搜索引擎获得大量优质的流量，把网站要实现的最终效果最大化。有些人把网站运营的内容也纳入到 SEO 范围，不过这篇文章不讨论网站运营的东西，尽管我是很建议大家把网站运营和 SEO 结合。

在“SEO作弊与反作弊”里，我把SEO和心理学家类比过。其实他们有相同点也有不同点。

相同点就是：你研究的对象，有很多的秘密等着你去探索，对于研究它的人，经常都会有东西是你不知道的。 不同点就是：心理学是一门自然科学，自然界创造的东西，以人类现在卑微的探索能力，你永远无法说你有多接近真实。但是搜索引擎，却是完全由人自己创造的，所以理论上还是能完全把搜索引擎弄明白的。

这就谈到了一个真正的 SEOer 应该学习的第一大技能：弄懂搜索引擎相关技术和原理。

我是不太同意那样的说法的：做 SEO 不需要太技术化，考虑好用户体验就不用担心什么了。考虑用户体验，做好内容是绝对应该做的，这其实是在网站运营方面就要考虑的东西，但是技术化也是另一个非常重要的东西，它很多时候甚至是决定性的条件。

用常识想一想，我们在一个叫做 google 或者百度的平台上拉流量，但是我们对于这个平台内部是怎么运作的竟然不了解，这不是搞笑吗？？ 农民种菜还要了解天气和季节对农作物的影响呢。

不光要了解，而且要把这些知识应用到 SEO，这种了解还需要达到一定的深度才可以。 像迈克·摩尔，做了 20 多年搜索引擎开发，在搜索引擎领域有多项专利，他也在做 SEO。而现在有些 SEOER,仅仅知道怎么样排列关键字就觉得是在做 SEO 了。这个差距真的不是一般的大。当然，最后的结果也是很悬殊的。 迈克·摩尔说过：只有少数人能真正控制搜索引擎。我一直认为他自己就是那种能控制搜索引擎结果的人，记得 2 年还是 3 年前，他的团队就硬生生把一个网页在搜索“SEO”的时候排在了第四，那个网页，当时连一些资深的 SEOER 都看不明白怎么能排在第四的。

关于怎么在页面上排列关键字，一个报纸的排版人员其实更擅长，他们非常明白怎么兼顾阅读性和内容突出度。如果罗列关键字的技巧就是 SEO，那也难怪很多人局外人说 SEO 非常简单了。

至于怎么去了解，我推荐大家先去当当网搜一下，有很多的介绍搜索引擎原理的书籍。如果有条件，还需要自己做一个搜索引擎。如：可以用 Lucene 之类的自己搭建一个搜索引擎实践一下。 虽然 google 对自己的很多技术都很保密，但是放心好了，搜索引擎并非 google 和百度独有的东西。把类似的搜索引擎了解清楚了， 你再来看 google 和百度，发现绝大部分还是一样的。

了解得比较透后再来做 SEO，你就能从搜索引擎的角度出发来看待你在做的事情。你会非常的理解做搜索引擎的人，了解他们的短板在和痛苦在哪里。了解他们将来会怎么去改进他们。

第二大技能：了解网站制作相关的技术，至少能独立做一个静态网站。

了解完了搜索引擎，还要了解我们服务的对象 - 网站。能从头到尾自己做一个静态网站是最基本的要求，当然能做动态网站更好。这个就要求你懂动态网页开发语言，精通 HTML，基本的 CSS，javascript 等等。一个好的 SEO 人员，最好能帮助网页设计师改写和优化代码。这个技能，在你做内部优化的时候，能帮助你很多。

SEOer 每天做的事情，非常多和网站的技术相关的。 要配合搜索引擎的要求对网站进行调整，仅仅依靠工程师和网页设计人员是不行的，你要清楚里面的细节。不然你都不知道改动某个地方可以对 SEO 有利。

会网页开发后，还要知道网站架构相关的知识，服务器架设、CMS、还有数据库的性能调优等等都是需要了解的。比如，在稍微大一点的网站，提高单个页面的加载速度，对 SEO 都是非常有利的。但是哪些改进可以提高加载速度呢？ 依靠工程师，它给你提高 10%的加载速度你已经感恩戴德了，但是如果你自己懂的话，把页面冗余代码除去，js 外调合并压缩，图片实时压缩，页面 cache， 马上提高 400%的加载速度，这个效果是不一样的。

对这两大技术方面的了解，越详细越对自己有利。我在“锚文本在SEO当中的重要性”中提过：好的SEO，一切因素都是可控的。那要拿什么来保证一切因素都可控。 是首先你了解到了所有的因素，然后你具备了控制这些因素的能力。这样，你做的每一个改动，你非常清楚带来的效果会是什么。这个就是SEO的核心竞争力。

怎么样去学SEO（二）

在学习搜索引擎的相关技术和原理的时候，特别要注意研究爬虫。这也是从常识出发来想的：搜索引擎和网站之间，是爬虫把他们连在一起的。这就是我那么执着于研究爬虫的原因，所以博客里有好几篇都是介绍爬虫的。还写了一些与之相关的：nofollow， URL 静态化等等。

一个 SEO 同行也认可这种方法的。今年，我们另一个部门的领导在参加美国 SMX 大会的时候，碰到了一个以前在 google 工作了 8 年、现在辞职做 SEO 顾问的人。那个顾问给的意见就是：SEO 要站在搜索引擎的角度来看待问题；然后把网站的技术问题解决好；那些技术问题，不是可有可无的，而是不掌握就不能开始做 SEO 的。我听到这个转述，真的毫不怀疑他确实在 google 工作过 8 年的人。大家还可以在《google 网站质量指南》里多看看，无处不充斥着很多技术问题。

很多优秀的 SEOer,都在各自独立做 SEO，但是最后大家都殊途同归、都在朝正确的方向走的。这是因为他们对这些常识有了解，知道怎么走是不会错的。对常识的了解深到什么程度，你就能有优势到什么程度。其他一些一线的 SEO，在这些常识的基础上做得既大胆又创新，连我这个对手都不禁要为他们喝彩。

我基于对搜索引擎技术的学习，使我都非常想和 google 的人交流，因为我明白他们设计某些规则的思想，以及碰到的问题，有时候觉得自己说不定也能给他们提供一点解决方法。在 08 年 4 月厦门的 SMX 大会期间，我就和朱建飞单独聊了一个半小时，主要谈他的本行-anti spam。我相信他那时是非常愿意和我谈并且印象深刻的。

第三种技能：数据分析能力。

数据分析能力是做 SEO 应该具备的基本能力。很多影响 SEO 效果的重要因素，都可以从数据上反映出来。不管是前期的预测，还是流量波动后的事后分析，都是离不开数据分析的。SEO 数据分析需要做到三步：1，知道哪些因素可以数据化；2，建立适当的数据公式或模型；3，分析这些数据和流量之间的关系。这里的每一步做到什么程度，也就决定了你的整体能做到什么程度。比如“建立适当的数据公式或模型”这一步，有些国外优秀的 SEM 公司就做得很好，它根据这个公式得到的一个数据，能很准确的反应你在这一块和竞争对手的差距在哪里。
这个能力，是先要有正规的教材帮助你入门才可以的。有很多现成的分析方法需要你先掌握，然后再根据 SEO 数据分析的特点来变通。同样的数据，分析方法不一样，得到的结论也不一样。

这个数据分析中，要特别注意 LOG 日志分析。SEO 数据分析中的数据来源，很大一部分来自服务器 log 日志。这里记录了爬虫和用户访问网站的种种信息。如果你具备了相关的能力，可以把 log 日志里的任意数据合并拆分来分析的。比如从 log 日志里分析爬虫的到访的次数，每次停留的总时间，单个页面的平均停留时间…… 等等任意维度。

还有第四种技能：了解你要排名的那个搜索引擎。

可能有人很奇怪为什么这个能力可以和其他能力并列，并且好像和前面谈到的第一种能力是重复的。

是这样的：

如果有人问我为什么觉得自己能有信心在 google 上做好 SEO，我脱口而出的回答会是：“因为我非常的了解 google”。从 google 一开始为什么会做这个搜索引擎、一开始他们在技术上怎么考虑的，到 google 现在推出的各种各样产品的由来和现状，到 google 将来会对哪些产品做什么样的改进我觉得自己都能体会到。虽然了解得很粗糙，但是还是一直在努力探索。

记得也是 08 年 4 月在厦门，我跟一个人在极力解释 google 应该会推出自己的浏览器，他还是半信半疑的。不过我是非常肯定这个事情会发生的，而且这只是一个开始而已。google 那时和 firefox 的合约恰好快到期，浏览器这么重要的互联网入口，按照以往 google 的做事风格是绝对不会不理的。而且这个和 google 长期的战略目标非常吻合。再有，做这样的产品很合施密特（Google CEO）这个人的胃口。

还有，原创性是现在 google 排名因素当中一个非常重要的因素。了解 google 以前历史的人，都非常明白 google 会用什么算法来检查原创性。这个算法在 98 年 google 诞生之前就有了。google 这个网站的灵感来源于布林（google 创始人一）开始做的数字图书馆项目，在图书领域，也是存在很多的抄袭行为的。在这个算法基础上，之后合并一些算法应用到了搜索引擎。不过这种算法，在面对上百亿网页的时候，会产生很多的“噪音”。所以在判断原创性方面 google 现在的表现不完美，不过一直在努力，而改进办法之一就非常依赖 google 数据中心的效率。

不光这个算法依赖 google 数据中心的效率，google 拉开和竞争对手的距离，也是依靠数据中心的。很多人一直不明白这个才是 google 的核心竞争力之一。

关于这些以后都会有相关文章介绍的。

掌握这些技术知识，就有了一个非常好的基础。接下来就是长期的跟踪和实践。那么很多人会问：其他一些能力呢，那些很多人都强调的比如 SEO 关键词的选择、内链外链的分析技巧等等？

我觉得那些技巧都是在这些基础之上长期实践得出的常识性的东西。比如 关键词的选择好了，像在“SEO关键词的选择”中那样的技巧，其实是只要你有数据分析的意识就是可以发现的。要是再进化下去，还可以发现更复杂的技巧。比如，从另一个角度来分析，长尾关键词和热门关键词应该偏重于优化那种呢？我的结论就是：优化大型网站的时候，把资源偏重优化热门关键词，会让你丢失掉 50%以上你本来应该得到的流量。这个结论可以说颠覆了很多人的想法，但是数据分析可以为我们揭示其中的缘由。

怎么样去学SEO（三）

SEO 新手入门，学习资料我只推荐一本书、一个网站。

一本书是《搜索引擎营销:网站流量大提速》。

此书的作者之一就是我常说的 Mike Moran，研究搜索引擎技术 20 多年的人。这个书的中文版只有 2006 年那版的，后来的更新版本只有英文版。最新版本的购买链接http://www.mikemoran.com/ 。(后注：2009 年 10 月，这本 08 版的中文版也已经出版，淘宝上有售。)

这本书完整讲述了搜索引擎营销的过程和方法。对于有些人来说，这本书好像平淡无奇，那可能是因为受一些错误观点误导太久的缘故。真正的 SEM 方法，没有秘籍，一开始就是那么简单，但是要深入下去就很复杂。我当初拿到这本书的时候，是用一天一夜的时间连续不断看完的，之后又反复看了 20 几遍。因为我看到一些我辛苦总结的东西，作者一笔带过就讲清楚了。这本书很厚，即使 SEO 已经从业很久的人员，这本书相信精读下去你还是不断会有收获的。

这本书还有很特别的一点，就是专门花一个章节来讲述你如何向你的老板和你的同级部门去推销你的 SEO 方案。因为即使是现在，在一个大中型网站里要说服大家接受某些 SEO 改动还是很费精力的。

其实有很多好的 SEO 教程，本来我以为只有我力捧这本书的，没想到有一天拥有 11 年 SEO 经验 Stephen 也跟我极力赞扬这本书。他很少这么推崇一本书的，所以大家快买来看看吧。

一个网站是指《google网站质量指南》。http://www.google.com/support/webmasters/

相信大家也看到我多次引用里面的文章，那是因为好的 SEO 方法都在这里面的缘故。很少有人知道的是，《google 网站质量指南》里其实有几百篇文章，涉及到 SEO 的方方面面。里面提到的很多方法都是 SEO 的最终解决方法。

比如：如何去写 meta description。

主流的做法会告诉你：在保证语句自然流畅的时候，适当的重复几次关键字。

但是 google 会这样告诉你写：

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=35264#1

1，为不同网页创建不同的描述。（注意网站一级描述和网页一级描述的区别）

2，在描述中包含清楚标记的事实。 （这样的写法就很好： <meta name=" description" content="作者：A.N. Author，插图制作者：P. Picture，类别：图书，价格：$17.99，页数：784 页"> ）

3，程序生成的描述。（大中型网站都这么做）

4，使用高质量的描述。（要考虑排名的转化率，排在第一名的有时候不一定比第四名获得更多流量，怎么样让用户最先点你的网站而不是别人的，就需要你多多注意）

google 的这四点建议，我觉得就是写 meta description 的最终解决方案。一定要多实践，才能体会得到为什么。

《google网站质量指南》里尽管事无巨细写了很多，但是还有很多是点到即止的。像我在《Lynx浏览器在SEO上的应用》里写的Lynx就是。还有一些是要你自己去发掘的，像《我也谈一下nofollow》里写的nofollow属性，《google网站质量指南》是在很多篇文章里从不同的角度谈到了它的特点。很多条目是因为那件事情本无法仔细描述而没有详细写，还有一些是因为不能透漏更具体的信息以免被喜欢spam的人利用。

Google 之所以把这些方法公布出来，是想和这些给他提供内容的网站达成双赢的局面。搜索引擎应该明白的一个道理就是：搜索引擎的内容来自于其他网站，如果这些网站都按照搜索引擎提供的一个质量标准优化自己的网站，把自己有什么内容都告诉搜索引擎，哪些内容是重点都标示出来，就可以达到这样的局面：一，搜索引擎检索到高质量的内容给了用户。二，那些网站拿到了属于自己，比作弊得来的还更好更多的 SEO 流量。

google 其实就是用这样的一种策略化解了原本过度的 SEO 和搜索引擎之间的对立关系。这一点是国内的百度一直不会明白的。

这个质量指南其实google对外宣传的时候一直在强调的，也非常重视它的更新。告诉大家一个小技巧就是：一旦google针对某个算法有更新，google也会在第一时间更新《google网站质量指南》里相关的条目，所以请随时留意右下角的更新日期。有新的算法参与排名，也会在里面加上相关的内容，并给予你非常合适的指导。如在《“丰富网页摘要”，让你的网站与众不同》里说的微格式，google已经增加了这个内容的指导。

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=99170

再来谈一下网上的其他 SEO 资料。

因为 SEO 算法一直都是不公开的，所以 SEO 行业就没有一个自己的标准。这就使这个行业内充满了很多完全不一样的观点。对于新手，有时候一些观点会让你误入歧途的。很多道理都可以自圆其说，但是可能完全错误的。

我观察下来，发现还是应该去参考那些在一线的，有实践经验的人。有很多SEOer，即使是新手，因为他们有自己的网站练手，也已经做得很不错了。

做 SEOer，包括我，都要经历一个阶段：那就是做了很多自己认为是好的优化动作，但是流量就是不涨起来，排名也没有什么变化；感觉能用的都用上了，但是就是没有起色。这个就表示你对影响 SEO 的因素还没有了解清楚，还有就是不具备控制好这些因素的能力。一定要清楚的是：好的 SEO，一切因素都是可控的。

接下来我会写得勤快一点，把各类因素都揭示出来。

分词与索引库

分词是很多做 SEO 的人常听到的概念，为了让大家在这个方面不会有疑惑，现在要来讲一下分词以及索引库。这也是更深入的了解搜索引擎的开始。

搜索引擎每天都是在处理一个基本的需求：用户搜索一个关键词，搜索引擎马上找到相关的网页给用户。这个过程要怎么实现呢？ 下面就分步来了解这个过程。

首先搜索引擎要尽可能多的把互联网上的网页搜集下来，这样能提供大量的网页给用户查询。这一部分由爬虫来解决，顺着互联网上的链接一个个往下抓取。最后就有了一堆记录着网页各种信息的资料库。目前的现状，最后能使这个资料库里有大概 100 多亿个网页。资料库里记录了这些网页的 URL，整个网页的 HTML 代码，网页标题等等信息。

然后，搜索引擎拿到用户输入的这个关键词后，要从这个资料库里把相关的网页找出来给用户。这里就碰到好几个问题了：

1，要怎么快速的从上 100 亿个网页里找出匹配的网页的呢？

要知道这是从上百亿的网页里找符合这个关键词内容的网页，如果像用 word 里那种用 ctrl + F 轮询的查找方式的话，即使用超级计算机，也不知道要消耗多少时间。但是现在的搜索引擎，在几分之一秒里就实现了。所以一定是做了一些处理才实现的。

解决办法也倒简单，就是建立一份索引库。就像我们查《新华字典》一样，我们不会翻遍《新华字典》的每一页来查那个字在哪页，而是先去索引表那里找这个字，拿到页码后，直接翻到那页就可以了。搜索引擎也会为上百亿的网页建立一个索引库，用户查询信息的时候，是先到搜索引库里查一下要找的信息在哪些网页，然后就引导你去那些网页的。

如下图：

索引库

2，索引库里用什么样的分类方式？

我们知道，《新华字典》的索引表是用字母列表或者偏旁部首的分类方式的。那么搜索引擎的索引库里是怎么分类的？是不是也可以用字母列表的方式？

搜索引擎如果以字母列表的方式排列索引库，那么平均每个字母下要查询的网页数量是 100 亿÷26＝3.85 亿 ，也还是一个很大的数字。而且搜索引擎上，今天是 100 亿个网页，过不了多久就是 300 亿个网页了。

最后，终于找到一个解决办法：索引库里用词语来分类。

因为尽管互联网上的网页是不断激增的，但是每一种语言里，词语的数量都是相对固定的。比如英语就是一百多万个单词，100 亿 ÷ 1 百万 = 1 万；汉语是 8 万多个词语，100 亿÷8 万=12 万 5 千。都是计算机很容易处理得过来的。

用词语来分类还有一个好处，就是可以匹配用户查询的那个词语。本来用户就是要查这个词语的，那我就按这个词语去分类就是。

所以，搜索引擎的索引库，最后就是这个样子的：

单词	URL
mp3	www.mp3.com, en.wikipedia.org/wiki/MP3, www.winamp.com, www.mp3raid.com, www.amazon.com/MP3-Music-Download www.last.fm, www.creative.com/products/mp3/ .....
.....	.....
player	www.bbc.co.uk/iplayer/,www.itv.com/ITVPlayer/ www.videolan.org/vlc/,www.winamp.com/,www.apple.com/quicktime/download/ , www.real.com , www.adobe.com/products/flashplayer/ .....

模拟的索引库

理论上，当用户输入关键词“mp3 player”搜索时，搜索引擎就从 “mp3”那行和 “player”那行里拿出同时都有的、交集的 url 来即可。

上图也是现在英文版的 google.com 上的真实排名情况，可以看到 www.winamp.com 这个网站在搜索“mp3”的时候排第 4 位，在搜索“player”的时候也排第 4 位。当搜索“mp3 player”的时候，因为没有其他网站比它更匹配这个词语，所以它排在了第一位。

排在第一

当搜索引擎把一个网站抓取下来后，接着要做的事情就是把网页里的词语分开放到索引库里。分词在这个时候就要应用到了，所谓的分词，其实很简单，就是把词语分开而已。

英语的分词好处理一点，因为英语的每个单词之间是用空格分开的，基本上就只要处理一些虚词、介词，还有一些词语的单复数，变形词等等。但是中文的分词就复杂很多了，句子中的每个字都连在一起，有时候即使是人来判断，都还有产生歧义的时候。中文的分词有很多方法，也很容易弄懂的，如正向切分法，逆向切分法等等，网上有很多相关的资料。

谷歌的中文分词方法是从国外一家第三方公司买的。百度的分词方法是自己创立的，可能在词库上面比谷歌有点优势。不过其他方面差了一些。

当爬虫找到一个网页的时候，在搜索引擎看来，这个网页就是一大堆词语的组合。基本流程如下：

用户看到的

```
<div style="CLEAR: both"></div>

<div class=tradeshowCycleItem style="DISPLAY: none"><A href="
http://tradeshow.alibaba.com/trade_shows/10071001/The_11th_Asian-Pacific_Int%2526rsquo%2530_Plastics_%2526amp%
2530_Rubber_Industry_Exhibition_.html" target=_blank>Plastics & Rubber Industry Exhibition</A><br>
<div class="column one" style="LINE-HEIGHT: 18px"><A class="sampleborder tradeshowmer" href="
http://tradeshow.alibaba.com/trade_shows/10071001/The_11th_Asian-Pacific_Int%2526rsquo%2530_Plastics_%2526amp%
2530_Rubber_Industry_Exhibition_.html" target=_blank><IMG src="
http://img.alibaba.com/tradeshow/images/20090820/750/7057976b79becbd4516e68f8073aa17_4f1600.gif" border=0
alt="Plastics & Rubber Industry Exhibition"></A> <A href="
http://tradeshow.alibaba.com/trade_shows/10071001/The_11th_Asian-Pacific_Int%2526rsquo%2530_Plastics_%2526amp%
2530_Rubber_Industry_Exhibition_.html" target=_blank>Rubber & Plastics </A><br><SPAN class=c>24-27 Nov 2009<br>
[China]</SPAN> </div></div>
<div style="CLEAR: both"></div>
        <div>
                <ul class="tabs" style="margin:3px 0;float:left;">
                        <li class="tradeshowCycleTab current">1</li>
                        <li class="tradeshowCycleTab">2</li>
                        <li class="tradeshowCycleTab">3</li>
                        <li class="tradeshowCycleTab">4</li>
                        <li class="tradeshowCycleTab">5</li>
                            <li class="tradeshowCycleTab">6</li>

                </ul>
                <a class="more" style="margin:3px 2px;float:right;" href="http://tradeshow.alibaba.com/">More</a>
        </div>
</div>

<script language=javascript type=text/javascript>
(function(){
var gConfig={
sTabClass:"tradeshowCycleTab",
sItemClass:"tradeshowCycleItem",
eParentNode:get('tradeshowCycle'),
sSwitchEvent:"mouseover",
```

网页源代码

<title>Alibaba.com - The world's largest online B2B marketplace.</title>

Help

electric scooter

<h3>Browse by Category</h3>

Popular Searches:

alt="digital photo frame"

title="Manufacturers"

<h1>digital photo frame</h1>

搜索引擎关注的

搜索引擎的处理过程

看完这个流程图，应该能给大家在做内部优化的时候有所启发的。

我建议大家再去看一下《把Web标准化进行得更彻底一点》这篇文章，还有《丰富网页摘要”，让你的网站与众不同》以及《SEO案例：锚文本、关键字、nofollow、Web标准化（一）》和 《SEO案例：锚文本、关键字、nofollow、Web标准化（二）》。那些文章和这篇文章一样，都是在讲同一个问题。

一定要站在搜索引擎的角度，把它的这些原理了解清楚了，才会让你明白哪些因素才是你应该关注的重点。

有人说：SEO 就是重在细节。这应该是经验之谈。但是不知道大家有没有想过的是：是不是可能原本这些看似细节的东西，其实就是应该注意的重要的东西呢？如果你不能控制好你的排名，有没有想过可能你以前特别在意的一些 SEO 因素，其实有些并不是重点？；而只是你把影响排名的部分因素弄错了？

上面的很多知识，其实在《搜索引擎营销:网站流量大提速》里都有提及的。那本书要去精读的原因之一就是它讲了很多看似很普通的原理，但是都是有用的。

比如在选关键词的时候，也可以参考一下这个词语的索引量。从上面的原理可以看出，这个索引量反应了这个词语在这种语言当中人们使用的流行程度。所以国外有些计算关键词 KEI 指数的公式里，也把这个关键词的索引量加入了进来。

有兴趣再追溯下去的朋友可以看看google黑板报上的这篇文章
http://www.googlechinablog.com/2006/05/blog-post_10.html

google Caffeine(咖啡因) 更新了什么

很多人很关心 google Caffeine 的更新，有些猜测说是为了应对 bing 的突起而做的改动。

前面讲了很多理论，那这次我们来实践一下，从搜索引擎的角度，来判断一下 google Caffeine 到底更新了什么。

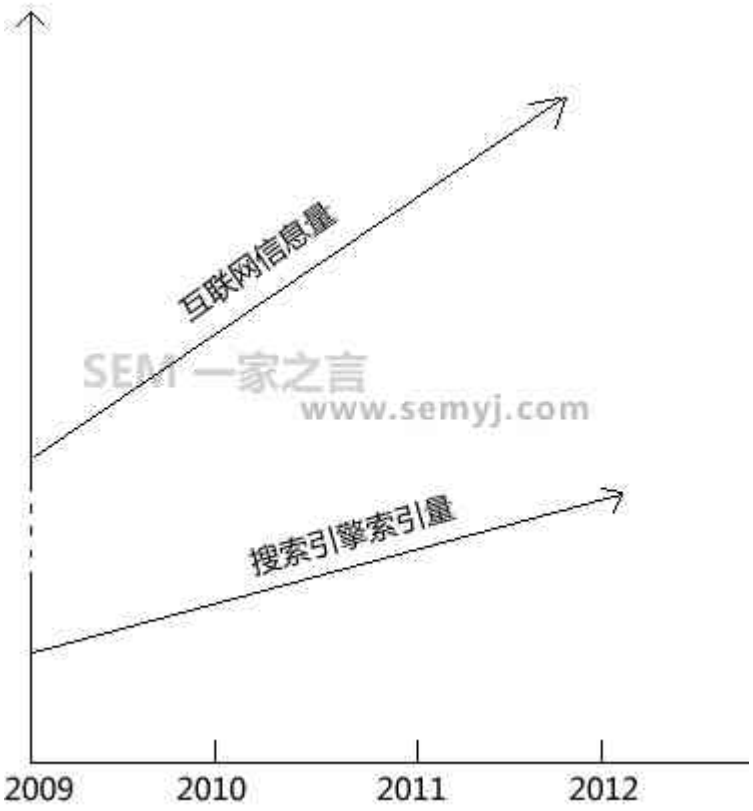
先看google官方的解释： http://googlewebmastercentral.blogspot.com/2009/08/help-test-some-next-generation.html （需要翻墙）

我觉得，这次改动主要是重写了搜索引擎的底层架构，涉及到爬虫、索引库、排序规则等等很多方面。提升的方向主要是 google 一直以来追求的速度，而速度的提高会进一步带来准确性和全面性的提升。

Google 是一个对速度的追求达到了变态的地步的公司。他们的价值观之一就是“速度为王”。举好几个例子：你可以在 google 首页看到，所有的代码被压缩成几行，因为这样能提高加载速度，甚至在代码变量的命名上，都是坚持能用 1 个字母的就不用 2 个字母的原则；google 非常多的产品大量应用 AJAX 技术，就是为了在速度上更快一点；在 google 的历史上，曾经想把搜索结果首页的默认条数从 10 条增多到 30 条，用户也乐意接受这个改动，但是测试下来，发现这样会拖慢 0.5 秒的速度就放弃了。

追求速度不光是为了用户打开页面快而考虑的。我相信 google 在 98 年就开始意识到这样一个瓶颈问题：摩尔定律描述了每隔数年计算机的硬件水平就翻倍。而互联网上的信息，也是这样一个规律。有人甚至说是每隔 9 个月互联网上的信息量就翻倍。搜索引擎要保证一个基本的信息查全率，就需要能跟上这种信息暴增的速度。

现在搜素引擎的索引量和互联网上的信息量是这样的一种关系：



互联网和搜索引擎

理论上来说，有越来越多的信息是搜索引擎找不到了的。如：现在百度在收录速度上落后于谷歌，所以谷歌上能找到比百度更多更新的结果。

有这样一个现状在前面摆着，我想搜索引擎想不在意速度都难。google 其实从一开始就知道如去做的。首先是有条不紊得增加数据中心的服务器数量，现在 google 所有数据中心的服务器加起来应该超过一百万台了，目前还在不断的修建数据中心。二是提升这些数据中心的效率。效率的提升有硬件上的也有软件上的。硬件上的就如：自己制造服务器，然后想办法提高每台服务器的速度和稳定性。所以 google 在服务器硬件上有很多自己的专利；软件效率上的升级也是一直都有的，但是近年来主要的精力应该是放在算法的调整上。我相信这么多年下来，google 已经积累了很多底层架构上需要改进的地方，代号“咖啡因”的升级就由此应运而生了。所以不管有没有 bing 的发布，google 都会做这样一个升级。

“咖啡因”的首要的改变会是改进爬虫的效率和提高索引库的速度。从表现上来说，“咖啡因”的第一个表现就是整个搜索引擎的索引量增加了。如果输入单词搜索，每个词语的索引量都增加了很多。搜索的速度也增加了，这是索引库也升级了的缘故。

还有一个我自创的方法，可以来看搜索引擎的整体索引量的。那就是在 google.com 输入“*a”去搜索。这个搜索的意思是把只要一个网页上有字母 a 或网页上某个单词里含有字母 a 的网页都找出来。当然一个网页在 99.999%的情况下都有字母 a 的，所以这个符号的索引量约等于整个搜索引擎的索引量。

“咖啡因”刚发布的时候，用这个符号去搜索，发现 http://www2.sandbox.google.com/ 和 google.com 的索引量差距有 80 多亿左右。 而现在你去搜索，发现都是一样的数量，大概有 254 亿。

索引量对比

所以现在有一个结论是可以确定的：“咖啡因”抓取的那些页面，现在已经列入到 google.com 的索引库里了。

只要排序规则不变，有更多的网页参与排名，这对谁都好的，所以 google 马上就应用了。

索引量增加后，还有另一个最直观的感受应该是：搜索一些长尾词，会看到很多以前不在首页的网页冒了出来。

“搜索引擎的速度跟不上互联网信息的增长速度”这听起来很让人觉得沮丧。不过其实搜索引擎并不一定要追求把互联网上所有的信息都抓取下来的。只要把有价值的信息都能抓取下来即可。那么如何判断一个信息是有价值的呢？这也要依靠数据中心的速度。

现在搜索引擎上的主要问题，不是信息太少了，而是原创的、用户需要的信息太少了。想一想我们自己在搜索引擎上找信息，哪一次不是找遍大量的网页后才找到想要的信息的呢？ 要让这些信息很容易被用户找到，基础就是数据中心的效率要很高。如:判断原创性的算法中，爬虫的效率和数据计算的速度提高了，判断原创性就更准确了。还有排序规则里很重要的链接因素，现在的 google 之所以能比其他搜索引擎更能给用户想要的搜索结果，来自于它 3 天就可以更新一次数百亿网页的速度，能计算这些网页彼此之间的关系。现在效率提高了，如果 1 天就可以 update 完一次，那计算出来的排序就更符合用户的需求了。

这次“咖啡因”的升级应用起来以后，那些依靠采集的垃圾网站会越来越没什么流量。搜索引擎已经索引了 40%以上重复的垃圾信息了，而还有那么多有价值的信息等着去索引，如果你是搜索引擎，也会把原创性高的网页的重要性越排越高的。有时效性的网页也是。当然依靠人为制造大量外部链接在做排名的效果也会大打折扣。

不过我觉得，google 还是会用更多的时间来测试这次改动。虽然本质上这次升级就是强化以前的一些理念。但是在一个这么大的系统里，这么一次脱胎换骨的改动会产生什么样的影响也还是无法预料的。

可以看到，爬虫、索引库、排序规则，无一不需要数据中心的速度更快。所以我在《分词与索引库》中说：google的数据中心，才是它的核心竞争力之一。google 也把速度快归结为自己成功的原因。

google 一直以来都在拼命拉大和竞争对手的距离，已经形成了牢不可破的竞争壁垒。bing 这个搜索引擎非常清楚这点，所以只有剑走偏锋，做一些 google 目前无法部署的事情。但是以后 google“咖啡因”完善并上线后，一定又可以为 google 拿下几个百分点的市场份额。

百度如何优化

已经有很多网友问我百度如何优化了，不过我一直不清楚如何来写更合适。

有好几个原因。首先是，很多知识，我不先介绍一下的话，到时候我写出来大家不容易理解。就如我博客刚开的时候，我就想写《SEO案例：锚文本、关键字、nofollow、Web标准化》这篇文章，但是如果一开始不讲一下为什么要重视锚文本、如何选关键字、nofollow的作用、为什么要标准化的话，很多人可能会觉得我只是讲了一个特别注意细节的案例。

所以我会先把 google 的优势讲明白，把搜索引擎是怎么运作的讲明白，然后才能把如何做百度优化讲明白。我博客里的很多文章都是从 google 的角度出发来讲 SEO 的，但是你都可以思考一下同样的事情要是百度来处理的话，会如何去解决。

第二个原因就是百度优化和 google 优化在技术上有 80%是相似的，所以不用特别的区分是百度优化还是 google 优化。而且，做一个网站的 SEO 工作，那些流程和方法都是一样的。说起相似性，你就可以看到：一个在 google 上排名很强势的网站，在百度上也会有不错的排名。还有，百度前几年不是一直宣称 google 侵犯了百度在超链分析法上的专利吗，虽然很荒谬，但是可以看出百度也是注重外部链接的。而且现在还有一个趋势就是，百度在算法上越来越模仿 google。

另一个原因就是其他方面的。互联网的圈子其实很小，百度排名确实有一些技巧，但是我今天在这里说明的技巧，明天大家就不能用了。

总体来说，百度的优化要比 google 英文的优化容易得多。现在我主要从事的英文优化，同时也在做部分百度优化的工作。在百度上，基本上现在只要特别注意去优化的词语，不是非常热门的话，非推广的搜索结果都排在第一了。（但是对于一个大型网站的话，这样特别注意优化热门词语，会让你丢失掉原本属于你的 50%以上的流量。）而我用的一些技巧，都是在英文 SEO 领域用滥了的技巧。

如果能把英文的 SEO 做透的话，做百度的优化感觉很容易。很早以前，在一些 SEOer 眼里，百度优化，非推广的搜索结果排在第一不是难事，要保持第一就很伤脑筋了。

接下来谈一下影响你做好百度优化的几个因素。

1，百度的搜索技术很糟糕。

不管是在爬虫，还是索引，还是排序算法上都有很多缺陷。比如搜索一个词语，同一个网站占据前几十个搜索结果，就是一个很低级的错误。而很多大型网站，也应该被百度的爬虫把服务器“攻击”得不行了吧。反作弊措施也很初级，所以一大批作弊的网站照样可以活的很好，对于一贯不作弊来做 SEO 的人来说，这点无可奈何。这都是技术上的，还有的就是人为的。比如很多百度认为“影响业务”

的网站，无论从抓取率还是排序上都是被分配了特殊的权重的，导致这些网站的SEO 流量怎么也做不起来。这样被区别对待的网站有很多，国内好几家著名网站包括阿里巴巴中文站都是这样，所以阿里巴巴中文站的 SEO 流量中，google 来的流量反而比百度高很多。 而百度自己的业务，却被赋予越来越高的权重，甚至自己网站上很多明显的垃圾页面，照样有好的排名。如：

百度搜索结果

尽管我是不主张 SEO 上作弊的，但是如果是小网站在百度上作 SEO，在注意百度的一些反作弊措施下，就适当用一些方法吧。

2，服务器效率低

《google Caffeine(咖啡因) 更新了什么》揭示了数据中心的效率是搜索引擎的命脉。而百度的数据中心无疑是没什么亮点的。经常性的当机和数据丢失，使国内的很多做SEO的人以为是自己的网站有了作弊问题而从搜索引擎消失了。一个大型的服务器集群，管理起来特别考验一家公司的技术实力。管理数据中心很像“打地鼠”的游戏，这边几个当掉的服务器刚刚解决，那边又有服务器当机了。google在这方面也遇到过很多问题，得益于他们的技术能力，现在都得到了不错的解决。虽然也有数据丢失，但是平均每份数据都有 2 份备用，所以在百度上出现的那种突然大量收录消失了的状况，在google上是很难见到的。

我经常看到很多人对于网站在百度上的收录量患得患失，其实很多时候，都不是因为百度处罚你。而是他们自己出了问题。

3，反作弊措施

无论百度还是 google，SEO 要做得好都要从这个搜索引擎的角度来看待你做的 SEO 优化工作。在百度上作 SEO，心里要时刻想着它有可能会用什么样的反作弊措施来检查你的网站。这个是在百度上做好 SEO 的秘籍。

举一个我操作过的案例：

以前给一个国内比较有名的网站做 SEO，因为比较遵守这条规则，把总的 SEO 流量从 6 万做到 267 万。（之所以流量翻了 40 多倍，还有一部分原因是他们以前喜欢用 ajax 技术，妨碍了收录。）后来流量一下子又降到了 30 万左右，因为那个公司有特殊的渠道，得到了百度内部的意见说网站优化过度，我当时特别纳闷怎么就优化过度了。后来就想，如果我是百度的话，我会如何判断一个网站优化过度呢？从国内那时的 SEO 现状来看，我会这么判断：因为每个做 SEO 的人，基本上都会去改 title, keywords, description。那我首先设置一个过滤条件，就是把那种每个网页上 title, keywords, description 都写了大量内容的网站特别对待，因为这些网站有了 SEO 的企图，所以也会在外部链接等等其他因素上作很多优化的动作的。这种被列入嫌疑的网站，只要流量有异常的大量增长，就开始严格清理。那个网站，在其他方面都无异常，甚至没有在 title, keywords, description 重复一次关键词，唯独每个网页， keywords, description 都是写了很丰富的内容的，当时为了写这些内容还花费了大量的时间，因为每个网页都要写得不一样。

想清楚了这个，我就把那种能不写 keywords, description 的就不写。甚至做了一个很大胆的决定，就是所有的网页都不写 keywords。因为我如果是百度的话，我绝对不会考虑把 keywords 作为影响排名的因素的。description 之所以不抛弃，是因为在搜索结果里，人还是需要阅读到的。这样改动后一个月，流量就恢复了，直到我离开那家公司前，SEO 流量都还稳定在 200 万以上。

因为这篇文章实在是罗嗦了，以后再讲一些其他案例。不久前 Matt Cutts 在博客里也说 google 不把 keywords 作为排名的因素。如果你在第一线实践的话，这些东西早就知道了。我以前就想把这个写出来，没想到 Matt Cutts 抢先说了。

4，人工干预

百度是一个很仇视SEO的搜索引擎，在他们眼里，SEO妨碍了他们的收益。这与我在《怎么样去学SEO》一文中提到的google与网站共赢的策略完全相反。google也有人工审核，但是是奔着处理作弊网站而来的，而且尺度还很宽松的。百度的人工干预会让你在做一些热门词语的时候非常麻烦，这点就不多说，大家都经历过。

这 4 个影响你在百度上作 SEO 的因素都是客观的。还有一个因素就是竞价排名，但是这个在慢慢消退。

不过还是可以在百度上把 SEO 做得很好的，那就是整体把握一个网站的策略，方法。SEO 应该做的是给网站带来利益，要在各种条件和资源的限制下，把网站的利益最大化。这是才是一个 SEOer 应该追求的目标，而不是今天有多少收录，哪个热门词没有排上去等等。我还是慢慢一个个的讲吧。

热门还是长尾？大中型网站的关键词优化策略

接下来的两篇文章，会讨论网站是选热门关键词还是长尾关键词，以及应该注意内部链接还是外部链接。

相信很多 SEOer 都有这样经历：开始做一个网站的 SEO 的时候，都是先选一些计划中要排名的词语，希望藉由这些词语在搜索引擎上获得大量的流量。在这些词语中，有些人选的是些热门词；有些人明智一点，是一些在当前的能力下能做到的适当热门的词语。在接下来的过程中，会集中很多的“资源”来做这些这些词语的排名。包括在 title 中适当的重复这些关键词；突出这些关键词的密度；外部链接指向这些关键词页面等等。不过，在计划中的词语还没有排名的时候，很多人会发现网站其实已经有一些 SEO 流量了。而去分析流量就会发现，流量几乎都不是计划中的这些关键词带来的，而是各种各样奇怪的长尾词。等到计划中的关键词有好的排名的时候，根据网站的不同，有些网站可能大部分流量靠那些热门关键词贡献，而有些网站则不然。

那么，在做一个网站的 SEO 关键词优化的时候，是把精力偏重放在做热门关键词还是做长尾关键词呢？

为了看清楚这个问题，我们需要来看看用户使用搜索引擎的现状。

不需要太多的数据参考，可以回想我们自己或者周围其他人使用搜索引擎的情况，就会发现：

1，用户搜索时输入的关键词越来越长尾。

用户是能自我学习的。一个从来没有用过搜索引擎的人，让他去搜索几次，他也马上能明白如何才能找到自己想要的信息了。那就是输入几个关键的信息点，才能把想要的信息找出来。尽管每年还是很多新手用搜索引擎，但是看看大家的搜索历史记录就知道了，输入的关键词越来越长尾，不用长尾词你是很难找什么信息的。而热门词的搜索量没有大家想象的大。

2，搜索引擎的搜索结果还远远没有到结果过剩的时代。

所谓“搜索结果过剩”，是指你随便搜索什么词语，都有一大堆你要的信息在搜索结果里。而现在的实际情况却是：就算输入了很长尾的词语去搜索，还是找不到他们要的信息。那些排在前面的网页，可能只是页面中包含了用户查询的一两个关键词就排在了前面，不过不是要找的内容。但是实际上，用户要找的信息是存在的，不过因为很多网站要么没有 SEO 意识，要么听从了误导的 SEO 方法，没有把自己的内容突出来而已。

看过《分词与索引库》这篇文章，应该要明白的是：即使那个关键词在页面中只出现了一次，搜索这个关键词还是有机会让这个页面出现在搜索结果里的。假设当用户输入的长尾关键词中有 4 个词语，你的网页上所有的 4 个词语都有了，那

理论上来说，应该是你的这个网页排到前面才是。但现状是很多这样的网页被埋在很深的地方。因为搜索引擎不光要看你网页上是不是有这个关键词，还要看这些词在什么地方出现。

那些听从了误导的 SEO 方法的网站，是把一些重要的位置让给了他们事先选定的词语，没有放这些用户搜索的长尾词。导致在这些热门的词语上他们可能竞争不过其他更有优势网站，而那些本来应该属于自己的流量，又被那些在重要的位置放上了长尾词的网页给拿去了。

在一个大中型网站，把资源偏向于优化热门关键词，是非常不明智的。从网站流量最大化的角度来说，毫无疑问的，长尾关键词才是需要注意的。

两者操作的难易程度上也不一样。做热门词的优化是很辛苦的。就像开始说的那种很多 SEOer 的经历，到最后他们会发现，历经千辛万苦，还有绝大部分计划中的词语没有好的排名。长尾词的优化就不一样。做长尾词不需要什么资源也不需要什么技巧，你只要把页面中的关键内容点放到重要的位置就可以了。（只是要注意这种重要的位置绝不止 title、meta 和 H1 那么简单。）这样你在这些关键词上面就有好的排名，能恰好匹配用户查询的关键词。何乐而不为呢？

再说一下《怎么样去学SEO(三)》里面讲的那个description为什么要那么写。

Google 在《网站质量指南》中指导大家写 description 的时候，第二点要求就是：在描述中包含清楚标记的事实。

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=35264#1

还举了一个例子：

〈meta name=”description” content=”作者：A.N. Author，插图制作者：P. Picture，类别：图书，价格：$17.99，页数：784 页”〉

这个例子之所以好，就是因为它把关键的信息点都写清楚了。没有写热门关键词，它只是诚实地把反映页面内容的长尾关键词列在这里即可。当用户找信息的时候，会凭借各种各样完全无法预料的长尾关键词来找内容，这样写，马上就可以让他们找到这个网页。像这个列子中，用户就可能只记得这本书的插图作者和价格，那么用户输入这两个信息搜索的时候，这个网页就是用户很容易发现的。

从竞争程度来说，长尾关键词的优化也要比热门关键词容易得多。因为很多人都去抢热门词了，你只要注重优化长尾，你每做一个动作，流量就应声而涨。《百度如何优化》提到的那个从 6 万到 267 万的例子，也是应用长尾策略成功的。优化那个网站，那时我只有一个人+两个月的时间，要是偏重优化热门词，要到猴年马月才能涨这么多流量了。

客户的匹配程度、流量的质量、转化率，长尾关键词也不知道要比热门关键词好多少。从网站的利益出发，网站真正需要的不是你那些热门词语的排名，而是最后的效果。比如一个销售惠普笔记本电脑的网页，即使你在“笔记本电脑”这个关键词上排第一了，最后的效果就一定比关键词“惠普笔记本电脑”排第一的效果好吗？

这样做 SEO，还能让 SEO 部门和网站运营部门完美的融合。一定要相信网站运营部门出于平常的业务需要来突出的内容，也是搜索引擎的用户需要的。

这种优化长尾关键词的策略，会极大的帮助搜索引擎提高检索信息的全面性。如果你对搜索引擎说：我要选一批热门关键词来做这些词语的排名，搜索引擎会说：你这个动作其实多少都会涉嫌作弊了，是在拿一些不属于你的流量。如果你说：我要用这种优化长尾关键词的策略来做 SEO，搜索引擎会说：求你了，快点这样做吧。还是上面谈到的那个原因，就是搜索引擎的现在还没有到搜索结果过剩的时代。大部分人搜索信息的时候，都是要好几次搜索，才能找到自己需要的信息。网站长尾关键词的优化，能很大的改善这种情况。以前如果每 10 次搜索才能找到需要的信息的话，那么经过大家的长尾优化后，就可能只需要 5 次搜索就可以了。节约了用户的时间，提高了搜索引擎的效率，网站也受益。

严格说来，不是我们要选什么关键词去搜索引擎那里拿流量，而是搜索引擎通过我们的内容给我们相关的流量。

一个新站的 SEO 流量很多是长尾关键词贡献的现象，不是个别或偶然现象，而是必然的。一个正常的大中型网站，长尾流量一定占到所有 SEO 流量的 90%以上。而小型网站，也是长尾流量大的居多。因为网站中的词语如此之多，随意几个词语的组合都有可能给网站带来流量。所以这种情况是一个自然而然的规律，那么就不要去打破这种规律和平衡。

应该是把握这个大局，尽量的向搜索引擎展示你的内容。试图让搜索引擎收录网站所有的页面，然后通过你的结构突出重要的内容部分，再让搜索引擎来选择。至于用户搜索什么词语到达你的网站，其实不用特意去操纵，因为你永远都无法预测用户喜欢什么，会用什么词来搜索。

在《怎么样学SEO(二)》的结尾，我说：优化大型网站的时候，把资源偏重优化热门关键词，会让你丢失掉 50%以上你本来应该得到的流量。其实有时候小网站也何尝不是呢？

让我们抛弃那种传统的优化方法，侧重长尾，让流量“自由地生长”吧。

内部链接还是外部链接？

这篇文章承接上篇《热门还是长尾？大中型网站的关键词优化策略》。明白了长尾效应在一个大中型网站中的作用后，还需要明白内外部链接谁更重要。

在搜索引擎上，去获取流量的最基本单位就是网页。一个网页的外部链接因素，对这个网页的排名影响很大。这个网页的外部链接，既有同一个网站的其他页面给的站内链接，也有其他网站上的网页给的站外链接。下面文章里的内部链接是指站内链接，外部链接是指站外链接。那么在优化一个网站的时候，是特别注意优化内部链接还是外部链接呢？或者在分析一个网页排名的时候，是觉得内部链接贡献的价值大，还是外部链接的价值大？

长久以来，大家都非常重视外部链接。源于那么一个说法，那就是：一个网站你自己说你的网站里有什么是不算数的，要别的网站说你这个网页里有什么才算数；相对于你自己如何评价自己，别人的评价才更准确。 所以很多人在做一个页面的排名的时候，只做一件事情，就是疯狂的给这个网页做外部链接。而内部链接呢，很多人认为不重要或对排名影响不大。

这种说法应该是有人从PR值的计算方法发展而来的。因为在PR值的计算理论里，影响一个网页 PR 值的是这个网页的外部因素。当扩大到整个网站的时候，有人就认为影响这个网站整体排名的因素来自于其他网站。其实这个说法有一个明显的误区，就是没有明白网页和网站的区别。别说 PR 值和排名没有直接的关系，就算在 PR 值的计算理论里，向来也只有网页才是被计算的对象，而不是整个网站。列在搜索结果页面的，也是一个个的网页。

那如果一个网页同时有 10 个外部链接和 10 个内部链接，谁对排名的影响大呢。我们再来看那个“外部链接的评价更准确”的理论。其实这个理论要成立是要有一个前提的，那就是互联网上所有的网页都是不值得信任的，要靠这些网页彼此之间的关系才可以确定谁更重要、谁的内容更和什么关键词相关。这在一个搜索引擎建立的初期，是非常科学的方法。但是，搜索引擎发展到现在，积累了大量的数据，环境也不一样，那很多问题都要重新审视了。Matt Cutts 曾经说过 google 不会停止对 PR 值的改进，其实更多的其他改进也是如此。而百度的超链分析法，在面对如此泛滥的群建链接的情况下，也会做出相应的修改的。

实际上，最清楚那个网站里面讲什么内容、哪些内容是重点的，是那个网站自己本身。别人都没有那个网站那样清楚它自己。但是搜索引擎并不能确保那个网站会如实的标注自己的内容。所以才会借助别的网站对那个网站的评价来区分。但是如果那个网站是一个值得信任的网站呢？搜索引擎是不是可以相信那个网站自己对自己的描述？答案是肯定的。

如果一个经历过时间的考验、无论从各种渠道都表明那是一个值得信任的网站。它说自己的网站有什么内容的时候，那应该是很准确的。而外部网站对它的描述可能反而是不全面和不准确的。自己对自己的描述，就是内部链接，别人对自己的描述，就是外部链接。 这个时候是内部链接更重要还是外部链接更重要呢？

这个时候，很难说谁更重要，起码这两者都是一样重要的。因为都是值得信任的网页对另一个网页的描述。

所以在优化一个网站，特别是有点历史了的大中型网站的时候，不用那么特意区分外部链接还是内部链接。可以借鉴计算 PR 值的视角，把你要优化的网站看成是无数个网页的组合，每个网页以外的链接都是要注意优化的链接。有了这种视角，在分析很多网页的排名的时候也不会困惑了。很多人还一直不明白为什么有的网页只有 7 个内部链接就排得很好，而自己有几十个外部链接还是没什么好的排名。国内也有不少人在优化一些大中型网站，应该可以观察到一个现象，就是，只要你尽可能用白帽的手段优化网站，当积累到一定的程度，网站的 SEO 流量会有一个突发性的整体提升，这主要就是内部链接的功劳。

只要你优化的网站不是那种只有几十个网页的网站，不然都要先关注好内部链接再去关注外部链接。

不过有些人心里应该还是有疑问。 Matt Cutts 曾经用花钱和赚钱的说法比喻过内外部因素的重要性。他说：“考虑如何花好 100 元是一个好的网站结构的问题，但是对于大多数人来说，如何另外再赚 300 元更能使他们受益”。 这句话好像是在说外部链接要比内部链接更重要。实际上这句话没有错，从搜索引擎的角度和互联网全局的角度来看这是对的。但是在具体的操作过程中就不一样了。

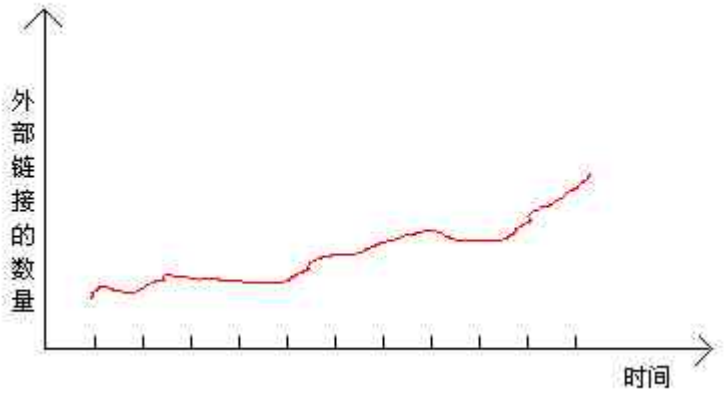
对于大中型网站，不知道有没有人去统计过自己网站外部链接的数量以及每天别的网站会给自己的网站增加了多少外部链接。（google webmaster tool 里可以看到一个网站的大部分外部链接）只要稍微有点名气的网站，每天别人给你贡献的外部链接的数量要远远超过你每天自己给自己加的外部链接的数量。而且从质量上来说，那些链接也要好得多。那你何必每天辛辛苦苦去加那么一小部分链接呢？而对于小网站，只要你的网站不是几十个页面，你要先分配好内部链接，才能更好的利用好外部链接。花钱和赚钱的那个比喻，这个时候真实的情况变成了：你每天都有几百万元进账，但是花钱的时候这些钱有 50%都浪费掉了，其他该花钱的地方要么只花了一少部分，要么完全没有钱。这个时候你说每天再去赚几万、几十万更重要还是先管理好钱怎么花更重要呢？只要你侧重做一个网站的 SEO，积累一段时间后，你都是那个需要去研究怎么花好 100 元的人，而不是大多数还要去赚 300 元的人。

（这里我很想顺便说一下如何对待搜索引擎的工作人员说的话。我相信他们在公开场合是不会故意误导你或者说假话的。你要看那个说话的人在这个公司是处于什么样的职位，是在什么时间、什么场合、出于什么样的目的说那样的话。很多时候都不会是他们的说法错误，而是你没有掌握到他们那么多信息，所以你无法理解他们的话。也不会推测出他们的潜台词以及他们没有说全的话而已。）

接下来想谈一下外部链接，因为有很多人来咨询，而且我也看到一些人还没意识到一些误区。由于篇幅过长只讲两点。

1，怎么样去做好外部链接。

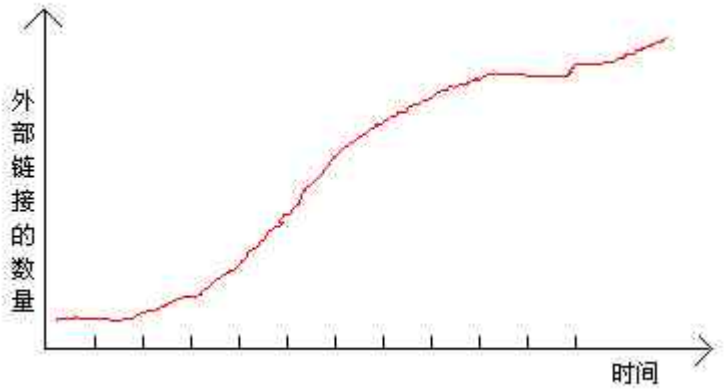
先放下这个问题。我们先来看看搜索引擎如何判断一个网站的外部链接是好链接。在互联网上，如果一个网页很受欢迎，就会有很多其他的网页链接或引用。那它的外部链接增长的速度，用一个曲线图来表示是这样的：



外部链接变化

虽然会有一些外部链接因为那个网页消失了或者删除了给你的外部链接，但是总体趋势还是上涨的。

如果碰到一个突然很热门的网页被大量转载，那它的曲线图应该是这样的。



外部链接变化

关于这种突然热门的网页，可以去看看最近“成都暴力拆迁引发自焚”的视频页面：

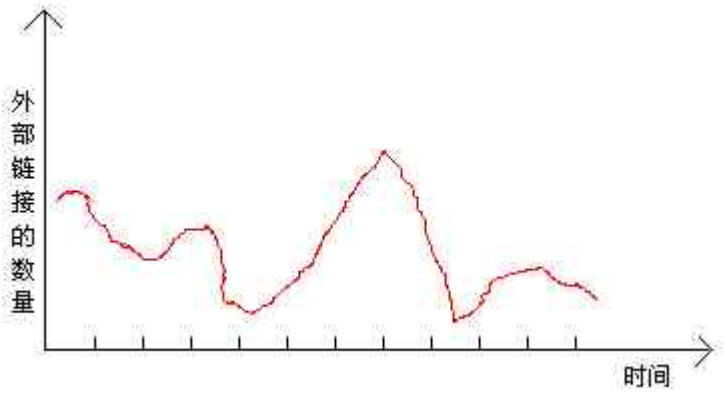
http://v.youku.com/v_show/id_XMTM1OTQ3MTcy.html

点击页面上的“全部视频信息”，往下拉，可以看到这个视频最近被转载的记录：

该视频的外站引用记录

这里的外部网站链接的增长幅度和那个曲线图的增长曲线是吻合的。

那如果一个网页的外部链接是群发或者群建的呢？曲线图就是这样的：



外部链接变化情况

因为群发和群建的链接，很多会被大量删除的。而且这个群建的网站，也没办法保证经常性的去群建和群发。这正应了那句老话：路遥知马力，日久见人心。搜索引擎对好的网站不会埋没你的好，差的网站最终也逃不过时间的检验。

这三种不同的网页的外部链接曲线图，用一些简单的数学公式就可以判断并描述出来，准确率应该至少 90%以上。

2，在哪些网站上做外部链接好？

太多的人还是执迷于博客群建一类的方法。博客群建和论坛群发，在百度和google 短时间都会有一些效果。不过从上面的曲线图可以得知，一旦你开始这样做，离被搜索引擎惩罚也不远了的。

搜索引擎依据外部链接的速率来判断一个网站是否作弊，在很多情况下也有失误的时候。所以尺度很宽松，只有一些实在是很明显的网站才会被处罚（百度不一定很宽松）。不过搜索引擎还有其他的检验方法，就是看你的网站经常出现在什么样的网页上。这种判断是非常容易实现的。

和上面的曲线图一样，一个正常的网站，外部链接出现在什么地方是有自己的规律的。运用数学的方法，也能把这些好的网站和不好的网站区分开来。

我以前和别人合作做过一个项目，就是要用爬虫把互联网上所有的企业网站都搜集下来。乍看一下这个是不可能完成的事情，因为企业网站的设计千差万别，代码有各种写法，怎么可能判断哪个网站是企业网站哪个不是呢？后来用排除法就轻松解决了主要问题。因为互联网上的动态网站，真正从头到尾自主开发的非常少（那些网站也不太会作弊）。大多是用一些开源的 CMS，如 wordpress，discuz 等等。有的网站依附于一些知名的网站系统，如新浪博客，51 空间等。这些系统都有自己的特征，只要根据这些特征设定好过滤规则，这些网站是很容易被排除掉的。后来检测结果，发现准确率有 80%以上。

上面的 2 个问题，我都没有正面回复。那如何做外部链接最好呢？

负责任的回答是：不去为了 SEO 而刻意地做外部链接最好。 你只要专注于你的内容，考虑别人如何才能主动链接你就好。即使主动去推送信息，也要在别人恰好需要你的地方出现。大家可以回忆一下，没有 SEO 之前，网站都是如何做外部链接的就明白了。上面那个视频，短短的 3 天时间就有 4000 多个外部引用，以后还会被大量引用。如果要这个网站依靠人力去做，大家觉得要用多少人、多少时间和多少钱才能达到同样的效果呢？

怎样形成一套非常科学系统的**SEO**方法

尽管 SEO 在中国已经不陌生，甚至都有形成一个行业的趋势，但是至今业内都还没有一套非常科学系统的分析方法。原因恐怕要归结于搜索引擎优化这个行业的特殊性。搜索引擎严格保守他们的算法，只公布一些大家很难去知道原因的指南。所以很多 SEOer 都在玩一个永远也不知道具体规则的游戏，这是这个行业混乱的根源。

我多次强调《google网站质量指南》的重要性，还因为这是搜索引擎告诉网站主的仅有的一些正确的规则，如果连这点规则都不好好掌握，那我还不确定大家能从什么地方得到更权威的指导。但是在实战中，尽管你熟读这个《指南》已经比很多人更了解搜索引擎的规则，不过仅仅知道这点东西是不够的，一套科学系统的分析方法能让你走得更远。

我想 SEO 经过了这么多年的发展，已经不应该再出现那种靠感性分析去做 SEO 的分析方法了。这种分析方法常用的语句就是：我觉得搜索引擎会如何如何。如：我觉得搜索引擎不会那么笨，这点一定能处理好；我觉得搜索引擎会把这个因素当作排名的因素之一……。如果你是依靠感性分析去做 SEO 的，那你的 SEO 流量的变化曲线也是很感性的。当然更不能去无根据的臆测和道听途说。如：没有理论基础的去猜想搜索引擎会怎么样或者每逢搜索引擎的相关人员以及什么权威人士发表什么演说，就去盲目听从。

既然搜索引擎不告诉我们具体算法，那我们怎么才能建立这套科学系统的分析方法？答案是：从你知道的确信一定正确的理论开始，慢慢在实践中进化。

在上一篇《网页加载速度是如何影响SEO效果的》中的那个分析过程，就是从一个确切知道的理论去分析，然后得到了另一个确切的影响SEO流量的因素。在这个过程中，确信没有错的理论是：搜索引擎爬虫一定要抓取过那个页面以后，才会有机会收录这个网页的。根据文章中那个接下来的数据分析，可以得到：网页加载速度会在很大程度上影响SEO流量。

那接着分析，什么措施能影响网页加载速度呢？ 网络环境、服务器硬件、CMS 本身都能影响网页加载速度。优化其中的任何一项，都能提升网页加载速度。那马上又可以得出：网络环境影响 SEO 流量、服务器硬件影响 SEO 流量、CMS 本身的速度影响 SEO 流量。

接着分析，CMS本身的优化可以做的事情有哪些呢？ 启用Gzip压缩、合并CSS和JS文件、减少DNS查询、启用缓存等等都能优化CMS本身的速度。……这些东西，看起来是这么的眼熟，那是因为在《google网站管理员工具》里的“网站性能”里，已经把这些建议都告诉你了。但是根据我们上面的这个分析过程，可以知道，“网站性能”里提到的这些优化，都是CMS本身的优化，并没有提到网络环境和服务器硬件的优化。只不过你确定这两个因素是确实影响SEO流量的。如果哪一天《google 黑板报》或者 google的官方博客（需要翻墙）上出现一篇文章，告诉你如何挑选一个好的服务器托管商，千万不要惊讶，因为你早就知道为什么了。

google一直以来都在用这种方式告诉你要如何去优化一些什么因素，只是站在他们的立场，不会详细向你解释为什么要这么做。

通过数据分析，还能知道谁影响的程度大一点，谁小一点。

很多的常识因素都可以这样一步步进化下去，这个分析过程，是非常科学的。不管是对你自己还是其他人，其中的原理你都可以解释得非常清楚。并且在这个进化的过程中，你会发现你越来越能控制好 SEO 流量了。每一步的进化，意味着你对搜索引擎的了解又多了一点、SEO 的知识结构又完善了一点，同时，对 SEO 流量的控制能力又变强了一点。同时，你发现你和网页设计师以及工程师的矛盾也越来越少，因为好的 SEO，是不会让 SEO 和网页设计师以及工程师的利益是矛盾的。

SEO知识结构

SEO流量的控制能力

与其他部门的利益

知识结构、SEO 可控性、部门关系

只要经历过非常多这样的分析过程，一定会颠覆很多人原有的SEO知识结构。因为以前很多流传的SEO方法，很多都是感性分析的居多，没有解释为什么要这么做，没有数据上的支撑，甚至没有理论上的支撑，所以没有抓住重点。我在《分词与索引库》说过，可能你以为是细节的东西，其实是重点，你以为是重点的东西，其实都可以忽略。

那么，在日常的 SEO 工作中，是一些什么能力支撑着你去进行这样一个分析过程呢？

不知道大家还是不是记得我在《怎么样学SEO》提到的那四种能力，在这个分析过程中：

1，弄懂搜索引擎相关技术和原理：可以从根本上了解搜索引擎，确定很多一定正确的理论，并可以找到很多值得去分析的线索。

2，了解网站制作相关的技术：能让你清楚网站上有哪些因素能影响搜索引擎的哪些方面，并用什么方法来解决问题。

3，数据分析能力：可以了解各种现有的因素如何影响 SEO 流量，并依靠这种能力挖掘更多的因素。科学系统的 SEO 分析过程，从头到尾都离不开数据的支撑。

4，了解你要排名的那个搜索引擎：不管你怎么努力，还是会有一些数据上和理论上都无法理解的问题。每个搜索引擎就像和人一样，是有一定的秉性的。可以通过你对这个搜索引擎的了解来得到答案。同时了解这个搜索引擎，也能让你获得更多的可以分析的因素。

最后说一下，这种从常识出发来科学系统的进行 SEO 分析的方法比了解部分搜索引擎的算法还更能控制 SEO 流量。

可能很多人会反驳这个观点，比如前段时间我朋友就和我说某外贸B2C网站的创始人是从谷歌出来的，那他们一定能做好SEO，我说那是不可能的。只有那些自己做过搜索引擎的人才会理解为什么。比如：alibaba的B2B网站也算是一个搜索引擎，我是知道其中的排序规则的，但是如果给我一个商家的网站，要我在alibaba上获得流量，在没有一套科学系统的方法之前，我是肯定做不好的。因为搜索引擎的算法不是加减乘除，不是这个因素加那个因素做好了就可以获得好流量的。搜索引擎的设计者，知道这个或者那个因素的权重大小，以及可能产生的大致结果，但是具体的结果是自己也不能控制的。要不然百度的人，不会每天搜索上千个词语来查看搜索结果的准确度了。而google的成功，也有一部分原因是当初yahoo采用了它的搜索技术，google借此积累了大量数据，实践并改进了算法。

而且，在搜索引擎内部，只有极少数的人知道各个因素的权重大小，绝大部分设计搜索引擎的工程师，都是负责某个具体的任务，优化和解决某个具体的问题，如负责爬虫的工程师解决提高爬虫效率这一块的工作，负责内容消重的工程师就去减少索引重复内容。 连设计这个搜索引擎的工程师都如此，更别提一个远在其他国家的分公司的人员了。要不然，百度和 google 这么多离职的工程师不早就把算法泄漏了。

如果能自己用开源的程序做一个小规模的搜索引擎，就更能理解这个问题。即使这个搜索引擎的算法都是你自己调配的，你都不能预料到后来的搜索结果。而且做搜索引擎是一回事，在搜索引擎上拉流量又是另一回事了。不然 google 不会后知后觉的知道原来网页加载速度影响 SEO 流量。

整体还是局部—如何制定好的SEO策略（1）

已经有好几个月没写点东西了，感觉还有很多东西可以写，而且现在经常有一些新的发现和感想。不过一直在忙着给一些大中型网站提供 SEO 顾问服务，时间都是优先花在给他们解决问题上。

已经给很多网站做过 SEO 顾问服务，其中有 SEO 流量才几千 UV 的中型网站，也有上百万 UV 的大型网站。发现有一个问题是非常突出的，就是很多网站都没有一个清晰的 SEO 策略，只是埋头做事，这导致了一些问题。

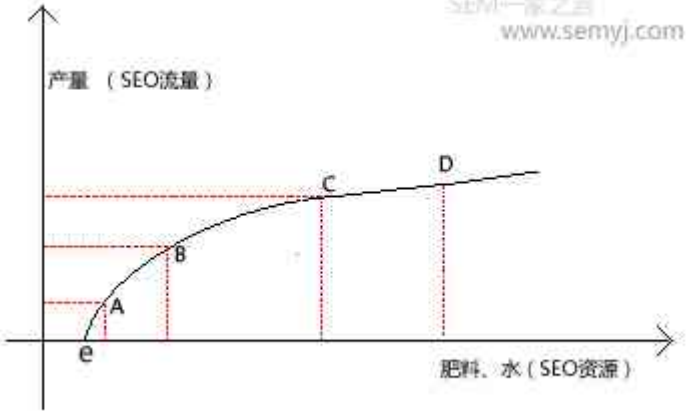
最明显的问题就是把手段当目标，把过程当结果。

现有的很多网站，他们常用的 SEO 做法是：搜集几十、上百个本行业里的热门词，把这些热门词加到一些网站中权重很高的网页上。然后不断的给这些页面增加外部链接，希望这些关键词能有不错的排名。这种典型的做法就是我说的把手段当目标，把过程当结果的做法。

让我们来看看一个网站为什么要做 SEO 吧。很多网站之所以要做 SEO，是因为 SEO 是目前性价比最高、效果最好的网络营销手段。他们的目的是希望 SEO 给网站带来直接或间接的收益。要达到这个目的，就需要有大量相关的 SEO 流量。“带来大量相关的 SEO 流量”才是给一个网站做 SEO 的首要目标。而上述网站做 SEO 的目标是：给某一批关键词做排名。 做关键词排名只是 SEO 过程中的一个手段，但是很多网站把它当作了目的。更别说还只是给少数关键词做排名了。

以前在《热门还是长尾？大中型网站的关键词优化策略》这篇文章评论中，有人问道：难道SEO流量不是由关键词排名贡献的吗？

要说明这个问题，我最喜欢用的比喻是一个经营果园的例子。假设你有一个网站，有一万个有内容的页面，目标是带来大量相关的 SEO 流量。就好比你经营一个果园，果园里有一万棵果树，而你的目标是提高果园的产量。如果是给某一批关键词页面做排名，就好比是你把提高整个果园产量的目标，放在了希望少数几十棵大树的产量提升 10 倍、20 倍的基础上。 这种做法是不容易达成目标的。因为一棵果树的产量，不是你投入资源加多少倍，它就能涨多少倍产量的。不管是一棵果树的产量，还是单独一个页面的 SEO 流量，如果去追踪他们的增长方式，会发现它们都是遵循“边际效益递减”的道理。 如：



边际效应递减

在 C 点之前，不停的投入资源是值得的，但是 C 点以后，投入资源的回报率显然不好了。 所以，少数几十棵大树的产量提升 10 倍、20 倍是很难的，就算达到了，代价就是每一棵树消耗的资源远远超过 D 点对应的资源。而且这几十棵大树以后增加产量需要的资源呈几何倍数增加。结果整个果园大部分资源都被这几十棵树消耗掉了。消耗了这么多资源，他们能为整个果园提升多少收益呢？ 可能是 50%都不到，而且果园再大一点，那能不能有 20%的增收都是问题；再大一点，10%也困难了……还有就是今年可能是有这么多产量，那明年、后年再用这种方式就越来越难增收了。

那我们换一种方法，关注点不要放在那几十棵大树上。而是我把果园看做一个整体，不管大树小树，都是我果园里能增加产量的来源。 我不要去给少数几十棵大树提升 10 倍、20 倍的产量，我只要把平均每棵树的产量提升 1 到 2 倍就可以了。这样整个果园的收益就提升了 100%到 200%。用这样的方式去做以后，那我就会去平均的分配我的资源。有时候还会“劫富济贫”，对于那种不缺资源的果树，我就克扣和节省那些资源，分给那种很缺资源、但是给一点资源就能产生很大收益的果树上。

如果体现在上图中，就是我确保每棵果树的资源，都是在 e 和 C 之间。但是因为资源有限，果树太多，是不可能给每棵果树的资源都能达到 B 和 C 之间的。 所以我就控制好资源的分配要在 e 和 A 之间。资源的投放控制在 e 和 A 之间还有一个原因，就是果树的数量随着时间的推移开始增加了，有些树还没突破 e 点，也就是还没有产量。那我就定一个标准：在还有很多树没有产量之前，每棵树投入的资源都不要超过 A 点的资源。 e-A、A-B、B-C 这个三个区域，把资源投在 e-A 之间的投入产出比也是最好的。而且等果园整体的产量超过 A 点后，后面还有很大的成长空间。

这是一个果园的例子，但是大家可以对上面的文字重新看一遍。只要把“产量”看成“SEO 流量”，“果园”看成“网站”，“果树”看成“网页”，“e 点”看成“有收录并开始有流量”就可以了。

然后再回答那个问题：难道 SEO 流量不是由关键词排名贡献的吗？ 首先 SEO 流量不光是由关键词排名贡献的，还是因为很多的网页被收录，才会有关键词的排名的。而且就算有了排名，也是需要有人点击才会有流量的。即使是关键词排名也绝不是少数页面、少数关键词的排名，是整个网站所有页面，所有关键词的排名。

前不久，有传言说 google 会停止 PR 值的更新。虽然我没看到 google 官方的声明，但是我觉得 google 停止 PR 值的更新没什么不可以，要真这么做就太好了。现在很多网站一做 SEO 就会提到 PR 值，经常关注自己的首页 PR 是多少，又把提高首页的 PR 值看成他们的目标了。我以前说过 PR 和排名关系不大，不过假使 PR 值跟排名的关系很大，为什么很多人又只喜欢看少数几个页面的 PR 值呢？

其实 google 一直都在强调整体考虑的重要性。如果用过老版本的 webmaster tools 的人，应该还记得这么一个数据。

pr 的分布

这个数据就表明了整个网站所有的页面的 PR 值分布状况。因为 PR 值虽然和排名关系不大，但也是一个对 SEO 流量有利的因素之一。webmaster tools 里面公布的这个数据是很想让大家注意网站整体的 PR 值分布。上图的这个数据看起来还算不错，至少大部分页面还有 PR 值。一个网站如果能达到这么一个状态，那 PR 值的分配就还很不错，对 SEO 流量的帮助也不小。我观察很多网站，都看到了 PR 值的正确分布对流量的正面影响。但是如果没有注意整体 PR 值提升的网站，“PR 值尚未分配”的部分就越来越大，就造成了只有少数页面有高 PR，那这几个高 PR 值的观赏作用就大于实际作用了。

Google 的本意如此，但是把这个数据长久放在这里，造成不好的影响就是大家更加注意 PR 值了。 所以后来 google 撤掉这个数据的理由之一就是不想让大家太注意 PR 值。不过 google 从来都没放弃告诉大家注意整体资源的利用。

这个 PR 值分布的数据，在 webmaster tools 中后来演变成了两个数据：一个是内部链接数据，一个是站外链接数据。从 PR 值的计算公式里可以看到，PR 值就是被内部链接和站外链接影响的，所以拆分后的数据更加有参考价值了。

不过这个时候，很多人还是没有意识到整体考虑的重要性。具体我们就来看一个例子。

这是某中型网站外部链接的数据：（大家也可以看看自己的数据）

站外链接分布

这是一个有近百万有效内容页面的网站，但是整个网站中，只有 333 个网页有站外的链接。 而且和绝大多数网站一样，网站首页的站外链接占到总链接数量的 95%以上。站外链接是一种比较有价值的 SEO 资源，大家可以想象那个经营果园的例子。这就是只注意局部忽视整体的结果。

给一个网站做 SEO，最重要的是 SEO 策略上的制定。只有策略才是统揽全局和整体的，而各种 SEO 的优化方法，只是局部和片面的。制定一个好的 SEO 策略，其实只要注意一个问题，就是：要整体的效果还是要局部的效果。

这篇文章太长，过几天继续讲这个话题。

以后会经常更新博客。成立公司后的这几个月，看过了很多网站的情况，感觉中国的 SEO 非常的缺人、缺方法，所以我希望能发挥自己的价值。

最近又新做了个论坛，就是希望能让大家在一起探讨各种有价值的问题。以前的基地论坛纯粹是为了做论坛而做论坛。现在做论坛的想法很简单，就是做一个我自己都经常想泡在上面的论坛，聚集一些同道中人，大家一起研究一点事情。论坛不光只会专注 SEM 这一块，以后会根据需要开一些其他实用的板块。总体是专注电子商务和 IT 领域。 希望也能成为大家喜欢去的论坛。

如何用好**nofollow**

我记得很多人知道我的博客是因为一篇关于nofollow的文章，恰好是在 2 年前写的。真没想到两年就这样转眼而过。现在我就来讲一下那篇文章中提到的那个nofollow做得好网站是哪个，以及他们如何做nofollow的吧。这个例子是我各种培训中都会讲到的，都已经讲得快起茧所以不想以后再说了。同时也为了让大家看看数据分析是如何指导SEO的。

这个网站就是曾经在外贸B2B领域做得很成功的网站 - Tradekey。这里先介绍一下这个网站的历史：这是一个完全依靠SEO起家的网站，总部在迪拜，现在是一家跨国公司，创始人非常年轻。我以前写过一篇《依靠SEO去打造一个成功的网站》，那这个网站就是经典案例。在外贸B2B领域，曾经有的网站一年都需要十来万费用，还拿不到多少询盘。但是曾经在Tradekey上，免费会员都能拿到很多不错的询盘。所以它依靠口碑在B2B领域慢慢流行起来，被誉为“B2B领域的一匹黑马”。在国内一些都比较懂网络的外贸人群中也很流行，以至于有段时间国内都出现了很多tradekey的伪代理。Tradekey是如此的强势，所以它后来基本关掉了免费会员，也就是所有在这个网站上发布产品的供应商，都要成为付费会员才可以。 在B2B领域，基本都是靠免费会员来拉人的，这样做真的需要底气。

Tradekey 的底气，就来源于它不错的 SEO 技术。很多做英文 SEO 的人都应该知道，在 google 上搜索很多的产品关键词，它都能有不错的排名，它的 SEO 流量非常的可观。加上和那些热门平台相比，它的供应商数量不是很多，所以大家的效果相对都能得到保证一些。

这个网站也是我唯一见过的收录量曾经达到 100%的大中型网站。我要讲的 nofollow 的应用，就从这个网站如何提升收录开始。

现在很多人都会把类似“注册”或“登录”这样的链接 nofollow 掉，这是因为 google 官方就建议这样做，当然 tradekey 也做了。

红色框内的是 nofollow 的链接

不过一个网站中可以加 nofollow 的地方还有更多。我们来看它的导航条，在主页上，只有“Member Area” 这个链接被加了 nofollow。 （如上图）

但是在这个网站的其他网页上，导航条上所有的链接都被 nofollow 了。

导航条更多的 nofollow

很少有人能敢把导航条上的链接 nofollow 掉，它这么做的原因，通过数据来分析一下也就明白了。

如果你经常使用《光年日志分析系统》这样的软件来分析日志，就会发现一个网站中有很多的链接在一天之内是能被访问很多次的。如：以下就是这个软件统计出来的某个网页一天内被搜索引擎爬虫访问的数据：

页面	总抓取量	蜘蛛	蜘蛛抓取量
/ Help.html	419	BaiDu Spider	166
		雅虎蜘蛛	97
		谷歌蜘蛛	63
		Sogou Spider	38
		有道蜘蛛	37
		msnbot/	7
		soso	6
		Alexa crawler	3
		Speedy Spider	2

一天内不同蜘蛛的访问次数

理论上来说，如果一个网页上的内容更新得不是很频繁，那这个网页一天被抓取一次就可以了。对于那种已经被收录而且内容一直不变的页面，一天被抓取一次都太多了。就算是更新很频繁的网页，一天被访问 50 次也完全够了。不过实际情况远比理论上的糟糕，就像上面的这个抓取数据，一个无关紧要的页面，百度爬虫一天都能抓取 166 次。大中型网站更糟，有一次我们分析完一个大型网站的数据，发现这个网站爬虫每天的抓取量虽然有 120 多万，但是其中有 16 万次抓取都是在抓首页这么一个网页，可以想象其他网页又有多严重。

为什么我们要这么在意一个网页被重复抓取的几率呢？ 这是因为一个网站中还有很多其他的网页，爬虫压根就抓取不到。哪怕你的网站只有几百个网页，都可能面临着这个问题。一个网站如果每个页面平均被重复抓取 10 次，尽管可能爬虫每天的抓取量有 100 万，那也只有 10 万个页面被抓取了。一天之内的情况是如此，时间拉长到一个月内、半年内，情况不会有多大改善。虽然搜索引擎也试图解决重复抓取的状况，但是由于各种原因，会导致今天重复抓取的页面，明天还是会重复抓取的。所以很多的大中型网站，一年下来，还有一半的网页，爬虫压根都没看到过。 如果不是分析了很多网站的数据，很多人都是无法想象情况有这么严重的。

在抓取量一定的情况下，适当减低一些页面的重复抓取量，那会有更多的其他页面会被抓取到。 一个网站中，最容易被过度抓取的页面，就是那些经常曝光的页面，导航条上的链接就是经常曝光的。所以 Tradekey 的解决办法很简单，就是在首页这么一个页面上，给爬虫留下入口去抓取导航条上的链接，但是在其他网页上，就把导航条上的链接 nofollow 掉。这样处理，会使导航条上链接的抓取量，从以前被抓取上万次降低到现在被抓取几十次。虽然不能达到理想中的状况，但是也比以前好了非常多。

Tradekey 就用这种思想处理了网站上的很多链接。如：

大量应用 nofollow

想象一下，当爬虫以这么一个页面作为访问的入口时，由于很多通用的链接都被屏蔽掉，这样就“逼着”爬虫去访问那些它以前从来没有看到过的页面。 整个网站能被爬虫访问到的页面就大大增加了。

在 google 咖啡因改版的很久以前、Tradekey 还只有英文版、产品信息只有 200 多万条的时候，它整个网站的真实收录量是两千多万。 所以基本认为这个网站做到了 100%收录。 （真实收录是指按一个网站的 URL 特征查询各自的收录量，再把所有 URL 特征的收录量加起来的数据。这个数据在 google 咖啡因改版以前是比较准确的。）

不过如果 Tradekey 只是这样来用 nofollow，那还是有点平淡无奇的。更能体现 Tradekey 用活了 nofollow 的是它其他的改动。

打开 Tradekey 的首页，可以看到 Tradekey 把网站最新发布的产品和推荐的产品信息给 nofollow 了。

产品信息都被 nofollow

相信那些经验丰富的 SEO 人会觉得这是不可思议的，因为它这样做可能犯了两个错误：一是把最新发布的产品 nofollow 掉，那这些最新发布的产品的收录会受到影响。二是影响了“首页效应”，会让一些关键词的排名消失。 所谓的“首页效应”我要解释一下，因为很多人第一次听说这个名词。但是对于做大中型网站的 SEO 人员来说应该会观察到这个现象。就是在很多的大中型网站上，要做一些关键词的排名其实是比较容易的，只要把这个关键词链接在首页上放一段时间，这个关键词的排名就上升了。这是因为大中型网站首页的权重（权重不是 PR）实在很大，首页上的链接分享了这个权重。如果这个放在首页的关键词是个长尾关键词，那基本会排前几位。大家去查一些大中型网站的首页链接，也都可以观察到这个现象。

对于 Tradekey 来说，它在首页 nofollow 掉的这些链接，由于都是一些长尾关键词，如果不加 nofollow，很多关键词都会有排名和流量的。如现在网页上的“Wheel Hub Centric Spacers” 这样的词语。 但是它为什么又不要这种词语的排名和流量呢？

这是因为它从整体角度考虑，要把网站的收益最大化。“首页效用”是有前提的，就是首页的链接越多，每个链接的“首页效应”的效果越弱。这和 PR 值的原理一样，只是这个效果不是由于 PR 值的被稀释造成的。在首页上，把一部分链接 nofollow 掉，另外一些链接的效果就会增强。此消彼长，总体的流量不一定会降低。这时候就是一个取舍问题，那一个 B2B 网站要增强哪些页面的效果而减弱哪些页面的效果呢？ 一个英文 B2B 网站中，用户在列表页面的转化率是产品页

面的 4 倍以上，（中文网站也差不多，转化率高的原因是由于用户在列表页有更多的选择。）所以在平常的优化中，列表页面是要重点照顾到的页面。Tradekey 要 nofollow 这些产品信息页，就是想增强其他列表页面的效果。 至于那个 nofollow 影响了新增加的产品的收录，要怎么解决呢？那就在其他页面上加一个“Latest Products”页面，专门可以解决这个问题。

如果我们来做一个数据分析，也能证明这样做是明智的。 这个首页上有 263 个链接，假设在加这些 nofollow 之前，这 263 个链接的流量总和是 1 万 IP，带来了 100 个询盘；那有可能加了 nofollow 以后，这 263 个链接（很多链接只是从首页 nofollow 了，只是不能沾“首页效应”的光，但是其他地方并没有 nofollow，所以依然会有流量。）的流量总和还是 1 万左右的 IP，带来了 120 个询盘； 从整体收益出发，不知道大家更喜欢哪个结果。

为了增强整体的效果，牺牲一些局部利益是完全可以的， 我在《整体还是局部—如何制定好的SEO策略（1）》一文中讲了这样一个道理。整体还是局部，是要经常注意的一个问题，很多老的SEO方法就在这方面出了很多问题。

上面是以 Tradekey 做为例子讲了两个 nofollow 的应用方法，让我们再回到主题，那要如何用好 nofollow 呢？ 其实重要的不是如何用好 nofollow 的问题，nofollow 永远只是一个手段，重要的是怎么知道用这些手段来达到你的目的。太多的人把手段当目的，把过程当结果了。还是以上面的例子来说，可能有些人马上会去模仿 Tradekey 的做法，我的建议先等一下。Tradekey 做得好的地方不是它的 nofollow 用得如何好，而是它背后那种依靠数据分析指导 SEO 的过程做得很好。应该先分析一遍自己网站的数据再来做决策，每个网站不一样，别人的方法不一定适合你的网站。 永远以数据分析来指导 SEO 的进行，就不会停留在那种永远只做表面优化的阶段，而让你知其然也知其所以然。只要你知道了为什么要这么做，那怎么做的方法可以一天想一个出来。

其实呢，Tradekey 的这两个改动，起码是好几年前就有了。对于那些在第一线的 SEO 人员来说，这不是什么新鲜的做法。我要介绍它是希望大家可以不要那么重视主流的 SEO 观点，如果有数据做支撑，那就要相信你自己，按自己的想法来做事情。不管是国内还是国外，有些名人博客只是为了说而说，水平说不定要低于那些在第一线的人员，很多一线人员是没那么多时间或者不愿意出来说，不然主流 SEO 的整体水平会更上一层楼。

Tradekey 也有很多做得不好的地方，这就是我为什么一开始介绍 Tradekey 的时候用了“曾经”这样的文字。这是因为它实在是太依赖 SEO，（Tradekey 的运营中心在巴基斯坦，SEO 人员有 28 人。）所以它白帽的方法也用，黑帽的方法也用。曾经有两次被 google 惩罚过，现在网站正在走下坡路。

Tradekey 在 09 年 4 月被惩罚了一次。不过它的处理方法也很巧妙。其实即使在 google，一个网站被惩罚过的话，如果想以后不受限制，最好就是直接换个域名。Tradekey 发展到 09 年的时候已经是一家有好几个语言版本的大网站，已经是一个品牌，不可能轻易换域名。 它的处理方法是启用新的二级域名 www1.tradekey.com 来替代 www.tradekey.com ， 然后

把 www.tradekey.com 302 跳转到 www1.tradekey.com 。这样既不需要换域名也相当于是个新网站，后来流量马上恢复。其实不用 302，还有一个解决办法就是用 cname 也可以。（SEO 人员不懂技术是很难做得好的。）

最近的一两年，Tradekey 的核心 SEO 人员不断流失。SEO 这块也就慢慢变弱。在今年的 google 内容农场事件中，Tradekey 又因为内容问题被惩罚了一次。（Tradekey 的内容一直都很差。） 直到现在流量还一直在跌，一个曾经 SEO 这么优秀的网站也就开始没落了。

如何规划好网站的URL（1）

URL 的问题是 SEO 过程中的一个基本问题，做一个新网站也好，优化现有的网站也好，都绕不开这一点。这两篇文章就来大体总结一下 URL 的规划应该怎么做。

在开始讲这些问题之前，需要先阅读完以下文档：

- 《优化网站的抓取与收录》 http://www.google.cn/ggblog/googlewebmaster-cn/2009/08/blog-post.html
- 《谷歌搜索引擎入门指南》第 7 页到 11 页。 点此下载
- 《创建方便 Google 处理的网址结构》 http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=76329

这些都是 google 官方的文档，讲述了各种各样的规则。这些对百度也是同样适用的，因为它是针对爬虫的特性提出来的，并不是只有某个搜索引擎才适用。

看完上面的那些这些规则，发现翻来覆去讲得都是怎么让爬虫能非常顺畅的抓取完整个网站。其实绝大部分网站都存在这样或那样的问题的，也包括我这个博客，在抓取方面也存在一些问题。但是看在每篇博文都能被收录的情况下，也就不去优化了。但是对于很多收录还成问题的网站（特别是大中型网站）来说，就要好好规划一下了。大家可以用HTTrack抓取semyj这个博客看看，就能发现为什么我这么说了。（谁能一天之内抓取完这个博客的人请告诉我。）

还是先从搜索引擎的处境讲起吧。正如 Google 在文章中写道的那样：

网络世界极其庞大；每时每刻都在产生新的内容。Google 本身的资源是有限的，当面对几近无穷无尽的网络内容的时候，Googlebot 只能找到和抓取其中一定比例的内容。然后，在我们已经抓取到的内容中，我们也只能索引其中的一部分。
URLs 就像网站和搜索引擎抓取工具之间的桥梁：为了能够抓取到您网站的内容，抓取工具需要能够找到并跨越这些桥梁(也就是找到并抓取您的 URLs)。

这段话很好的总结了搜索引擎所面临的处境，那么爬虫在处理 URL 的时候会遇到哪些问题呢？

我们先来看重复 URL 的问题，这里说的重复 URL 是指同一个网站内的不同页面，都存在很多完全相同的 URL。如：

http://www.semyj.com/archives/1097 和 http://www.semyj.com/archives/1114 这两个页面。

http://www.semyj.com/archives/1097

http://www.semyj.com/archives/1114

模板部分的 URL 是一样的

虽然页面不同，但是他们公用的部分，URL 地址是一样的。看起来如果不同的爬虫抓取到这些页面的时候，会重复抓取，从而浪费很多不必要的时间。 这确实是一个问题，不过这个问题搜索引擎倒是基本解决好了。实际上，爬虫的抓取模式不是像我们理解的那样看到一个网页就开始抓取一个网页的。

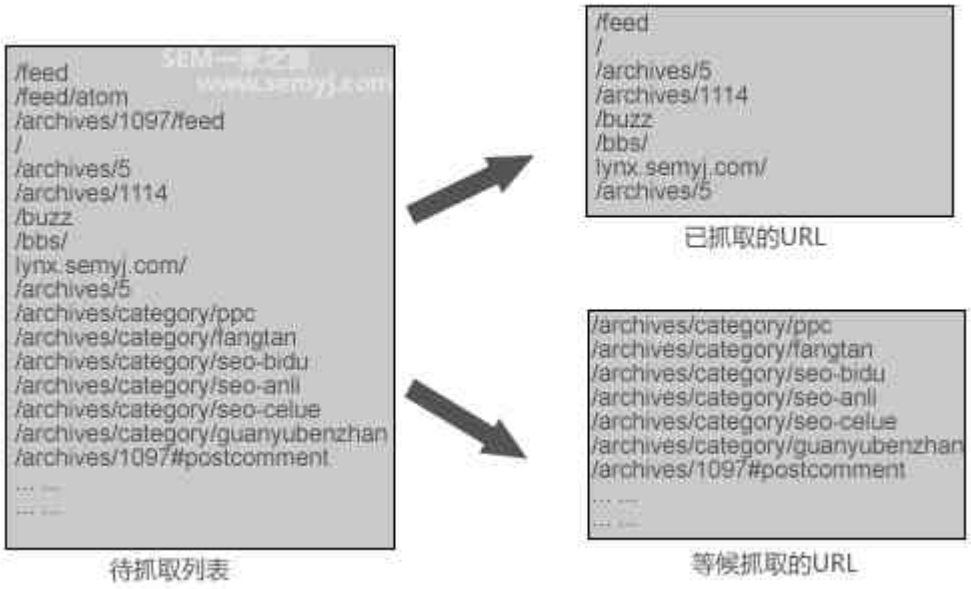
爬虫顺着一个个的URL在互联网上抓取网页，它一边下载这个网页，一边在提取这个网页中的链接。假设从搜索引擎某一个节点出来的爬虫有爬虫A、爬虫B、爬虫C，当它们到达semyj这个网站的时候，每个爬虫都会抓取到很多URL，然后他们都会把那个页面上所有的链接都放在一个公用的“待抓取列表”里。（可以用lynx在线版模拟一下爬虫提取链接。）

待抓取列表

这样一来，在“待抓取列表”里，那些重复的 URL 就可以被去重了。这是一个节点在一种理想状态下的情况，不过实际上因为搜索引擎以后还要更新这个网页等等一些原因，一个网站每天还是有很多重复抓取。所以在以前的文章中，我告诉大家用一些方法减少重复抓取的几率。

这里有一个问题，很多人肯定想问是不是一个网页上所有的链接搜索引擎都会提取的，答案是肯定的。但是在《google网站质量指南》中，有这样一句：“如果站点地图上的链接超过 100 个，则需要将站点地图拆分为多个网页。”有些人把这句话理解为：“爬虫只能抓取前 100 个链接”，这是不对的。

因为在“待抓取列表”里的 URL，爬虫并不会每一个链接都会抓取的。 链接放在这个列表里是没问题的，但是爬虫没有那么多时间也没必要每个链接都要去抓取，需要有一定的优先级。在“待访问列表”里，爬虫一边按照优先级抓取一部分的 URL，一边把还未被抓取的 URL 记录下来等待下次抓取，只是这些还未被抓取的 URL，下次爬虫来访问的频率就每个网站都不一样了， 每一类 URL 被访问的频率也不一样。



按优先级抓取

那么在“待抓取列表”里的 URL，哪些是能被优先抓取，哪些是被次要抓取的呢？

我们稍微思考一下都能明白这个抓取的优先级策略应该怎么定。首先，那些目录层级比较深的 URL 是次要抓取的；那些在模板部分的或重复率非常高的 URL 是被次要抓取的；那些动态参数多的 URL 是次要抓取的…..

这么做的原因，就是因为搜索引擎的资源是有限的，一个网站实际拥有的内容也是有限的，但是 URL 数量是无限的。爬虫需要一些“蛛丝马迹”来确定哪些值得优先抓取，哪些不值得。

在《谷歌搜索引擎入门指南》中，google建议要优化好网站的URL结构，如建议不要用“…/dir1/dir2/dir3/dir4/dir5/dir6/page.html”这样的多层嵌套。就是因为在待抓取列表里，在其他条件相同的情况下，爬虫会优先抓取目录层级浅的URL。如用Lynx在线版查看本网站的页面：

抓取优先级

如果说，在这 17 个链接里，爬虫只能选几个链接抓取的话，红色箭头所指的链接在其他条件相同的情况下是要优先的。

但是这里又有一个误区，有人在 SEO 过程中，把所有的网页都建立在根目录下，以为这样能有排名的优势。这样也是没有理解这个原因。而且爬虫在这个网站上先抓取哪些 URL 后抓取哪些 URL，都是自己的 URL 和自己的 URL 比，如果所有网页都是在同一个目录下，那就没有区别了。

最好的规划 URL 目录层级的方式，就是按照业务方的逻辑来规划，从内容上应该是什么从属关系就怎么规划 URL 就是。就像《谷歌搜索引擎入门指南》中举的那些例子一样。

（顺带说一下。我经常看到，一个网站中，很多人非 SEO 的人员，如工程师和网页设计人员或者网站编辑，都以为 SEO 和他们做的事情是相反的。这都是因为长期以来一些 SEOer 经常提交很多明显违反用户体验的 SEO 需求给他们，造成他们以为 SEO 就是和他们做的事情是有冲突的。实际上，SEO 和别的部门有非常少的冲突，只要你能用科学的方法去实践，就能发现以前有太多误导人的观点了。还有，对于其他部门的专业人员，他们专业领域的意见非常值得去考虑。）

爬虫有一个特点，就是它不能实时的比较它正在抓取的内容是不是重复的内容。因为如果要做到实时的比较，那它至少要把正在抓取的页面和那些已经在索引库的页面做对比，这是不可能短时间内可以完成的。 前面把所有 URL 统一放到一个待抓取列表中的方法只能避免那种 URL 完全一模一样的重复抓取，但是无法应对 URL 不一样、但是内容一样的抓取。

正如所有搜索引擎都强调的那样，动态参数是一个经常产生URL不一样、但是内容一样的现象的原因。所以搜索引擎建议大家用静态化的方法去掉那些参数。静态化的本质是URL唯一化，在《优化网站的抓取与收录》这篇文章中，曾经用的

“一人一票”这个描述就很贴切的表达了这个意思。静态化只是一个手段而不是目的，为了保证URL的唯一化，可以把URL静态化、也可以用robots.txt或nofollow屏蔽动态内容、可以用rel=canonical属性、还可以在webmaster tool里屏蔽一些参数等等。

而静态化也会有好的静态化和不好的静态化之别。我们这里不说那种把多个参数直接静态化了的案例，而是单纯来看看如下两个 URL：

http://www.semyj.com/archives/1097 和 http://www.semyj.com?p=1097

这两个 URL 中，这个静态化的是不是就比动态的好呢？ 实际上这两个 URL 的差别很小。首先这两种 URL 搜索引擎都能收录，如果说动态 URL “?p=1097”可能产生大量重复的内容让爬虫抓取，那这个静态的 URL “archives/1097”也不能保证不会产生大量重复的内容。特别是爬虫在抓取时碰到大量有 ID 的静态的 URL 时，爬虫无法判断这个网站是不是把 session ID 等参数静态化了才造成的，还是这个网站本来就有这么多内容。 所以更好的静态化是这样的：

http://www.semyj.com/archives/seo-jingli

这种 URL 就能保证唯一化而不会和其他情况混淆了，所以 URL 中要尽量用有意义的字符。这不是因为要在 URL 增加关键词密度而这么做的，是为了方便搜索引擎抓取。

以上是因为爬虫固有的特点造成的抓取障碍，而有时网站的结构也能造成爬虫的抓取障碍。这种结构在《优化网站的抓取与收录》一文中用的名字是“无限空间”。文中举了一个日历的例子：如很多博客上都会有一个日历，顺着这个日历的日期一直往下点，永远都有链接供你点击的，因为时间是无限的。

其实还有更多的“无限空间”的例子，只是“无限空间”这个名词没怎么翻译好，翻译做“无限循环”就容易理解多了。 举一个例子：

京东商城笔记本分类页面：
http://www.360buy.com/products/670-671-672-0-0-0-0-0-0-0-1-1-1.html

筛选条件

当点击“惠普”+“11 英寸”这 2 个条件后能出来一个页面，点击“联想”+“14 英寸”+“独立显卡”也能出来一个页面。那总共能出来的页面有多少呢？

这个页面中，品牌有 18 个分类、价格 9 个分类、尺寸 7 个分类、平台 3 个分类、显卡 2 个分类。 那么可以组合成的 URL 个数为：

按 1 个条件筛选： 18+9+7+3+2 = 39 。

按 2 个条件筛选：18×9+18×7+18×3+18×2+9×7+9×3+9×2+7×3+7×2+3×2=527 。

按 3 个条件筛选：18×9×7+18×9×3+18×9×2+18×7×3+18×7×2+18×3×2+9×7×3+9×7×2+9×3×2+7×3×2=3093。

按 4 个条件筛选：18×9×7×3+18×9×7×2+18×7×3×2+18×9×3×2+9×7×3×2=7776。

按 5 个条件筛选：18×9×7×3×2=6804。

总共可以组合出的 URL 数量为：39+527+3093+7776+6804=18239 个。

笔记本分类里总共才 624 个商品，要放在 18239 个页面中，而有的页面，一个页面就能放 32 个产品。势必造成大量的页面是没有商品的。如点击这几个筛选条件后，就没有匹配的商品出来了：

无结果

这样的结果，就是造成大量重复的内容以及消耗爬虫很多不必要的时间，这也可以认为是“无限空间”。 这类情况非常常见。如

某房产网的无限空间

上面举的京东商城的例子还是不怎么严重的，有的网站能组合出几亿甚至无穷无尽个 URL 出来。我在国内和国外看过那么多同类的网站，居然发现迄今为止只有两家网站注意到了这个问题。究其原因，还是因为很多 SEO 人员不太重视数据，这种问题稍微分析爬虫的日志就可以看出来的。直到现在，还有一些 SEOer 认为把这些以前是动态的页面静态化是个有积极意义的事情，没看到不好的一面就是这样的动作制造出了大量重复的页面，向来就是一个在 SEO 方面不好的改动。

Discuz论坛SEO优化指南

因为现在很多人在做自己的论坛，为了对他们有些帮助，我打算把我优化这个论坛的步骤写下来。 文章会分为好几篇来写，由于涉及的细节很多，我自己也是在边写帖子边给论坛做SEO优化，所以我也不知道会写到什么时候结束。

1，选择论坛程序和版本。

我选择的论坛程序是Discuz!x1.5，语言版本是 gbk 版。为什么选这个版本呢？

首先Discuz!x1.5的用户体验要比Discuz!7.2好很多，大家慢慢用这个论坛就会发现这一点。然后Discuz!x1.5的SEO基础也要比Discuz!7.2好。其实Discuz!7.2是有很多SEO上面的缺陷的，以前那个老论坛我想做一下SEO优化，但是发现要改的还真不少。但是Discuz!x1.5注意到了很多对SEO不友好的地方，如很多容易产生重复的链接就用JS调用等等。

显然 Discuz!x1.5 的开发团队做事非常用心，让我也对改这个论坛程序有信心很多。

那为什么要选GBK版本而不选UTF8版本呢？ 这是为了让中文搜索引擎第一时间知道我网站上的内容是中文版本。

爬虫在GBK编码的网页，看到的是：

```
1. <html xmlns="http://www.w3.org/1999/xhtml">
2. <head>
3. <meta http-equiv="Content-Type" content="text/html; charset=gbk" />
```

复制代码

而在utf-8编码的网页看到的是：

```
1. <html xmlns="http://www.w3.org/1999/xhtml">
2. <head>
3. <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
```

复制代码

Utf-8编码的网页，一时半会还真不知道这个网站里的内容是什么语言的，而且如果一个网页中有中文和有英文的时候，搜索引擎还要根据其他一些条件来判断网站的语言版本。而GBK版本一看就知道是中文的了。

大家如果去查看一下的话，Discuz官方论坛用的就是GBK版本。

那已经在用utf-8的中文discuz论坛怎么办呢？ 其实还是有方法解决的，可以定义一下xmlns 属性，把 lang="zh-CN" 加在里面就可以了。 所以utf-8版本的代码变为：

```
1. <html xmlns="http://www.w3.org/1999/xhtml" lang="zh-CN">
2. <head>
3. <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
```

复制代码

Discuz论坛很多文件都需要这么改，可以用Dreamweaver整站查找一下。很多其他网站也一样。这样改好后，搜索引擎能识别这个网页为简体中文版。

wordpress程序之所以SEO方面表现很好，就是因为这些细节它都注意到了的。 可以看看我的博客 http://www.semyj.com/，wordpress程序默认都会定义这个标签的。

2，选择服务器系统

我是很早就不想用windows做服务器操作系统了，只要体会过linux系统好处的人恐怕都是如此。其实，选择什么样的服务器系统也能影响SEO效果的。我最近给很多大中型网站做SEO顾问的时候据发现一个很有趣的规律：凡是用windows类系统搭建的网站，SEO方面的表现都是不太理想的，而且要优化起来难度也是大一些的。

原因是很多方面的，因为windows类主机不是很稳定，只要程序员不那么熟悉整个网站，要么被动的频繁当机、要么需要主动停机维护、要么数据库压力大以及运行的代码先天不足导致服务器速度非常慢。 我观察过很多网站的爬虫访问情况，在同等条件下，windows类主机的抓取量都是差一些的。

当然，这个问题在一个资深的技术人员手里都不是问题，但就是优秀的技术人员实在太难找到了。(顺便广告一下：我们公司招c++和PHP人员，有兴趣的联系一下我。本广告长期有效。)

3，优化网站的访问速度

网页的加载速度对SEO影响比较大，可以看我博客上的这篇文章来了解原因--（http://www.semyj.com/archives/969）。优化网站的加载速度，可以从以下几个方面来优化。

1）DNS
2）服务器网络环境
3）服务器硬件和系统
4）网站程序或CMS
5）前端代码

这些因素不用去记的，基本上就是看爬虫从发起一个请求到返回数据，中间需要经过哪些途径，然后优化这些相关因素即可。
现在这个论坛只优化了 2 个地方，就是是DNS优化和网页打开GZIP压缩。因为用的是现成的程序，其他地方都不太差，暂时先解决一些基本的问题。

DNS上的优化，就是启用了双线主机以及智能DNS。 为什么我要先做这个呢？ 因为我想优化百度爬虫访问我网站的速度。
因为这是中文论坛，做SEO优化肯定要以百度优先。

因为很多人还是没有养成先看数据再来做SEO的意识，所以在优化速度的过程，有个问题没注意到的。这就是没有看看爬虫到底是从什么地方来访问的。 对于大部分中文网站来说，爬虫可能 90%以上都是从北京联通（网通）访问过来的。这个时候就要特别优化北京联通（网通）的访问速度。

所以我用的双线机房有 2 个IP，一个电信的IP和一个联通（网通）的IP。有了个 2 个IP，还要做智能DNS，这样当电信的用户访问论坛的时候，就解析到电信的IP上，联通的用户访问论坛的时候就解析到联通（网通）IP上。 这样，百度爬虫从北京联通访问我论坛的时候，速度就快很多了。 我用的智能DNS服务是DNSPod（http://www.dnspod.com/）提供的，设置的界面如下：

主机记录	记录类型	线路类型	记录值
@	A	默认	202.91.246.211
@	A	电信	202.91.246.211
@	A	联通	202.91.234.147
www	A	默认	202.91.246.211
www	A	电信	202.91.246.211
www	A	联通	202.[illegible]

我在 DNSPod 里面的账户是免费账户，收费账户应该速度更好一点，但是 DNSPod 对于收费账户还要审核，我就一直没升级了。

设置好了以后，还要检查一下到底优化的效果如何。 可以用监控宝（http://www.jiankongbao.com/）的工具检测一下。以前北京联通的响应速度是 1831 ms。经过优化，速度确实会提高很多，如：

这里还列出了是哪方面影响速度的因素大。最好是长期监测这个响应速度，因为这个因素的变化能比较大的影响到 SEO 效果。可以注册成为这个网站的付费用户，就可以每隔几分钟去检测一下网页的响应时间等等。

为了加快前端的速度，我启用了论坛自带的 gzip 压缩。Discuz!x1.5 后台现在还没有启用 gzip 压缩功能的地方，需要手动设置：

打开 /config/config_global.php 文件，把

```
1. $_config['output']['gzip'] = '0';
```

复制代码

改为

```
1. $_config['output']['gzip'] = '1';
```

复制代码

即可启用 gzip 压缩。

Discuz!x1.5 后台还可以做一些速度上的优化如启用 memcache 等等，但是这个相对麻烦点，留着下次来做。

4，静态化 URL

Discuz!x1.5 后台自带了一个静态化 URL 的功能，而且默认也写好了静态化的规则。但是这里有一个问题，就是帖子页面的静态化规则没有写好。

如默认的帖子页面规则是：

```
1. thread-{tid}-{page}-{prevpage}.html
```

复制代码

即规则为：

```
1. thread-{帖子 ID}-{帖子翻页 ID}-{当前帖子所在的列表页 ID}.html
```

复制代码

问题就出在“当前帖子所在的列表页 ID”这里，因为在论坛板块中，当一个帖子是最新发表或最新回复的时候，“当前帖子所在的列表页”是第一页，url 中的数字是 “1”。当这个帖子很久没人回复沉下去的时候，“当前帖子所在的列表页”就不知道是几了，可能出现在第二页，也可能在第十页。这样，每个帖子的 url 经常在变化。会产生很多的重复页面，而且 url 经常变化，当前帖子积累的权重会丢失。

为了解决这个问题，可以重写 url 静态化规则。当然修改页面代码也能解决这个问题，但是不方便维护，因为修改后的文件以后可能会被升级文件覆盖，而且会丢失部分功能。

论坛用的是 linux+apache，而且论坛是作为一个虚拟主机放在服务器上。 Url 静态化的过程就这么操作：

新建一个文本文件，文件名为“.htaccess”，然后用 UltraEdit 编辑这个文件，写入的规则为：

```
1. # 将 RewriteEngine 模式打开
2. RewriteEngine On
3. # 修改以下语句中的 RewriteBase 后的地址为你的论坛目录地址,如果程序放
   在根目录中，为 /，如果是相对论坛根目录是其他目录则写为 /{目录名}，如:
   在 bbs 目录下，则写为 /bbs
4. RewriteBase /
5. # Rewrite 系统规则请勿修改
6. RewriteCond %{QUERY_STRING} ^(.*)$
7. RewriteRule ^topic-(.+)\.html$ portal.php?mod=topic&topic=$1&%1
8. RewriteCond %{QUERY_STRING} ^(.*)$
9. RewriteRule
   ^article-([0-9]+)-([0-9]+)\.html$ portal.php?mod=view&aid=$1&page=$
   2&%1
10. RewriteCond %{QUERY_STRING} ^(.*)$
11. RewriteRule
   ^forum-(\w+)-([0-9]+)\.html$ forum.php?mod=forumdisplay&fid=$1&page
   =$2&%1
12. RewriteCond %{QUERY_STRING} ^(.*)$
13. RewriteRule
   ^thread-([0-9]+)-([0-9]+)\.html$ forum.php?mod=viewthread&tid=$1&ex
   tra=page\%3D$3&page=$2&%1
14. RewriteCond %{QUERY_STRING} ^(.*)$
```

```
15. RewriteRule
    ^group-([0-9]+)-([0-9]+)\.html$ forum.php?mod=group&fid=$1&page=$2&
    %1
16. RewriteCond %{QUERY_STRING} ^(.*)$
17. RewriteRule
    ^space-(username|uid)-(.+)\.html$ home.php?mod=space&$1=$2&%1
18. RewriteCond %{QUERY_STRING} ^(.*)$
19. RewriteRule ^([a-z]+)-(.+)\.html$ $1.php?rewrite=$2&%1
```

复制代码

用 UltraEdit 写好规则后，按 F12，在文件另存为的窗口上，有个“格式”选项，选“utf-8-无 BOM “保存。然后把“.htaccess”上传到论坛根目录。

然后在进入后台 --> 全局-->优化设置-->搜索引擎优化 。 其他保持不变，就把“主题内容页”规则改为：

```
1.  thread-{tid}-{page}.html
```

复制代码

如：

保存设置再更新一下缓存就可以了

4，解决重复 URL 的问题和屏蔽垃圾页面

Discuz! X1.5 还是不可避免的出现重复 url 的问题。（希望有渠道的朋友能把这些问题反馈给 Discuz 相关人员）
这些重复的 url 即浪费了爬虫大量的时间，又使网站的原创性受到损害。所以一定要屏蔽很多重复页面。

另外还要干掉一些垃圾页面，所谓垃圾页面就是一些没什么 SEO 价值的页面，也帮助爬虫节约时间。

解决这个问题，最好是用 robots.txt 文件来解决。因为里面的规则是最强势的，所有爬虫第一次访问一个域名，第一个动作都是下载这个 robots.txt 文件并读取里面的规则。 其他一些 nofollow 和 rel=canonical 等标签适当的时候再用。

虽然 Discuz 默认写了一些 robots 规则，但是还是不够理想。

根据从首页的代码中发现的问题，需要在 robots.txt 里增加的规则有：

```
1. Disallow: /forum.php$     （这条规则在第 3 节中去掉了）
2. Disallow: /search-search-adv-yes.html
3. Disallow: /space-username-*
4. Disallow: /forum.php?gid=
5. Disallow: /home.php?mod=space&username=
6. Disallow: /forum.php?showoldetails=
7. Disallow: /home-space-do-friend-view-online-type-member.html
8. Disallow: /space-uid-*
```

复制代码

根据在板块帖子列表页面发现的问题，需要在 robots.txt 里增加的规则有：

```
1. Disallow: /search.php$
2. Disallow: /forum-forumdisplay-fid-*
```

复制代码

根据在帖子详细信息页面看到的问题，需要在 robots.txt 里增加的规则有：

```
1. Disallow: /forum-viewthread-tid-*-extra-page%3D.html$
2. Disallow: /forum.php?mod=viewthread&tid=
3. Disallow: /forum-viewthread-tid-*-page-*-authorid-*.html
4. Disallow: /forum-viewthread-tid-*-extra-page%3D-ordertype-*.html
5. Disallow: /forum-viewthread-action-printable-tid-*.html
6. Disallow: /home-space-uid-*
```

复制代码

至于为什么要写这些规则，由于描述起来实在啰嗦，所以大家自行到源代码里查看为什么。

robots的写法是很灵活的。
可以看一下百度的robots写法指南：
http://www.baidu.com/search/robots.html
以及google网站管理员中心的说明：
http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=156449

robots.txt写到这里并不是结束，还有两件事情要做。

1，因为robots.txt和nofollow是不同的意思，所以robots.txt并不能代替nofollow。以上这些需要屏蔽的地方还需要用nofollow标注一下。 不过因为要改的源码太多，暂时先不动。需要用nofollow，还有一个原因是某些搜索引擎并不遵守自己所定下的robots规则。

2，因为只看过论坛中的三类主要页面，还有很多页面没查看过，难免会有漏掉的地方，所以需要以后经常到日志中查看爬虫的轨迹，看看爬虫还有哪些抓取问题。

5，修改网页的头部信息

Discuz x 1.5 的<head>部分有一些信息是我们不想要的，所以需要修改。

打开 \template\default\common\header_common.htm 。(在第 2 篇中加那个 lang="zh-CN" 也是在这个文件)

找到 <title> 部分， 去掉 Powered by Discuz! 以及最后一个中划线“-”。

然后删除 <meta name="keywords" 这一行。 因为keywords已对SEO没有任何用处，所以去掉也没什么。

如果说要列SEO行业的落伍行为，那写"keywords"算是其中一项了。

其他的一些选项如 ： <meta name="generator" 和 <meta name="author" 等等按理可以去掉，但是很多开源系统存在很多年了，搜索引擎能适当的判断出来一个网站用的是什么CMS，所以暂时保留。因为Discuz 在SEO方面还是存在其他一些不合理性的，让搜索引擎知道这个网站是用Discuz做的会有好处。

6，修正系统本身的一些静态化错误。

Discuz x 1.5 升级到 11.25 补丁后，因为系统默认的首页都是 forum.php，即使访问 index.php也会 301 重定向到 forum.php 。那么在《Discuz论坛SEO优化指南（2）》中用robots.txt文件屏蔽的forum.php需要重新放开了。 也可以在模板设置里把这个forum.php的文件名改成其他文件名，不过暂时先这样。

还有一个问题，就是“主题内容页面”（详细帖子页）的静态化规则，很多规则都和版块列表页面的URL都不统一。如：帖子的翻页地址从第二页起都为：

forum-viewthread-tid-220-extra-page%3D-page-2.html 这样的形式，但是实际上URL应该为 /thread-220-2.html 这样的形式。 还有就是在面包屑中，论坛版块的URL为：/forum-forumdisplay-fid-45-page.html 这样的形式，而实际应该为：/forum-45-1.html 。 如下图：

所以打开 /source/module/forum/forum_viewthread.php

找到 第 108 行

```
if(!empty($_G['gp_extra']))
```

在上面加一行：

```
$_G['gp_extra'] = !empty($_G['gp_extra']) ? rawurlencode($_G['gp_extra']) : '';
```

7，让搜索引擎收录图片

图片 SEO 过来的流量也会不少的，但是 Discuz x 1.5 默认的设置是 游客无法看到图片的。也就意味着搜索引擎也收录不了帖子中的图片。

打开 后台 - 用户 - 用户组 - 系统用户组 - 游客 - 编辑 - 附件相关，在“允许下载/查看附件”上选“是”。

但是如果光这样设置了，那游客也可以下载其他附件了。 所以在 后台 - 全局 - 积分设置 - 积分策略 里，把下载附件设置需要 1 个金币就是。这样下载其他附件还是需要注册成为会员的。

论坛的 SEO 优化还只是开始，后面还会有很多其他的优化。

如何做好外部链接

这篇文章写在《光年外部链接挖掘系统 2.0》即将发布的时候，希望能普及好的外部链接理念，同时也希望大家能认可这种软件。

在目前的很多 SEO 团队中，外部链接人员都是人数最多的部分。很多团队的外部链接人员多到了外人无法想象的程度。我亲眼所见的外部链接团队，最多的只有一百多人的规模。但我听说过的外部链接团队，有达到一千多人规模的。据说还有另外一个团队特别有意思，他们有 300 多个外部链接人员，工作时间是每 8 小时一班也就是工厂里经常用的“三班倒”，真正的把外部链接做成了“链接工厂”。之所以这样，是因为现在依靠大量外部链接确实可以把某些关键词的排名做上去。（把某些关键词的排名做上去并不意味着给网站带来很多的流量，有时候甚至是相反的。）

在写过《内部链接还是外部链接？》以后，有人问我是不是外部链接不重要。实际上我从来都没有否认过外部链接的价值，只是有一点，我眼中的外部链接是以一个网页为单位的，也就是一个网页以外给这个网页的链接都是外部链接。而很多人追捧的是以一个网站为单位，大家更喜欢把一个网站之外给的链接称为外部链接。不过建议现在还有这个想法的人，可以去看看PR值的计算公式也好，去看看一个关键词的搜索结果也好，他们都是以网页为基本单位的。（并不是说PR值很重要，而是从它的计算方法上可以看出搜索引擎是怎么看待网页和网站的）

不过今天要讲的部分是怎么做一个网站外的外部链接（以下把网站外的外部链接都称为外部链接）。我觉得这一部分实在是浪费了太多 SEO 团队的资源，其实也造成了很多社会资源的浪费：在我们招聘员工的过程中，不知道遇到过多少那种刚毕业、在实习过程中只是人肉做链接的毕业生；SEO 这个领域也不知道还有多少那种上百人的外部链接团队。

回到主题，那如何做一个网站的外部链接呢？ 我的建议是：第一，为用户来做外部链接的方法是最好的方法；第二，要追求效率，尽可能自动化地解决问题而不是人海战术。

首先第一点：为用户来做外部链接。其实 SEO 在国内流行之前，很多网站也要做外部链接的，如证券之星这样的网站在 1996 年就有了，他们也会投入很多的资源去做外部链接。（互联网在中国其实还蛮早的，SEO 也不只是 2003 年以后中国才有的，实际上至少是在 2000 年的时候，我就看到过一些中文的讲搜索引擎排名的资讯。）但是那时候去做外部链接的目的和现在不一样，那时候的目的比较简单，就是希望通过外部链接给网站带来直接的流量。后来随着大家知道外部链接能提升关键词的排名，这种简单的目的反而没有了。

为什么说现在这样的方法是最好的呢？答案很简单，这样做既能带来更多的流量，也能带来更好的排名。

在 SEO 没有泛滥之前，大部分网站做外部链接的原则是：一定会找相关的网站，还有就是看在对方网页上留下这个链接后，会有多少人直接点击进来。现在依然可以用这样的思维做外部链接的第一个原因是：不管这个外部链接对 SEO 会产生什么效果，至少这些直接的流量已经进来了，而且是相关的流量。我曾经给某个小网站这样操作过，尽管之前这个网站所有的流量才 2 千多 UV，但是由于做了大量相关网页的链接，仅第二天额外增加的直接流量就有 3 千多 UV。 现在很多上百人的外部链接团队，一天能做的外部链接达到两三万个，如果都是做在一些相关的有流量的网页上，那直接流量也非常可观的，只是实际情况差很远。

那这样操作为什么也会带来更好的排名呢？原因就是当初做外部链接的那些原则：因为是在相关的网页上做外部链接，因为链接放在这些网页上会有人直接点击进来。

很多新手经常会问的问题就是：为什么某个网页没多少外部链接或者没有什么 PR 值，排名会很好呢。网页的排名在他们眼里变成了比拼外部链接的数量和 PR 值。由于我很多次都说过不要这么重视 PR 值，那现在可以举个我常用的例子来讲透一下 PR 值和排名的关系。

如果把人比作网页，那么“芙蓉姐姐”的 PR 值相当的高，可能达到 9 或者 10。但是当有人找会计的时候，一定不会去找“芙蓉姐姐”吧，因为她和会计一点关系都没有，哪怕她全身上下都写着“会计”两个字。搜索引擎也是一样的，当搜索引擎分析完整个互联网上几百亿个网页之间的关系，它会发现“物以类聚，人以群分”，判断一个网页讲了什么内容的时候，最科学的方法是看这个网页的外部链接讲了什么内容。所以在 SEO 上来说，做外部链接最重要的标准就是相关性、相关性、和……相关性。

这也是《“锚文本”在SEO方面的重要性》一文中，为什么搜索一个中文词、一个完全是英文的网页排在第一的原因。因为在外部链接的相关性中，锚文本是一个非常直接清楚的信号，说明了这个链接的相关性。其实不仅在google上搜索这个词语排第一，在百度上长期以来也都是这样的。如：

百度的搜索结果

非常在意外部链接数量和 PR 值的 SEO 人员，也可以这么思考一番：用户去搜索引擎是按关键词找相关的网页的，而搜索引擎会根据关键词帮用户找到相关的网页，但是 PR 值高的网页或外部链接数量很多的网页和用户找的相关网页之间没什么直接关系。所以排名和 PR 之间也没什么直接的关系。只要这个网页外部链接的内容相关度高，不管 PR 高不高，在相关的关键词上排名会很有竞争力。就算某个网页 PR 值很高，外部链接的相关度不高，那排名也没什么竞争力。实际上，搜索引擎不光会考虑直接的外部链接的相关性，连外部链接的外部链接的相关性都会考虑。

还有，弄清楚了 PR 值的那个计算公式的人，应该知道一个 PR 值高的网页只能代表这个网页的外部链接多以及外部链接的 PR 值也高而已。（最初始的 PR 值由搜索引擎选网站来指定。）

那从大量的外部链接页面带来很多直接流量，这个在 SEO 上又有什么好处呢？这主要是和那种垃圾链接区分开来。在一个搜索引擎专家的眼里，互联网上各种类别的信息，无论是好的还是差的，都有自己的特征。垃圾链接的特征是这样的：这种链接的数量非常庞大，但是就是没什么人点击。如果一个网页的有大量的外部链接，但点击率只有万分之一或者其他更低比例，那这种链接就是垃圾链接。这个判断会非常的准确，问题是在于搜索引擎能不能知道用户的行为呢？

如果用过 google adplanner 的人应该是不会怀疑这点的。它能知道互联网上每一个稍微有点流量的网站的用户行为。如：网站大概的流量，用户的停留时间等等。

查看数据： 中国

流量统计信息　　所有流量统计信息都是估算值

	区域	全球
唯一身份访问者（Cookie 估算）	120K	140K
唯一身份访问者（用户）	77K	84K
覆盖面	0.0%	0.0%
浏览量	820K	900K
总访问次数	280K	310K
每个 Cookie 的平均访问次数	2.3	2.3
网站平均停留时间	8:00	9:50

Google 预估的网站数据

这种数据是由 google 的工具栏搜集的，然后再根据 Google Analytics 里共享了数据给 google 的样本和统计学上的算法来修正数据。现在 Google 工具栏的装机量有 18%，也就是每 100 个网民中，有 18 个人的行为能被 google 追踪到，日复一日的积累，可以说几乎不会出错。无非就是不同的网站因为用户不一样有不同

的误差而已。 对于百度，它的工具栏也能搜集很多数据，就算目前没有这种算法也会紧随其后吧，因为外部链接对百度排名的影响更大。

综上所述，外部链接收益最大化的方式是要忘记掉 SEO 再来做外部链接，也就是说为用户来做外部链接，而不是为搜索引擎来做。在确实满足了这个外部链接质量的前提下，用人来做链接还是用软件来群发都是可以的，无非就是一个效率的问题。搜索引擎不可能因为外部链接影响排名就不让那么多网站连正常的外部推广也不做了吧。

由于中国的人力非常的便宜，很多人想到的就是“人海战术”。过去的一年多，我以外部顾问的方式接触过很多的SEO团队，凡是有点规模又很重视SEO的公司，很多都有大量的外部链接人员。但是我也发现他们外部链接人员做的每一件事情，没有一样是电脑不可以替代的。哪怕涉及到内容的分析，电脑都可以做得比人有效率。

这就是我要讲的第二点：要追求效率，自动化地解决问题而不是人海战术。

现在很多“人海战术”做外部链接的团队，他们做的效率其实非常低的，效果还不可控。在这些团队中，有的团队会稍微注意一下质量，所以大部分用人手去找链接、分析链接和发布链接的；（只是质量标准和上面说的标准不一样）有的团队软件用得多一点，然后借助人手解决一些软件不能解决的问题。大部分团队都是混用各种方法的。一般用人手去做外部链接，一个熟练的外部链接人员一天最多只能做一百多个外部链接。借助了软件去，数量可以很多，不过由于软件的设计理念本来就是错的，所以做的大部分还是垃圾链接，有质量的链接非常少，甚至有很多用软件后反而被搜索引擎惩罚了的案例。 效果也不可控，同一个外部链接团队，连操作两个基本相同的网站都有可能一个效果好而另一个效果差。换了另一个网站，那就更不可预测效果了。

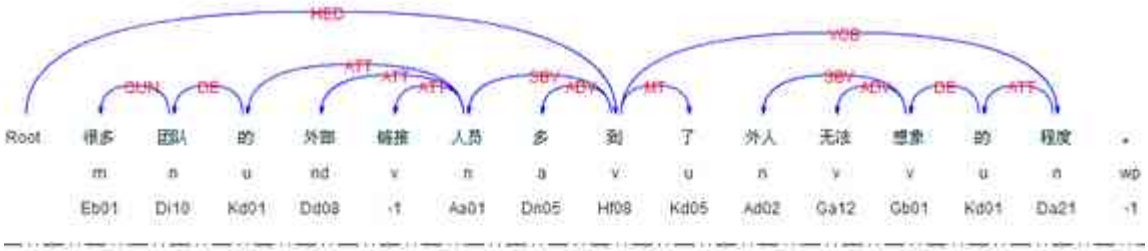
如果深入调查原因就会发现人海战术的过程中有很多问题会导致这种结果。先不说找链接的标准有问题，就说具体地执行某个标准的过程中也不能很好得执行到位。这是一种非常人肉、非常简单枯燥的工作，时间久了没有人真正喜欢。如果没有很好的控制和效果检测，外部链接质量会离当初的标准越来越远。而后面去检测那些外部链接的效果，也无非就是看是否成功的留下了外部链接，以及还是看 PR 值等这种信息。

其实在做这种简单又重复的工作上，无论从什么方面人都没有电脑可靠。从搜集与分析链接，到发送链接等等，只要软件设计得好，电脑更有效率。那如果判断一个网页的相关性呢？电脑只要有相关的技术，人在判断内容的相关性上也没有电脑准确。比如现在我写的这篇文章，到时候由人来判断内容的时候，只会觉得这篇文章讲了外部链接，要人去给这篇文章提炼关键词就只有一个“外部链接”了。如果是用电脑来分析呢？ 那就是下面这样的：

词性（n）	频率	首次出现位置
外部链接	80	7
网页	43	392
软件	30	53
搜索引擎	27	541
网站	26	293
团队	19	69
问题	18	800
流量	14	299
电脑	14	3202
排名	13	269
相关性	11	1825

判断出现的名词以及频率、位置

随便拿一句话，如文章开头的句子，电脑都能非常准确的精简句子、识别主谓宾、并做语义分析等等：



人员.....到.....程度

很多 团队 的.....人员.....到 了 外人.....想象 的 程度

很多 团队 的 外部 链接 人员 多 到 了 外人 无法 想象 的 程度

ATT ——定中关系
DE ——“的”字结构
QUN —— 数量关系
ADV ——状中结构
MT ——语态结构
HED ——核心
SBV ——主谓关系
VOB ——动宾关系

分词、语义分析、识别主谓宾

从效率上来看，要一个人一天分析完一万个网页讲了什么内容的话，这个人不光准确度会越来越差而且估计早已累成痴呆状，而电脑一天可以轻松并精确地分析完几百万个网页。

那这种判断网页内容的方法和搜索引擎判断内容的方法是不是一致的呢？ 可以说有 90%-95%以上的相似度，在搜索引擎这样的领域里，大家差的只是最后的几个百分点，很多算法都相差不大的。

既然人海战术效率不高，而机器又可以解决所有的人的操作问题，而且效率可能还是人的几百倍左右，那完全可以多用机器而不是全用人来做好外部链接。当然我们还是不会违反第一点就是要为用户来做外部链接，只是借助一些设计得很好的软件来提高效率。

在今年年初，看到我们的客户效率那么低后，我们开发出了一个外部链接挖掘与发布软件，也就是开头提到的《光年外部链接挖掘系统 2.0》，这个软件是基于上面提到的那些理念来设计的。

在分析和总结了很多团队的外部链接经验的基础上，借鉴了搜索引擎的一些相关技术，可以把做外部链接的效率提高至少几十倍以上。

软件首先从很多途径搜集到大量潜在的可以做外部链接的网页，一台电脑 24 小时内可以收集到几百万个；然后分析这些网页，在分析网页内容相关性的时候，借用了上面提到的一些搜索引擎的算法，相关度判断得比较准确而且这几百万个网页一天之内也能分析完成；在批量发送的过程中，由于在分析网页的时候已经自动判断出了这个网页是用什么网站系统做的，以及这些页面上是不是能做外部链接，所以可以调用相关的发送模块有针对性的来发送链接。

（软件的截图请看公司网站：http://www.hz-gn.com/software.html ）

这只是一个粗略的过程，很多细节才是这个软件独特的地方。如：在判断网页的相关性上，会判断网页标题中有没有关键词，会用语义分析来分析页面上的内容，会判断关键词的密度、频率、位置等等信息…… 你会发现这个软件的分析行为很像搜索引擎。实际上这个软件的目标就是要像搜索引擎那样思考，百度和 google 这样的搜索引擎认为某个页面怎么样，那这个软件也能基本判断这个网页怎么样；甚至接下来还会做一个爬虫像搜索引擎那样来抓取互联网上的网页。这些技术不要求和搜索引擎完全一样，只要能和现有的搜索引擎有 90%的相似度就可以，毕竟我们只是借用搜索引擎的相关技术来找优质的外部链接。做得更进的一步是结合各种特征去判断网页是由什么网站系统生成的并判断是不是可以做链接等等。（我们之所以可以做这样的事情是因为对搜索引擎甚至百度和 google 的一些技术细节都比较了解。）

使用这个软件，在效率上，5 个人用 20 台电脑就能至少超过现在的那种 300 个人的外部链接团队。一套软件至少可以替代掉 50 个外部链接人员。这不知道可以促使多少人停止浪费他们宝贵的青春去做那么无聊的外部链接工作。这也是要发布这种软件的初衷之一，因为不应该在已经有了电脑的情况下，还用逻辑能力和计算能力这么好的电脑去做那么单调重复的工作。这就像时下流行的一个段子：买来一个诺基亚手机，却拿来砸核桃用一样。发布这个软件后，我也想再一次说明，SEO 是一个有技术含量的、也需要有技术才能做好的工作。

但是技术并不能解决所有的问题，如管理的问题，还有就是使用软件的人本身的能力问题等等。这个软件在过去的半年来一直都在给我们的SEO顾问客户使用，有的网站用它在一个多月内涨了 4 倍的流量，有的网站好像效果没那么好（不过只有一个人在操作）。对于外部链接人员的能力，以前我的想法和很多人一样，觉得只要一个会上网、不那么笨的人就可以。但经过了很多案例，我才发原来和其他任何岗位一样，外部链接这么一个很人肉的岗位也需要找那种领悟力好、学习能力强、又有一定工作能力的人。即便使用上了这个软件都是一样的。

不过软件本身也还是有很多问题等待完善，因为还只出来半年的时间，所以还没有达到一个理想的状态。目前在找链接和分析链接这块的问题不大，找到的链接资源已经用不完，分析得也比较准确。但是在发送链接的阶段，还是有一些问题的，如验证码的问题，目前没有自动识别验证码的功能，只有当碰到验证码后自动弹出再人工输入。这一部分拖慢了速度，不过幸好有验证码的网页还是只有少部分的。在以后的版本中会加入自动识别验证码的功能并引入第三方的打码接口。 还有就是目前能自动发链接的网站系统比较少，还在不停的增加中。

现在很多外部链接团队用的一些 SEO 软件，由于开发这些软件的人对 SEO 的理解不够深入，所以很多软件的功能是比较鸡肋的，然后效率也低。开发软件，编程技术不一定是问题，最重要的是决定开发一个什么样的软件，那个软件运行的指导思想很重要。还有一点，大部分 SEO 软件都喜欢用作弊的方法，纯粹在为搜索引擎做外部链接，不小心被惩罚也是理所当然的。《光年外部链接挖掘系统》是一个白帽软件，我们不是要作弊，只是用电脑来代替人工，借助搜索引擎的技术来处理复杂的人工问题。

软件会在随后的公司网页上公布详细信息，现在开始接受预订： http://www.hz-gn.com/software.html

希望以后的外部链接工作进入一个智能化时代，希望业界以后尽可能少出现那种泛泛而谈的“找外部链接的五大秘籍和十个技巧”之类的文章，大家来光年论坛上讨论如何用电脑来提高生产力吧。

SEO必读

依靠**SEO**，去打造一个成功的网站

说了很多 SEO 相关的东西，但是从来没有说说如何去对待 SEO。我想这篇文章比很多篇讲如何去做 SEO 的文章都还对大家有用一些。

王通曾经写过一篇《阿里巴巴B2B必然走向衰落》，虽然这篇文章一塌糊涂，但是这篇文章当中说阿里巴巴依靠SEO成功的观点我觉得是对的，只是没有他说的那么简单。《SEO是如何依赖技术分析的》一文的结尾我也提到：“从某方面来说，是SEO成就了阿里巴巴”。为了让大家更明白一点，我可以说一些已经公开了的信息：阿里巴巴是先有英文站，才有中文站的。而在 08 年以前，阿里巴巴英文站的收入都是占到阿里巴巴整个收入的 70%。可能有人会说，阿里巴巴英文站的收入都是来自于国内那些做外贸的中国人，但是，在早期，那些外贸企业之所以肯爽快的付费，是因为在阿里巴巴上确实有效果。而这种效果，来自于大量优质的国外买家的流量。

我看到有些人用 alexa 分析阿里巴巴的流量构成，说绝大部分流量都是中文的流量，其实是分析有误的。因为阿里巴巴中文站的域名是 china.alibaba.com ，和英文站 www.alibaba.com 是同一个主域名。所以 alexa 把中文站的流量也算进了英文站的流量里。而 alexa 上的数据本来也就不准的。（这个 alexa 其实可以抛弃不用了，可以用 adplanner 代替）

那么这些大量优质的买家流量如何来的呢？ 大家可以想一想，一个中国人做的网站，不能去国外那么多国家的电视上做广告，不能搞一些类似“赢在中国”的活动。大家也不会听你一个明星般的企业主“忽悠”。那还能有什么办法？无非就只有在线营销。

而SEO不管是在一个网站的什么时期，都是最有效的在线营销手段。早期阿里巴巴大量优质的买家流量，就是通过SEO优化后，十个搜索结果当中有六个是阿里巴巴的页面这样的局面来实现的。

接下来再来看看一个网站一般是通过一种什么运作方式盈利的。说一说我觉得很多网站盈利的本质是什么。

我们可以想一想我们平常生活中很熟悉的超市，菜场，咖啡馆等等实体经济是通过什么方式盈利的。他们盈利的本质是什么呢？其实没有什么复杂的，无非就是低价买进某些商品，或者加工或者转手，然后高价卖出去，赚取中间的差价即可。

而高价能高到什么程度，能卖出去多少，很大一部分原因来源于卖的东西质量怎么样。所以实体经济玩的那个游戏、它们的本质总结起来就是：“低买高卖，注意质量”。

其实网站也是在玩一个这样的游戏，而买卖的东西就是网站的流量。不管是新浪、百度、腾讯、阿里巴巴、google 这样的平台性网站，还是卓越、当当、京东、凡客（VANCL）这样的电子商务网站，还是像一些 SEOer 的喜欢做的垃圾站。都是先通过一定的成本“买进”一些相关的流量，然后“卖出”这些流量具有的价值。本质上都是这么回事，而差别就是每个网站流量“买进”的方式不一样，“卖出”的方式也不一样。

如果以一个网站的营业额来计算，除去成本，各个网站在买卖流量这个生意上差别体现在：

1，你“买进”的价格有多低。2，你“买进”了多少。 3，你“卖出”的价格有多高。 4，你“卖出”了多少。

比如腾讯，因为有 QQ 这个客户端，可以在上面捆绑很多服务，加上知名度，流量“买进”的价格是很低的；“买进”的数量也很大；而流量的“卖出”，是通过它的一系列产品体现出来的，“卖出”的价格其实不高；但是它“卖出”的数量非常大。所以腾讯一个季度的营业额是 4 亿多美金。

腾讯这种平台性质的网站，买卖流量的痕迹还不那么明显。像凡客（VANCL）这种电子商务网站才非常明显的反应出了这种买卖流量的事实。只要价格合适，凡客（VANCL）在互联网一切能低价买流量的地方都购买流量，这是真金白银的直接买进，当然“买进”的价格还是要比腾讯高；买进的数量也不少；但是“卖出”的价格也比腾讯的高了很多；而“卖出”的数量不如腾讯。所以凡客的营业额比腾讯低，但是估计一年也有十几亿人民币以上了。

现在的互联网，各家推出的产品，其实已有越来越同质化的趋势，卖什么是不太重要的，怎么去卖才是竞争力所在。凡客（VANCL）以前卖衬衫很成功，现在卖鞋也非常成功，就是因为从另一个角度来说它卖的不是衬衫也不是鞋，是流量。

更多的其他经营性网站，都是在 4 个方面各有特点，所以才造就了各种不同的网站。比如很多 SEOer 做的垃圾站：“买入”的价格很低；数量比较多；但是靠挂 adsense 这样的“卖出”方式的话，“卖出”的价格奇低；“卖出”的数量其实不少的。但是一年的营业额也就是几万元而已。而当年盛极一时的 PPG 衬衫，倒是知道流量只要能“卖出”，通过一定的成本大量“买入”是很值得的。只是我一直不明白为什么 PPG 当年选择电视广告投入这么贵的买入方式。

而 SEO 在上面谈到的那 4 点里面的作用是什么呢？ 好的 SEO，能给你带来大量、免费、优质的流量。

早期的阿里巴巴英文站，因为有了SEO，流量“买进”的价格很低，甚至有时候可以忽略这个价格；“买进”的数量很多；还因为这些流量非常的优质，所以能

“卖出”的价格也很高；也因为优质，“卖出”的数量也很大。 所以才有了今天大家看到的这个阿里巴巴。

很多 SEOer 都不明白自己掌握的是一种什么技能。更多的人浪费了自己拥有的这种技能。所以大家现在都换种思路去经营网站吧。利用 SEO，其实可以做出更好的成绩的，甚至可以成就一番事业。

现在有很多的网站开始进军国际市场。但是他们首先要面临的问题，就是如何大量低价的“买入”优质的流量。在国内市场，SEO 的重要性还不这么强烈，但是一旦你开始进入国际市场，就发现 SEO 是你海外推广的一个必要的选择。

当然，除了 SEO，很多网站也不惜在其他“买入”流量这个方面投入很大的资金的。如国内某刚刚崛起的外贸 B2C 平台，投 adwords 广告，都是几十万词语的数量。因为到时他们“卖出”的价格会很高，所以这个投入其实是很划得来的。还有一个外贸 B2C 平台，除了 adwords，甚至不惜用人肉在国外论坛发帖的方式去推广网站，而这种方式也占到他们第二大非直接流量的来源。

这些网站可能也尝试过 SEO，但是应该是不理想的。因为现在国内的 SEO 理论水平真的是比国外落后很多。国外同行，不管是 SEO 意识在网站中的普及程度，还是竞争力远远要比国内的企业高得多。我现在通过 hitwise 可以看到国内很多英文网站在海外的流量，一些国内很牛的英文网站，在国外拿到的 SEO 流量其实很少的。实际的关键词排名也不理想。（很多人在查排名的时候都没有用国外的 IP 去查，结果往往会查到自己网站的排名很好，而实际上，可能在国外的前几十页也找不到他们的网站。）

这也是促使我写博客的原因之一。其实很多人，只要有好的基础，再坚持实践多年，是一定能摸索出正确的方法的。但是阻碍他们的，还有很多误导人的言论。

真希望有更多的网站能依靠 SEO 成功。

大中型网站如何推行**SEO**

现在有很多的大中型网站都有专职的 SEOer，相信很多 SEOer 在执行 SEO 项目的时候会碰到各种各样的问题。这篇文章就探讨一下如何在一个大中型网站推行 SEO。

做 SEO 的人都清楚：SEO 能给网站带来大量免费的流量。不过在一个大中型网站中，是由很多部门配合一起来做这个网站的，SEO 所要改动的很多东西基本上涉及到各个部门，那就需要他们来配合你的改动。这样一个过程是很庞杂的，对做 SEOer 的水平要求很高。

很多 SEOer 在公司遇到的各种问题就不描述了，相信很多人都经历过。接下来我想说的是 SEOer 为什么会碰到这些问题，还有怎么去解决。

首先最要命的是 SEO 看起来没有一个固定的标准。而一个网站中的其他岗位，大家都有标准，该怎么做和不该怎么做不会有太大的异议。所以平常做事，凡是涉及到 SEO 的，好像都是 SEO 团队的人说了算，别的部门觉得 SEO 就是一个黑盒子，不明白你的下一个改动是什么样的理由。加上如果一个 SEO 团队中有好几个人，水平又参差不齐的话，先别说和其他部门沟通了，自己团队内部都沟通不好。自己团队内有各种不同意见，即使你有统一的说法传递到其他部门，但是因为很多原因，传到其他部门的信息也是非常混乱的。由此产生的其他各种问题在很多大中型网站屡见不鲜。

要解决这个问题还是很简单的。其实 SEO 是有一个标准的，长久以来很多 SEOer 都忽视了它的存在，那就是《google 网站质量指南》。我之所以老是提这个，是因为很多人就是不去看过里面讲了什么，谁能真的仔仔细细读完那几百篇文章，就能意识到这个《指南》做一个 SEO 标准绰绰有余。还因为它是 google 出品的，所以也没有比它更权威的标准了。这让很多针对 google 做 SEO 的人有了一个很好的依据，在你说服团队内的成员和其他部门的人的时候，都是很有说服力的。涉及到很多细节的问题，到《指南》里找到相关的条例给他看就是。用这个《指南》做标准，也能迫使你抛弃掉很多以前错误的 SEO 方法。相信即使是有多年 SEO 经验的人看完这个，也会发现你在其它地方得到了一些错误的观点。

至于百度的SEO标准，就像我在《百度如何优化》中提到的那样，百度优化和google有 80%是一样的，把《google网站质量指南》里面的内容整理好，完全可以做好百度的SEO。额外你只需要注意百度的特殊性，如他们自己都控制不好自己的搜索结果，他们为了自己的利益干涉算法等等（见注 1）。即使是这样，你能从整体上把握百度的SEO，在优化百度的时候也不会失败。因为决定一个网站SEO流量的其实不是百度，而是用户，百度再怎么烂，也不能阻扰用户找到他们想要的信

息。至于如何从整体上把握，以后还有专门的描述，不过从《SEO关键词的选择》里就可以看出一点从整体考虑的思想。

看完上一篇《google 的良苦用心：网站管理员工具》，请一定要相信：《google 网站质量指南》是来帮助你的，而不是来限制你能做什么不能做什么。

光有了这个标准，还需要做SEO的人的自身的能力要达标，不然在一个大中型网站中推行SEO会困难重重。我之所以在《怎么样去学SEO》中要让大家去了解搜索引擎，了解与网站相关的技术，不光为了让你自己很好的理解SEO，还为了让大家在一个大中型网站中推行SEO的时候更顺利一点。

一些主流的说法会说：“SEO 做好内容和用户体验，其他是不用担心的”。但是在一个大中型网站，做什么内容不是你一个 SEOer 能决定的，这个在网站创立之初就定下来了。而用户体验，那是运营部门和 UED 部门要做的事情，和一个 SEOer 不太相关。既然引进了 SEO 这个岗位，它的职责很明确，这个岗位就是负责在现有的基础上去搜索引擎拉流量的。

在大中型网站中，SEOer 最常打交道的部门是技术部门和 UED(design)部门，很多的 SEO 改动都要依赖他们进行。在一个 SEO 推行得不好的网站，这些部门最烦的就是 SEO 部门的很多需求都是解释不清楚原因的，一个需求为什么要那么做，那么做有什么效果，效果有多大都说不清楚。这在哪里都是一个非常不靠谱的事情，更何况一个技术性那么强的网站。而最后的结果也加深了这种不靠谱的印象，运气好的情况下，做的 SEO 项目可能可以看到一点效果，但是运气不好的情况下，SEO 流量反而下降了。这样永远也得不到其他人的尊重和理解的。作为一个老板，要看你这些改动带来的效果，而作为其他部门的同事，也是希望自己做的事情是有意义的。只要一两次这样的不靠谱的 SEO 项目，就足以让 SEO 在网站推行不下去了。

其实很多 SEOer 其他都不缺，就是缺少对网站和搜索引擎技术性的东西的了解。以及在这个基础上的实践经验和独立的想法。和一些人的说法不一样，我是不相信一个 SEOer 不懂技术可以做好 SEO 的，不管是在大中型网站还是小网站。

首先 SEOer 不懂技术的话，就不能明确掌握一个网站中所有影响 SEO 效果的因素。模拟一个例子：

一个 SEOer 接到一个网站，第一个想改的就是网站的标题和 meta，这个 SEOer 会认为改完这么多 title 和 meta，网站流量会有一个很大的增长。当然他认为这个项目是一定会很成功的。 但是等他把网站全部的标题和 meta 优化好了，真的会有 SEO 流量的增长吗？实际上，这个项目一上线，SEO 流量可能反而下来了。为什么呢？

因为就算这些 title 和 meta 改得很完美，搜索引擎没有特别对待，网站的 SEO 流量也只涨了 10%。而这个 SEOer 没注意的是：这个期间，技术部门调整服务器，使网页的响应时间增加了 100%。这个 SEOer 不明白这个和 SEO 流量之间有什么关系。而实际上，服务器响应时间的增加，使网站损失了 20%以上的 SEO 流量。

总的 SEO 流量从数据上看还是损失了 10%。这个 SEOer 查了很多他知道的原因，都觉得没有什么问题，那他会觉得是这次 title 和 meta 改动的问题。如果他把这次 title 和 meta 回滚的话，可以预见到，流量又下降了 10%。

很多人正在经历着类似的情况，不一定是服务器响应速度变慢，而是其他很多 SEOer 还不知道的细节问题。究其原因，是因为很多人学到的所谓 SEO 知识，有很多都是没有技术作为支撑的很虚的理论。SEO 绝不是几个简单的“注意关键词突出度”，“注意外部链接的数量和质量”就能概括的。首先，为什么要注意这些东西而不是别的？要知道任何高深的知识都是可以从最基本的原理出发，逐渐推导出来的；还有，知道了这些东西，接下来会涉及到这些方面的细节有哪些呢？ 一个大中型网站，去拉流量就要靠科学的办事方法，不能依靠那些虚的东西。需要把每一个影响 SEO 流量的非常细节的因素都掌握，还要对这些因素从多大程度上影响 SEO 流量有了解。

可以保证的是，只要按《怎么样去学SEO》里提到的那 4 点去学习和实践，是一定可以形成一套非常科学的SEO理论和方法的。能大体上确保SEO因素的可控性。这些是真正的知识，而有些其他的所谓SEO知识，只是常识而已。

懂得这些相关的技术的话，在与其他部门沟通如何解决问题的时候是非常顺畅的。

现在很多 SEOer 和技术部门以及 UED 部门沟通不好，就是因为不了解工程师和 UED 在做的事情是什么，不了解他们看待网站问题的角度。每个人职责不同，但是没有一个人不希望自己做的事情是对网站好的。很多 SEOer 和其他部门协作出现问题主要是由于：你要么说不出做这个改动的理由或者理由牵强，说不清带来多大的收益；要么你的 SEO 方案会牺牲其他部门的利益；要么找不出解决问题的办法。

还是举刚才那个例子，就算你知道了流量下降是由服务器响应速度慢引起的，那么怎么解决这个问题呢？

首先你可能要告诉他们为什么服务器响应速度慢会引起 SEO 流量下跌。这个说服过程，你要讲得让一个没有接触过 SEO 的（如你的老板）都能听懂。从收录量如何影响 SEO 流量，再从网页加载速度如何影响收录量，再从服务器响应速度如何影响网页加载速度，就可以让每个人都能听明白。然后，接下来改进这个问题的时候，技术部门的人员会说，以他们的技术，现在服务器只能做到这么快了。那么你怎么办呢？ 不可能说让他们把改动回滚，或者跟你老板说就只能这样了。如果你有相关的技术背景，可以协助他们尽量把服务器响应时间加快，同时就算工程师那边实在没有办法了，也可以另辟蹊径，让 UED 部门改进那些影响网页加载速度的冗余的 JS 代码或者 CSS 文件，从前端加速就是。

当你能和工程师讨论哪几家网站的的架构有什么优缺点，和 UED 部门讨论什么样的 JS 代码其实可以换种写法的时候，你推行 SEO 已经完全不成问题。

注 1： 请相信百度现在一定控制不好自己的搜索结果。因为搜索引擎的结构比一个普通的网站结构要复杂很多。搜索引擎分为：下载、分析、索引、查询四大系统。一个关键词的搜索结果，是经过这 4 个系统共同协作完成的。一个单纯的服务器集群中，当掉几台服务器，这些服务器可以在几分钟内重启，而且其他服务器可以接替工作。但是一个有几万台服务器，又那么复杂的搜索引擎，还要经常更新数据的话，情况就不一样了。google 以前就在这些方面有过很多失误，现在即使有强大的数据中心和先进的文件系统，也还存在部分问题。google 有时也会出现一些和百度一样的错误的，只是看到的几率很小。这些 google 曾经的错误，百度正在经历。当然你的网站因为作弊被处理的情况，不在这个范围。

那些藏在《google网站质量指南》里的SEO技巧

写这篇文章是因为最近收到很多人的咨询，感觉很多人还是没有去看最基础的东西-《google 网站质量指南》。有些人是不知道怎么看，有些人是不屑于看吧。所以再啰嗦的写一下如何看这个《指南》。

其实很多的 SEO 技巧，在这里都写得清清楚楚的，很多技巧是连一些目前的 SEO 专家都还不知道的。虽然我说新手要去看《指南》，但是很多 SEO 多年从业人员都可以看看。

SEO 如果算是一门学问的话，那它和其他所有的学问一样。都要从最基本的东西学起，要下得了苦功夫的。最近翻《读者》的时候看到一个故事，觉得应该和大家分享一遍，文章抄录如下：

陆宗达曾拜国学大师黄侃为师。见过先生，黄侃一个字也没给陆宗达讲，只给他一本没有标点的《说文解字》，说：“点上标点，点完见我。”陆宗达依教而行。

再见老师时，黄侃翻了翻那卷了边的书，说：“再买一本，重新点上。”

第三次见老师时，陆宗达送上点点画画得已经不成样子的《说文解字》。黄侃点点 头，说：“再去买一本点上。”

三个月后，陆宗达又将一本翻得很破的《说文解字》拿来，说：“老师，是不是还要再点一本？我已经准备好了。”

黄侃说：“标点三次，《说文解字》你已经烂熟于心，这文字之学，你已得大半，不用再点了。以后，你做学问也用不着再翻这书了。”黄侃将书扔进书堆里，这才给陆宗达讲起了学问的事。

后来，陆宗达终于成为我国现代训诂学界的泰斗。他回忆说：“当年翻烂了三本《说文解字》，从此做起学问来，轻松得如庖丁解牛。”

看完这个故事，我觉得对于我们 SEOer 来说，《google 网站质量指南》就是那本要去翻烂的书。里面那些最基本的东西，是构建整个 SEO 理论依据的基石。以后所有的技巧什么的都是从这些基本的东西发展而来的。下苦功夫研究完以后，就不太会有让你困惑的东西。我非常相信上文中陆宗达说的：从此做起学问来轻松得如庖丁解牛。以前我把《google 网站质量指南》里的几百篇文章都翻来覆去看完以后，就有类似的感觉。

而读《google 网站质量指南》，不像研究搜索引擎的基本原理一样，有很多晦涩难懂的东西。《google 网站质量指南》里非常偏重于直接告诉你如何去做是

最好的。下面我分析一些《google 网站质量指南》里的文章，看看里面向我们展示了多少技巧。

《google网站质量指南》要从这里开始阅读：http://www.google.com/support/webmasters/

这里只是一个总的目录，以后的几百篇文章都没有一个清晰的列表，要顺着一个个的链接下去才能读完。

先来看这一篇：《我的网站在搜索方面表现不佳》 网址：

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=34444

首先，文章一开始就讲述了搜索引擎的目标和搜索结果是怎么出来的：

我们的目标是为每个查询返回高度匹配的结果。搜索结果是通过我们的搜索索引返回的。

我们都知道搜索引擎一定会遵循为用户返回高度匹配的结果的。在谈到搜索结果是如何出来的时候，google 这里用的字眼是“通过搜索索引返回的”。 “搜索索引”有一个链接，那篇文章里面解释了搜索引擎的工作流程。如果仔细看的话，会发现很多有意思的东西。如提到索引的时候，它写到：

Googlebot 会处理所抓取的每个网页，以便将其找到的所有字词和这些字词在每个网页上的位置都汇编到包含大量索引的列表中。

这里就说了建立索引的时候，关键词的位置，搜索引擎是会记录的。建议大家也看一下我的那篇《分词与索引库》，就应该知道google其实告诉了你蛮多的知识点，如果你自己深究下去的话，是会很有收获的。

再回到《我的网站在搜索方面表现不佳》这篇文章，里面接着写到：

如果您的网站已与网络上的其他网站建立起可靠的链接，那么，我们很有可能会在下次抓取时再度添加您的网站。

用“可靠的链接”来描述优质的外部链接非常合适，其中包括了：链接你的网站的权重、链接存在的时间长久度、链接页面内容的合适度等等。

接着讲了用什么样的步骤提高你的排名。文中写到：

查看您的网站是否被 Google 编入了索引
确保 Google 能够找到并抓取您的网站
确保 Google 能够将您的网站编入索引
确保您的内容实用且具有相关性

这 4 点是这篇文章内容的提纲。我不知道大家看到这 4 点的时候看到了什么。我看到的内容是：1、2、3 点说的是收录量，第 4 点说的是排名。而关于收录量，又分为三步：先查询网站有多少页面被收录，然后再确保有没有爬虫抓取过你的网站，最后看看网站的收录量有多少。这个步骤恰好是我优化大型网站的时候的步骤。我还会用很多的数据来查看这些方面都做到了什么程度。

看一篇文章也好，一本书也好，一定要看它的内容结构。为什么作者会那么安排内容都是有原因的。我就经常感觉我以前的文章很少有人看懂。而如果看过《搜索引擎营销-网站流量大提速》这本书的人，不知道有没有人能回想起整本书的结构，以及很多文章的结构？

在谈到“查看您的网站是否被 Google 编入了索引”这一点的时候，google 写了一个简单的方法判断你的网站有没有被惩罚：

在 Google 上搜索 www.[您的域名].com。如果您的网站未出现在搜索结果中，或在搜索结果中的排名不佳，那么，这说明，您的网站可能由于违反了网站管理员指南而受到了处罚。

这里要注意的是，如果搜索你的域名，排名不佳的话也可能是你的域名受到了惩罚。

在“确保 Google 能够找到并抓取您的网站”这一项内容中，google 稍微写了一下爬虫在页面上的抓取模式：

我们的抓取过程是根据网页网址的列表进行的，该列表是在之前进行的抓取过程中生成的，且用网站管理员提供的站点地图数据进行扩充。在 Googlebot 访问每个网站时，它会检测每个网页上的链接，并将这些链接添加到它要抓取的网页列表中。

搜索引擎爬虫到达了一个页面后，这个页面上的所有链接都是会收集的。但是很多链接不一定会被爬虫接着访问，而是放在一个网址列表里，等着下次来访问。至于下一次什么时候来访问，访问了是不是会被收录，就看其他因素了。这里写得不详细，所以不是很好理解。要更深入的理解整个过程，可以查看搜索引擎原理之类的书籍就可以理解了。以后我会写一下爬虫的具体访问过程，其中分为单个爬虫如何处理、多个爬虫如何协同处理的。

google 还有写：

如果您最近调整了您的网站结构，或将网站移到了新的域中，那么，以前排名较高的网页现在可能会排名不佳。为避免出现这种情况，请在您的 .htaccess 文件中使用 301 重定向（”永久重定向”）来灵活地重定向用户、Googlebot 和其他信息采集软件。

这里要着重看“灵活地重定向”几个字。很多人在做 301 重定向的时候是不能灵活的处理的，因为他们的 URL 没有规划好，所以只能简单的把所有的某类 URL

重定向到同一个 URL。其实为了保证效果，最好是用正则表达式继承前面 URL 的特征来跳转。另外顺便提一下，百度这样的搜索引擎对于不是用.htaccess 文件做的跳转是识别不好的。所以有些网站用 PHP 代码做 301 跳转后，百度依然不识别。

另外还写到：

即便您的网站已经编入索引，站点地图仍是向 Google 提供有关您的网站和您认为最重要网址的信息的一种方法。

这里强调了 sitemap.xml 文件不光是帮助收录的，更是让搜索引擎了解你的网站的。特别注意 sitemap.xml 文件里权重的设置。

在“确保您的内容具有相关性且实用”这一项里，写了两点平常大家不去注意的内容。如：

通过查看热门搜索查询页来了解用户到达您网站的方式。第一个列表会显示您的网站最常出现在哪些 Google 搜索中。第二个列表则显示用户通过点击哪些 Google 搜索来进入您的网站。此信息非常有用，因为它能使您深入了解用户搜索的内容（第一个列表），以及哪些搜索内容可吸引用户点击您的网站（第二个列表）。

了解 Google 查看您网站的方式。关键字页会显示其他网站链接到您网站时所使用的关键字和短语。了解其他用户查看您网站的方式可帮助您弄清如何最有效地定位您的受众。

我在《google 的良苦用心：网站管理员工具》中说过，webmaster tools里面的每一个功能都是对SEO有用的。“热门搜索查询”和“关键字”都是webmaster tools里面的功能。这里的“热门搜索查询”有一个链接，里面解释了各种数据代表什么意思以及如何应用好这个数据。看完那里的内容应该可以解决很多人的疑问。如：以前有人问我“热门搜索查询”那里显示的排名是不是不准，其实是没有理解那个排名的意思，那里的排名是过去几天的平均最高排名。 “热门搜索查询”这里虽然只提供了这么一个简单的工具，但是大家应该学会的是这个工具的一种思想。在优化很多网站的时候，有排名的关键词和实际带来流量的关键词是有差别的。那么接下来就可以做一点什么事情来改善这个情况了。

“关键字”这里就可以查看你的外部链接的语义相关性。如果能常看这个数据，很多人应该就不会去到处乱发链接了。

接下来还写了一个让你增加流量的方法：

优秀的图片内容可能是增加点击量的绝佳方法。我们建议您在发布图片时，仔细考虑如何尽可能为用户带来最佳体验并遵循我们的图片指南。

图片的SEO流量，在国外SEO流量中是能占到很高的比例的。特别是在一些B2B, B2C , C2C网站，基本能达到所有SEO流量的 20%左右。“图片指南”有一个链接，里面都是告诉你如何提高图片SEO流量的方法。 和我在《怎么样去学SEO（三）》中讲的一样，这些方法都是终极的解决方案。不过有些方法还是要你带着思考去看的。如：里面为什么要你指定所有图片的宽和高呢？不光了为网页速度，还因为搜索引擎的图片搜索结果只会返回适当大小的图片，一个只有 10 x 8 像素的按钮图片是不可能出现在搜索结果里的。有些东西《指南》里不能明着告诉你，但是你可以自己推测出来。

回到这篇《我的网站在搜索方面表现不佳》，文章最后还讲了两点：

您的竞争对手基本上无法破坏您网站的排名，也不可能将您的网站从我们的索引中删除。

这里顺便说一下，最应该担心的不是对手陷害你的网站，而是看自己的 SEO 方法有没有违反搜索引擎的质量指南。这点在百度尤为重要，因为一些 SEOer 觉得正常的方法，百度都列为作弊行列，很多“正常”的网站被 K，如果不是百度自己系统出问题的话，都是因为作弊的原因。

我们数据中心之间的差异偶尔会导致搜索结果排名出现波动。当您执行 Google 搜索时，您的查询会被发送到 Google 数据中心以便检索搜索结果。我们有多个数据中心，决定将查询发送到哪个数据中心的因素（例如，地理位置和搜索访问量）有很多。由于我们的数据中心并不是都可以进行同步更新，因此，处理您查询的数据中心不同，所产生的搜索结果排名也可能会有所偏差。

这里已经说得非常明白了。如果去了解搜索引擎的原理，就还能发现对于搜索引擎来说，这种多数据中心有很多好处，不过一个很大的坏处就是同步数据很麻烦。但是为了给用户最好的搜索结果，这种牺牲是值得的。这里是让很多人明白，有时候排名的波动可能什么异常也没有，仅仅是你的查询被定位到了不同的数据中心。

我这篇文章篇幅太长，不一一赘述了。如果哪天我也做 SEO 培训的话，我也会要求培训的人员先看完这个《指南》，才能开始上课的。另外，最近太忙，有非常多的邮件和 MSN 上的咨询都来不及回复，望见谅。

SEO案例

SEO案例：SEO是如何依赖技术分析的

我前面的文章，都是从技术角度出发来做 SEO 的。这篇文章就再举几个例子，来说明一下做 SEO 为什么要依赖技术分析的。另外写这篇文章还出于我一直以来的一个想法，就是我一直都很想赞扬一下 07 年之前阿里巴巴某些做 SEO 的同事，他们很早就在 SEO 领域做出了非常多好的实践，也给网站做出了很大的贡献。

07 年以前的阿里巴巴，经过几年的努力，已经把 SEO 做到了一个很高的境界。大家那时可能还只关注国内中文版的阿里巴巴，称“google 是阿里巴巴的站内搜索引擎”。其实阿里巴巴国际站在国外同行当中的表现要更加优秀。当时很多产品类词语，排在首页的 10 个结果当中就可能会有 6 个是阿里巴巴国际站的。

当时领导 SEO 团队的人员是做技术出生，所以大家大量借助技术手段来分析和解决 SEO 当中出现的很多问题，取得了很好的效果。

因为涉及到现有的业务，只能说几个不那么敏感的例子。

Google 网站管理员工具刚出来的时候，我们网站有很多频道都验证不了那个 google 需要你上传的文件。工程师那边帮助查了很多问题，以为是什么跳转之类的没有做好。查了很多资料，也没有找到特征吻合的相关解决办法。而 meta 验证的方法因为技术上有一点问题做不了。

所以我们 SEO 团队就帮工程师去找问题。我同事瞿波不一会就找出问题出在什么地方了，原来问题出在泛解析上。

具体的过程是这样的：

用了泛解析的 url，无论你把 url 组合成一个什么样子，都会有一个正常的页面给你的。比如：如果你网站的根目录下用了泛解析，
http://www.xxxxxx.com/a.html 这个 url 是你网站本来正常的 url。那么你随意的输入一个本来不存在的 url 如
http://www.xxxxxx.com/adasdsadw.html 甚至
http://www.xxxxxx.com/@####￥Ｙ.html ，网站 CMS 返回的都是一个正常的页面。

这在一个大型网站中，很多地方出于业务需要，都是这么处理的。但是这样做，在“网站管理员工具”的验证方面就一定不能通过。为什么呢？

因为这样谁都可以把这个网站加到自己的网站管理员工具中。比如：www.made-in-china.com 根目录如果用了泛解析，我把这个网站添加到我的“网站管理员工具”里，系统要我验证一下 http://www.made-in-china.com/google15c03c9b508311f6.html 这个文件是不是存在的时候，因为有泛解析，这个文件是一定存在的，那么我就成功把这个本不属于我的网站加到我的“网站管理员工具”里了。我可以随意更改里面的很多设置。

而实际上这样的情况是不会发生的，因为 google 不光会验证你上传的文件存不存在，还会验证一个不应该存在的文件是不是不存在。google 验证完你上传的文件后，接着会模拟一个叫做 google404errorpage.html 的页面是不是不存在。google 觉得你网站根目录下恰好存在一个名叫 google404errorpage.html 的几率是零，所以如果检测下来发现你这个页面也存在的话，那就不能验证通过。google 这个时候已经知道你这是因为泛解析导致的缘故。出于保护你的网站，google 不会让这个验证通过。

上面的这个分析过程，在公开的渠道里是找不到的。现在在《google 网站质量指南》里也只是让你给不存在的页面返回 4xx 状态码而已。

http://www.google.com/support/webmasters/bin/answer.py?hl=cn&answer=35638

而且这个规则也是最近加进去的。以前，根本找不到相关的资料来参考。

那我的同事为什么一下子就找到问题在哪里了呢？那是因为服务器的 log 日志里一定会记录 google 验证的这个过程的，把相关目录下、某个时间段的 log 日志调出来查看就可以看到了。

如果没有 LOG 日志分析，谁能想得到还有这么一个过程在里面呢？ 至今，还有很多网站验证不了这个文件的，现在就可以看看有没有这个泛解析的问题，或者去分析 log 日志看看。

还有一次，网站改版后，网站流量骤然下降了。我们知道影响 SEO 流量的因素有很多，那到底是什么原因导致流量下降呢。我以前的主管 BEN 通过自己的分析，觉得是 url 出了问题。

当时的 url 是这样的： http://www.alibaba.com/bin/buyoffer/mp3.html

我想很多人都不会觉得这个 url 有什么异常。但是在当时，这个 url 有一个致命问题的。

在 02 年 google 的爬虫还不是很成熟的时候，为了避免陷入死循环，爬虫不光会对那些有多余参数的 url 抓取量减少，还会对某些特定的目录不抓取的。这样的目录中，就有 /cgi-bin/ 以及类似的 /bin/ 这样的目录。学过 CGI 语言的人都知道，/cgi-bin/这个目录下是放置 cgi 程序的地方，这种目录下进行抓取是

没什么意义的。/bin/这个目录也是其他很多系统或者语言默认的文件夹名称，这些目录下都不存在 google 应该抓取的页面，所以搜索引擎就屏蔽了这样的目录抓取。而偏偏我们定义的文件夹名称就是/bin/，google 是不会抓取这个目录的。

之后，把这个目录名称改为/trade/，流量马上就恢复了。如今，百度也在robots文件的用法中，就拿/cgi-bin/这个目录做了例
举。 http://www.baidu.com/search/robots.html

我相信这样的问题即使放到现在，也没有人敢怀疑是 google 本身出了问题。有些人还会从上百个因素里找一个看似很合理的原因，导致真正的原因被掩盖了。但是 ben 通过技术分析并实践，却得出了让人信服的结论。类似的事情，我后来也碰到过好几回，因为有他们的经验在鼓舞我，使我也做了一些让别人不能理解，但是却给网站带来很大流量的事情。

技术分析在和竞争对手抢流量的时候，也是竞争力之一。举一个不那么恰当的例子：

sitemap.xml 刚出来的时候。我们自己制作好了 sitemap.xml 文件，但是毕竟这么大型的 sitemap 文件谁也没有做过，特别是里面权重的设置在一个大型网站来说是很有讲究的。所以我们就想参考一个国外主要竞争对手的文件。一开始通过一个方法拿到了他们的文件地址，但是怎么也打不开那个链接，老是返回 404 错误。通过国外的代理服务器去访问也是这样。最后，通过模拟 google 爬虫才能正常的访问这个文件。 原来同样非常重视 SEO 的这个对手，为了让自己的 sitemap.xml 文件不被其他人看到，只有对那种 user-agent 是 google 爬虫的访问才显示这个文件，由于浏览器的 user-agent 是很容易判断出来的，就拦截掉了浏览器的访问。

《怎么样去学SEO》一文，讲述了学SEO要从了解网站和搜索引擎相关的技术开始。而这篇文章就是让大家看看具体是如何应用的。阿里巴巴最早做SEO的那批人，早在国内还不知道SEO是什么的时候就已经涉及到了诸多技术问题，并马上取得压倒性的优势。虽然现在他们因为某些原因都没有在做SEO了，但是他们给网站的贡献是非常大的。我个人的观点：从某方面来说，是SEO成就了alibaba。

网页加载速度是如何影响SEO效果的

“谷歌中文网站管理员博客”刚刚发表了一篇新文章，介绍了一下《google 网站管理员工具》中推出的新功能 -“网站性能”。这个工具是通过 google 工具栏记录了用户访问你网站的速度，并给出了很多加快你网站速度的建议。

http://www.googlechinawebmaster.com/2009/12/blog-post_30.html

而最近也传闻 google 将会把网页加载速度作为影响排名的一个因素。那么网页打开速度是不是能影响 SEO 效果？如果能影响，那是怎么影响的呢？

在揭示其中的道理之前，我希望大家能把上面那些传闻或“网站性能”的功能都忘掉。让我们追本溯源，来看看网页打开速度和 SEO 流量之间的关系。

做SEO有时候不需要听从别人给你的信息和意见，你只要专注于研究搜索引擎，同样也能成功的。这种关系的发现，也得益于早期我非常重视数据分析，所以我在《怎么样去学SEO（二）》中把数据分析能力列为SEOer应该具备的四大能力之一。由于有很多数据做支撑，现在我来给大家分析其中的联系，大家就容易看懂很多。

到了后期，连 google 也认识到了网页速度和 SEO 流量之间的关系，所以在这个“网站性能”以前就推出过相关的工具来帮助网站管理员。

要说明这种关系，就要从搜索引擎爬虫说起。不知道大家对于搜索引擎爬虫在一个网站上的行为有没有概念，我现在发一下某个网站（不是 alibaba）的一些数据出来，大家就能意识到一些爬虫的特性了。下面是从服务器 LOG 日志中分析出的数据。

网络蜘蛛排行

	网络蜘蛛	点击	占全部的 %	访问
1	MSN Spider	384,513	47.68%	591
2	Google Spider	168,996	20.95%	206
3	Yahoo Spider	148,556	18.42%	130
4	AdsBot-Google (+http://www.google.com/adsbot.html)	80,759	10.01%	58

图 1：爬虫访问次数

——

爬虫停留时间

访客	国家/地区	爬虫停留时间	访问
72.30.78.227 yahoo.com (Yahoo Spider)	美国	10:21:56	3
66.249.73.81 google.com (Google Spider)	美国	01:59:17	8
66.249.73.109 google.com (Google Spider)	美国	23:59:16	1
130.60.169.80 unizh.ch (Google Spider)	瑞士	08:59:11	14

图 2：爬虫停留时间

从上面图 1 中可以看到 google 访问这个网站 206 次，这 206 次里面是由很多个不同的爬虫访问的。图 2 显示：有的爬虫一天之内来了 8 次，一共停留了 2 小时左右，有的爬虫来了 1 次，停留了 20 多个小时以上。所以这个网站是被很多个爬虫在不间断的访问的。为了计算方便，可以把 google 所有的爬虫停留在这个网站的总时间加起来。虽然一天只有 24 个小时的，但是 google 的爬虫这一天花在这个网站上的时间可能是成百上千多个小时。这里真实的数据是：在这个网站中，google 所有爬虫那天在这个网站上花费的实际总时间是 721 个小时。

而服务器 LOG 日志里同样可以分析出爬虫在一个网页上的停留时间。如：

搜索引擎排行

	搜索引擎	访问	占全部访问的%	带宽	占全部流量的%	每访问停留时间
1	Google	32,769	13.33%	1.53 GB	7.18%	0:35
2	Yahoo	9,638	3.92%	1.04 GB	3.73%	0:43
3	MSN	757	0.31%	99.29 MB	0.35%	0:18
4	Live Search	473	0.19%	56.19 MB	0.20%	0:37
5	Ask Jeeves	438	0.18%	48.84 MB	0.17%	0:24
6	AOL	199	0.08%	15.00 MB	0.05%	1:18

图 3：爬虫停留时间

得到了这两个数据以后，用所有爬虫总的停留时间除以单个页面的停留时间，就是搜索引擎爬虫这天所抓取的页面总量。

721 小时× 3600 秒 ÷ 35 秒 = 74160 页

那么得到这个搜索引擎爬虫这天所抓取的页面总量有什么用呢？

对于一般的小网站来说，一天能被抓取 74160 页是一个很不错数据。但是我上面给出的是一个中型网站，它整个网站的页面总量有 800 多万有 SEO 价值的页面。那么，在最理想的情况下，这些页面被搜索引擎抓取完要花费的时间为：

800 万 ÷ 74160 页 = 108 天

这是理论上最理想的情况，实际上真实的情况为：

由于有多个爬虫在抓取网站，有很多的页面在一天之内是会被爬虫重复抓取的。有的页面一天之内被抓取 20 多次，有的页面一天之内只被抓取 1 次。通过 “停留总时间 ÷ 单个页面停留时间” 得到的页面数量，是没有去除那些重复抓取的页面的。所以搜索引擎一天之内实际抓取的不重复页面没有 74160 页那么多，而是 40000 页的样子。还有一点，爬虫今天抓取的页面，到了明天还有很多会被重复抓取。所以爬虫不光在同一天内会重复抓取很多页面，而且到了下一天还是会重复抓取前一天抓取过的很多页面。

这样下来，爬虫平均每天抓取不重复的页面数量就只有 10000 页的样子了。那么，要爬虫把这个网站所有的页面抓取完需要的时间为：

800 万 ÷ 1 万页 = 800 天

其实，这个 800 天抓取完整个网站还是太理想化了。实际的情况是很多网站由于结构的原因，有些网页创建后 4-5 年，都还没被爬虫抓取过。

上面的这个分析过程，都没用到什么惊天动地的计算方法。只要你有一点数据分析的意识，就能看清楚事实的。当一个网站收录量不理想，大家就应该去看看那些页面是不是被搜索引擎爬虫浏览过。如果一个页面都没有被搜索引擎爬虫浏览过，是不可能会被收录的。一个网站的收录量没有上去，那 SEO 流量的提升就会有很大的一个瓶颈。

根据我们刚才上面的那个分析过程，要提升网站的收录量，首先要解决的就是搜索引擎每天抓取网站的抓取量。而：

抓取量 = 爬虫总的停留时间 ÷ 单个页面的停留时间

一个网站，爬虫总的停留时间在某个期间是保持相对固定的。当然，有方法提高爬虫总的停留时间，不过这不是这篇文章要讲的内容。我们先通过减少个页面的停留时间也可以增加爬虫的抓取量。

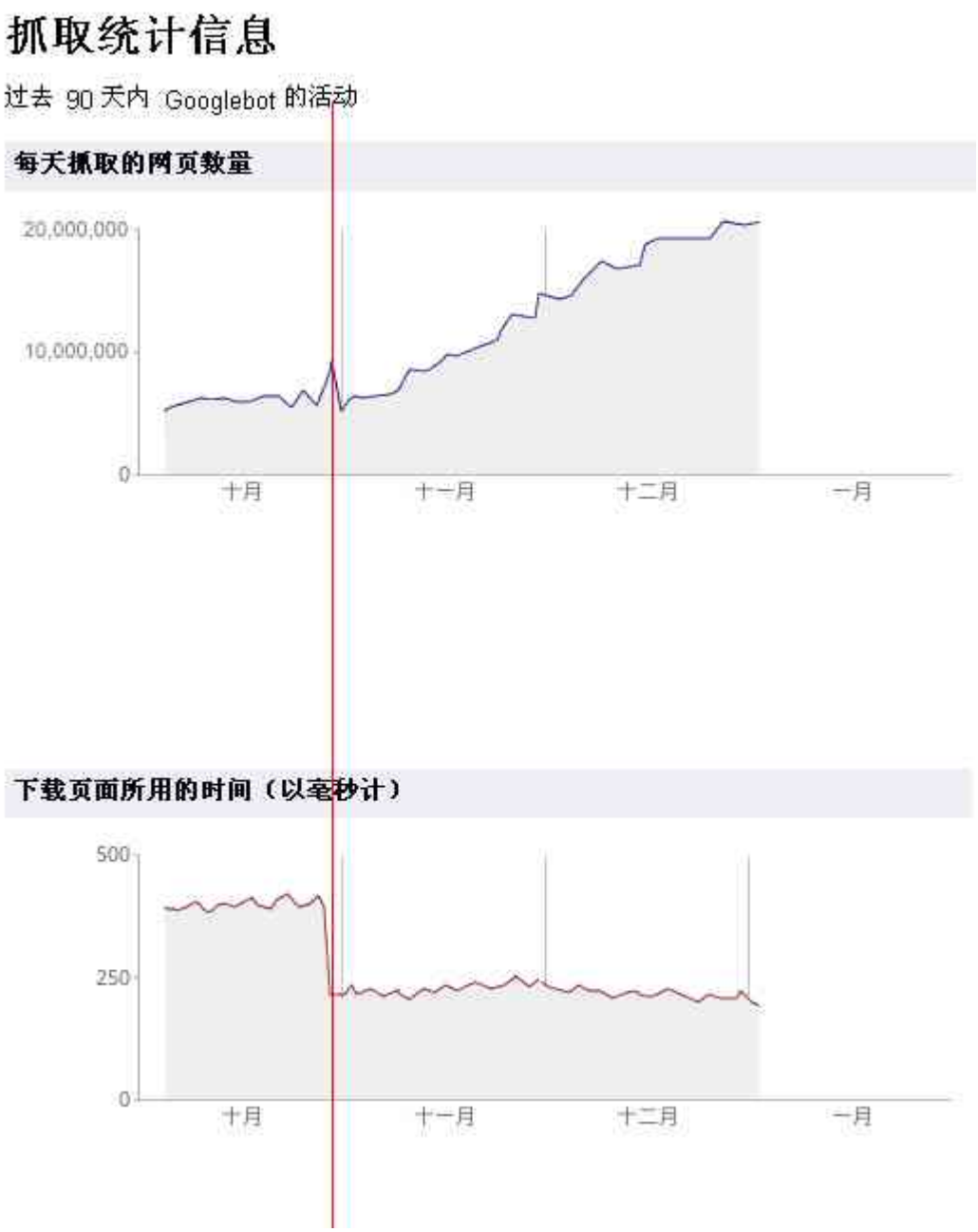
减少爬虫单个页面的停留时间，可以简单的认为提高网页加载速度就可以了，虽然实际上这两个因素之间还存在着一个对应关系，但是这里先不讲。这个时候，网页的加载速度就和抓取量之间有了一个正比的关系，网页加载速度越快，爬虫

整个的抓取量就越大。抓取量越大，有效收录量就会增加，从而促使SEO流量增加，因为给一个大中型网站带来流量的，90%以上都是长尾词。大中型网站，有时甚至是小网站，只要收录量增加一定的百分比，SEO流量也会增加一定的百分比。网页的加载速度，就和SEO流量之间建立了一个这样的关系。

其实，google 也知道网页速度和 SEO 流量之间的关系，所以早在“网站性能”这个功能推出以前，就推出了另一个功能让大家去用，那个功能就是“google 网站管理员工具”里的“抓取统计信息”。那里把爬虫对你网站的抓取量，和你网页下载的平均时间都列了出来。

如果大家平常注意观察这里面的数据就会发现这样的规律：一旦网站下载时间减少了，那抓取量就会增大一点。

一般小网站，这样的规律还不是太明显。因为有其他影响这个规律的因素存在，而且小网站页面数太少了，爬虫随便多增加几百页的抓取量就影响了 50%以上的抓取量。但是在大中型网站，这个规律是非常明显的。如：



抓取统计信息

当这个网站加快了网页加载速度，爬虫的抓取量就稳步增加了。这个图表能很有利的证明上面提到的理论。

这个图表也再次证明了我在《google 的良苦用心：网站管理员工具》里说的：“google网站管理员工具里的每个功能都是和SEO相关的”。

关于网页速度和排名，google 否认曾经将网页速度列为提高排名的因素。而 Matt Cutts 最近也说：In a nutshell - while slow page load times won't negatively impact your rankings, fast load times may have a positive effect. 有人把它翻译成：网页加载速度慢，不会影响 Google 排名，但是网

页加载快却对排名有积极作用。 其实这样翻译是没有理解这句话，Matt Cutts的意思是：网页加载速度慢，不会影响Google排名，但是网页加载快却有积极作用。拿掉以前那个翻译中的“对排名”三个字即可。至于其中的原因，我想通过这篇文章大家都理解了吧。

这也是为什么我要在《内部链接还是外部链接》里强调一下的：有时候是因为“你没有掌握到他们那么多信息，所以你无法理解他们的话。也不会推测出他们的潜台词以及他们没有说全的话而已。” 我其实很少关注Matt Cutts说什么，但是我看到那篇翻译的文章，就断定Matt Cutts不会那么说。

（2011.2.23日注：日志分析软件请访问： http://www.semyj.com/archives/1539 ）

SEO案例：锚文本、关键字、nofollow、Web标准化（一）

前面谈到了做 SEO 需要注意的好几个因素。但是因为工作上的原因，好多因素没有讲透的。（不过其实有些东西我给我们团队的人都没有讲过的。）我看到一些人的回复，对有些 SEO 因素有误解。

还有，我看到很多人都没怎么关注“Web标准化”这一篇文章。其实这篇文章不是在解释为什么要web标准化，而是这样的：几乎所有的SEO站内优化，最终都要体现在网页代码里，而在网页代码里，“结构层”和“内容层”里的东西如何写是很重要的。

所以接下来讲一个具体的案例，让大家了解一下一个 SEO 同行是怎么应用这些基本的因素的。

这个案例就是 Globalsources，是一个非常重视 SEO 的网站。它在细节上的考虑，帮助它获得了很不错的 SEO 流量。

先看它的页面：

http://www.globalsources.com/manufacturers/A-C-Motor.html

Globalsources

首先，它很重视锚文本，因为锚文本描述了被链接的页面的内容。所以全站内所有指向首页（首页统一用顶级域名 www.globalsources.com）的链接，锚文本都是两个选好的关键字。

但是一般从网站设计的角度考虑，有很多指向首页的链接，是不能都用文字的。可以看看它网页的上部分，至少就有 2 个：

1，链接指向首页的 logo 就不是文字。遇到这样的情况，它就退而求其次，给这个链接的 title 属性和 logo 图片的 alt 文本都用和锚文本一样的文字。代码如下：

<img title=”globalsources.com - manufacturers” src=”
http://…../SITE/I/GS2.GIF” border=”0″ alt=”globalsources.com - manufacturers” />

这些文本也起到了和锚文本一样的效果。

2，有时候，指向首页的锚文本是一些” Home”、“back”之类的文字，这些文字削弱了对首页的描述。它的处理方式就是把这些文字图片化，然后再和那个 logo 的处理方式一样，在 title 和 alt 里面加文本。 大家看到的那个“Global Sources Home”其实不是文字，是图片而已。代码如：

再来看它的锚文本是怎么选的。这个就涉及到 SEO 关键词的选择了。

它就给首页选了两个关键字：manufacturers （产品型搜索），globalsources（导航型搜索）

manufacturers 这个词语自然不必说，这个词语准确了描述了这个网站的内容，也是一个转化率很高的词语，本身的搜索量也是非常的大。给首页用这个词语，还有一个好处，是给这个网站“定了性”。所以它的其他很多页面在 manufacturers 的长尾关键词上都排得很好。关于“给网站定性”这一说会在以后的文章中解释。

而选“globalsources”这个词语，有些人可能有点惊讶的。其实，当你拥有自己独有的一个品牌或产品名，它就成了你自己的“导航型搜索”关键词。在看那篇“SEO关键词选择”的文章中，有些人可能只想着怎么把别人的“导航型搜索”流量导到自己的网站上来，却忽视了属于自己的“导航型搜索”。就象我在回复中说的那样，你有一个自己的品牌，你的客户可能是从朋友那里听到或你的广告上看到，然后才来搜索这个词语的。但是一搜之下，发现找不到你的网站，或者只有你的代理商排在前面。而这个客户可能是有购买意向，那这个损失是蛮大的。这种丢失了自己的“导航型搜索”关键词的情况，在很多中小网站中比较多。有时候，竞争对手抢了你的“导航关键字”，在把本来属于你的客户转化成自己的。这一点，在PPC中，是公开的策略了。

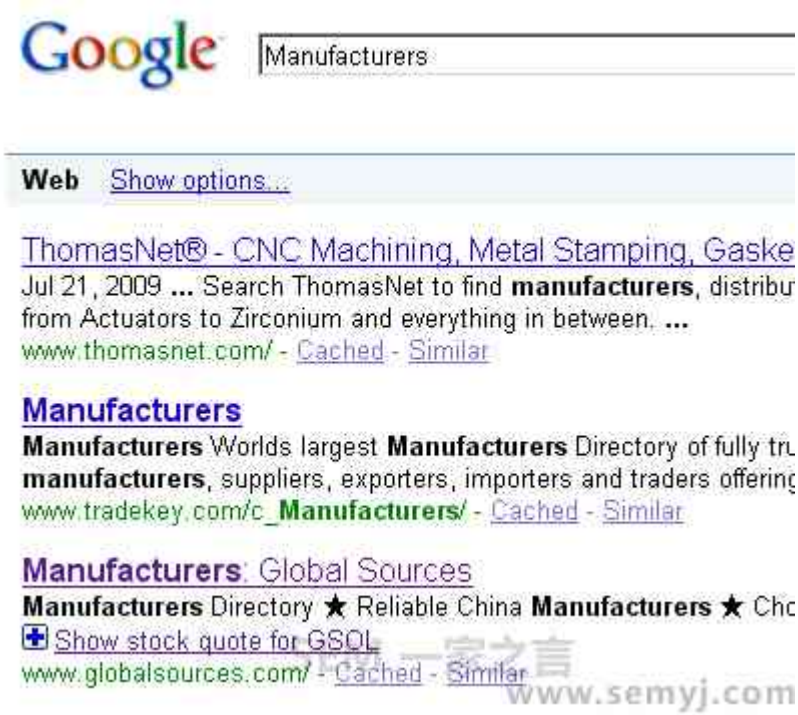
而 globalsources 这个网站，即使“globalsources”这个词不用在锚文本里，其实搜索这个词语它还是能排第一的。因为它是一个大站，一定有很多网站介绍

它的时候，锚文本是这个关键字。而且它的域名就是它的品牌名，在外部链接的很多锚文本中，就会包含这个词语。如： 很多外部链接的锚文本就是它的域名本身—www.globalsources.com。所以外部链接的权重就可以让这个词语排第一。

不过，外部因素是很难控制的。它还要靠内部的锚文本来加强一下，加强这个词语的绝对排名优势，确保一定可以出现 sitelink。

这样一操作，现在它的两个目的都达到了：

1，用美国 IP ，在英文版的 google.com 上搜索 “Manufacturers”，它排在第三。



一定要用美国 IP ，在英文版的 google.com 查询

2，同样，搜索“globalsources”，排名第一，而且有 sitelink。

sitelink

这篇博文写得有点啰嗦，所以要分成两篇来讲这个案例。其实，还可以给globalsources的做法改进一下的。大家可以先想一想怎么改进。

SEO案例：锚文本、关键字、nofollow、Web标准化（二）

上篇谈到了 Globalsources 是如何应用“锚文本”和“关键字”的。这篇还是讲 Globalsources，我们来给它的做法改进一下，看看如何应用“nofollow”和“Web 标准化”。

Globalsources的这些SEO优化，全是Stephen在 06 年指导他们完成的。之后，他们就一直延续了那时候的改造而没有进一步优化。

为了处理“Global Sources Home”这个文字干扰主页内容的表达，它用的是把文字转化成图片的方式。其实就像在上篇的评论中“cm”说的那样，可以保留这个文字，而用nofollow把这个锚文本屏蔽掉的。 不知道大家还记得吗，在“锚文本的重要性”中提到过：一个链接被nofollow, PR和锚文本是不会被传递的。所以，用了nofollow,也是一种不错的解决方法。这样，在大网站的话，和网页设计人员也好交代了。不然，文字图片化在他们看来是很奇怪的做法。这样做不会削弱锚文本对首页的影响力。

在那篇和ZAC的nofollw争论后，有很多人都问我为什么不用robots.txt文件或者meta robots标签来控制爬虫的抓取。其实，nofollow是这两种方式无法替代的。具体的原因，等你真正打算用nofollow，来查看一个网页上所有的链接的时候就知道。

nofollow 的应用，在另一个国外大网站上已经用到了炉火纯青的地步。这个或许以后可以说。

再来看“web标准化”，SEO要在这方面做文章就要时刻想着“web标准化”中的那个网页模型。那个模型不是给网页设计人员看的。

Globalsources 在处理 logo 指向首页的链接中，用的是退而求其次的方法。其实不用退而求其次的，有一种更好的方法，这种方法就是“图片替换”。有比较多的国外 SEO 人在用。

先看 Globalsources 的代码，是这样写的：

```
<a name=” top” href=” http://www.globalsources.com” >
<img title=” globalsources.com - manufacturers” src=”
http://…../SITE/I/GS2.GIF” border=” 0″ alt=” globalsources.com -
manufacturers” /></a>
```

而“图片替换”的代码就是这样的：

1，

```
<div id=”logo”>
<a href=”http://www.globalsources.com/”
title=”globalsources.com - manufacturers”>globalsources.com -
manufacturers</a>
</div>
```

然后在CSS文件里这样写：

2，

```
#logo a {
background: url(http://…../SITE/I/GS2.GIF);
height:80px;
width:300px;
text-indent:-1000px;
}
```

这样的做法，是先做一个完美的“内容层”和“结构层”给搜索引擎看。代码1中，既有锚文本，又有title属性。而不像原来一样是没有锚文本的，只有title属性和alt文本。

代码2是一个“表现层”，text-indent:-1000px; 的意思把文字偏移到屏幕的一千个像素以外。用户看不到那些文字的。然后把那张logo图片作为这个 的背景。把原来那个的logo图片放在CSS里来了。这样在外观上和原来是一模一样的。这种做法完美的平衡了用户体验与SEO。

这里有一个问题，就是google认不认为这个是作弊的方法。因为这个好像是隐藏内容，用户看到的和搜索引擎看到的不一样。这个担心确实有点道理。不过暂时来看，这个方法要应用得当，是没什么负面影响的。

因为这个方法其实一开始是网页设计人员为了更好的web标准化而做的。此方法由来以久而且比较多的网页设计人员在用。在08年11月份，有人咨询过google的工程师，这个方法，在技术上不会被列入作弊的范围的。但是如果你恶意应用，还是会被认为作弊。比如在里面写一大堆东西，或者用来隐藏一篇文章等等。

这种“图片替换”的思想就是把“内容层”的无关因素移到“表现层”。然后想办法怎么在“内容层”和“结构层”更好的表达信息给搜索引擎看。我上次说的web标准化会有很多技巧，这就是其中一个。还有更多技巧大家可以去发掘。

下篇博客我会讲一个三年前就预测到google会使用的排序方法。以及讲述一下我为什么能预测到这个。

SEO访谈

与SEM专家的对话（一） - 回忆录

Phyllis 发表在 SEM访谈 于 2009 年,八月 11th

我接触到 PPC 纯属机缘巧合，Gordon 可以说是我的领路人。Gordon 不仅在 SEM 非常专业，他的管理方式，也完全改变了我对“老板”的印象。与 Gordon 一起做 PPC，是段愉快充实的时光。

Gordon Choi 的经验包括：

7 年半的 SEM/SEO 经验，为许多大小公司制定 SEM/SEO 策略和方案

海外和国内的网站，包括谷歌 Adwords，雅虎 Search Marketing，微软 Adcenter 和百度。

拥有管理 SEM 团队和 SEM 培训经验

协助大型公司的实施 SEM 系统自动化

详细经历：http://www.gordonchoi.com/about
或 http://www.linkedin.com/in/gordonchoi

以下内容来自于回忆与 Gordon 的对话，希望会对一些 SEM 领域的新手有启发，有帮助，许多新手，可能想进入这个领域，也可能想规划自己下一步怎么走。希望有帮助。

如果大家有新的问题，也可以抛给我。有些问题我可以在后面的博文中予以解答。

Phyllis：什么是成为一个好的 SEMer 需要具备什么样的素质？

Gordon：要成为一个好的 SEMer 最重要的条件是信誉（credibility）。 其次是自我学习能力和对数据的敏感度（分析能力）。 要成功做好海外 SEM，以上 3 个条件再加上了解当地用户习惯和语言能力。

Phyllis：自我学习能力为什么这么重要？

Gordon：很大原因就是加入互联网行业的人本身就必须要有这么条件。 互联网行业比其他所有的行业变化都要快。 从 2005 年底到 2008 年底 Google Adwords 就曾经对质量分（Quality Score）进行 6 次大更新，其中并不包括小更新或变动：

http://www.gordonchoi.com/google-adwords-quality-score-for-beginners-20090518

Phyllis：SEMer 应该怎样选择加入什么公司?

Gordon：SEMer 的选择应该有 4 类公司：

（1）SEM 代理公司

最好加入一家技术上比较全的 SEM 代理公司。在里面边学边做，跟同事多交流经验和心得，

和很快接触到很多不同类型的网站，SEM 技术会进步得很快。

缺点是可能每涉及面会比较窄，比较容易慢慢会对互联网其他的技术不太了解。

（2）Online Marketing / Digital Marketing 代理公司

跟只做 SEM 代理的公司差不多，不过学到的也会有其他网络营销的技巧。

不过学到的 SEM 本领可能不会有专做 SEM 代理的公司精。

（3）传统广告公司转型的代理公司

有些代理原本是传统广告公司，因客户对在线广告或 SEM 的需求每年增加，

代理就在技术上不成熟的情况下中途转型为 digital marketing 代理公司或 SEM 代理公司。

如果你是个新手，在这类公司会比较难掌握到深入的，全面的 SEM 技术。

（4）自己拥有好几个大型网站的公司，如阿里巴巴

最大的问题是大公司有大公司的规模和想法。每个员工的分工都会分得很细，

加上做 SEM 的员工可能就只有 3-4 个，技术精的可能一个都没有。

如果你是个新手，在这类公司会比较难掌握到深入的，全面的 SEM 技术。

Phyllis：SEM 会不会是个太狭窄的行业?

Gordon：SEMer 人员和专家需要很好的将自己定位。可以这样说，Online Marketing 专家、SEMer 和 SEOer 的定位既不是全技术人员，也并不是传统市场人员。 自己需要先将自己好好定位，这样才能确立以后的发展方向。

Phyllis：5 年后的 SEM 将会是怎样?

Gordon：雅虎会玩完，其本上就剩下 Google 和微软。但是对国内 5 年内影响不大，国内还会是百度和谷歌的天下。SEM 在国内还没有成熟，还有很多空间。不过应该是有 2 个方向：

第一：国内的公司扩展中国市场，搭建中文网站，需要做国内 SEM

第二：国内的公司扩展国外市场，搭建英文（或其他语种）网站，需要做国外 SEM

互联网公司大多都类似。一时生意会非常好就发展得快，遇到客户的广告预算减少也会很快关门。

所以在其中的 SEMer 需要时常居安思危。

答复**SEM Watch** 的采访内容

插篇 SEM Watch 的采访内容。回答得比较仓促，不过还是不想修改，原文登出吧。

————————————————————————-

1.前不久的点石北京茶话会上提出了 SEM,SEO 2.0 的概念，认为 SEO 的关注点应该从之前排名和流量的关注，更远更高的看向转化率和营销。你对目前国内 SEO 行业发展情况有什么看法呢？

其实早就应该看转化率和营销，只要往深处再想一点就谁都明白的。另外不应该说 SEO2.0 的概念，SEO 就是 SEO。
我现在还不清楚国内的 SEO 是不是可以成为一个行业，因为至今还没有一个规范出来。或者说有现成的规范但是很多人还没认识到，如 google 的《网站质量指南》。有些人认为 SEO 前景堪忧，我觉得如果 SEO 还是给人一种忽悠的感觉的话，那确实堪忧。但是如果能有一个标准和规范去操作，很多方法是能让局外人也能信服的话，那就前景很大。毕竟现在大家都用搜索引擎找信息，而 SEO 还是效果最好的网络营销手段。

SEO 就算短时间不能有标准，也要形成一套公认的非常科学的 SEO 方法。推行这套方法也相当于有了半个标准了。

2.你对那些没有能力组建自己公司 SEO 部门的中小企业有什么样的 SEO 建议？

先做好内容，从用户的角度提供更多有价值的内容出来。内容稍微一多，就需要做 SEO 方面的优化了。把 SEO 外包可以的，但是不要事先自己来定义应该怎么做 SEO。比如：给几十个关键词，要一个 SEO 代理公司把这些词语做上去。那 SEO 公司就不得不用不正规的方法。做 SEO 的最终目的还是为了营销，所以只要能带来效果，就不要拘泥于形式。

即使是小网站，也应该用整站优化的思维来做 SEO。如果有专职的技术人员或网页设计人员，可以让他们参考《google 网站质量指南》来不断地改进自己的网站，那在 SEO 方面也会做得很不错了。

3.从你多年的经验来说，SEO 的工作应该在网站运营中占怎样的地位，算得上是不可或缺的吗？

只要人们还是用搜索引擎在找信息，那就需要 SEO。而且越大的网站，越需要 SEO。它在一个大中型网站中是不可或缺的。如果说 UED（User experience design）是为了让网页对用户友好，那么 SEO 就是为了让网页对搜索引擎友好。SEO 在一个网站中基本和 UED 是同样的地位。但它更好的地方是它能带来非常直接的效果。

4.SEO 的从业者应该具有什么样的素质？

一定的技术能力，如了解搜索引擎，会做网站，以及其他相关的技术技能。
数据分析能力，能从错综复杂的数据中找到规律性的东西和本质的东西。
多年互联网从业经验，能从 Marketing 角度考虑问题。
一定的悟性、热衷于实践。创新、有韧性、懂策略性思维，擅于直达本质的思考问题。
另外，很好的人品。

5.你认为现在很多人做 SEO 的人只是四处交换链接，做群发和伪原创的状况正常吗？你看到的现状是什么样的？

这种情况可以理解的，因为 SEO 确实能带来不错的经济效应。但是我觉得这些方法都用错了。我了解很多黑帽方法，但是从来没看到过哪个黑帽方法能有做白帽的方法好的。

群发的坏处大大多于好处，有点经验的人应该都不会做的。伪原创有很多人在做，但是我觉得他们做的事情回报太低了，其实在做一件无意义的事情。至今应该还没有靠做伪原创做成一家公司的，但是靠做其他事情成立了一家公司的比比皆是。
6.你认为 SEO 行业的市场规范应该由谁来主导确定？或者说谁来定 SEO 的黑或白？另外，现在 SEO 能够寻找到与搜索引擎直接对话的机会吗？与搜索引擎有直接的交流吗?

只能由众多的 SEO 从业者自发的规范。如果是 google，它其实已经带头推出了《网站质量指南》。这对这个行业的规范起到了很好的促进作用，就看 SEO 从业者愿不愿意遵守了。
界定 SEO 的黑与白，当然是搜索引擎。但是我们能看到一个这样的规律：在拿不属于自己的 SEO 流量的，一般都是黑的。在拿属于自己的 SEO 流量的就一般是白的。
与 google 的对话，就去 google 的网站管理员论坛。这里面提出的很多问题都可以被 google 的人看到。不过 google 貌似没有推广好，很多人忽视了它而更愿意去一些 SEO 论坛。
百度几乎没有对话的窗口。

7.百度和 Google 在中国的 SEO 行业的规范中现在能起到多大的作用？

google 起到的作用很大，不过它的这个《网站质量指南》的标准应该大力推广的。百度在这些方面没有什么建树。来源于它长期把 SEO 当敌人看，要等百度意识到 SEO 可以和搜索引擎双赢还需要时日。

phpwind访谈记录

今天在 phpwind 接受了一次 QQ 群访谈，可能我的回答用词不当，有些人情绪不好。现在把访谈的内容发出来：

一、很高兴能请到 SEO 专家张国平，请您先向网友们做下自我介绍吧。

不敢称专家，在现在 SEO“专家”满天飞的环境下，我觉得 SEO 专家是个不好的头衔。其实在早期我是一个站长，我从 2000 年开始零零碎碎的做些小网站，那时候开始接触 SEO，只是那时候还没有 SEO 这个时髦的名字，总觉得这个是一些旁门左道的推广网站的方法。从 2002 年起，自己做了一个比较大的网站后，就主要靠 SEO 推广网站，是国内比较早做 SEO 的人吧。后来有机会去了阿里巴巴国际站做专职的 SEO，多年的积累、不错的平台以及周围很多优秀的同事，就有机会去沉下心来慢慢总结系统化的 SEO 方法。

后来，觉得网上很多不科学的言论影响我的工作，同时觉得，如果把自己总结的写出来可以帮助很多人，就开始写博客来分享一些心得。从去年起，创办了杭州光年，主要帮一些国内外的大中型网站做 SEO 顾问，同时自己也做一些其他有意思的事情。

二、“SEM 一家之言”（http://www.semyj.com/）是网上非常有影响力的 SEO 博客，大家可以去看看。那么您为什么会提出“科学的 SEO”的说法？

因为 SEO 本来就可以做得很科学的。但是由于国内的大部分 SEO 人员都是从网上获得一些资讯来学 SEO，很多资讯非常滞后又没有根据。甚至有些方法基本上是十年前的 SEO 方法。所以从他们学习的一开始就被带到沟里去了。大部分人都是继承的某些“名人”的那种套路在做 SEO，很少怀疑过是不是可能还有更科学的方法。我觉得自己很幸运的是在一开始做 SEO 的时候没有那么多漫天飞舞的资讯，可以坚持独立思考来做事情。

提出“科学的 SEO”的说法，就是要来抵抗那种“猜谜语式的 SEO”。让大家意识到一个看似无法科学化的 SEO 其实是可以做得很科学的。也希望号召大家能更科学化的来做这个事情，不然 SEO 永远只是一个边缘领域。

三、 那么，科学的 SEO 有哪些特点？科学的 SEO 同常见的 SEO 有什么不同呢？

“科学的 SEO”的特点就是事实求是，看数据、重技术、系统化。不知道大家有没有发现很多的 SEO 方法，从来都只是告诉你要怎么去做，但是很少或者不能解释为什么要这么去做。科学的 SEO 就能告诉你为什么要这么去做，而且这么做能有多大的效果。科学的 SEO，基本的思维就是一切从常识出发，从一些不需要证明的“公理”开始。用数据来一步步细化每个步骤，把影响 SEO 流量的因素发掘出来，再把这些因素用数据监控起来，然后通过大量的实践，来观察数据和数据之间的关联关系。

科学的 SEO 方法，一定会让大家一看就知道一定是对的，因为都是一些基本的事实。但是如果不看数据，就看不清楚这些事实。重技术，就是因为一个网站和一个搜索引擎，从头到尾都涉及到大量的技术领域的知识，能多了解这些知识，就能更让你控制好后面的效果。系统化，是因为影响 SEO 流量的因素非常多，而且影响强弱程度不同，系统化的思维能让你明白轻重缓急。

科学的 SEO 能告诉你为什么要做某个事情，而且能告诉你能有多大的效果。常见的 SEO 做不到这点。科学的 SEO 会有大量的时间是在做数据，常见的 SEO 只关注几个简单的数据。科学的 SEO 先从整体上把握了所有的事情，再钻到细节里去研究。而常见的 SEO 一开始就只关注了细节，结果整体效果不好。重要的是常见的 SEO 方法到现在已经走到瓶颈了，无法继续下去。而科学的 SEO 就可以越走越远。

四、那您认为哪些常见的说法其实是不太正确，或者值得进一步讨论的？

很多常见的说法都不太正确，如：很多人都过分重视首页，以为首页能代表整个网站，提升首页的 PR 就能提高整个网站的权重。实际上首页也只是一个页面而已，首页和整个网站的关系就是一棵树和整个森林的关系。

还有过分重视 PR 值，其实也是一个没必要重视的东西，要重视的话，也要起码重视 PR 值在整个网站如何分布的。Google 都是建议大家看整站的 PR 值的，结果都只盯着首页看。就算是 PR 值，在影响 SEO 流量的过程中的作用都很小。PR 值和排名也没有直接的关系，

很多人以为做排名就是发外部链接，其实排名绝不是大家要来拼外部链接数量，实际过程中大家也能发现 PR 和排名没有直接关系。引起这些问题的原因就是我前面说的，一开始就钻到细节里了，没有先从整体上把握。还有就是不喜欢做数据挖掘，然后也没有重视各种技术细节对 SEO 的影响或者没有能力从技术上分析 SEO 的各项因素，所以才会有这些言论。

我觉得如今的 SEO 现状用一个成语形容最合适，就是“盲人摸象”。大家还记得这个成语的话发现这个成语再适合不过了。

五、 听到这里，感觉很多习以为常的“SEO 常识”其实是错误的，我们要重新去了解 SEO，那么科学的 SEO 应该如何做起呢？

我建议大家忘记很多 SEO 知识，去寻根问底的学习 SEO。如果去寻根问底学习 SEO，那就要了解搜索引擎是怎么运作的，了解得越清楚越好。然后了解一个网站是怎么运行起来的，有非常多的细节知识，也了解得越透彻越好。然后再了解，搜索引擎和网站之间发生了一些什么事情。这些知识当然不是三天两头就可以知道的，但是了解得越清楚越好。如果能自己做一个简单的搜索引擎（注：原意是用开源程序搭建一个搜索引擎），以及从服务器架设开始一个人从头到尾做一个简单的网站出来，那了解就差不多了，但是越更了解细节，越能帮助你很多。有了这些知识，就能发现 SEO 所涉及的范围比之前想象的都大，也能发现有太多的因素在影响着 SEO 效果。

接下来就是尽可能把影响 SEO 效果的因素都挖掘出来。这个时候不是要求大家去追搜索引擎的算法，而是从一些最基本的常识出发，来看看在流量达到你的网站之前，发生了哪些事情。然后就是大量的数据挖掘，把影响 SEO 的因素都数据化，并长期观察和研究数据和数据之间是怎么关联的。

举一个例子：很多人都抱怨网站的收录不理想，但是如果他们去看看搜索引擎爬虫在网站上的行为，就会发现很多的网页，搜索引擎一年到头竟然都没抓取过，你要搜索引擎怎么收录这些网页呢？特别有些大中型网站，一年下来可能还有 50%的网页搜索引擎爬虫一次都没抓取过。至于为什么没有抓取到，一部分是爬虫有它的局限性、网站结构有局限性还有一些其他因素等等。

弄清楚了这些基本事实以后，也只是明白了搜索引擎和网站上发生了什么事情。接下来还要研究的是使用搜索引擎的人。要通过数据挖掘去看看他们的搜索习惯和为什么要有那样的搜索习惯。比如：还有很多人都在追一些热门词语的流量，结果是造成搜索引擎上 80%的人在抢 20%的流量，还有 80%的流量没什么人要，原因就是没有从整体上看看到底整个搜索引擎上流量是怎么分布的。

科学的 SEO 说到底其实只有一个关键词，就是实事求是。一切从常识做起，重新来学习 SEO。

六、 不管哪种类型的 SEO，SEO 流量都是我们最终的追求目标，那么您认为 SEO 流量受到哪些因素的影响呢？

套用我们刚才说的，实事求是地来做事情。要去追求 SEO 流量，那就要去看看 SEO 流量是怎么来的，去看看在 SEO 流量到达你的网站之前发生了什么事情。

在 SEO 流量到达你的网站之前，发生了三件事情，就是网站的整体收录、整体排名和整体点击。从常识出发就能知道，一个网站要追求 SEO 流量，那就要确保网站所有有价值的页面都要被收录，然后这些被收录的页面整体要有不错的排名，有了排名也不一定有流量，因为所有使用搜索引擎的人都会选择他们最想要的搜索结果点击，所以最后还要有人能点击你那些有排名的网页。这样子去思考问题，就会发现已经能从整体上把握 SEO 流量这个事情了。当然这三大因素里面还有很多的细节在分别影响他们。那么我们再顺着这个思路去研究那些细节。

这里，常见的 SEO 方法做得不好的是，很多人都不知道其实自己的网站存在收录问题，哪怕那些看似好像收录不错的网站，其实是查收录的数据出了问题，整体的收录说不定很差。在整体排名这块，太多的人只注意他们觉得重要的词语的排名，没有关注那些潜在客户都在搜索什么词语，结果竟然认为一个页面中有所谓的核心关键词的，还有就是 80%的人在抢 20%的流量。没有注意整体排名，只注意少数页面的排名。实际上看看数据就知道，很多人认为的重要的页面，其实带来的流量都不大。如：很多人重视的首页，在绝大部分中型以上的网站上，加起来也没有带来超过 1%的流量。

最常忽视的就是整体点击，没有点击，什么收录和排名都是白搭。那会发生没有点击的情况吗？在我们的实践中，发现同一个网站，点击率好的时候和点击率差

的时候，有 10 到 20 倍的差距，这就意味着其他因素不变的情况下，流量有 10-20 倍的差距。因为不看数据，不清楚更多的细节，所以常见的 SEO 方法错过了太多的 SEO 流量。

七、 网站的页面收录数量也是一个 SEO 最重要的因素，那么页面收录数量是不是越多越好呢？如何才能提升收录数量？

那种有效页面的收录量越多越好。所谓有效页面，就是指对用户获取信息有价值的页面。一个网站上是存在很多页面对从搜索引擎上来的用户没什么价值的，如反馈表单，登录页面等等。

提升一个网站的收录量，也是从常识出发来思考问题，要清楚的知道在一个网站的网页被收录之前发生了什么事情。以及技术上要怎么改进那些页面。具体的内容在可能不好透漏。因为之前有很多人付费在我的培训中听这些内容的，现在公布出来对他们不公平。

八、 那么，您建议大家要如何去做来提升网站在搜索引擎上的效果？ 如何让搜索引擎抓到网站上有价值的页面。

互联网上，内容为王，在搜索引擎上也是这样。要尽可能的提供给用户最喜欢的内容。

对于内容还不够好的网站，Google 和百度都有关键词查询程序，可以通过挖掘上百万关键词找到用户需要的信息。每个行业都可以挖掘很多用户在找的内容，当然这样去挖掘也能看到搜索引擎上有很多别人不要的流量。那你可以创造一些相关的内容去获得这些流量。

对于已经有很多内容的网站，当然也要去看看用户在找什么信息。但是更要紧的是要去挖掘爬虫在你网站上的行为，看看爬虫在抓取你的网站的时候遇到了什么问题。把你现有的内容更好的展示给搜索引擎是第一要紧的事情，然后再去想办法获得那些你应该能拿到的流量。

我也不会否认外部链接的作用，在中小网站，外部链接还是很重要的。但是要注意很多传统的只注意外部链接数量的方法不可取，那种滥做链接的效果是不可控的。滥做链接不一定是指群发，而是那种碰到有链接可以做就去做的行为。另外，在判断外部链接的质量的标准上，也压根就不能用 PR 值来做判断。其实在我们的 SEO 方法里，压根就不会去管 PR 值。我们不会去做那些不能确定效果的事情，PR 值这个东西至今没有一个人能用数据说明它能从多大的程度上影响 SEO 效果。而我们的其他方法可以说明白，那就是去做我们能确定效果的事情。

大型网站，基本可以不做外部链接，当然有的话也可以的。只是少数的外部链接对大型网站的影响有限了。大型网站要注意的是怎么很好的展示自己的网站内容，和业务结合把内容上的轻重缓急定义好。

九、 很多网站喜欢占据热门关键词的排名，这种做法是不是有价值呢？

这种行为可以理解，但是就是没有看过数据就来做 SEO 的行为。就回答一个简单的问题好了，大家觉得 100 个热门关键词加起来的流量大，还是一万个适中的长尾关键词加起来的流量大。不要想当然的来回答，大家去调用搜索引擎公布的数据，精确匹配每个词语的搜索量，就会发现原来表面上看起来的热门词语其实搜索量不是那么大。所以当然一万个适中长尾关键词加起来的流量大。因为哪怕是一个刚学会上网用搜索引擎的人，都会发现用热门词语找不到他要的信息的，所以都是趋于用长尾关键词。我们自己平常估计也很少用热门关键词找信息吧。

那需要投入的资源有多大的差别呢？十年了，我至今没见过有 100 个热门词语都排得很好的网站，有一万个适中的长尾关键词排得很好的网站还是有很多的，比如现在有些做淘宝客的网站。那他们的转化率相比如何？想一下同样有 100 个人在找黑莓手机，是“手机”这个词语的转化率高还是“黑莓手机 8700”这个关键词的转化率高。

不过我们提倡的方法是既不刻意追求热门关键词的排名，也不刻意追求长尾关键词的排名。应该追求的是—“把相关的人带到相关的页面”。搜索引擎也在做这样的一件事情，就是把相关的人带到相关的页面。而作为一个网站，就是用相关的内容去获得相关的流量。好的 SEO 是能让网站和搜索引擎双赢的。

这么多网民每天在搜索引擎上搜索的关键词是无所不包甚至是很难预测的。不可能凭空想一想就能知道网民在搜索什么关键词。做热门关键词的排名是一个明显的拍脑袋做决策的行为。网民搜索的关键词的大概分布可以看看我博客上的 2 篇文章，从不同的维度来分析的结果。http://www.semyj.com/archives/776 、http://www.semyj.com/archives/188

十、 受到堆积关键词的思路的影响，很多 SEO 入门者会认为，做好 SEO 就是堆积关键词，同时，也有很多站长认为，堆积关键字会导致不好的用户体验，于是得出一个印象：“SEO 优化和用户体验冲突”，您如何看待这种说法？

这就是我说的十年前的 SEO 方法，准确的说是 11 年前的方法了。在百度还没有流行的时候，大家最常用的是当时三大门户（网易、新浪、搜狐）的搜索引擎，当时就是用这样的方法在上面做排名的。 到了现在，还有人要堆砌关键词的话就实在太落伍了。

我们还是用数据说话好了。经常会有这样的操作方法：有些人，会给一个页面定义一个所谓的核心关键词，然后想办法适当重复这个核心关键词，再用各个方法拼命加外部链接等。因为很可能这个核心关键词也恰好是个热门词语，努力了一番以后还是发现这个词语的排名没有上去。但是如果你去查这个页面的流量，发现这个所谓核心关键词虽然没有排名，但是不代表这张页面没有 SEO 流量，说不定 SEO 流量还很多，只是那些带来流量的关键词不是你预先设定的那个核心关键词罢了。 那这个时候大家觉得那个堆砌所谓核心关键词的方法是不是很蠢呢？

堆砌关键词和我说的那个做热门关键词的原因是一样的，没有去看过数据。这里我建议大家去看一个数据，就是统计一下：你的网站上每个网页分别带来了多少流量，以及每个页面上带来流量的关键词是什么。GA 里就可以看到这些数据。

相信任何一个有常识的人看完数据就明白了，但是以前的很多 SEO 方法是压根不看这种数据的。

生搬硬套的 SEO 方法才会破坏用户体验。这也是判断一个 SEO 方法好坏的标准。SEO 和用户体验面对的都是同一群人，SEO 和用户体验都是要为用户着想，怎么会有冲突呢？只能说明一个问题，就是现在很多常见的 SEO 方法，源于那种作弊的 SEO 思维，所以才会造成这样的结果。

SEO 和用户体验在代码的写法上和网页的排版上也不会有什么冲突，很多用户体验友好的做法在 SEO 上面也是效果非常好的。只要不是那种让爬虫看不到内容的方法，其他都没什么问题。

十一、从论坛产品来说，在 SEO 方面存在哪些问题，有哪些解决方法。以 phpwind 的产品为例，有哪些改进的办法。

论坛的产品也经常会有很多 SEO 上的问题，根源就是做 SEO 的很多人对相关领域的技术了解太少，而做技术的人要么不懂 SEO 要么被一些不好的 SEO 观点误导。他们之间产生了一个断层，所以才导致很多不管是开源的程序和系统还是自己开发的网站上都有各种各样的 SEO 问题。

论坛产品的常见问题有：
1， 一个页面的 URL，从有些入口进入 URL 是动态的，而从另一些入口进去又是静态的。造成了同一个页面有多个 URL。如： 论坛板块和帖子页面经常既有动态，也有静态的 URL。

2， 即使是同一个内容的 URL，也可能经常变化。如：由于论坛从功能上要定位一个帖子所在的翻页，所以 URL 中的参数就会随着帖子所在的翻页经常变化。这个现在 phpwind 已经改进好了，但是 Discuz 还是有。

3， 还在写每个页面的 meta keywords，这个标签到现在是一个完全可以扔掉的东西。好几个网站由于扔掉了 meta keywords，在百度的流量反而涨了很多。

4， 但是 meta description，反而写得很简单。这个不能影响排名，但是能很大程度上影响点击率的，所以需要认真批量写。

5， 代码上太多丢失和没有闭合的标签，会影响搜索引擎抽取内容。

。。。。。。

解决办法就是把这些问题都修补好。不过上面这些问题仅仅是代码层面的，还有就是服务器上的设置也能影响 SEO 效果，还有网站的结构等等。

一般都是解决整体的问题再来看细节。太多人把细节看得太重要，如考虑 h1 有多少个等等。 在全局问题没有解决之前，考虑再多的细节也没用，很多细节其

实也没那么重要。如：你一个网站有 50%的网页搜索引擎还没看到过的时候，你哪怕把这个网页的 SEO 做到一朵花一样了，对这 50%的网页来说，一切都是白搭。

当然这些都还只是表面的东西，更深层的有整个 SEO 的策略如何制定。怎么样拿到一个预期的效果等等。不能为了做 SEO 而做 SEO，要考虑的最后的成果。

Admin5.com版聊记录

这是上周接受A5 版聊的记录，感谢光年论坛网友 cx69 的整理。由于很多人不逛论坛的，所以发到这个博客上。 我觉得这个访谈的内容对大家有帮助。

———-

问：嘉宾你好，嘉宾的是科学的 SEOer，首先问，什么样的 SEO 才是科学的?平时站长用的 SEO 方法都是不科学的吗?科学是相对什么而言的?
嘉宾说，提出“科学的 SEO”的说法，就是要来抵抗那种“猜谜语式的 SEO”。让大家意识到一个看似无法科学化的 SEO 其实是可以做得很科学的。那请嘉宾说说科学的 SEO 做法的一些方式吧。

答：现在很多人用的方法，当然有很多是不科学的。 因为很多方法都是没有来由的，不能解释清楚里面的原因，或者解释得很牵强。
科学的 SEO，就是事实求是的来做事情。每个改动都知道为什么，以及清楚将会有什么结果。 要做到这点就是把所有影响 SEO 流量的因素数据化，长期观察数据和数据之间的联系。
只要你去追寻 SEO 流量到达你的网站之前，具体发生了一些什么事情，那就会发现很多影响 SEO 流量的因素。具体可以后面回答问题的时候举例一下。

问：请问下嘉宾
1. 百度权重及 GOOGLE PR 对网站哪个重要些?
2. SOSO 现在对网站的重要性?
3. 购买外链的话,要注意的重点是?
4. 最快的 SEO 操做可行方式有没?

答：1、针对百度的优化，百度给予这个网站的权重很重要。针对 google 的优化，PR 值不重要，PR 和排名以及 SEO 流量没什么直接的关系。
2、SOSO 的 SEO 流量也需要重视。
3、不管是买还是通过其他方式去做外部链接，最重要的是相关性。PR 和是不是 nofollow 也没那么重要。买链接的话，不要和那种垃圾链接呆在同一个网页上。
4、有很多快速获取流量的方法，但是基本只在很小的圈子流传。黑猫白帽都有，很多白帽的方法也能获得大量流量，这些方法都是数据分析和实践得来的。

问：A5 版聊活动最近几个话题都是 SEO 优化的，今天又是大师来分享了。相信嘉宾最近也关注 PR 的更新了，一个多月时间更新了四次，这比以往算是有些变态了，谷歌动作频频。你对这个有什么看法。

答：不要管 PR 值，我们做 SEO 在意的应该是能不能获得大量相关的 SEO 流量，而不是首页的 PR 值是多少。提高 PR 值不是目的，同时也不是提高 SEO 流量过程中要用的方法。所以可以完全不要看 PR 值，如果要看的话，可以去看看 PR 值在整个网站是如何分布的，而不是盯着首页看，首页只是一个页面而已。

建议看看 google 黑板报上的最新文章《不要局限于 PageRank：逐渐选择其它可操作性指标》。

问：网页代码对 SEO 有什么影响，div+css 和 table 对 seo 的好处好坏处分别是什么?

答：不要一开始就钻到这种细节里。他们的区别很小，关键是在你的内容质量怎么样。先整体再局部，先把一些大方向影响 SEO 流量的因素发掘出来，再去细化各自的更多的细节。 这样才能知道轻重缓急，也能保证可控性。其实有很多其他大家都不关注的因素比这个因素重要多了。 如速度问题，服务器性能问题，http 头信息等。
特别是 http 头信息，很多网站都在这方面出了很多乌龙的事情，但是公开的资料里都没人知道这个对 SEO 有什么影响。所以先整体再局部非常的重要，很多公开的资料里至少还有一半的影响 SEO 的因素都没提到。

问：最近百度变化很大，就说排名跟快照，我的一个站快照明明是隔天的，但是一查排名发现不是，以前的快照跟实际快照相差差不多一个月吧，你分析一下这是怎么回事?现在好多的网站都出现这种情况了。

答：网页快照的本来功能是当一个网页不存在的时候，你还能用网页快照看到这个网页的信息。我不知道大家为什么要去看这个快照时间，这个快照时间只能说明搜索引擎在这个时间访问过你的网站。 快照信息不一致，是因为你查询的方式不一致吧。 这个和排名没什么关系。

问：怎么才能做好百度排名?

答：排名不等于 SEO 流量。
我们的目的始终是大量相关的 SEO 流量。 就算要说排名的话，也不是几十个关键词的排名， 是所有潜在用户可能搜索的所有关键词的排名。
就算有排名，如果没什么人点击也没有流量。有的网站的整站点击率有 5%，有的只有 0.5%不到。也就是说，这两个网站整体排名一样的话，流量差十倍。
而真实情况中也有很多这样的例子。只能说，很多方法不科学，所有会有这种问题。

问：百度和 google 在 seo 操作上是否有不同?

答：绝大部分是相同的。至少策略都是一样的。
不一样的地方是，很多作弊方法 百度非常有效，但是 google 没有效果。 google 优化的一些方法，在百度容易被看作作弊的方法。
有些人做习惯了百度的 SEO，用同样的方法去做 google 的 SEO，很多都被惩罚了。
有些人做习惯了 google 的 SEO，用同样的方法做百度的 SEO，被惩罚了或效果不大。
当然也有些作弊方法在百度是没什么效果的，但是因为 google 非常的宽容，所以段时间能获得一些流量。但是时间长了，还是会被降权。
还有百度的排序规则越来越在抄袭 google。

问：嘉宾你好，新站怎么做 seo，介绍几个重要的方面。

答：新站第一要紧的是怎么去做一些用户需要的内容。 这个工作会一直持续在做 SEO 的整个过程。

问：战群怎么做 seo 呢?

答：站群先想好怎么做内容。
其实更重要的是要想一想为什么做站群。作弊的事情，哪怕技术上成功了，策略上也会失败的。
纯粹的做站群是一个没有创造什么价值的事情，除非它是你的一个过程而已。

问：想知道下网站用 xxx.com/bbs 还是 bbs.xxx.com 好，好在哪里?

答：大部分情况下 xxx.com/bbs 要好。

问：企业网站怎么做 seo。

答：关注你的潜在客户，看他们喜欢找什么内容，然后在你的网站上满足他们的需求。
这个需要通过大量的数据挖掘收集用户的需求。
这也是做大型网站要做的，小企业站也一样。
用户的需求摆在那里，争取到了这些流量的网站就会很不错。

问：做 gg 优化要注意什么?做外链有没有数量上的要求?

答：注意外部链接的相关性，注意外部链接的增加速率。
外部链接不是用多少判断的，是看你增加的速率如何。

问：一个单页如何 seo?

答：大量相关的外部链接。
相关性的重要性远远超过 PR 值和是否 nofollow。
还有：一个网页以外的所有链接都是外部链接，排名的最基本单位是网页而不是网站。

问：现在百度的排名很乱 真搞不明白他是怎样排的 ，有一些内容页都排到第一位了，说到外链接啊等都没有主域名的网站强吧，那他的排名怎样排到第一位了呢?

答：一个网页的排名，不是在拼外部链接数量。
想一个简单的问题：
难道，外部链接多的网页就是用户需要找的网页吗?
所以这个问题也是不应该存在的。

问：最近谷歌的排名我发现主要还是搞 PR 在支掌着，我现在的 PR 站排名还算可以大部分的流量都是谷歌来的，百度方面很少。做谷歌的你认为是以什么为主，比如外链，是高 PR 的还是数量呢?

答：中文更容易，因为很多好的方法没有普及，正因为大部分人还在用 N 年前的方法，所以流量很容易做。
不过英文的更可控,因为 google 和 bing 这种搜索引擎不会自己犯一些低级错误。

问：请问嘉宾 ，百度再次被站群占领了， 百度的变化我应该如何应付!淡定就不用说了。

答：其实我一向反对作弊的。不过，如果一个搜索引擎自身不成熟的话，可以适当作弊。
要适应这个环境就是这样。
不过更要是要关注做 SEO 的目的，永远不要忘记当初为什么做 SEO，看现在做的是不是在实现当初的目的。

问：前段时间很火的狼雨的网站，我分析了一下，他能在那么短的时间作上去是靠的是外链。有几百个首页优质的外链，这些外链都是跟 SEO 有关的，是靠优质外链才能做上去的，所以得出的结论是百度的排名核心是外链?
首页的外链越多越好，还如 mbaobao 的也是靠外链排上去的， 它购买了很多的外链，把他的排名拉得很高，就连淘宝这个词都在首页前面。你是这样认为的吗?谈谈你的看法吧?

答：一个网页以外所有的链接都是外部链接。
外部链接现在在百度还是很重要。在 Google 依然很重要，但重要性正在逐步降低。
做 SEO 的话，还是要关注整体，很多人在某个关键词是可能是排在第一了，但是流量不一定很多的。甚至因为太关注这些词语的排名，整体流量其实比那些在这个词语上没什么排名的网站还要低。很多时候都是这样。
很多依靠 SEO 成功的网站，可不是某某热门关键词排得很好就在 SEO 上很成功了的。

问：seo 最重要是哪几点?

答：最重要的是怎么满足你的那些有可能给你网站带来收益的潜在用户的需求。他们在找大量的资讯，而你就是考虑用什么样的内容去满足他们。
在现阶段，仅仅考虑这个问题并解决好，后面的很多 SEO 方面的技术只要不出大问题，都能让你在 SEO 上获得很大收益。
因为现在没多少人考虑这个问题，所以很多的流量是没有人去要的。

问：SEM 是网络营销，跟他 seo 是否是可以分开的?seo 是帮助网站在搜索引擎中一个有排名的策略，sem 是推广以及推销，可以是这么的去理解吗?

答：SEM=搜索引擎营销 = SEO+PPC 类的付费推广
SEO 和 PPC 在一个网站要互相配合，因为他们都是从同一个平台获得相同的收益。
所以很多资源都可以共享，如关键词库等等。
SEO 和 PPC 的目的都是一样的，只是一个投入技术资源，一个投入钱。

问：网站有大量图片，是 B2C 商城。有没有好的办法优化图片(除了写 ALT 标签)，让图片为网站增加权重，带来流量?

答：在《google 网站质量指南》中有详细的 SEO 图片流量的优化方法，去那里搜索一下即可。
同样，那里有大量 SEO 知识， 很多知识是那些所谓专家都不知道的。

问：还是外链的问题，没有友情连接怎么最快的把一个竞争不大的词做上来

答：很多网站不依靠外部链接，网站的很多网页都能获得不错的排名。
主要是依靠权重。 权重不是 PR，很多 SEO 文章把这个基本概念都搞错了。
权重是由一个网站的大量正面因素累加的、一个搜索引擎对网站的重视程度。

问：一个站要发展是不是只有做好 SEO 才有用?如果是那大概是怎样子的呢?

答：SEO 只是一个网站在网络上推广的一个方式而已。

对很多 SEO 能力不好的网站来说，去买流量或做 PPC 还划得来一些。因为与其花大量的力气，SEO 流量还是原地打转，不如去花钱买点流量。

问：网站内部优化和外部优化，哪个更重要一点?

答：这个问题也是不应该存在的。
排名的最基本单位是网页。 没有一般 SEO 教程中说的外部和内部之分。
大家去搜索一个关键词，出来的结果是一个个的网页而不是一个个的网站。 很多 SEO 教程总把网站当作个体来做事情。

问：我想问 关于链接的问题。
1：外部链接 是首叶好，还是全站好?为什么?
2：内部链接 怎么样做比较好?

答：一个网站的首页，是最不缺少外部链接的。
要明白这个意思，就要把网站的首页当作一个网页来看，网站的首页不能代表整个网站，它只是一个页面而已。
这个网页之外所有给他的链接都是外部链接。所以首页不缺链接。
很多人还是不理解这句话，就是“排名的最基本单位是网页”。当你明白“排名的最基本单位是网页”，自然就知道怎么去做外部链接了。

问：一个关键词，本来有排名，只是排名在第一页 5-6 位置
推一些外链，就没有排名了，是什么原因

答：很多问题，我不看一个网站的数据，都不能真正的知道原因的。
都只是基于经验来猜。
科学的 SEO 就不是来猜，是用数据说话。

问：您从 2002 年就开始做 SEO 了，你是怎么想到提出“科学 SEO”这个理论的?

答：因为本来可以做得很科学，其实很多一线的 SEO 人员也在做着很科学的 SEO。可惜很多人没有动力出来分享，所以由着一些 N 年前就转载泛滥了的资讯充斥着网络。
一个事情不实事求是的做是肯定做不好的，大家不会觉得猜来猜去能做好什么事情吧。

问：嘉宾您好，我目前有几个网站是之前一直挂着出售页的老域名，最近才开始做成网站的。
到现在大概都有一个月的时间，百度收录和快照都很良好，但是 gg 出现了 K 站的情况，应该不存在过度优化的状况，因为新站都没有做过外链之类的，每天更新的文章数量也是比较有规律的。
这是不是说 gg 对网站改版(从单页面的出售页改为内容增多的网站)看得很重呢?这种情况还有可能被 gg 重新收录么?

答：应该是内容质量出了问题。
还有可能是一些基本的技术问题，我说了，其实影响 SEO 效果的因素非常多的。举一个例子，有一个英文网站，在返回的 http 头信息里申明自己的语言编码是 zh-cn,那在英文的搜索是不会有什么流量的。
这个因素没什么人注意，但是确实影响了 SEO 流量。 关于收录也是，你还有很多因素没有去看，比如是否返回码错误，是否搜索引擎知道你在伪原创等等。

问：真正把 seo 做到可控 需要坚持几年? 另外个人做外贸应该怎么入手?

答：不是时间的问题，是方法的问题。
有技术基础的人，如果同时会做数据挖掘，教给他一个系统的方法，可能半年就可以。

问：现在有好多 SEO 的培训个个都称自己为大师，还没有报名时说得很好，说包教包会，包能找到工作，保证毕业出来后能赚多少多少钱等等，但是进去了往往是失望，赚不到钱的就说不认真学，不舍得花钱投资去项目之类的。
为了我们自己的利益我们在去报名培训的时候应该注册哪些情况呢?

答：有一个简单的判断标准： 凡是声称保证赚钱的，基本可以不去。
确实很多人用 SEO 赚了钱，但是学会 SEO 压根不等于赚钱。 别说是不是能学到有用的 SEO 知识不说，就是很多 SEO 流量很大的网站，和赚钱也没扯上什么关系。一个培训机构这种逻辑都没有的话，可以不去。

问：嘉宾您好。我是做医院网站优化的。
目前我的网站有 7 个月了，外链有 5000 多条，请问这个基数的外链，如果现在

开始使用软件群发外链(论坛博客、分类信息群发等)，会不会有危险?会不会因为外链增加过猛而被 K 站?

答：作弊都有危险。
很多出问题的网站，如果是因为作弊出问题的话，那很难怎么做好的。
SEO 可以不作弊也能拿到很多流量的。

问：我的一个站半年了，收录外链什么的都很好，可 IP 就很少，请问网站流量和哪些因素有直接或者间接的关系呢?

答：和 SEO 流量最相关的 3 个因素是 收录，排名，点击
一般的网站都存在很多问题。
很多网站收录其实不好的，因为 site 出来的数据不是真正的收录。一个网站网页的数量比大部分人预计的多。

问：国平一直是业内难得一见的 SEO 大师，看过很多他的文章。感觉文章都从技术的角度实验。今天问个，就是针对 google panda 对国内 SEOER 有些什么好的建议

答：panda 算法是迟早会来的，google 不可能坐等被伪内容垃圾站毁了。
做 google 的 SEO 可以这么说：最好的黑帽方法就是做白帽。
只要你和 google 的方向是一致的，你会经常得到很多实惠。这次 panda 算法后，有些大型网站涨了一倍流量，绝对值在几十万 UV，这就是做白帽 SEO 的好处，而黑帽如果好不容易搞几十万 UV，每天都在担惊受怕。

问：我比较赞成科学的 SEO 手段，因为这中方法很务实，切合实际
在数据分析、观察、技术的使用都有哪些技巧呢?
我们大家都知道数据的观察和分析但是大家的方法都不是很合理，嘉宾可以介绍介绍吗?

答：首先，影响流量的三大因素要数据化：整体收录量数据化、整体排名数据化、整体点击率数据化。
一般人只把收录数据化了，而且还是不准确的数据，所以需要拿到靠谱的数据。
还有，其实整体排名和整体点击这两项都可以数据化的，这就需要自己具备数据挖掘的意识和方法。
所有影响 SEO 流量的因素都不会超过这三大块。
接下来就是把这 3 大块再细分，比如收录量是由什么引起的，也可以知道很多因素在影响收录量，如知道有 20 多个因素在影响收录量，那就把这些因素数据化。
同理把整体排名各个因素数据化，整体点击率各个因素数据化。
然后长期观察这几十个数据的变化是如何互相影响、以及如何影响 SEO 流量的。
再加以经验的积累，就很容易做到可控了。
我博客里有一些文章提到了这个。强烈建议看一下，如：
网页加载速度是如何影响 SEO 效果的 http://www.semyj.com/archives/969
怎样形成一套非常科学系统的 SEO 方法 http://www.semyj.com/archives/1032
不过，其实博客里写的还只是入门，还能继续细化下去。

很多人总问我的 SEO 培训内容是什么，其实最后一个回答可以总结我的 SEO 培训的一半内容，就是用数据和案例详细阐述各个因素之间是如何互相影响的，还有一半内容是很多基于这些知识总结出来的一些 SEO 方法。

另外最近发现一本书，虽然是在讲管理的，但是和我的 SEO 方法和策略惊人的相似，很多东西真的是一通百通。我现在是个创业者，在管理这块要从头学起。没想到能看到这么一本好书，既能让我看到我以前的探索 SEO 方法的影子，又能一下子解决目前的管理瓶颈。推荐大家也阅读：

http://www.gnbase.com/thread-3547-1.html

SEO工具

利用Google Search Appliance 服务器做SEO

昨天，Stephen 留在中国的 Google Search Appliance 服务器到了。这次 Google Search Appliance（简称 GSA）和去年用的那个 google mini 不一样，这个 GSA 基本上可以看做是 google mini 的升级版本吧。

Google 推出 GSA 的目的是让那些信息量暴增的企业和机构能用它建立自己的搜索引擎。它支持的格式有 220 多种，你可以用它来抓取和收录自己的博客、网站、数据库和网络文件夹等等。它是机器和软件全部打包在一起的。

Google官方网站介绍：http://www.google.com/enterprise/search/gsa.html

Google Search Appliance 正面

主要特点有以下这些，随意看看就好：

- 有一个连接器管理工具，可以让你收录和那些非WEB格式的文档。如：word,pdf,flash 等等
- 如果使用 Feed API 和元数据搜索功能，可以为自己的论坛建立搜索功能。
- 还提供了强大的 Onebox 编程接口，可以让你在搜索结果中展现一些定制的信息。
- 强大的安全搜索功能支持多种身份认证方式，使用户在搜索结果中只见到自己有权限访问的文档。
- 为小规模文档设计了专门的网页排序算法。
- 用户可以定制搜索结果界面，甚至以 XML 格式的形式，来整合到您自己的应用中去。

这个对SEO也是非常有用的。为什么这么说呢？

你可以把这个GSA看做是google 的微缩版，它有爬虫，有索引库，有排序算法。它的的硬件和软件都是现在google.com这个网站正在用的东西。所以两者之间相似程度非常的高。我在过去操作google mini的时候已经证实：至少它的抓取机制和现在的google.com几乎是一摸一样的。

它的排序算法，我觉得也会有很大的相似度。当然不会一摸一样的，因为现有的GSA好像是依据 06 年的搜索技术改变而来，而以google每年 450 次的算法调整频率，到现在也相差比较多了。但是至少和现在的排序算法会有相当大的相似度。

还有一些有意思的东西，大家看上面提到的两个特点：“Onebox 编程接口”和“在搜索结果中只见到自己有权限访问的文档”。这其实就是现在google的一些应用。

Onebox在搜索引擎现有的排序算法中享有优先级，它的数据来源就是Google Base或其他google产品。这次GSA也提供了这种Onebox的编程接口，现在自己亲手给自己的GSA添加Onebox，一定会对你如何利用Onebox拿到更多流量有帮助的。

“在搜索结果中只见到自己有权限访问的文档”，这个就类似igoogle 。还有google其他一些产品，在搜索结果页面，你登陆了gmail看到的和不登陆看到的是不一样的。

还有更多的细节，会在以后详细讲述，会把GSA后台的操作也讲述一下。到时候你会对google webmaster tool这个工具有更深一层的理解。

下面直接上图，机器顶部：

GSA 机箱上印有大大的 LOGO

机箱非常的重，可能有 40 多公斤。
为了保护里面的数据和硬件，机箱要用专用的螺丝刀才能打开。Google 在服务器硬件上有很多专利的。

GSA 正面

Google Search Appliance 背面 1

背面和一般的服务器没太大差别，但是注意它有一进一出两个网线口。设置它的时候需要用另一台电脑辅助。

Google Search Appliance 背面 2

来一张 google 机柜里的图：

机柜

google mini（只有 GSA 一半大小）

利用 GSA 服务器做 SEO 测试，可以做出完美的 SEO 网页。

gsa 侧面

HTTrack 在SEO上的应用

这纯粹是一条个人喜好，我经常拿 HTTrack 模拟搜索引擎爬虫用。

HTTrack 是一个网站镜像工具，本来是用来抓取网站做离线浏览用的。但是我发现它的爬虫特性和搜索引擎爬虫非常的像，逐渐应用到了自己的 SEO 工作中。其实这两种看似不同的爬虫做的都是同样的工作，就是复制网站并存储下来（搜索引擎的网页快照就是被存储下来的内容）。以下是这个软件的界面：

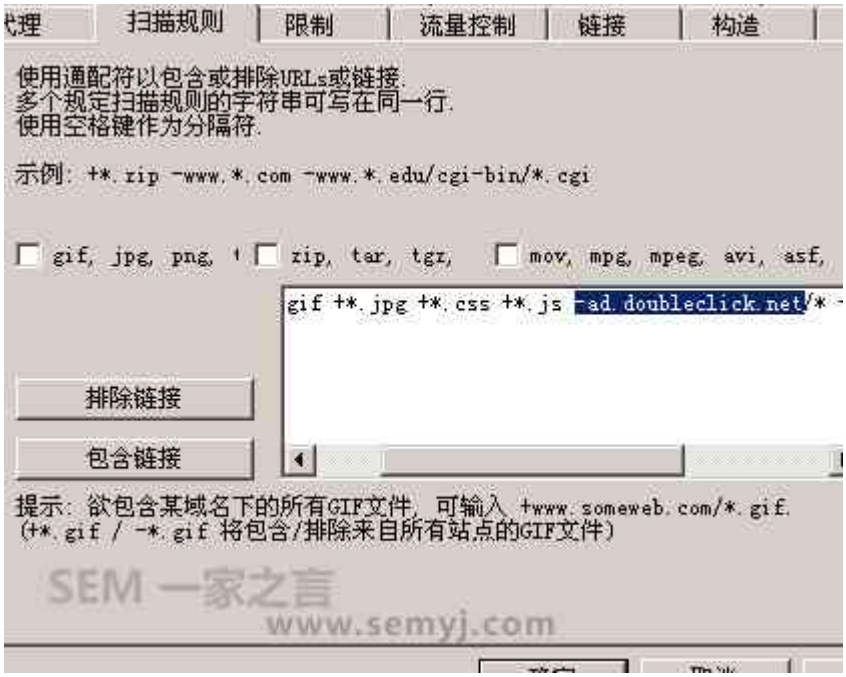
HTTrack 界面

软件的官方网站是：http://www.httrack.com/ 软件安装后可以换成中文界面。

一般用它来检测网站的坏链接和测试搜索引擎对这个网站可能面临的抓取问题。另外用它也可以探知一些 SEO 做法的由来。

软件的使用方法非常简单，在“Web 地址”里填上 URL 就可以了。然后点“选项”，

先看“扫描规则”



扫描规则

这样的扫描规则搜索引擎也一定会有的，比如不收录.exe 文件,zip 文件等等。然后不收录一些特定的跟踪链接， 如 ad.doubleclick.net 。你需要把一些搜索引擎爬虫不收录的特征加进去。

然后在“搜寻”里面，很多的特征都是现在搜索引擎爬虫的特征：

搜寻

搜索引擎不会接受 cookie,所以取消“接收 cookie”。

至于“解析 java 文件”，google 爬虫也会去解析 java 文件的。这是一个像 HTTrack 这样的通用爬虫都可以做到的事情。可能很多人还不知道，google 会去试图解析 javascript 代码。如果你的页面上放很多 javascript 代码，就会使爬虫的停留时间增加，进而影响爬虫效率。这也可以算是为什么要把 javascript 代码外调的另一个原因。

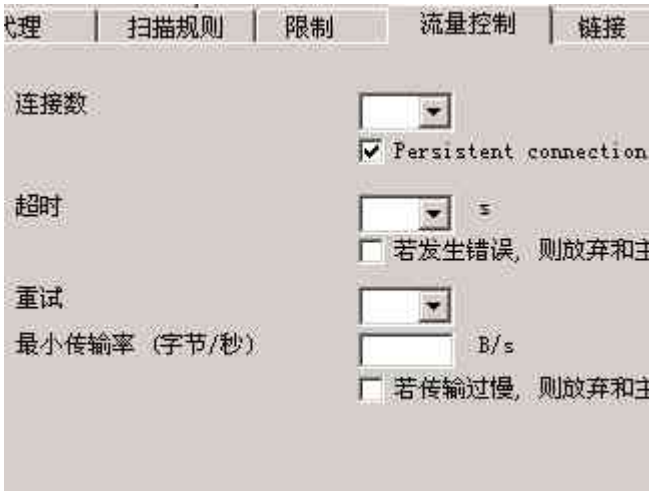
还有，有些 javascript 代码里面的 URL，google 爬虫是可以收录的，原因不明。这样做可能是因为有些内容很好的网站，很多链接就是喜欢用 javascript 来做的缘故吧。但是不代表你的链接可以用 javascript 来做。

HTTrack 也同样能识别并遵守 robots.txt 文件。

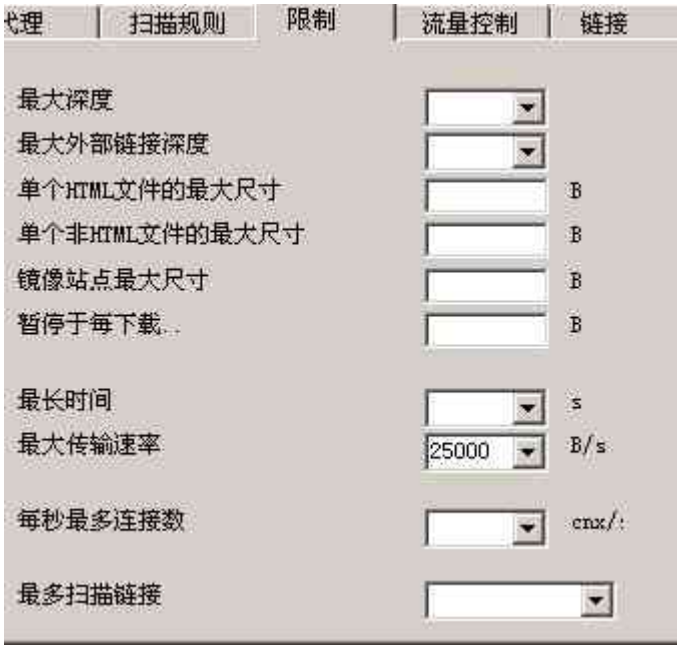
至于 url hacks ，就是让那种带 www 和不带 www 的网址，如 www.***.com 和 ***.com。以及有斜杠和无斜杠的网址，如 http://www.***.com 和 www.***.com 能统一。

这种网站上 URL 不统一的状况爬虫程序其实能很简单的处理好。至于 google 为什么要网站所有者在 webmaster tool 后台指定一下“首选域”，是因为有些网站 www.***.com 和***.com 指向不同的内容。所以 google 不能那么武断的就认为 www.***.com 和***.com 是同一个网站。

至于“流量控制”和“限制”，



流量控制



限制

里面可以设置“连接数”和“深度”什么的。我相信 google 也有这些设置，不然，google 的《网站质量指南》里不会这么写“如果站点地图上的链接超过 100 个，则需要将站点地图拆分为多个网页。”
至于深度，有报告说，google 抓取的最大深度是 12。超时时间可以设为 10 秒。

还有其他“浏览器标识”和“预存区”也和搜索引擎爬虫一样的。

其他设置

下面用它来抓取一个网站，看看会有什么样的情况。

首先爬虫会去网站根目录下访问 robots.txt 文件，如果碰到该网站的二级域名，还会去二级域名下访问 robots.txt 文件。这个和搜索引擎是一样的。

在抓取的时候，是多线程的，你可以实时的看到哪些 URL 正在被抓取以及速度怎么样。

很多人用它抓取完一个网站后会惊讶的发现有很多没什么 SEO 价值的页面在被抓取。而这些“垃圾链接”竟然还是最先被抓取到的。可惜这个爬虫不支持 nofollow 属性，不然更加能模拟 google 爬虫。你还会用它发现很多死链接和超时的页面。

要是经常使用，你还会发现这个软件的一个规律，就是在抓取那些动态 URL 的时候，经常会产生重复抓取的现象，抓取 URL 类似 www.***.com/index.asp?=12345 这样页面会陷入到死循环当中。这个和早期的 google 爬虫又是一样的。由此判断，这应该是爬虫天生的一个弱点，可能它没办法实时的比较多个页面的内容，如果加上网页程序在处理 URL ID 的上遇到什么问题，就会重复抓取。也由此得出为什么要有 URL 静态化了。 URL 的静态化与其叫静态化不如叫唯一化，其实只要给网页内容一个唯一的、结构不容易陷入死循环的 URL 即可，这就是静态化的本质。

google 最新的声明不要静态化，是不希望爬虫从一种重复抓取陷入到另一种重复抓取才这样说的。其实 google 举例的那几种不好的静态化一般是不会发生的。只要你明白那些 URL 中的参数代表什么，还有不要把很多个参数直接 rewrite 到静态化的 URL 里即可。

用这个软件，能让你直观的感受一个爬虫是怎么工作的。对于让一个新手正确认识爬虫有帮助。

这个软件的功能也差不多就这么多，要逼真的模拟搜索引擎爬虫，就要用《google 网站质量指南》里提到的 Lynx。但是 Lynx 是一个页面一个页面检查的。以后会写一篇应用 Lynx 的文章。

更好的模拟google爬虫就要用GSA了。不应该说是模拟，而应该说它就是google爬虫。

用 HTTrack、Lynx 和 GSA，再配合服务器 LOG 日志里面的爬虫分析，会让你对爬虫的了解到达一个更高的水平。分析爬虫会让你得益很多的。很多都以后再讲。

Lynx浏览器在**SEO**上的应用

曾经有朋友问我怎么才能判断一个 SEOer 是不是高手。我就出了一个主意，就建议他问那个 SEOer 是不是知道 Lynx 在 SEO 上的应用方法。这么来提问，其实能从一个侧面反映这个 SEOer 对 SEO 研究有多深的。

现在 SEO 行业，虽然有很多以讹传讹的言论，但是如果自己经常实践，还是能找到很多真正有用的操作方法。实践久了，也能判断谁的说法正确，谁的说法有问题，这样的 SEOER,可以放心的让他去操作一些比较重要的网站了。再进一步的给网站各个细节优化过程中，就会发现很多以前别人没有谈到过，也很难在优化一些小网站的过程中注意到的细节。这些细节，在别的地方很难找到相关的参考资料，或者根本就找不到。但是在 google 的《google 网站质量指南》、《google 黑板报》、《google 中文网站管理员博客》，基本上都可以找到关于这些细节的只言片语的。只不过那里面也只是给出了一个方向，更具体的细节还是要靠你自己再去实践。

在《google网站质量指南》 的第一页 ，就已经建议大家去用Lynx这个工具区检测你的网站：

使用诸如 Lynx 的文本浏览器来检查您的网站，因为大多数搜索引擎信息采集软件查看您网站的方式与 Lynx 几乎一样。如果诸如 Javascript、Cookie、会话 ID、框架、DHTML 或 Flash 等复杂功能造成您无法在文本浏览器中看到整个网站，则搜索引擎信息采集软件在抓取您的网站时可能会遇到问题。

这里提到了“Lynx 查看网站的方式和搜索引擎几乎一样的”。一个 SEOer，如果真的到了很多细节都无法从别人那里获取参考的程度，那这段话相信他很难忽视掉的。

我用了一段时间的 Lynx，发现这个曾经的文本浏览器和搜索引擎爬虫很像的。你所听过的爬虫特性，在这里面都能找到一点影子。

比如检测隐藏链接，我们只知道搜索引擎是不喜欢的，但是具体的检测方法是怎么样的呢？ 如果你用熟了 Lynx，就发现一个非常简单的命令就搞定了。

首先要搭建一个 Lynx 的运行环境。Lynx 不能用那种编译过的在 windows 下运行的版本，有很多功能是不能用的。建议在 XP 下装一个虚拟机，然后在虚拟机里装一个 linux 系统来运行 Lynx。

虚拟机软件用 VirtualBox 或者 VMWare，具体的安装方法大家 google 之。Linux 系统推荐用 Ubuntu，它可以在图形界面上安装 lynx。

在装了 Lynx 的 Linux 系统的命令模式下输入:lynx -dump www.alibaba.com 并回车,这个页面上的隐藏链接就一览无余了。如：

检测出了隐藏链接

然后再进一步的分析一下，是哪些链接 Lynx 会认为是隐藏链接呢？

可以看到，至少以下的一种链接是会被 Lynx 认为是隐藏链接的。代码为：

这个链接，即没有文字作为锚文本，也没有图片或其他作为链接的对象。如果不去加载 CSS 文件或 JS 文件，光就这个代码，在网页上是看不到这个链接的存在的。当然这就是隐藏链接，毫无争议的。

这是 Lynx 认为的情况，搜索引擎也是一样的。从整个互联网来看，这种检测方法在 99%的情况下都不会冤枉一个网站的。对于 google 来说，一个检测方法，如果能有 40%以上的反作弊效率，那是非常好的一个方法。

一个非作弊的网站，产生这种情况的原因，是因为网页设计人员的一些“奇怪”的代码写法。如果你去检测你的网站，说不定也能看到这些隐藏链接。

当然，Lynx 的作用不止这个。它首先是能以一个可视化的角度来展现爬虫看到了什么内容。用它可以挨个检查你的网页给搜索引擎爬虫展现了怎么样的内容。如：

Lynx 看到的内容

然后才是其他的一些功能：

- 可以检测网页代码的完整性。如果提示有“Bad HTML”就要注意一下。
- 可以和 IE 一样查看源文件。 命令在附录中。
- 对 cookie 的跟踪是特别对待的。会提示你是不是跟踪 cookie。
- 对框架和表单的处理和爬虫是一样的。
- URL 太多参数，会造成浏览困难。
- 可以查看网页返回的 http 头信息

……

你会看到很多似曾相识的东西。

Lynx 的出现时期，恰好是第一个爬虫程序诞生的时候。有相当大的理由相信他们的是一样的理念。而且现在维护和更新 Lynx 的人员，有些也在维护其他开源的爬虫程序。你其实也可以把 Lynx 看成一个可视化的爬虫。

HTTrack 是一个比较宏观的爬虫模拟器。而Lynx就更细节一些，也更实用一点。

附录 Lynx 的简要使用说明：

移动命令：
下方向键：页面上的下一个链接(用高亮度显示)。
上方向键：页面上的前一个链接(用高亮度显示)。
回车和右方向键：
跳转到链接指向的地址。
左方向键：回到上一个页面。

滚动命令：
+,Page-Down,Space,Ctrl+f：
向下翻页。
-,Page-Up,b,Ctrl+b：
向上翻页。
Ctrl+a： 移动到当前页的最前面。
Ctrl+e： 移动到当前页的最后面。
Ctrl+n： 向下翻两行。
Ctrl+p： 往回翻两行。
)： 向下翻半页。
(： 往回翻半页。
#： 回到当前页的 Toolbar 或 Banner。

文件操作命令：

c： 建立一个新文件。
d： 下载选中的文件。
E： 编辑选中的文件。
f： 为当前文件显示一个选项菜单。
m： 修改选中文件的名字或位置。
r： 删除选中的文件。
t： Tag highlighted file。
u： 上载一个文件到当前目录。

其他命令：

?,h： 帮助。
a： 把当前链接加入到一个书签文件里。
c： 向页面的拥有者发送意见或建议。
d： 下载当前链接。
e： 编辑当前文件。
g： 跳转到一个用户指定的 URL 或文件。
G： 编辑当前页的 URL，并跳转到这个 URL。
i： 显示文档索引。

j： 执行预先定义的“短”命令。
k： 显示键盘命令列表。
l： 列出当前页上所有链接的地址。
m： 回到首页。
o： 设置选项。
p： 把当前页输出到文件，e-mail，打印机或其他地方。
q： 退出。
/： 在当前页内查找字符串。
s： 在外部搜索输入的字符串。
n： 搜索下一个。
v： 查看一个书签文件。
V： 跳转到访问过的地址。
x： 不使用缓存。
z： 停止当前传输。
[backspace]：
跳转到历史页(同 V 命令)。
=： 显示当前页的信息。
： 查看当前页的源代码。
!： 回到 shell 提示符下。
_： 清除当前任务的所有授权信息。
*： 图形链接模式的切换开关。
@： 8 位传输模式或 CJK 模式的切换开关。
[： pseudo_inlines 模式的切换开关。
]： 为当前页或当前链接发送一个“HEAD”请求。
Ctrl+r： 重新装如当前页并且刷新屏幕。
Ctrl+w： 刷新屏幕。
Ctrl+u： 删除输入的行。
Ctrl+g： 取消输入或者传送。
Ctrl+t： 跟踪模式的切换开关。
;： 看 Lynx 对当前任务的跟踪记录。
Ctrl+k： 调用 Cookie Jar 页。
数字键： 到后面的第 n 个链接。

google 的良苦用心：网站管理员工具

2005 年的 google 做了大量的调整，因为到了 05 年，很多 SEO 的方法慢慢泛滥了起来。同时很多网站主对 google 如何对待他们的网站一直没有明确的途径去了解。google 应对这个局面的方法非常的开放，也非常聪明，就是希望和网站主达成一种双赢的局面。所以有了 google webmaster tools （网站管理员工具）这个工具。

这个工具从推出到现在，经历了很多次的增增减减，它努力追求让这个工具越来越对站长有利。一直以来，我都看到很多人对它的认识还不够深刻，所以单独来讲一讲这个工具是很有必要的。

《利用Google Search Appliance 服务器做SEO》一文中，我曾经说：“会把GSA后台的操作也讲述一下。到时候你会对google webmaster tool这个工具有更深一层的理解。” GSA就是一台把google整个硬件和软件打包在一起的服务器。这台服务器就是一个小型的google搜索引擎，它以前的版本的名字就叫 google mini ， 能形象的说明这个服务器的性质。

黄色的是 GSA，蓝色的是 google mini

现在我就把 GSA 后台的截图发出来，大家一定能发现点什么。

GSA 后台

这个后台对很多人来说一定有似曾相识的感觉，因为在 google webmastel tools 里，不光界面和这个相似，里面的很多功能其实都已经有了。

google webmastel tools 的前身是 google sitemaps，以前主要的用途是让网站主解决爬虫的抓取故障和提交 sitemap。这两大功能其实只解决了 google 爬虫抓取的局限性，这主要只解决了 google 自己的问题。而那时 SEO 越来越流行，很多网站甚至用作弊的方法来做 SEO。大家这么忙活，无非是想从 google 上面多拉一点流量，这个是广大网站主需要解决的问题。

本来，SEO 看起来和搜索引擎是矛盾的。百度对 SEO 的认识就是这样，所以它仇视 SEO，把自己和很多做 SEO 的网站主搞得处于对立的局面。

但是 google 不这么认为的。因为搜索引擎需要大量的网站来供应内容，它的期望是内容主次分明，越优质越好。而网站主希望能从搜索引擎获取流量，期望值是流量不光越多越好，还要越匹配越好的。那两者之间其实可以达成双赢的局面。

我做了很多年 SEO，虽然从 google 获得了大量的流量。但是也越来越发现我是在给 google 打工的。因为我把一个网站的结构理顺了，把重要的内容突出了，google 就知道了我网站有些什么内容，也知道了这些内容中的重点。这样，至少在判断我这个网站讲了什么内容的时候，google 是很有把握的。而当很多网站都这么做的时候，google 的内容质量整体就上升了一个等级。用户从 google 搜索到的内容更符合他们的需求了。同时，网站主凭借着主次分明的内容拿到的流量也是匹配网站主需求的优质的流量。

google 从一开始就会这样说：“好吧，网站主，既然你想得到你想要的流量，那你提供相关的内容给我。你如果不知道什么是相关的内容，那么我来告诉你，还告诉你怎么来突出重点。”

所以《google 网站质量指南》里的几百篇文章，以及 google webmastel tools 都是来告诉你要如何提供什么内容给搜索引擎。在我看来，google webmastel tools 是 google 提供的最好的 SEO 工具，里面的每一个功能都是和 SEO 相关的，google 在里面告诉你了要如何做 SEO。

GSA 的硬件和 google 现在用的服务器是一样的，包括传说中的自带电源和从没向外界说过的几公斤重的散热片。

google 的专利-自带电源

这个 GSA 的软件部分，后台应该是 google.com 的老版本的一个子集。所以你可以简单的认为 google 的后台也是这样的。为了能尽量为网站主着想，google 陆陆续续的把后台的一些功能都放进了 google webmastel tools 里。对于 google 来说，只要不泄露自己的核心机密，很多的数据和工具，如果能对网站主做好 SEO 有帮助的话，就把它开放出来让大家使用。

我就不一一说明每个功能在 SEO 上的作用，因为这个里面的很多细节都繁琐得可以写成一篇文章。以下就讲几个最近在 google webmastel tools 增减的功能，看看 google 是出于什么目的来调整的。

1，google webmastel tools 里有个控制爬虫“抓取速度”的选项，以前只能控制三个速度，就是“更快”、“正常”、“更慢”。而在我以前操作的 google mini（GSA 的老版本）中，也有这样一个调节爬虫抓取速度的选项，但是是一个拉动的滑块，可以调节出非常精确的抓取速度来。某一天，当我验证完一个新站的时候，发现 google webmastel tools 也已经是这样的了。

调节抓取速度

这个对于很多网站来说是很有好处的，因为那些网站不怕你爬虫来得多了把服务器爬死，就怕你不经常来。

2，最近增加的“像 Googlebot 一样抓取”的功能，在 GSA 的后台也是有的，只是不是这种表现形式。为什么要加一个这样的功能呢？这是因为 google 在抓取很多网站的时候碰到的一些问题越来越多才加这个功能的。

模拟 google 爬虫

同一个网页，搜索引擎爬虫看到的和你用浏览器看到的其实是不一样的。因为浏览器有很大的容错性，你少一个 ，或者 html 标签没有关闭，只要不错得离谱的话，网页看起来还是很正常的。 但是搜索引擎爬虫是不会这样的，你在 CMS 上或者代码上有什么问题，会让爬虫看不到你的页面。所以你需要模拟搜索引擎爬虫去访问你的网页，看看搜索引擎访问到的源代码是不是你真实的源代码。这个 google 模拟爬虫，还可以查看一个网页的返回码。这是因为一个网页，如果你给搜索引擎返回多个返回码，根据各种情况，搜索引擎要么不收录你的网页，要么你的用户从搜索引擎点进你的网站会出现 500 错误（服务器内部错误）。

我在《Lynx浏览器在SEO上的应用》一文中建议大家用Lynx去访问你的网页也有这样的原因， 那个工具也可以实现这样的功能。可惜的是尽管google在《网站质量指南》里力荐这个浏览器，但是还是很多人不会去用，所以google干脆把这个功能放在google webmastel tools里了。

3，我还是低估了 google 的良苦用心程度。当我前几天看到新推出的这个“参数处理”的功能的时候，几乎要感动得哭了。

参数调节功能

大家可能对URL静态化是有一些疑虑的。因为很久以前google说URL要静态化，而google年初的时候又说不要静态化了。为什么会有这么截然不同的说法呢？ 其实URL静不静态化根本不是问题的核心，核心问题是出在URL的参数上。如果有人仔细去看《HTTrack 在SEO上的应用》一文，并不断地去使用这个工具的话，就会发现：是因为URL上的参数复杂，才导致了爬虫陷入死循环的。现在即使你把URL静态化，如果没有处理好的话，和没有去静态化是一样的效果。关于这点以后还要写一篇文章才能说得清楚。

google 当然清楚是由参数引起的，所以在以前，它都是有一套自己的方法过滤参数的。但是，这个过滤方法并不一定很准，可能你觉得不是参数的重要的页面，google 把你过滤了，那就不会收录了。所以 google 就干脆让你自己来调节，先自动过滤一些参数，然后让你看看哪些参数过滤错了，或者还有哪些参数没有过滤，就由你来告诉 google。

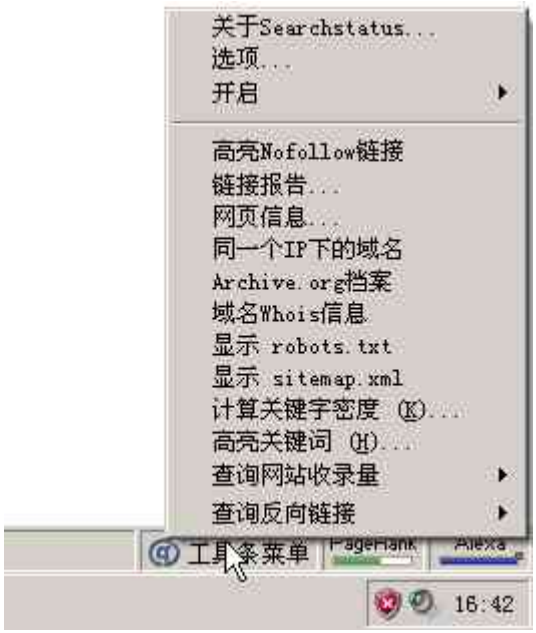
SEO工具条-Searchstatus汉化增强版

Searchstatus 是一个 Firefox 上的 SEO 插件，是一个非常好用的 SEO 辅助工具。不过因为它的官方版本只有英文版，所以普及程度还不高。我最近花时间把这个插件汉化了一下并增减了上面的一些功能，现在提供给大家使用。

点击下面的图标就可以安装。或者把文件下载下来后，把这个文件拖到 Firefox 的窗口上也可以安装。

文件URL：http://www.semyj.com/upload/searchstatus-1.33-zh-cn.xpi
在 Firefox 3.0/3.5/3.6 下测试通过。 Firefox 3.6 下载地址：
http://download.mozilla.org/?product=firefox-3.6.12&os=win&lang=zh-CN

安装后，工具条默认显示在右下角的浏览器状态栏里。



软件菜单

鼠标左键点击“工具栏菜单”，是更新所有的 Rank 值。鼠标右键点击“工具栏菜单”，是弹出工具栏菜单。

在“选项”里可以设置这个工具条的位置，有四大位置可以调。

选择位置

这四大位置分别对应浏览器的如下位置：

工具条位置

在这些显示 Rank 值的按钮上，鼠标左击是更新 Rank 值。特别注意 PR 值有时候因为 google 服务器的问题是不会显示的，所以要多点几下。每个按钮，鼠标右击都能弹出相关的功能菜单。如：

每个按钮都有一个菜单

compete 是一个类似 alexa 的统计信息，但它是统计用户在一个网站上的停留时间的百分比。

mozRank 是 www.seomoz.org 自己推出的一个类似PR值的统计数据。

建议长期打开着“高亮Nofollow链接”，当然要先看一下《我也谈一下nofollow》这篇文章。

“计算关键词密度”这个功能里，只有关键词的显示次数是准确的，总字段和百分比由于是用英语的习惯按照空格分词的，所以不准确。

“查询网站收录量”里，我在原版的基础上增加了查询百度的收录量这个功能。百度的反向链接查询功能已经失效，所以没加。

另外注意一点的是:每次查询一个网页的信息,最好先刷新一下那个页面再来查。

最后，用它去欣赏一下 google 的 robots.txt 文件和 sitemap.xml 文件吧。


```
-<urlset>
  -<url>
     <loc>http://www.google.com/</loc>
     <priority>1.000</priority>
   </url>
  -<url>
     <loc>http://www.google.com/3dwh_dmca.html</loc>
     <priority>0.5000</priority>
   </url>
  -<url>
     <loc>http://www.google.com/a</loc>
     <priority>0.5000</priority>
   </url>
  -<url>
     <loc>http://www.google.com/a/cpanel/domain</loc>
     <priority>0.5000</priority>
   </url>
```

google 自己的 sitemap.xml 文件

如果有什么 Bug，请反馈给我。

2010.11.21 日更新：
1，更新了该版本，能够兼容 Firefox3.6 版本；
2，增加了 lynx 查看网页的功能；
3，增加了到光年论坛的链接。

Lynx 在线版以及浏览器插件

最近还是太忙，所以关于内外部链接的文章还没开始写。现在给大家一个 Lynx 在线版以及相关的浏览器插件。

我在《Lynx 浏览器在 SEO 上的应用》一文中介绍过这个工具。不过有些人说在 linux 下没有安装好或者有乱码的出现。后来有人给我看了国外的一个 Lynx 在线版，但是那个在线版也存在着一些问题。我的这个 Lynx 在线版把那些问题都解决了，使用起来还不错。

如果正在仔细阅读《google网站质量指南》的朋友应该能注意到，在这个《google网站质量指南》里，至少十几篇文章中都出现了要你去用Lynx检测网站的提示。而且是一到具体的做法的时候，都说：请用Lynx去检测你的网站，因为它和爬虫看到的内容几乎一样。这个工具在 05 年就有一些一线的SEOer在用了。

使用方法非常简单，你只要填入你要查看的 URL，点击“查看”就可以了。网址要以 http://开头。

URL:

http:// 查看

如，我输入 http://www.baidu.com/ ，查看到的界面如下：

[1]登录

[2][USEMAP:baidu_logo.gif]

[3]新 闻网 页[4]贴 吧[5]知 道[6]MP3[7]图 片[8]视 频

_____ 百度一下[9]设置
[10]高级

[11]空间 [12]hao123 | [13]更多>>

[14]把百度设为主页

[15]加入百度推广 | [16]搜索风云榜 | [17]关于百度 | [18]About Baidu

(c)2009 Baidu [19]使用百度前必读 [20]京ICP证030173号 [gs.gif]

References

1. http://passport.baidu.com/?login&tpl=mn
2. LYNXIMGMAP:http://www.baidu.com/#mp
3. http://news.baidu.com/
4. http://tieba.baidu.com/
5. http://zhidao.baidu.com/
6. http://mp3.baidu.com/
7. http://image.baidu.com/
8. http://video.baidu.com/
9. http://www.baidu.com/gaoji/preferences.html
10. http://www.baidu.com/gaoji/advanced.html
11. http://hi.baidu.com/
12. http://www.hao123.com/
13. http://www.baidu.com/more/
14. http://utility.baidu.com/traf/click.php?id=215&url=http://www.bai
15. http://e.baidu.com/

以 lynx 查看百度

输出的结果分为两部分：

第一部分，就是搜索引擎爬虫看到的内容，这个内容和别的查看方式都不一样。不仅显示了文字信息，还显示了网页的结构信息。去了解搜索引擎的原理就会知道，这种结构信息也是搜索引擎会储存下来的。 并且在分析你的网站讲了什么信息的时候，这些结构信息就是判断的依据。每个锚文本旁边还标上了这个链接的序号。

第二部分就是网站中所有爬虫能够识别的链接。有些网页这里会显示隐藏的链接。经常有人问我这个隐藏链接要不要紧，我这里统一回答一下：其实不是太重要，当你网站的 SEO 优化是正规的方法的话，可以忽视掉这个；但是当你用了很多黑帽的方法，这个隐藏链接就是让你“罪加一等”的地方。所以在 alibaba 的首页虽然也检测出几个隐藏链接，但是都没改过来。

刚接触这个工具，可能很多人不觉得这个工具有什么用的。建议大家用这个工具前，先看完以下几篇文章：

《分词与索引库》

《Lynx浏览器在SEO上的应用》

《把Web标准化进行得更彻底一点》

《“丰富网页摘要”，让你的网站与众不同。》

《“锚文本”在SEO方面的重要性》

这个工具支持绝大部分编码，日文、韩文、俄文等等都没问题的。

还有两个浏览器插件，一个是给 Firefox 的，一个是给 IE 的。

1，下载Lynx 在线版 for IE　　IE卸载文件

2，下载Lynx 在线版 for Firefox

装上了插件后，在你浏览一个网页的时候，在网页上点击右键的弹出菜单里，会有“以 Lynx 方式查看”的选项。这样非常方便平常大家查看网页。

IE 右键菜单

Firefox 右键菜单

浏览器右键菜单

由于这个工具放在国外的虚拟主机上，可能速度有点慢的。我还不知道有多少人会用这个工具，到时候可能有短暂的时间会使用不了。

不过大家可以先慢慢用着，以后还有讲述如何更好的应用这个工具的文章。

为了更好的推广这个工具，大家可以在自己的网站上，加上这个工具。

代码为：

<form action=”http://lynx.semyj.com/lynxview.php” enctype=” application/x-www-form-urlencoded” method=”get” target=”_blank”>

URL:

<input id=”url” style=”width:300px;” name=”url” type=”text” value=” http://” /><input type=”submit” value=”查看” />

</form>

另外 Lynx 的发音为：[liŋks] 点此听发音：http://www.103.net/dictzh/content/pronzh/000073867585.mp3

光年SEO日志分析系统

为了能让 SEO 的分析与决策更加的科学化，我们推出了这个《光年 SEO 日志分析系统》。

常用的统计系统如 Google Analytics 等是在网页中加载一段 JS 代码来统计数据的。而一旦用户的网页没有打开或者浏览器不能执行 JS 代码，那就没有统计到这个用户的数据。所以日志分析是一个网站数据分析中的必要补充。 而且有很多的数据用 JS 代码是不能统计到的。如：网站上出现的各种各样的错误，搜索引擎爬虫在网站上的行为等， 而这些对 SEO 的分析与决策都很重要。

软件操作视频的下载地址为：http://www.semyj.com/upload/gnanalyzer.rar

（2011.2.23 注：软件已经升级到 2.0 版本， 更详细的使用说明请访问：http://www.semyj.com/archives/1539 ）

《光年 SEO 日志分析系统》与其他的日志分析软件有什么不同呢？

1，这是第一个专门为 SEO 设计的日志分析软件。

以前的很多日志分析软件，都是顺带分析一下 SEO 方面的数据，而这个软件里面分析的每一个指标都是为 SEO 设计的。而且很多的分析维度，都是其他日志分析软件没有的。这能让你看到很多非常有用、但是以前获取不了的数据。

2，它能分析无限大的日志，而且速度很快。

很多的日志分析软件，在日志大于 2G 以后，都会越来越慢或者程序无响应。而这个软件能分析无限大的日志，并且每小时能分析完 40G 的日志。这对于那种需要分析几个月内的日志、以及要分析几十 G 的大型网站的日志都非常有帮助。

3，能自动判断日志格式。

现在很多的日志分析软件，对 Nginx 或者 CDN 日志都不支持，而且对日志记录的顺序都要格式要求。而这个软件就没有这么多的限制，它能从日志中自动检测到哪个是时间、哪个是 URL、哪个是 IP 地址等等。

4，软件容量小、操作简单、绿色免安装版。

这个软件不会动不动就几十 M，现在软件还不足 1M，可以用邮件附件非常方便发出去。 软件的操作也很简单，三个步骤就可以。 还有就是软件不需要安装，是绿色免安装版。

软件的缺点：

目前因为在解决软件的效率问题上花了很多时间，所以现在日志分析的维度还太少，以后会逐步增加很多功能。还有就是数据的准确性虽然还可以，但是还有很大的改进空间。

可以在论坛里讨论和反馈各种信息。 http://www.gnbase.com/thread-15-1.html

另外更新了 Lynx 在线版 以及 改进了 SEO status 汉化增强版。

下一步会推出《光年网站日志分析系统》，侧重分析网站的综合数据。

补充说明：

- 软件在 XP sp3 和 WIN7 旗舰版下测试通过。
- 目前还不支持 GZ 等压缩格式，需要解压缩以后才能分析。
- 目前发现的一个问题就是：当日志文件太小的时候，分析结束时会报错。

希望有其他问题的用户能加 QQ 11435092 详细描述问题。

SEO利器-Google GSA虚拟机版本

在所有的SEO工具中，能够被称为利器的工具不多，但Google GSA虚拟机版本绝对算是一个。去年我介绍了《利用Google Search Appliance 服务器做SEO 》，不过这个正式版实在太昂贵而且根据美国的某条法律不销售给中国，所以很多人都没办法用来做SEO应用。 而Google GSA虚拟机版本就很好的解决了这个问题。

GSA

先讲讲这个 Google GSA 虚拟机版本怎么应用到 SEO 上面吧。

如我以前所说：

你可以把这个 GSA 看做是 google 的微缩版，它有爬虫，有索引库，有排序算法。它的的硬件和软件都是现在 google.com 这个网站正在用的东西。所以两者之间相似程度非常的高。我在过去操作 google mini 的时候已经证实：至少它的抓取机制和现在的 google.com 几乎是一摸一样的。

其实何止爬虫抓取机制，连绝大部分排序的算法都是一样的。虽然这个 GSA 内置了更多给离线文档（如 pdf\word\）排序的算法，但是在给网页排序这块的算法和 google.com 如今正在用的算法是非常接近的。因为这个 GSA 的本意是给某些需要搜索的企业用户来索引他们自己的信息，是希望用 google 的技术能力来帮他们索引最相关的信息，不然就没有必要非得用 google 的产品了。 开发过小规模搜索引擎的人都知道，对于小型搜索引擎，其他东西大家都能基本做到，GSA 值钱的地方就是这个排序算法，这是大家选择 GSA 的首要原因。

另外，这个方法是一个有着 11 年 SEO 经验并且在美国 google 做过 2 年产品经理的人强烈推荐使用的方法，他自己就买了 2 台正式版。

不过排序算法总还是有差别的，根据我使用了 2 年多 GSA 的经验，对于网页的排序算法 90%以上是一样。

GSA 在 SEO 方面至少有以下几个应用。

第一个应用就是检查搜索引擎爬虫在你网站上可能遇到的问题。

因为这是一个真正的搜索引擎，而且对于 google 来说，GSA 和 google.com 的爬虫是一模一样的，所以检查到的问题都是真正的搜索引擎爬虫会遇到的问题。

操作方法为：

点击“抓取并编制索引” —> “抓取网址”，按如下格式输入你要检查的网址，按后点击“保存要抓取的网址”。

设置待抓取 URL

在“状态和报告”—>“ 抓取状态”里，点击“恢复抓取”。

等一段时间以后，如果一切正常，在 GSA 的前台就可以开始搜索到网站的内容。

在“状态和报告”—>“抓取状态”里，就可以看到爬虫遇到的一些问题。如：

检索错误

点击出错的部分，会列出哪些 URL 因为什么原因出错。

出错的 URL

这个虚拟机版本已经内置了一些数据。每次测试前，都需要把数据清空。 在“管理”—>“重置索引”里， 点击“立即重置索引”可以清空所有已经抓取的数据。

看着这些似曾相识的界面，应该能明白我以前写那篇《google 的良苦用心：网站管理员工具》的依据了。

第二个应用就是可以探测到 google 大部分基本的排序规则。

以前很多人在做 SEO 过程中对很多排名因素都是不能确定的。 如：到底 h1 放多少个好、有没有必要加导出链接等等。

有了 GSA，就可以做大量的 A/B 测试，就能把很多因素都测试出来。 如：测试〈h2〉 和〈strong〉这两个标签哪个对排名的影响更大一点。

那就可以设计 A 和 B 两个网页，其他部分都一模一样，只是某个特定的关键词，A 网页是用〈h2〉加粗的，B 网页是用〈strong〉加粗的。 用 GSA 只收录这两个网页，并且只对这两个网页进行排名。在前台搜索这个关键词，看哪个网页排在前面，这样就可以知道哪个因素对排名的影响大一些了。

有点需要注意一下的是，那些被测试的网页，如果上面有链接而你又没做限制的话，爬虫会顺着这些链接把很多网页都收录进来，那会对测试结果造成干扰。

A/B 测试

类似的测试方法还有很多，只要你想得到都可以去测试。 这样能把google宣称的 200 多项排序规则中的一大半规则都可以测试出来。 不过要明白一点的是：即使能把所有的规则测试出来，也不一定能做好SEO，在《怎样形成一套非常科学系统的SEO方法》中我说过：做搜索引擎是一回事，在搜索引擎上拉流量又是另一回事。 等大家把很多排序规则都测试出来了再来做SEO就明白了。

只是知道了这些规则，那就不需要听那些毫无来由的 SEO 规则了，很多事情你自己完全能确定是怎么回事。还有就是就算要向你老板交代你的 SEO 做法的时候也可以理直气壮一点。

另外，由于百度也在不停的“学习”google 的算法，所以这里的很多规则对百度也适用。（其实大部分搜索引擎的很多做法，甚至开发语言都是一模一样的。顺便广告一下： 杭州光年已经能开发搜索引擎及其很多应用。如小型搜索引擎、网站站内搜索、基于搜索的舆情监控系统、公司内部文档搜索等等。不是用开源程序开发。）

第三个应用就是可以查看内部链接的结构，看哪些网页被内部链接推荐得多一点。

在《内部链接还是外部链接？》一文中，讲述了内部链接的重要性。 但是极少有网站知道自己的每个网页内部链接的分布情况，有了 GSA，这个就很容易办到了。

在“状态和报告”—>“ 抓取状态”里，输入刚才 GSA 收录的网址，“网址状态”选 “已抓取”，就可以查询到已经被收录的网页的 PR 在站内有多高。

站内 PR 分布

这是在没有任何外部链接的情况下，网站纯依靠自身的内部链接造就的网站内部的 PR 值分布情况。 点击具体的 URL，还可以查看详细的信息如：

每个 URL 的信息

当然还有其他一些应用，如：只收录自己的网页和竞争对手的网页并进行排序，如果你自己的网页排在后面，就不停的改进直到超过对手的网页。其他更多的应用还是靠大家慢慢发掘吧，都写出来就没什么意思了。 用它确实是可以做出一个完美的 SEO 网页。

这个虚拟机版本是运行在Vmware上的，Vmware7.1.3 的下载地址是：http://download.pchome.net/system/sysenhance/redirectsrv-4673-1.html

初次使用虚拟机的同学最好装个 Vmware7.1.3 的汉化补丁。

GSA 虚拟机版本的下载地址放在光年论坛上：　（需要论坛会员才能看到下载地址）

http://www.gnbase.com/thread-13-1.html

Vmware 的安装过程略过，不过注意一下 Vmware 在安装过程中会安装几个虚拟网卡，如果电脑上的防火墙提示你的时候，一定要允许共享或通过。

要使用 GSA，google 官方建议的电脑配置为：

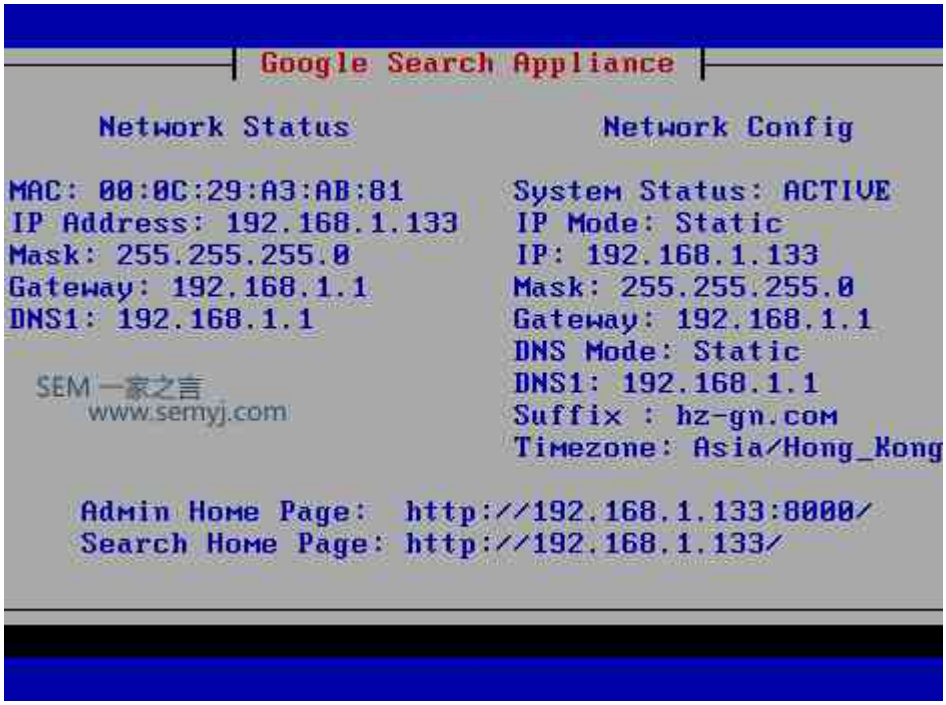
- Intel Pentium D 处理器 915（双核）或同级别的处理器
- 4 GB 内存
- 40 GB 可用硬盘空间，且硬盘转速为 7200 RPM 或更快
- SATA 或更佳存储接口

不过我在双核 1.8G 、2GB 内存的笔记本上运行也不是太慢。之所以建议用 40GB 的硬盘空间是因为这个虚拟机版解压缩以后的大小是近 35GB。

Vmware 安装好以后，直接导入解压缩以后的那个 vgsa.vmx ，然后打开虚拟机电源，接下来就是一段漫长的等待。

虚拟机导入

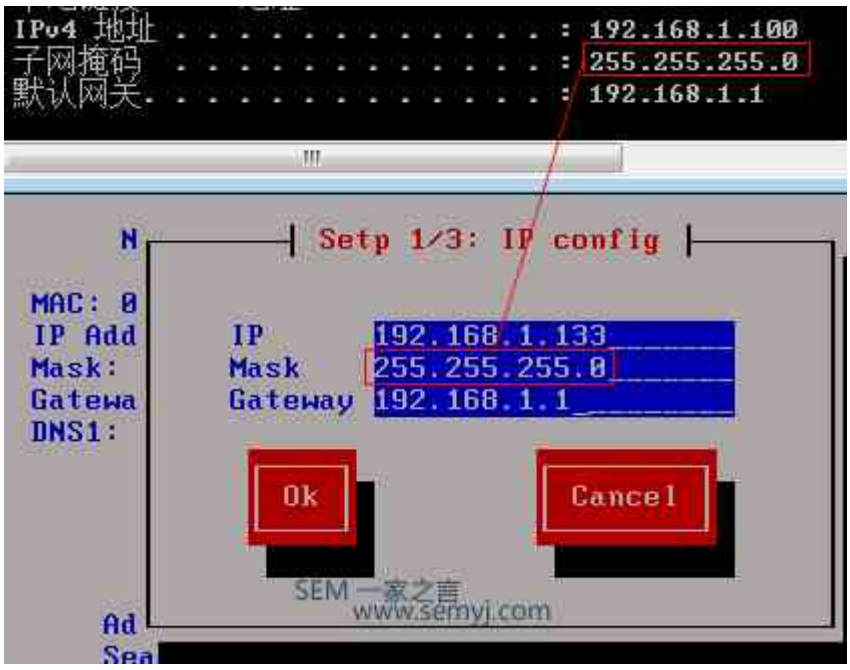
等出现配置界面的时候，就开始配置。



配置界面

大部分情况下，都需要手动配置 GSA 的设置，不然在后台抓取网页的时候会出现“DNS 错误”。　先用 ipconfig 命令查看你本机的 IP ，再来配置 GSA。

按 Ctrl＋G 进入虚拟机，再按 F2 ，按键盘上的 -> 选择 Manual ， IP 地址要填和你的电脑在同一个网段的其他 IP 地址。 其他和你本机的一样。 DNS 就填和 Gateway 一样的地址，DNS Suffix 可以随便填。



配置，按 TAB 键换行

即使配置好了，可能还不能马上使用，需要等待一会。 我记得 GSA 正式版从启动到能使用是需要等 20 分钟的，虚拟机版本也需要稍等一会。

GSA 的后台访问地址是： http://{刚才配置的 IP}:8000 ， 前台访问地址是：http://{刚才配置的 IP}。后台的登录密码压缩包里有。

软件的使用细节还有很多，GSA 的使用交流可以到光年论坛。 因为现在大家都可以用它来探测 google 的排序算法，我相信会有很多的成果能发布在论坛上的。

自从去年我发布那篇介绍 GSA 的博客 2 个月后，因为有 google 的朋友也看这个博客，我猜 google 可能意识到了某些问题，所以 google 中止了 GSA 虚拟机版的更新，我这个版本是最新的一个版本。 我放这个版本出来是希望 SEO 行业从此不要道听途说或人云亦云，SEO 是可以做得很科学的，SEO 要长久发展下去就必须走规范化和标准化的道路。

详解《光年SEO日志分析系统 2.0》

《光年 SEO 日志分析系统》刚才升级到了 2.0，有朋友在论坛里提问说不知道怎么用数据分析来指导业务，我就趁新版本发布的时候顺便说明一下各项功能。

《光年SEO日志分析系统》的介绍：http://www.semyj.com/archives/1309
2.0 版本的下载地址还是以前那个：http://www.semyj.com/upload/gnanalyzer.rar

第二版增加了更多的分析维度，还增加了日志拆分的功能。

下面先来看几个固定的分析维度，下面的数据是 semyj.com 我这个博客的日志分析数据。

首先是“概要分析”：

概要分析

	蜘蛛名	访问次数	总停留时间(小时)	总抓取里	占比 %
1	谷歌蜘蛛	223	153.290	16484	43.552
2	msnbot/	3	14.212	13342	35.251
3	BaiDu Spider	1868	209.959	2968	7.842
4	Sogou Spider	49	47.559	2131	5.630
5	雅虎蜘蛛	180	112.603	1677	4.431
6	Google Feedfetcher	303	51.981	666	1.760
7	Speedy Spider	112	6.053	220	0.581
8	Googlebot-Image	15	36.052	166	0.439
9	有道蜘蛛	3	3.230	43	0.114
10	Alexa Spider	11	0.331	41	0.108
11	DoCoMo Spider	10	0.941	33	0.087
12	soso	13	[illegible]	[illegible]	0.087
13	Gigabot Spider	6	[illegible]	[illegible]	0.055
14	Alexa crawler	3	0.336	20	0.053
15	YodaoBot Spider	1	0.000	4	0.011
	总计	2800	637.522	37849	100.000

这里有各个爬虫“访问次数”、“总停留时间”和“总抓取量”的统计。从上面这个数据可以看出，百度爬虫的抓取深度是不高的：访问 1868 次，抓取量是 2968，平均每次抓取 1.59 页。这其实是百度爬虫普遍的抓取特征，在绝大部分网站上都是这个规律。抓取深度不高的话，会造成很多层级很深的页面不会被抓取到；以及造成少数页面被反反复复在抓取，浪费了爬虫的时间。这样，很多网站想要

在百度上获得收录就成了问题，特别是大中型网站。我所接触的所有大中型网站，在刻意去优化之前，一年下来很多网站至少还有一半的网页没有被百度爬虫抓取到，部分网站甚至更严重。相比之下 Google 的抓取深度就好很多，总的抓取量也大一些。

这里面比较重要的数据是那个“总抓取量”，因为它影响网站的收录量，进而影响网站的SEO流量。在《网页加载速度是如何影响SEO效果的》一文中说明过抓取量和SEO流量的关系。这个“总抓取量”的数据是好还是坏，是要根据每个网站的实际情况来看的。就semyj.com这个网站来说，它现在有 53 篇文章，300 多个网页，而现在google每天有 16484 个抓取量，百度有 2968 个抓取量。如果光看这个数据，那看起来这 300 多个网页基本上在一天之内应该是能被抓取到的。但是很多大中型网站就不一样。

这里我先要说明一个有些人会混淆的问题。为什么我上面会刻意说明一下文章数量和网页数量呢，这是因为文章数量肯定是不等于网页数量的。不过有些人去查收录量的时候就忽视了这个常识。如某网站的文章量（或称单个资讯数量）是 30 万，去搜索引擎用 site 等语法去查询收录量是 29 万，就觉得自己的收录量差不多了，而实际可能差得很远。

因为单个页面都会派生出很多其他页面的。如果打开某一个文章页面，去数一下里面的 URL，除去那些模板上重复的，还是有那么一些 URL 是只有当前这个页面上才有的，也就是这个页面派生出来的。而一个 URL 对应一个页面，所以一个网站上拥有的页面数量是这个网站的信息量的好几倍，有时甚至是十几二十倍。

所以在看这个“总抓取量”之前，需要把自己网站内可能拥有的页面数量统计一遍。可以用lynx在线版把每一类型的页面上的URL都提取出来看一看。网页总的数量知道了，再和“总抓取量”做对比，就可以知道这个数据是好还是差了。我觉得基本上，google爬虫的抓取量要是网站页面数量的 2 倍以上，抓取量才算及格，baidu爬虫就需要更多了。因为实际上这个抓取量里面还有很多是重复抓取的；还有和上一天相比，每天的新增的页面抓取不是很多的。

这三个数据：“访问次数”、“总停留时间”和“总抓取量”，都是数字越高对网站越有利，所以需要想很多办法提高他们。大多数时候看他们绝对值没什么用处，而要看现在的和过去的比较值。如果你能每天去一直追踪这些数据的变化情况，就能发现很多因素是如何影响这些数据的。

以下其他数据也是如此：某个当前数据的值有时候不一定有意义的，但是长期跟踪这个数据的变化就能发现很多因素之间是如何互相影响的。

然后是“目录抓取”的数据：

目录抓取

	蜘蛛名	目录	爬取量
1	谷歌蜘蛛	/bbs/	15947
2		/archives/	370
3		/wp-content/	27
4		/tl/	16
5		/tl/archives/	16
6		/hl/	9
7		/page/	9
8		/uploads/	7
9		/cs/	6
10		/lv/	5
11		/lv/archives/	5

目录抓取统计

这个“目录”抓取的数据是对“总抓取量”的一个细分。一个网站当中，一定是有重点页面和非重点页面的，这个数据就可以让你看看哪一类型的页面被抓取的多，及时做一些调整。

还有就是可以去搜索引擎按 URL 特征查询一下各个目录下的页面的收录情况，再来和这个目录下的搜索引擎的抓取数据做一个对比，就可以发现更多的问题。对于 semyj.com 来说，看完这个数据就知道，可能那 300 多个网页在一天之内还是不能全部被抓取一遍的，因为原来大部分抓取都在 bbs 这个目录下。（有时候就是有很多这样意外的情况发生，bbs 这个目录早已经做了 301 跳转，没想到还有这么大的抓取量。——看数据永远能知道真相是什么。）

接着是“页面抓取”的数据：

页面抓取

	页面	总抓取里	蜘蛛	蜘蛛抓取里
1	/feed	639	BaiDu Spider	501
			Google Feedfetcher	121
			Gigabot Spider	7
			雅虎蜘蛛	6
			谷歌蜘蛛	2
			msnbot/	1
			Sogou Spider	1
2	/	162	BaiDu Spider	36
			Sogou Spider	31
			soso	19
			DoCoMo Spider	14
			Speedy Spider	14
			谷歌蜘蛛	14
			DoCoMo Spider	13

这个数据把一个网站中那些被重复抓取的页面统计了出来，并分别统计是哪些爬虫分别抓取了多少次。大家多分析几个网站就会明白，百度爬虫经常是过度抓取的常客。这个数据也验证了前面的数据：因为它平均每次抓取 1.59 页，也就是每次来抓取都停留在表层，但是又经常来抓，所以势必导致少部分页面是经常被百度抓取的。因为有重复抓取的存在，所以一个网站光看抓取量大不大是没什么用的，还要看有多少不重复的页面被抓取到了。还有就是要想办法解决这个问题。

在“蜘蛛 IP 排行”数据里，统计了每个爬虫 IP 的访问情况：

66.249.68.143 (谷歌蜘蛛)	查询IP	49	22
66.249.67.142 (谷歌蜘蛛)	查询IP	49	27
66.249.67.141 (谷歌蜘蛛)	查询IP	46	24
66.249.68.141 (谷歌蜘蛛)	查询IP	42	22
66.249.68.142 (谷歌蜘蛛)	查询IP		
66.249.67.143 (谷歌蜘蛛)	查询IP	38	22
222.186.24.59 (谷歌蜘蛛)	查询IP	28	4

IP 排行

如果分析过很多网站，就会发现爬虫对某一个站的访问，特定时间内的 IP 段都会集中在某一个 C 段。这是由搜索引擎的原理决定的，感兴趣的朋友可以查询相关书籍。知道这个特征有时候可以用得着。

报表里有个查询 IP 地址的功能，可以查询那些爬虫 IP 是不是真的，如上图红框内的 IP，就是一个伪装成 google 爬虫的采集者。

这个数据和上面的所有数据都一样，前后对比就可以发现更多的信息。

以下是“关键字分析”的数据：

关键字分析

百度 (464次)关键字下载

	关键字	类型	页数	访问里	占比%	上次用关键字
1	SEM		[1 (168)]	168		[SEO (11)] [SEM (4)] [个人二手车 (2)] [SITE:WWW.77BDF.COM (2)] [网络营销 (2)] [搜索引擎 (1)] [包包 (1)] [b-PHP (1)] [ADMIN5 (1)] [信鸽商务群发软件 (1)] [青芒果旅行网 (1)] [豆丁 (1)] [SEO分享 (1)] [互联网产品研讨会 (1)] [SE, (1)] [搜索引擎 营销 (1)] [SEM PPC SEO (1)] [MKT (1)] [百度网址构建器 (1)] [AEM (1)] [网站联盟 (1)] [MADE IN CHINA. (1)] [蒲地蓝消炎胶囊 (1)] [北京西站附近有什么吃的吗 (1)] [搜狗推广 (1)] [安徽经济管理学院 (1)] [不想上班

关键词分析

“类型”这里是说明这个关键词是从网页搜索还是图片搜索或视频搜索里来的 SEO 流量。而“上次用关键字”，是统计用户搜索当前的关键词进入网站之前，

是在搜索什么词语。这个功能只有百度有效，因为百度在 url 中记录了用户上次使用的关键词。 这个地方的界面还需要修改，下一版本中会完善。

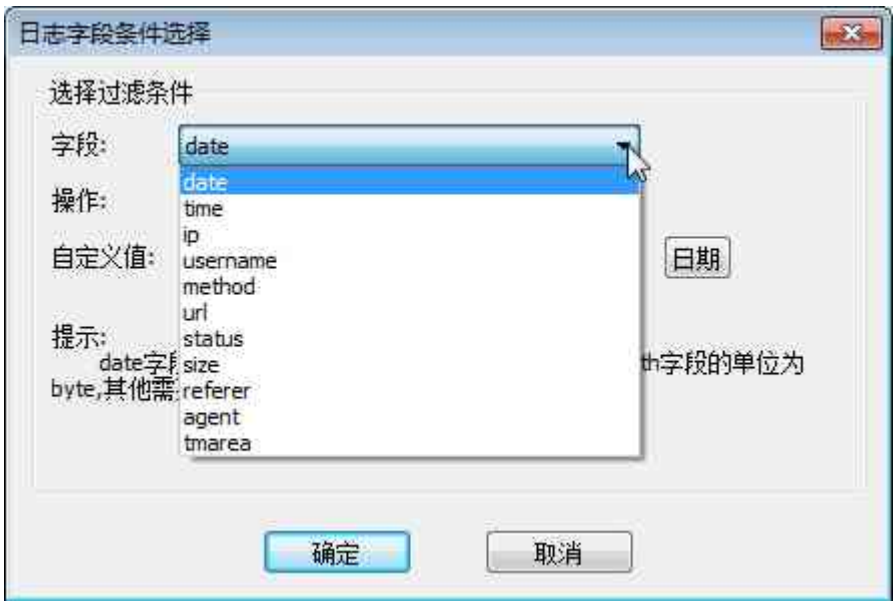
“状态码分析”报告中，现在把用户碰到的状态吗和爬虫碰到的状态码分开了，其他没有什么改变：

状态	URL	访问里	占比%
404		5548	14.853
	/ MIRSERVER.RAR -	442	7.967
	/ APPLE-TOUCH-ICON.PNG -	218	3.929
	/ APPLE-TOUCH-ICON-PRECOMPOSED.PNG -	218	3.929
	/ FORUM.PHP HTTP:/ / WWW.SEMYJ.COM/ FORUM.PHP	125	2.253
	/ WWWROOT.RAR -	49	0.883
	/ WEB.RAR -	48	0.865
	/ MT/ ARCHIVES/ 5/ HTTP:/ / WWW.SEMYJ.COM/ MT/ ARCHIVES/ 5/	48	0.865
	/ SAVE.ASP -	44	0.793
	/ WWW.RAR -	44	0.793
	/ \XD0\XC2\XBA\XF3\XD7\XBA -	36	0.649
	/ WWWROOT.ZIP		

状态码

这里每一行数据都分为两个部分，第 1 部分是表示哪个文件出现了这个状态码，第 2 部分是表示发生在哪个网页。从上面的数据可以看出，这个网站在被一些黑客工具扫描。

在《光年 SEO 日志分析系统》第二版中，最重要的升级是增加了“日志拆分”功能。有了这个功能，就可以用任意维度去分析网站日志了。
以下是可以拆分的日志字段：



拆分字段

只要你的网站日志是齐全的，有了日志拆分功能这个功能就相当于有了一个数据仓库。这个时候查看网站的数据，就：只有你想不到，没有它查不到的。

如：我们要查看上面那个伪装成 google 蜘蛛的 IP 采集了哪些网页，就把拆分条件定义为：ip 等于 222.186.24.59，agent 等于 googlebot，就可以把日志拆分出来了；还有要看是哪些 IP 在用黑客工具扫描网站时，就把拆分条件定义为：url 等于 MIRSERVER.RAR 或等于 WWWROOT.RAR 等等就可以看到了。

我还建议大家多去拆分爬虫的抓取轨迹，把某一个爬虫 IP 的抓取路径拆分出来，观察它的抓取路径，再和网站上的 URL 对应，就能明白爬虫抓取的很多规律。

其实本来还应该开发一个日志合并的功能，但是这个功能实在太简单，一般我们用 DOS 里面的 copy 命令就可以解决这个问题:

Copy 命令

这样，你可以把网站一星期内的、一个月内的甚至半年来的日志合并起来分析。《光年 SEO 日志分析系统》是支持分析无限大的日志的，只要你有时间。

在“设置”-“性能设置”里，有两个地方要注意。一个是那个“蜘蛛计算间隔”，这里表示一个蜘蛛多少时间内没有活动就算它离开了。这里要注意对比分析的时候每次都要是同一个时间，因为这里的时间按改变了，那计算爬虫来访的次数就变了。还有一个是“分析显示条数”，现在你可以自己定义在报表中要显示多少行数据，默认只有 5 条。