

SEOer: 了解搜索引擎

作者: Zac

个人博客: [SEO每天一贴](#)

欢迎发放、传播本文件，但不得向任何人收取任何费用。本文件必需免费传播。
不得对本文件之任何内容进行更改。
引用本文件的内容必需标明原作者及原始出处。

©版权所有

2010年11月

说明: 本电子书是 Zac 所著《SEO 实战密码》一书的第 2 章“了解搜索引擎”。《SEO 实战密码》于 2010 年 11 月由电子工业出版社出版。

Zac 于 2009 年出版了畅销书《网络营销实战密码》，出版一年，重印 10 次。“实战密码”系列读者交流专用网站: www.zacode.com



SEOer: 了解搜索引擎

一个合格的 SEO 必须了解搜索引擎基本工作原理。很多看似令人迷惑的 SEO 原理及技巧，其实从搜索引擎原理出发，都是自然而然的事情。

为什么要了解搜索引擎原理？

说到底，SEO 是在保证用户体验的基础上尽量迎合搜索引擎。与研究用户界面及可用性不同的是，SEO 既要从用户出发，也要站在搜索引擎的角度考虑问题，才能清晰地知道怎样优化网站。SEO 人员必须知道搜索引擎要解决什么问题？有哪些技术上的困难？有什么限制？搜索引擎又怎样取舍？

从某个角度来说，SEO 人员优化网站就是尽量减少搜索引擎的工作量和降低搜索引擎的工作难度，使搜索引擎能更轻松快速地收录网站页面，更准确地提取页面内容。不了解搜索引擎工作原理，也就无从替搜索引擎解决一些 SEOer 力所能及的技术问题。当搜索引擎面对一个网站，发现要处理的问题太多，难度太大时，搜索引擎可能就对这样的网站敬而远之了。

很多 SEO 技巧是基于对搜索引擎的理解。举几个例子。

我们都知道网站域名和页面权重非常重要，这是知其然，很多人不一定知其所以然。权重除了意味着权威度高、内容可靠，因而容易获得好排名外，获得一个最基本的权重，也是页面能参与相关性计算的最基本条件。一些权重太低的页面，就算有很高的相关性也很可能无法获得排名，因为根本没有机会参与排名。

再比如很多 SEO 津津乐道的“伪原创”。首先，抄袭是不道德甚至违法的行为，把别人的文章拿来加一些“的、地、得”，段落换换顺序就当成自己的原创放在网站上，这是令人鄙视的抄袭行为。理解搜索引擎原理的话，就会知道这样的伪原创也不管用。搜索引擎并不会因为两篇文章差几个字，段落顺序不同就真的把它们当成不同的内容。搜索引擎的去重算法要先进准确得多。

再比如，对大型网站来说，最关键的问题是解决收录。只有收录充分，才能带动大量长尾关键词。就算是有人力、有财力的大公司，面对几百万几千万页面的网站时，也不容易处理好充分收录的问题。只有在深入了解搜索引擎蜘蛛爬行原理的基础上，才能尽量使蜘蛛抓得快而全面。

上面所举的几个例子，读者看完搜索引擎原理简介这一节后，会有更深入的认识。

1 搜索引擎与目录

早期 SEO 资料经常把真正的搜索引擎与目录放在一起讨论，甚至把目录也称为搜索引擎的一种，这种讲法并不准确。

真正的搜索引擎指的是由蜘蛛程序沿着链接爬行和抓取网上的大量页面，存进数据库，经过预处理，用户在搜索框输入关键词后，搜索引擎排序程序从数据库中挑选出符合搜索关键词要求的页面。蜘蛛的爬行、页面的收录以及排序都是自动处理。

网站目录则是一套人工编辑的分类目录，由编辑人员人工创建多个层次的分类，站长可以在不同分类里提交网站，目录编辑在后台审核所提交的网站，将网站放置于相应的分类页面。有的时候编辑也主动收录网站。典型的网站目录包括雅虎目录、开放目录、好 123 等。

目录并不是本书中所讨论的 SEO 所关注的真正的搜索引擎。虽然网站目录也常有一个搜索框，但目录的数据来源是人工编辑得到的。

搜索引擎和目录两者各有优劣。

搜索引擎收录的页面数远远高于目录能收录的页面数。但搜索引擎收录的页面质量参差不齐，对网站内容和关键词提取的准确性通常也没有目录高。

限于人力，目录能收录的通常只是网站首页，而且规模十分有限，不过收录的网站通常质量比较高。像雅虎、开放目录、好 123 这些大型目录，收录标准非常高。目录收录网站时存储的页面标题、说明文字都是人工编辑，所以比较准确。

搜索引擎数据更新快，而目录中收录的很多网站内容十分陈旧，甚至网站可能已经不再存在了。

雅虎目录、搜狐目录等曾经是用户在网上寻找信息的主流方式，给用户的感觉与真正的搜索引擎也相差不多。这也就是为什么目录有时候被误称为搜索引擎的一种。但随着 Google 等真正意义上的搜索引擎发展起来以后，目录的使用迅速减少，现在已经很少有人使用网站目录寻找信息了。现在的网站目录对 SEO 的最大意义是建设外部链接，像雅虎、开放目录、好 123 等都有很高的权重，可以给被收录的网站带来一个高质量的外部链接。

2 搜索引擎面临的挑战

搜索引擎系统是最复杂的计算系统之一，当今主流搜索引擎服务商都是有财力和人力的大公司。

即使有技术、人力、财力的保证，搜索引擎还是面临很多技术挑战。搜索引擎诞生后的十多年中，技术已经得到了长足的进步。我们今天看到的搜索结果质量与十年前相比已经好得太多了。不过这还只是一个开始，搜索引擎必然还会有更多创新，提供更多更准确的内容。

总体来说，搜索引擎面对几方面的挑战。

页面抓取需要快而全面

互联网是个动态的内容网络，每天有无数页面被更新、被创建，无数用户在网站上发布内容、沟通联系。要返回最有用的内容，搜索引擎就要抓取最新的页面。但是由于页面数量巨大，搜索引擎蜘蛛更新一次数据库中的页面要花很长时间。搜索引擎刚诞生时，这个抓取周期往往以几个月计算。这也就是为什么 Google 在 2003 年以前每个月有一次大更新。

现在主流搜索引擎都已经能在几天之内更新重要页面，权重高的网站上的新文件几小时甚至几分钟之内就会被收录。不过，这种快速收录和更新也只能局限于高权重网站。很多页面几个月不能被重新抓取和更新，也是非常常见的。

要返回最好的结果，搜索引擎也必须抓取尽量全面的页面，这就需要解决很多技术问题。一些网站并不利于搜索引擎蜘蛛爬行和抓取，诸如网站链接结构的缺陷，大量使用 Flash，JavaScript 脚本，或把内容放在用户必须登录以后才能访问的部分，这都提高了搜索引擎抓取内容的难度。

海量数据存储

一些大型网站一个网站就有百万千万页面，可以想象网上所有网站的页面加起来是一个什么数据量。搜索引擎蜘蛛抓取页面后，还必须有效存储这些数据，数据结构必须合理，具备极高的扩展性，写入及访问速度要求也很高。

除了页面数据，搜索引擎还需要存储页面之间的链接关系以及大量历史数据，这样的数据量是我们用户无法想象的。据说 Google 有几十个数据中心，上百万台服务器。这样大规模的数据存储和访问必然存在很多技术挑战。

我们经常在搜索结果中看到，排名会没有明显原因地上下波动，甚至可能刷新一下页面，就看到不同的排名，有的时候网站数据也可能丢失。这些都可能与大规模数据存储的技术难题有关。

索引处理快速有效，具可扩展性

搜索引擎将页面数据抓取和存储后，还要进行索引处理，包括链接关系的计算、正向索引、倒排索引等。由于数据库中页面数量大，进行 PR 之类的迭代计算也是耗时费力。要想及时提供相关又及时的搜索结果，仅仅抓取也没有用，还必须进行大量索引计算。由于随时都有新数据

新页面加入，索引处理也要具备很好的扩展性。

查询处理快速准确

查询是普通用户唯一能看到的搜索引擎工作步骤。用户在搜索框输入关键词，点击搜索按钮后，通常不到一秒后就会看到搜索结果。表面最简单的过程，实际上牵扯了非常复杂的后台处理。在最后的查询阶段，最重要的难题是怎样在不到一秒钟的时间内，快速从几十万几百万，甚至几千万包含搜索词的页面中，找到最合理、最相关的一千个页面，并且按照相关性、权威性排列。

判断用户意图以及人工智能

应该说前四个挑战现在的搜索引擎都已经能够比较好地解决，但判断用户意图还处在初级阶段。不同用户搜索相同的关键词，很可能是在寻找不同的东西。比如搜索“苹果”，用户到底是想了解苹果这个水果？还是苹果电脑？还是电影《苹果》的信息？没有上下文，没有对用户个人搜索习惯的了解，就完全无从判断。

搜索引擎目前正在致力于基于用户搜索习惯及历史数据的了解上，判断搜索意图，返回更相关的结果。今后搜索引擎是否能达到人工智能水平，真正了解用户搜索词的意义和目的，让我们拭目以待。

3 搜索结果显示格式

让我们先来稍微深入地了解一下搜索结果的展现形式。

3-1 搜索结果页面

用户在搜索引擎搜索框中输入关键词，点击搜索按钮后，搜索引擎在很短时间内返回一个搜索结果页面。下图所示是 Google 的搜索结果页面，也是比较典型的搜索结果页面排版格式。



减肥方法

Google 搜索

高级

网页 打开百宝箱

搜索 减肥方法 获得约 19,600,000 条结果, 以下是第 1-10 条。(用时 0.10 秒)

.com

1周瘦10斤懒人减肥秘诀

laiba.tianya.cn 12种小吃越吃越瘦 吸走多余脂肪 天涯来吧_减肥吧

广告

赞助商链接

减肥方法网 - 寻找最好的瘦身方法!

减肥方法网-寻找最好的瘦身方法!... 把豆浆当减肥方法... 健康减肥不要过分依赖... 范冰冰从胖妹变美女的... 洗澡减肥方法详细介绍... 女人处处小心减肥雷区...

减肥食谱 - 运动减肥 - 瘦身方法 - 局部减肥

www.lpwcn.com/ - 网页快照 - 类似结果

减肥方法-如何快速减肥瘦身及最有效的减肥方法-17瘦身网

减肥瘦身网-提供最全面的减肥方法, 最有效的快速减肥方法, 专业介绍如何减肥的知识及瘦身方法技巧等。

www.17shoushen.cn/ - 网页快照 - 类似结果

减肥方法 减肥食谱 如何减肥-悠悠减肥网

悠悠减肥网是介绍各种健康的减肥方法, 减肥食谱, 探讨如何减肥, 最有效的减肥方法的专业性网站, 收录大量减肥食谱, 减肥操, 明星减肥, 针灸减肥, 减肥药, 减肥茶等资讯, ...

www.jianfeiuu.com/ - 网页快照 - 类似结果

Google 资讯: 减肥方法



中西减肥方法大盘点: 减肥药PK减肥茶 - 59 分钟前

内容摘要: 前面介绍过, 目前市场上的减肥药主要有西布曲明和奥利司他胶囊两种, 分别来看看它们各自的适用人群和禁忌人群: 总结经过5个回合的PK, 减肥茶以3:2的比分...

中国网 (新闻发布)

减肥瘦身: 9个减肥偏方 最有效的减肥方法 - 慧聪网 - 2 篇相关文章 >

减肥必知的4大黄金定律 - 新民网 - 5 篇相关文章 >

减肥方法 39健康减肥 39健康网

介绍各种运动减肥, 饮食减肥, 手术减肥, 中医减肥, 药物减肥, 减肥生活小窍门, 减肥心理调节, 另类的减肥方法等。

fitness.39.net > 减肥 - 网页快照 - 类似结果

【减肥】

太平洋女性网减肥频道减肥方法栏目包括排毒减肥、减肥食谱、减肥运动、减肥偏方、医疗减肥、减肥误区、减肥产品等, 给您最好最科学的减肥方法。

fitness.pclady.com.cn/jf/ - 网页快照 - 类似结果

网易女人频道减肥库--运动减肥

准妈妈在减肥, 美体方面有哪些心得, 让我们一起来看看! 1在所有减肥方法中, 运动减肥是最健康的, 不妨选个自己感兴趣的运动坚持做。2要根据自己想减什么部位来选择...

lady.163.com > 女人频道做更好的自己 - 网页快照 - 类似结果

快速减肥方法, 有效减肥产品, 怎样减肥安全健康【不胖啦减肥网】

快速减肥方法, 有效减肥产品, 健康减肥食谱, 专业减肥顾问, 尽在不胖啦减肥网, 还想减肥吗?

www.bupangla.com/ - 网页快照 - 类似结果

减肥方法 针灸减肥 安定门中医院

北京安定门中医医院针灸科治疗减肥方法。世界针灸联合会委员、全国中医针灸首席专家亲自坐诊, 一疗程即可见效, 对治疗减肥、偏头痛、面瘫、祛斑、祛痣、痛经等效果显著...

yuhua.adm999.com/acupuncture/acupuncture-23.html - 网页快照

吸脂减肥方法的常见问答

为什么说吸脂减肥是最有效的减肥方法, 减肥方法: 吸脂减肥是通过减少皮下多余的脂肪细胞达到局部减肥的目的。而且, 人体各个部位的脂肪细胞数量是恒定的, 而且脂肪细胞...

www.aimei100.com/zhengxmr/jianfs/zhengxmr_236.html - 网页快照

减肥产品|减肥方法---北京仁和盛科贸有限公司

40分钟瘦腰3公分, 手臂、大腿2公分, 微乐能量雕塑减肥, 风靡欧美的09年新品。热线: 010-62223033 62223066。

www.bjscar.net/ - 网页快照 - 类似结果

Google 博客: 减肥方法

打击肥胖超有效的减肥方法-女人时尚-女人时尚-1 天前

世界公认10大健康减肥方法, 完美天使_新浪博客-广西总队整形30天丑女变形-3 天前

适合12星座的减肥方法! 狮王勇者-狮王勇者-6 小时前

减肥方法的相关搜索

减肥食谱

减肥瘦身方法

减肥瘦身

最有效的减肥方法

健康减肥方法

最快的减肥方法

运动减肥方法

最好的减肥方法

腹部减肥方法

最有效的快速减肥方法

← 相关搜索

相关服务: 到天涯问答提问 减肥方法

1周瘦10斤懒人减肥秘诀

laiba.tianya.cn 12种小吃越吃越瘦 吸走多余脂肪 天涯来吧_减肥吧

广告

赞助商链接



← 自然搜索结果

图 1 Google 的搜索结果页面

页面主体有两部分最为主要：一是广告，二是自然搜索结果。如图所示，页面右侧 8 个结果以及左侧最上面的一个结果，都标注为“赞助商链接”，这就是广告。绝大部分网民都比较清楚右侧显示的是广告，所以右侧赞助商链接没有加特殊底色。页面左侧上部的广告链接使用浅黄色底色，可以和下面的自然搜索结果清楚的分开。右侧广告最多有 8 个，上部广告可以多至 3 个。

搜索广告在网络营销行业经常称为 PPC，是由广告商针对关键词进行竞价，广告显示广告商无需付费，只有搜索用户点击广告后，广告商才按竞价价格支付广告费用。PPC 是搜索营销的另一个主要内容。

搜索结果页面左侧广告下面，占据页面最大部分的就是自然搜索结果。通常每个页面会列出 10 个自然搜索结果。用户可以在帐户设置中选择每页显示 100 个搜索结果。每个搜索结果的格式下面再做介绍。

页面最左上角是垂直搜索链接，用户点击后可以直接访问图片、视频、地图等搜索结果。

搜索框右下方显示满足搜索关键词的结果总数，如图中所显示的 19600000 条结果。这个搜索结果数是研究竞争程度的依据之一。

自然搜索结果下面显示相关搜索。搜索引擎根据用户搜索数据，列出相关的其他搜索词。

页面最下面又是一个赞助商广告，与页面顶部的广告相同。页面左侧顶部及左侧底部的广告，并不是每次搜索有广告商竞价时都会出现，只有点击率和质量分数达到一定水平的广告才会出现在左侧顶部或底部。

SEOer 最关注的是占据页面主体的自然搜索结果。统计数据显示，自然搜索结果总点击访问数要远远大于广告点击数。但是企业花费在 SEO 上的费用却远远低于花费在搜索广告上的费用。这既是 SEO 的尴尬，也是最大的机会。掌握了 SEO 流量，才掌握住最大搜索流量。

我们再来看百度搜索结果页面。



图2 百度的搜索结果页面

百度搜索结果页面与 Google 大致相同，区别在于广告部分的显示方法。如图所示，右侧也是最多 8 个广告，不过并没有标注为赞助商链接或加其他提示文字。左侧最上面标注为“推广链接”的结果也是广告，这是百度启用凤巢系统后显示的广告。不过这几个广告只加了非常浅的灰色背景，不注意看几乎无法与下面的自然搜索结果分辨开来。

有的关键词搜索没有触发凤巢系统广告，还会继续显示传统百度广告，如下图所示。



图3 传统百度搜索广告

传统百度左侧广告既没有明确标注为推广链接，也没有使用任何背景颜色，与下面的自然搜索结果更不容易分辨。唯一能分辨出上面3个是广告的地方是，结果列表最后一行最右侧标有“推广”两个字。百度广告结果在背景颜色、文字标注上，都比较难于与自然结果区分。SEO人员当然很清楚这两者之间的区别，普通网民却比较难以察觉，尤其是百度传统左侧广告。这也就是百度搜索结果常为人诟病的原因之一。

3-2 经典搜索结果列表

我们再来看看每一个搜索结果页面的展现格式。如下图所示是百度的搜索结果列表，主要分三部分。

[SEO每天一贴 - Zac的SEO博客, 正在写书, 暂时每月两三贴](#)
 SEO每天一贴研究搜索引擎优化SEO技术,网络营销及电子商务思考。正在写一本SEO方面的书,所以不能每天一贴了,会尽快恢复。
www.chinamyhosting.com/seoblog/ 2010-2-24 - [百度快照](#)

图4 最常见的搜索结果列表格式

第一行是页面标题，通常取自页面 HTML 代码中的标题标签（Title Tag）。这是结果列表中最醒目的部分，用户点击标题就可以访问对应的网页。所以页面标题标签的写法，无论对排名还是点击率都有重要意义。

第二、三行是页面说明。页面说明有的时候取自页面 HTML 中的说明标签（Description Tag），有的时候是从页面可见文字中动态抓取相关内容。所以显示什么页面说明文字是用户查询时才决定的。

某些与日期有明确联系的页面，Google 会在说明文字最前面显示日期，省略号后再显示页面说明。如博客帖子这类有明确发布日期的页面。



URL网址规范化-搜索引擎优化SEO每天一贴
2006年4月10日 ... URL网址规范化- 作者: Zac, 首发于搜索引擎优化SEO每天一贴。
www.chinamyhosting.com/.../04/.../url-canonicalization/ - 网页快照 - 类似结果

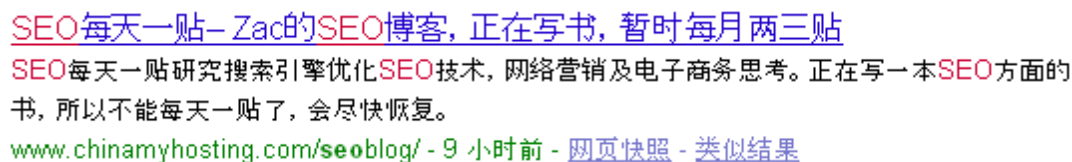
图5 Google 搜索结果列表中显示日期

第四行显示三个信息。最左侧是网址，用户可以看到页面来自哪个网站，以及目录、文件名信息。

中间是百度数据库中页面最后更新的日期。

然后是百度快照链接，用户可以点击快照，查看存储在百度数据库中的页面内容。有的时候页面被删除或者有其他技术问题不能打开网站时，用户至少还可以从快照中查看想要的内容。

用户所搜索的关键词在标题及说明部分都用红色高亮显示，用户可以非常快速地看到页面与自己搜索的关键词相关性如何。如上图中的 SEO 三个字母。



SEO每天一贴- Zac的SEO博客, 正在写书, 暂时每月两三贴
SEO每天一贴研究搜索引擎优化SEO技术, 网络营销及电子商务思考。正在写一本SEO方面的书, 所以不能每天一贴了, 会尽快恢复。
www.chinamyhosting.com/seoblog/ - 9 小时前 - 网页快照 - 类似结果

图6 搜索词在 Google 中文搜索结果列表中高亮显示

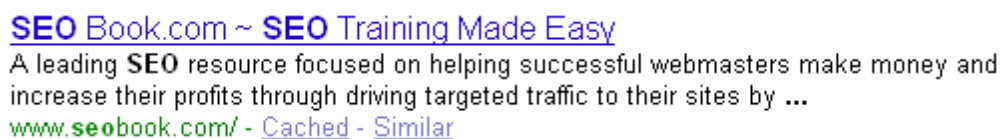
Google 结果列表与百度大致相同，几处小的区别包括：

搜索关键词在 URL 中出现时加粗显示，如上图中的 SEO 三个字母。

URL 右侧的页面最近更新时间不是按日期显示，而是显示为几小时前。

网页快照链接后面有时还有一个“类似结果”链接。用户点击类似结果后可以看到与这个页面相似的其他网页。

Google 英文结果列表与中文还有一个区别，就是搜索关键词在标题及说明部分不是红色高亮显示，而是加粗显示。



SEO Book.com ~ SEO Training Made Easy
A leading **SEO** resource focused on helping successful webmasters make money and increase their profits through driving targeted traffic to their sites by ...
www.seobook.com/ - [Cached](#) - [Similar](#)

图7 搜索词在 Google 英文搜索结果列表中加粗显示

红色高亮显示应该是 Google 为适应中国用户搜索习惯所做的变化。

2010年4月份，本章初稿完成后，我又发现百度也在实验将 URL 中的关键词加粗，但不是所有 URL 中的关键词都会加粗，如下图所示：



SEO论坛,搜索引擎优化论坛
提供**SEO**资讯、教程,搜索引擎优化经验交流与资源下载。
www.seobbs.net/ 2010-4-7 - [百度快照](#)

SEO每天一贴 - Zac的SEO博客,正在写书,暂时每月两三贴
SEO每天一贴研究搜索引擎优化**SEO**技术,网络营销及电子商务思考。正在写一本**SEO**方面的书,所以不能每天一贴了,会尽快恢复。
www.chinamyhosting.com/seoblog/ 2010-3-29 - [百度快照](#)

图8 搜索词在百度列表中 URL 目录部分加粗显示

上面讨论的是最经典的结果列表格式。搜索引擎近几年也在不停尝试不同格式的搜索结果格式，尤其是 Google，推出了很多特色结果列表，下面挑几个主要的简单介绍。

3-3 整合搜索结果

前面的 Google 搜索结果页面抓图中，大家就能看到两个整合搜索结果，中间的资讯结果和页面

底部的博客搜索结果。根据搜索关键词的不同，Google 还经常把其他垂直搜索结果混合在正常网页搜索结果中，比如图片结果、视频结果。

Google 图片:seo图片 - 举报图片



图9 图片搜索结果

Google 视频:zac seo 视频



图10 视频搜索结果

百度也有类似的整合搜索结果，主要以新闻内容为主，其他则很少见到。

seo的相关新闻

- [ASP.NET的SEO:HTTP报头状态码--内容重定向](#) 天极网 1天前
- 这些状态码和SEO又有什么关系呢? 每次当用户代理(可以理解为就是IE和Firefox)向Web站点请求一个URL地址,服务器都会给予回复,回复内容包括两部分:...
- [SEO在编辑过程中的一些应用](#) 鞭牛士 1天前
- [黑客充分利用SEO技巧 社会工程渐成主流](#) 赛迪网 1天前

图11 百度的新闻搜索结果

3-4 缩进列表 (Indented Listing)

当搜索结果页面上应该出现两个来自同一网站的页面时，比如第一位和第八位，按照正常排名算法是来自同一网站的两个页面，Google 不是把它们排在第一和第八位，而是把两个结果连在一起排在第一和第二位，第二位的结果向右侧缩进三个字的空间。这样的列表格式非常有助于提高点击率。

[SEO每天一贴- Zac的SEO博客, 正在写书, 暂时每月两三贴](#)

搜索引擎优化SEO每天一贴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 ...

www.chinamyhosting.com/seoblog/ - 9 小时前 - [网页快照](#) - [类似结果](#)

[SEO技术- 搜索引擎优化SEO每天一贴](#)

搜索引擎优化SEO每天一贴. 介绍和研究搜索引擎优化SEO技术, Zac的SEO优化及网络营销, 电子商务, 网站推广运营思考。订阅Feed: RSS · 评论RSS ...

www.chinamyhosting.com/seoblog/category/seo-tips/ - [网页快照](#) - [类似结果](#)

图 12 缩进列表

3-5 全站链接 (Sitelinks)

对某些权重比较高的网站, 当用户搜索一个关键词, 这个网站的结果是最权威的内容来源时, Google 除了正常结果列表外, 还会显示最多四行两列八个内页链接, 称为全站链接。

[SEO每天一贴- Zac的SEO博客, 正在写书, 暂时每月两三贴](#)

SEO每天一贴研究搜索引擎优化SEO技术, 网络营销及电子商务思考。正在写一本SEO方面的书, 所以不能每天一贴了, 会尽快恢复。

www.chinamyhosting.com/seoblog/ - 9 小时前 - [网页快照](#) - [类似结果](#)

网络赚钱	Google实时搜索
我的书	英文SEO论坛和博客推荐
英文网站SEO	电子商务
SEO技术	网站内容

图 13 全站链接

这无疑为权重高的网站提供了多几倍的访问入口, 视觉上的醒目也大大提高了点击率。

3-6 迷你全站链接 (Mini Sitelinks)

权重高的网站在某些情况下还会显示迷你全站链接, 不是四行八个, 而是一行四个链接。显示的内页与上面说的全站链接是一样的, 取其中前面四个。

[SEO每天一贴-Zac的SEO博客, 正在写书, 暂时每月两三贴](#)

SEO每天一贴研究搜索引擎优化SEO技术, 网络营销及电子商务思考。... 介绍和研究搜索引擎优化SEO技术, Zac的SEO优化及网络营销, 电子商务, 网站推广运营思考。...

[网络赚钱 - 我的书 - 英文网站SEO - SEO技术](#)

www.chinamyhosting.com/seoblog/ - 9 小时前 - [网页快照](#) - [类似结果](#)

图 14 迷你全站链接

3-7 One-box

某些关键词会触发 Google one-box 结果，直接在搜索结果页面上显示相关信息，用户不用点击到其他网站上查看。如下图显示搜索“北京银行”时显示的股价 One-box:



图 15 Google 的 One-box

3-8 富摘要 (Rich Snippet)

某些使用 RDFa 或 Microdata 格式标签的页面，Google 可能还会在标题下面以灰色文字加一行富摘要，如下图的论坛帖子页面还显示出帖子个数、作者数以及更新日期。

[ZAC的书咋还没出来急了- SEO咨询- 点石论坛](#)

2 个帖子 - 2 个作者 - 新贴子: 5 天之前

[点石论坛](#) » [SEO咨询](#) » [ZAC的书咋还没出来急了 ...](#) 本论坛支付平台由支付宝提供携手打造安全诚信的交易社区, Powered by Discuz! ...

www.dunsh.org/forums/thread-68014-1-1.html - [网页快照](#)

图 16 Google 的富摘要

这样的排版格式无疑也会提高关注度和点击率。在富摘要中显示合适的信息，有助于说服用户点击结果，比如显示产品价格、用户评分、用户评论数目等。

百度也有类似显示方式：

[2007年点石最后一次线下活动南京茶话会专题 - 点石线下茶话会 - ...](#)

14个回复 - 发帖时间: 2007年11月28日

[点石论坛](#)...[点石茶话会](#)一直秉承交流分享的宗旨，会在中国SEO技术爱好者和从业人员比较密集... 17: 00~17: 50 [互动问答](#) 本环节由演讲嘉宾帮助大家回答疑问，请参会...

www.dunsh.org/forums/thread-16295-1-1.html 2010-3-29 - [百度快照](#)

图 17 百度的富摘要

3-9 面包屑导航

Google 最近又在结果列表中大规模使用面包屑导航。原本显示一个网址的地方，改为面包屑导航格式，其中的每一个分类链接都指向网站上相应的分类页面。

[药物减肥](#) [39健康减肥](#) [39健康网](#)

介绍各种[减肥产品](#)，[减肥药品](#)知识，服用[减肥药](#)的注意事项，[减肥药](#)减肥的基本原理，[减肥药](#)的主要成分及副作用等。

fitness.39.net > [减肥](#) > [减肥方法](#) - [网页快照](#) - [类似结果](#)

图 18 Google 显示的面包屑导航

用户不仅可以点击标题访问产品页面，还可以直接从搜索列表的面包屑导航中点击上级分类链接访问分类页面。

3-10 说明文字中的链接

一部分使用了页面内锚链接的页面，Google 有时也尝试在说明文字中显示链接，用户可以跳到页内锚链接部分。

[Jeff Dean-程序员百科全书-科技中国——欢迎光临全球最大的互联网博物馆 ...](#)
跳到[谷歌自爆数据中心基础设施](#): 2009-06-29: 了Google一向很少对外透露其数据中心的工作，但5 ... 一窥Google数据中心自行定制的40台服务器机柜。...
[www.techcn.com.cn > 科技人物百科全书 - 网页快照 - 类似结果](#)

图 19 Google 显示的说明文字中的链接

这种显示方式目前还比较少见。

上面介绍的一些搜索列表变化形式有逐渐增多的趋势，不过它们的基本形式与经典搜索结果列表相差不大，最经典的结果列表还是最常见的。

4 搜索引擎工作原理简介

搜索引擎工作过程非常复杂，接下来几节我们简单介绍搜索引擎是怎样实现网页排名的。这里介绍的相对于真正的搜索引擎技术来说只是皮毛，不过对 SEO 人员已经足够用了。

搜索引擎的工作过程大体上可以分成三个阶段：

- 1) 爬行和抓取 - 搜索引擎蜘蛛通过跟踪链接访问网页，获得页面 HTML 代码存入数据库。
- 2) 预处理 - 索引程序对抓取来的页面数据进行文字提取、中文分词、索引等处理，以备排名程序调用。
- 3) 排名 - 用户输入关键词后，排名程序调用索引库数据，计算相关性，然后按一定格式生成搜索结果页面。

4-1 爬行和抓取

爬行和抓取是搜索引擎工作的第一步，完成数据收集的任务。

蜘蛛

搜索引擎用来爬行和访问页面的程序被称为蜘蛛（spider），也称为机器人（bot）。

搜索引擎蜘蛛访问网站页面时类似于普通用户使用的浏览器。蜘蛛程序发出页面访问请求后，服务器返回 HTML 代码，蜘蛛程序把收到的代码存入原始页面数据库。搜索引擎为了提高爬行和抓取速度，都使用多个蜘蛛并发分布爬行。

蜘蛛访问任何一个网站时，都会先访问网站根目录下的 robots.txt 文件。如果 robots.txt 文件禁止搜索引擎抓取某些文件或目录，蜘蛛将遵守协议，不抓取被禁止的网址。

和浏览器一样，搜索引擎蜘蛛也有标明自己身份的代理名称，站长可以在日志文件中看到搜索引擎的特定代理名称，从而辨识搜索引擎蜘蛛。下面列出常见的搜索引擎蜘蛛名称：

- Baiduspider(+http://www.baidu.com/search/spider.htm) 百度蜘蛛
- Mozilla/5.0 (compatible; Yahoo! Slurp China; http://misc.yahoo.com.cn/help.html) 雅虎中国蜘蛛
- Mozilla/5.0 (compatible; Yahoo! Slurp/3.0; http://help.yahoo.com/help/us/ysearch/slurp) 英文雅虎蜘蛛
- Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html) Google 蜘蛛
- msnbot/1.1 (+http://search.msn.com/msnbot.htm) 微软 Bing 蜘蛛
- Sogou+web+robot(+http://www.sogou.com/docs/help/webmasters.htm#07) 搜狗蜘蛛
- Sosospider(+http://help.soso.com/webspider.htm) 搜搜蜘蛛
- Mozilla/5.0 (compatible; YodaoBot/1.0; http://www.yodao.com/help/webmaster/spider/;) 有道蜘蛛

跟踪链接

为了抓取网上尽量多的页面，搜索引擎蜘蛛会跟踪页面上的链接，从一个页面爬到下一个页面，就好像蜘蛛在蜘蛛网上爬行那样，这也就是搜索引擎蜘蛛这个名称的由来。

整个互联网是由相互链接的网站及页面组成的。从理论上说，蜘蛛从任何一个页面出发，顺着链接都可以爬行到网上的所有页面。当然，由于网站及页面链接结构异常复杂，蜘蛛需要采取一定的爬行策略才能遍历网上所有页面。

最简单的爬行遍历策略分为两种，一是深度优先，二是广度优先。

所谓深度优先指的是蜘蛛沿着发现的链接一直向前爬行，直到前面再也没有其他链接，然后返回到第一个页面，沿着另一个链接再一直往前爬行。

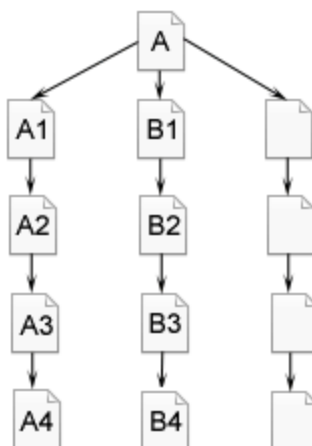


图 20 深度优先遍历策略

如上图所示，蜘蛛跟踪链接，从 A 页面爬行到 A1, A2, A3, A4, 到 A4 页面后，已经没有其他链接可以跟踪就返回 A 页面，顺着页面上的另一个链接，爬行到 B1, B2, B3, B4。在深度优先策略中，蜘蛛一直爬到无法再向前，才返回爬另一条线。

广度优先是指蜘蛛在一个页面上发现多个链接时，不是顺着一个链接一直向前，而是把页面上所有第一层链接都爬一遍，然后再沿着第二层页面上发现的链接爬向第三层页面。

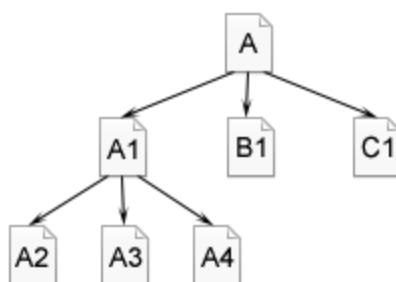


图 21 广度优先遍历策略

如上图所示，蜘蛛从 A 页面顺着链接爬行到 A1, B1, C1 页面，直到 A 页面上的所有链接都爬行完，然后再从 A1 页面发现的下一层链接，爬行到 A2, A3, A4.....页面。

从理论上说，无论是深度优先还是广度优先，只要给蜘蛛足够的时间，都能爬完整个互联网。

在实际工作中，蜘蛛的带宽资源、时间都不是无限的，也不可能爬完所有页面。实际上最大的搜索引擎也只是爬行和收录了互联网的一小部分。

深度优先和广度优先通常是混合使用的，这样既可以照顾到尽量多的网站（广度优先），也能照顾到一部分网站的内页（深度优先）。

吸引蜘蛛

由此可见，虽然理论上蜘蛛能爬行和抓取所有页面，但实际上不能也不会这么做。SEO 人员要想自己的更多页面被收录，就要想方设法吸引蜘蛛来抓取。既然不能抓取所有页面，蜘蛛所要做做的就是尽量抓取重要页面。哪些页面被认为比较重要呢？有几方面影响因素：

- 网站和页面权重。质量高、资格老的网站被认为权重比较高，这种网站上的页面被爬行的深度也会比较高，所以会有更多内页被收录。
- 页面更新度。蜘蛛每次爬行都会把页面数据存储起来。如果第二次爬行发现页面与第一次收录的完全一样，说明页面没有更新，蜘蛛也就没有必要经常抓取。如果页面内容经常更新，蜘蛛就会更加频繁地访问这种页面，页面上出现的新链接，也自然会被蜘蛛更快跟踪，抓取新页面。
- 导入链接。无论是外部链接还是同一个网站的内部链接，要被蜘蛛抓取就必须有导入链接进入页面，否则蜘蛛根本没有机会知道页面的存在。高质量的导入链接也经常使页面上的导出链接被爬行深度增加。
- 与首页点击距离。一般来说网站上权重最高的是首页，大部分外部链接是指向首页，蜘蛛访问最频繁的也是首页。离首页点击距离越近，页面权重越高，被蜘蛛爬行的机会也越大。

地址库

为了避免重复爬行和抓取网址，搜索引擎会建立一个地址库，记录已经被发现还没有抓取的页面，以及已经被抓取的页面。

地址库中的 URL 有几个来源。

1. 一是人工录入的种子网站。
2. 二是蜘蛛抓取页面后，从 HTML 中解析出新的链接 URL，与地址库中的数据对比，如果是地址库中没有的网址，就存入待访问地址库。
3. 三是站长通过搜索引擎网页提交表格提交进来的网址。

蜘蛛按重要性从待访问地址库中提取 URL，访问并抓取页面，然后把这个 URL 从待访问地址库中删除，放进已访问地址库中。

大部分主流搜索引擎都提供一个表格，让站长提交网址。不过这些提交来的网址都只是存入地址库而已，是否收录还要看页面重要性如何。搜索引擎所收录的绝大部分页面是蜘蛛自己跟踪链接得到的。可以说提交页面基本上是毫无用处的，搜索引擎更喜欢自己沿着链接发现新页面。

文件存储

搜索引擎蜘蛛抓取的数据存入原始页面数据库。其中的页面数据与用户浏览器得到的 HTML 是完全一样的。每个 URL 都有一个独特的文件编号。

爬行时的复制内容检测

检测并删除复制内容通常是在下面介绍的预处理过程中进行，但现在的蜘蛛在爬行和抓取文件时也会进行一定程度的复制内容检测。遇到权重很低的网站上大量转载或抄袭内容时，很可能不再继续爬行。这也就是为什么有的站长在日志文件中发现了蜘蛛，但页面从来没有被真正收录过。

4-2 预处理

在一些 SEO 材料中，预处理也被简化称为索引，因为索引是预处理最主要的步骤。

搜索引擎蜘蛛抓取的原始页面，并不能直接用于查询排名处理。搜索引擎数据库中的页面数都在数万亿级别以上，用户输入搜索词后，靠排名程序实时对这么多页面分析相关性，计算量太大，不可能在一两秒内返回排名结果。因此抓取来的页面必须经过预处理，为最后的查询排名做好准备。

和爬行抓取一样，预处理也是在后台提前完成，用户搜索时感觉不到这个过程。

提取文字

现在的搜索引擎还是以文字内容为基础。蜘蛛抓取到的页面中的 HTML 代码，除了用户在浏览器上可以看到的可见文字外，还包含了大量的 HTML 格式标签、JavaScript 程序等无法用于排名的内容。搜索引擎预处理首先要做的就是从 HTML 文件中去除标签、程序，提取出可以用于排名处理的网页面文字内容。

比如下面这段 HTML 代码：

```
<div id="post-1100" class="post-1100 post-hentry category-seo">
<div class="posttitle">
```

```
<h2><a href="http://www.chinamyhosting.com/seoblog/2010/04/01/fools-day/"  
rel="bookmark" title="Permanent Link to 今天愚人节哈">今天愚人节哈</a></h2>
```

除去 HTML 代码后，剩下的用于排名的文字只是这一行：

今天愚人节哈

除了可见文字，搜索引擎也会提取出一些特殊的包含文字信息的代码，如 Meta 标签中的文字，图片替代文字，Flash 文件的替代文字，链接锚文字等。

中文分词

分词是中文搜索引擎特有的步骤。搜索引擎存储和处理页面，以及用户搜索都是以词为基础。英文等语言单词与单词之间有空格分隔，搜索引擎索引程序可以直接把句子划分为单词的集合。而中文词与词之间没有任何分隔符，一个句子中的所有字和词都是连在一起的。搜索引擎必须首先分辨哪几个字组成一个词，哪些字本身就是一个词。比如“减肥方法”将被分词为“减肥”和“方法”两个词。

中文分词方法基本上有两种，一是基于词典匹配，另一个是基于统计。

基于词典匹配的方法是指，将待分析的一段汉字与一个事先造好的词典中的词条进行匹配，在待分析汉字串中扫描到词典中已有的词条则匹配成功，或者说切分出一个单词。

按照扫描方向，基于词典的匹配法可以分为正向匹配和逆向匹配。按照匹配长度优先级的不同，又可以分为最大匹配和最小匹配。将扫描方向和长度优先混合，又可以产生正向最大匹配、逆向最大匹配等不同方法。

词典匹配方法计算简单，其准确度很大程度上取决于词典的完整性和更新情况。

基于统计的分词方法指的是分析大量文字样本，计算出字与字相邻出现的统计概率，几个字相邻出现越多，就越可能形成一个单词。基于统计的方法优势是对新出现的词反应更快速，也有利于消除歧义。

基于词典匹配和统计的两种分词方法各有优劣，实际使用中的分词系统都是混合使用两种方法，达到快速高效，又能识别生词、新词，消除歧义。

中文分词的准确性往往影响搜索引擎排名的相关性。比如在百度搜索“搜索引擎优化”：

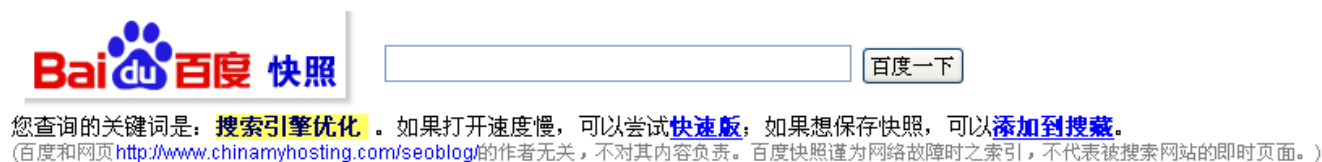


图 22 百度快照显示的对“搜索引擎优化”的分词结果

从快照中可以看到, 百度把“搜索引擎优化”这六个字当成一个词。

而在 Google 搜索同样的词:

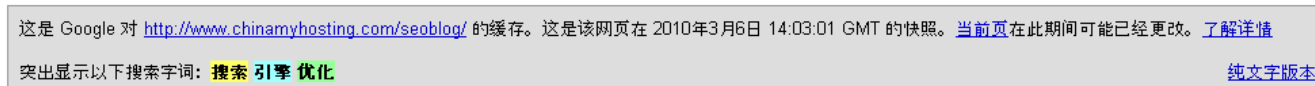


图 23 Google 快照显示的对“搜索引擎优化”的分词结果

快照显示 Google 将其分切为“搜索”，“引擎”和“优化”三个词。显然百度切分得更为合理，搜索引擎优化是一个完整的概念。Google 分词时倾向于更为细碎。

再举一个更明显的例子。在 Google 搜索“点石互动”四个字:

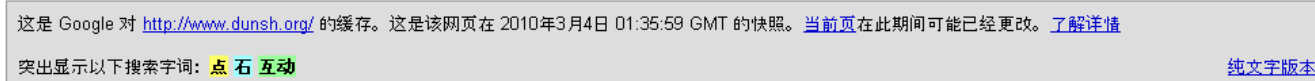


图 24 Google 快照显示的对“点石互动”的分词结果

快照显示 Google 将其切分为“点”，“石”及“互动”三个词。“点石互动”这个中文 SEO 领域最知名的品牌，显然并没有进入 Google 的词典中。在百度搜索“点石互动”时会发现，百度将“点石互动”当作一个词。甚至在百度搜索“点石大会报名”，可以发现百度把“点石大会”都当成一个词:

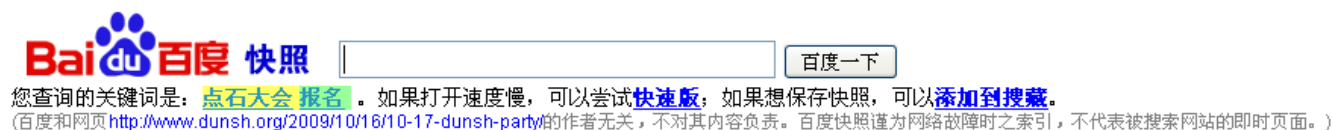


图 25 百度快照显示的对“点石大会报名”的分词结果

这种分词上的不同很可能是一些关键词排名在不同搜索引擎有不同表现的原因之一。比如百度更喜欢搜索词完整匹配地出现在页面上，也就是说搜索“点石互动”时，这四个字连续完整出现更容易在百度获得好的排名。Google 就与此不同，不太要求完整匹配。一些页面出现“点石”和“互动”两个词，但不必完整匹配地出现，“点石”出现在前面，“互动”出现在页面的其他地方，这样的页面在 Google 搜索“点石互动”时，也可以获得不错的排名。

搜索引擎对页面的分词取决于词库的规模、准确性和分词算法的好坏，而不是取决于页面本身如何，所以 SEO 人员对分词所能做的很少。唯一能做的是在页面上用某种形式提示搜索引擎，某几个字应该被当作一个词处理，尤其是可能产生歧义的时候，比如在页面标题、h1 标签以及黑体中出现关键词。如果页面是关于“和服”的内容，那么可以把“和服”这两个字特意标为黑体。如果页面是关于“化妆和服装”，可以把“服装”两个字标为黑体。这样，搜索引擎对页面进行分析时就知道标为黑体的应该是一个词。

去停止词

无论英文中文，页面内容中都会有一些出现频率很高，却对内容没有任何影响的词，如“的”，“地”，“得”之类的助词，“啊”，“哈”，“呀”之类的感叹词，“从而”，“以”，“却”之类的介词。这些词被称为停止词，因为它们对页面主要意思没什么影响。英文中的常见停止词如 the, a, an, to, of 等。

搜索引擎在索引页面之前会去掉这些停止词，使索引数据主题更为突出，减少无谓的计算量。

消除噪声

绝大部分页面上还有一部分内容对页面主题也没有什么贡献，比如版权声明文字、导航条、广告等。以常见的博客导航为例，几乎每个博客页面上都会出现文章分类、历史存档等导航内容，但是这些页面本身与“分类”、“历史”这些词都没有任何关系。用户搜索“历史”，“分类”这些关键词时仅仅因为页面上有这些词出现而返回博客帖子是毫无意义的，完全不相关。所以这些区块都属于噪声，对页面主题只能起到分散作用。

搜索引擎需要识别并消除这些噪声，排名时不使用噪声内容。消噪的基本方法是根据 HTML 标签对页面分块，区分出页头、导航、正文、页脚、广告等区域，在网站上大量重复出现的区块往往属于噪声。对页面进行消噪后，剩下的才是页面主体内容。

去重

搜索引擎还需要对页面进行去重处理。

同一篇文章经常会重复出现在不同网站以及同一个网站的不同网址上，搜索引擎并不喜欢这种

重复性的内容。用户搜索时，如果在前两页看到的都是来自不同网站的同一篇文章，用户体验就太差了，虽然都是内容相关的。搜索引擎希望只返回相同文章中的一篇，所以在进行索引前还需要识别和删除重复内容，这个过程就称为去重。

去重的基本方法是对页面特征关键词计算指纹，也就是说从页面主体内容中选取最有代表性的一部分关键词（经常是出现频率最高的关键词），然后计算这些关键词的数字指纹。这里的关键词选取是在分词、去停止词、消噪之后。实验表明，通常选取 10 个特征关键词就可以达到比较高的计算准确性，再选取更多词对去重准确性提高的贡献也就不大了。

典型的指纹计算方法如 MD5 算法（信息摘要算法第五版）。这类指纹算法的特点是，输入（特征关键词）有任何微小的变化，都会导致计算出的指纹有很大差距。

了解了搜索引擎的去重算法，SEO 人员就应该知道简单地增加“的，地，得”、调换段落顺序这种所谓伪原创，并不能逃过搜索引擎的去重算法，因为这样的操作无法改变文章的特征关键词。而且搜索引擎的去重算法很可能不止于页面级别，而是进行到段落级别，混合不同文章、交叉调换段落顺序也不能使转载和抄袭变成原创。

正向索引

也可以简称为索引。

经过文字提取、分词、消噪、去重后，搜索引擎得到的就是独特的、能反映页面主体内容的、以词为单位的内容。接下来搜索引擎索引程序就可以提取关键词，按照分词程序划分好的词，把页面转换为一个关键词组成的集合，同时记录每一个关键词在页面上的出现频率、出现次数、格式（如出现在标题标签、黑体、H 标签、锚文字等）、位置（如页面第一段文字等）。这样，每一个页面都可以记录为一串关键词集合，其中每个关键词的词频、格式、位置等权重信息也都记录在案。

搜索引擎索引程序将页面及关键词形成词表结构存储进索引库。简化的索引词表形式如下表所示。

文件 1	关键词 1, 关键词 2, 关键词 7, 关键词 10.....关键词 L
文件 2	关键词 1, 关键词 7, 关键词 30.....关键词 M
文件 3	关键词 2, 关键词 70, 关键词 305.....关键词 N
.....	
文件 6	关键词 2, 关键词 7, 关键词 10.....关键词 X

文件 x	关键词 7, 关键词 50, 关键词 90.....关键词 Y
------	---------------------------------

每个文件都对应一个文件 ID，文件内容被表示为一串关键词的集合。实际上在搜索引擎索引库中，关键词也已经转换为关键词 ID。这样的数据结构就称为正向索引。

倒排索引

正向索引还不能直接用于排名。假设用户搜索关键词 2，如果只存在正向索引的话，排名程序需要扫描所有索引库中的文件，找出包含关键词 2 的文件，再进行相关性计算。这样的计算量无法满足实时返回排名结果的要求。

所以搜索引擎会将正向索引数据库重新构造为倒排索引，把文件对应到关键词的映射转换为关键词到文件的映射。如下表所示：

关键词 1	文件 1, 文件 2, 文件 15, 文件 58.....文件 1
关键词 2	文件 1, 文件 3, 文件 6.....文件 m
关键词 3	文件 5, 文件 700, 文件 805.....文件 n
.....	
关键词 7	文件 1, 文件 2, 文件 6.....文件 x
.....	
关键词 Y	文件 80, 文件 90, 文件 100.....文件 x

在倒排索引中关键词是主键，每个关键词都对应着一系列文件，这些文件中都出现了这个关键词。这样当用户搜索某个关键词时，排序程序在倒排索引中定位到这个关键词，就可以马上找出所有包含这个关键词的文件。

链接关系计算

链接关系计算也是预处理中很重要的一部分。现在所有的主流搜索引擎排名因素中都包含网页之间的链接流动信息。搜索引擎在抓取页面内容后，必须事前计算出页面上有哪些链接指向哪些其他页面？每个页面有哪些导入链接？链接使用了什么锚文字？这些复杂的链接指向关系形成了网站和页面的链接权重。

Google PR 值就是这种链接关系的最主要体现之一。其他搜索引擎也都进行类似计算，虽然他们并不称之为 PR。

由于页面和链接数量巨大，网上的链接关系又时时处在更新中，链接关系及 PR 的计算要耗费很长时间。关于 PR 和链接分析，后面还有专节介绍。

特殊文件处理

除了 HTML 文件外，搜索引擎通常还能抓取和索引以文字为基础的多种文件类型，如 PDF、Word、WPS、XLS、PPT、TXT 文件等。我们在搜索结果中也经常会看到这些文件类型。但目前的搜索引擎还不能处理图片、视频、Flash 这类非文字内容，也不能执行脚本和程序。

虽然搜索引擎在识别图片以及从 Flash 中提取文字内容方面有些进步，不过距离直接靠读取图片、视频、Flash 内容返回结果的目标还很远。对图片、视频内容的排名还往往是靠与之相关的文字内容，详细情况可以参考后面的整合搜索部分。

4-3 排名

经过搜索引擎蜘蛛抓取页面，索引程序计算得到倒排索引后，搜索引擎就准备好可以随时处理用户搜索了。用户在搜索框填入关键词后，排名程序调用索引库数据，计算排名显示给用户，排名过程是与用户直接互动的。

搜索词处理

搜索引擎接收到用户输入的搜索词后，需要对搜索词做一些处理，才能进入排名过程。搜索词处理包括几方面：

中文分词

与页面索引时一样，搜索词也必须进行中文分词，将查询字符串转换为以词为基础的关键词组合。分词原理与页面分词相同。

去停止词

和索引时一样，搜索引擎也需要把搜索词中的停止词去掉，最大限度地提高排名相关性及效率。

指令处理

查询词完成分词后，搜索引擎的缺省处理方式是在关键词之间使用“与”逻辑。也就是说用户搜索“减肥方法”时，程序分词为“减肥”和“方法”两个词，搜索引擎排序时缺省认为，用户寻找的是既包含“减肥”，也包含“方法”的页面。只包含“减肥”不包含“方法”，或者

只包含“方法”不包含“减肥”的页面，被认为是不符合搜索条件的。当然，这只是极为简化的为了说明原理的说法，实际上我们还是会看到只包含一部分关键词的搜索结果。

另外用户输入的查询词还可能包含一些高级搜索指令，如加号、减号等，搜索引擎都需要做出识别和相应处理。有关高级搜索指令，后面还有详细说明。

拼写错误矫正

用户如果输入了明显错误的字或英文单词拼错，搜索引擎会提示用户正确的用字或拼法，如下图所示。



图 26 输入的错拼、错字矫正

整合搜索触发

某些搜索词会触发整合搜索，比如明星姓名就经常触发图片和视频内容，当前的热门话题又容易触发资讯内容。哪些词触发哪些整合搜索，也需要在搜索词处理阶段计算。

文件匹配

搜索词经过处理后，搜索引擎得到的是以词为基础的关键词集合。文件匹配阶段就是找出含有所有关键词的文件。在索引部分提到的倒排索引使得文件匹配能够快速完成。

关键词 1	文件 1, 文件 2, 文件 15, 文件 58.....文件 1
关键词 2	文件 1, 文件 3, 文件 6.....文件 m
关键词 3	文件 5, 文件 700, 文件 805.....文件 n

.....	
关键词 7	文件 1, 文件 2, 文件 6.....文件 x
.....	
关键词 Y	文件 80, 文件 90, 文件 100.....文件 x

假设用户搜索“关键词 2 关键词 7”，排名程序只要在倒排索引中找到“关键词 2”和“关键词 7”这两个词，就能找到分别含有这两个词的所有页面。经过简单计算就能找出既包含“关键词 2”，也包含“关键词 7”的所有页面 - 文件 1 和文件 6。

初始子集的选择

找到包含所有关键词的匹配文件后，还不能进行相关性计算，因为找到的文件经常会有几十万几百万，甚至上千万个。要对这么多文件实时进行相关性计算，需要的时间还是比较长。

实际上用户并不需要知道所有匹配的几十万几百万个页面，绝大部分用户只会查看前两页，也就是前二十个结果。搜索引擎也并不需要计算这么多页面的相关性，而只要计算最重要的一部分页面就可以了。常用搜索引擎的人都会注意到，搜索结果页面通常最多只显示一百个。用户点击搜索结果页面底部的“下一页”链接，最多也只能看到第一百页，也就是一千个搜索结果。

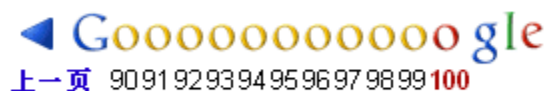


图 27 Google 显示 100 页搜索结果

百度则通常返回 76 页结果。

提示: 限于网页篇幅, 部分结果未予显示。

[上一页](#) [\[66\]](#) [\[67\]](#) [\[68\]](#) [\[69\]](#) [\[70\]](#) [\[71\]](#) [\[72\]](#) [\[73\]](#) [\[74\]](#) [\[75\]](#) 76

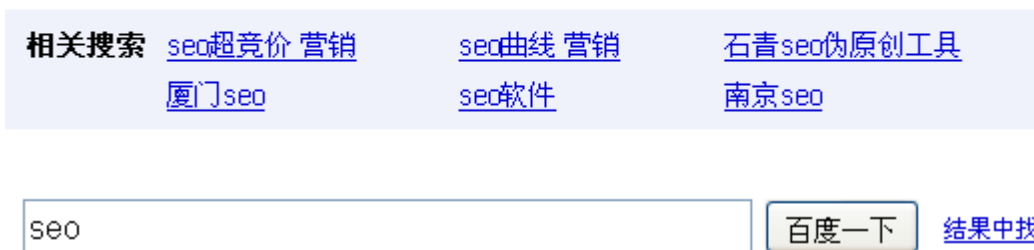


图 28 百度显示 76 页搜索结果

所以搜索引擎只需要计算前一千个结果的相关性, 就能满足要求。

但问题在于, 还没有计算相关性时, 搜索引擎又怎么知道哪一千个文件是最相关的? 所以用于最后相关性计算的初始页面子集的选择, 必须依靠其他特征而不是相关性, 其中最主要的就是页面权重。由于所有匹配文件都已经具备了最基本的相关性(这些文件都包含所有查询关键词), 搜索引擎通常会用非相关性的页面特征选出一个初始子集。初始子集的数目是多少? 几万个? 或许更多, 外人并不知道。不过可以肯定的是, 当匹配页面数目巨大时, 搜索引擎不会对这么多页面进行计算, 而必须选出页面权重较高的一个子集, 再对子集中的页面进行相关性计算。

相关性计算

选出初始子集后, 对子集中的页面计算关键词相关性。计算相关性是排名过程中最重要的一步。相关性计算是搜索引擎算法中最令 SEO 感兴趣的部分。

影响相关性的主要因素包括几方面。

关键词常用程度

经过分词后的多个关键词, 对整个搜索字符串的意义贡献并不相同。越常用的词对搜索词的意义贡献越小, 越不常用的词对意义贡献越大。举个例子, 假设用户输入的搜索词是“我们冥王星”。“我们”这个词常用程度非常高, 在很多页面上会出现, 它对“我们冥王星”这个搜索词的辨识程度和意义相关度贡献就很小。找出那些包含“我们”这个词的页面, 对搜索排名相关性几乎没有什么影响, 有太多页面包含“我们”这个词。

而“冥王星”这个词常用程度就比较低，对“我们冥王星”这个搜索词的意义贡献要大得多。那些包含“冥王星”这个词的页面，对“我们冥王星”这个搜索词会更为相关。

常用词的极致就是停止词，对页面意义完全没有影响。

所以搜索引擎对搜索词串中的关键词并不是一视同仁地处理，而是根据常用程度进行加权。不常用的词加权系数高，常用词加权系数低，排名算法对不常用的词给予更多关注。

我们假设 A、B 两个页面都各出现“我们”及“冥王星”两个词。但是“我们”这个词在 A 页面出现于普通文字，“冥王星”这个词在 A 页面出现于标题标签中。B 页面正相反，“我们”出现在标题标签中，而“冥王星”出现在普通文字中。那么针对“我们冥王星”这个搜索词，A 页面将更相关。

词频及密度

一般认为在没有关键词堆积的情况下，搜索词在页面中出现的次数多，密度比较高，说明页面与搜索词越相关。当然这只是一个大致规律，实际情况未必如此，所以相关性计算还有其他因素。出现频率及密度只是因素的一部分，而且重要程度越来越低。

关键词位置及形式

像在索引部分中提到的，页面关键词出现的格式和位置都被记录在索引库中。关键词出现在比较重要位置，如标题标签、黑体、H1 等，说明页面与关键词越相关。这一部分就是页面 SEO 所要解决的。

关键词距离

切分后的关键词完整匹配出现，说明与搜索词最相关。比如搜索“减肥方法”时，页面上连续完整出现“减肥方法”四个字是最相关的。如果“减肥”和“方法”两个词没有连续匹配出现，出现的距离近一些，也被搜索引擎认为相关性稍微大一些。

链接分析及页面权重

除了页面本身的因素，页面之间的链接和权重关系也影响关键词的相关性，其中最重要的是锚文字。页面有越多以搜索词为锚文字的导入链接，说明页面的相关性越强。

链接分析还包括了链接源页面本身的主题，锚文字周围的文字等。

上面简单介绍的几个因素在本书中都有更详细说明。

排名过滤及调整

选出匹配文件子集、计算相关性后，大体排名就已经确定了。之后搜索引擎还可能有一些过滤算法，对排名进行轻微调整，其中最主要的过滤就是施加惩罚。一些有作弊嫌疑的页面，虽然按照正常的权重和相关性计算排到前面，但搜索引擎的惩罚算法却可能在最后一步把这些页面调到后面去。典型的例子是百度的 11 位，Google 的负 6，负 30，负 950 等算法。

排名显示

所有排名确定后，排名程序调用原始页面的标题标签、说明标签、快照日期等数据显示在页面上。有时搜索引擎需要动态生成页面摘要，而不是调用页面本身的说明标签。

搜索缓存

用户搜索的关键词有很大一部分是重复的。按照 2/8 定律，20% 的搜索词占到了总搜索次数的 80%。按照长尾理论，最常见的搜索词没有占到 80% 那么多，但通常也有一个比较粗大的头部，很少一部分搜索词占到了所有搜索次数的很大一部分。尤其是有热门新闻发生时，每天可能有几百万人搜索完全相同的关键词。

如果每次搜索都重新处理排名可以说是很大的浪费。搜索引擎会把最常见的搜索词存入缓存，用户搜索时直接从缓存中调用，而不必经过文件匹配和相关性计算，大大提高排名效率，降低搜索反应时间。

查询及点击日志

搜索用户的 IP 地址，搜索的关键词，搜索时间以及点击了哪些结果页面，搜索引擎都记录形成日志。这些日志文件中的数据对搜索引擎判断搜索结果质量，调整搜索算法，预期搜索趋势等都有重要意义。

上面我们简单介绍了搜索引擎的工作过程。当然实际搜索引擎的工作步骤与算法是非常非常复杂的。上面的说明很简单，但其中有很多技术难点。

搜索引擎还在不断优化算法，优化数据库格式。不同搜索引擎的工作步骤也会有差异。但大致上所有主流搜索引擎的基本工作原理都是如此，在过去几年以及可以预期的未来几年，都不会有实质性改变。

5 链接原理

在 Google 诞生以前，传统搜索引擎主要依靠页面内容中的关键词匹配搜索词进行排名。这种排名方式的短处现在看来显而易见，那就是很容易被刻意操纵。黑帽 SEO 在页面上堆积关键词，或加入与主题无关的热门关键词，都能提高排名，使搜索引擎排名结果质量大为下降。现在的搜索引擎都使用链接分析技术减少垃圾，提高用户体验。这一节就简要探讨链接在搜索引擎排名中的应用原理。

在排名中计入链接因素，不仅有助于减少垃圾，提高结果相关性，也使传统关键词匹配无法排名的文件能够被处理。比如图片、视频文件无法进行关键词匹配，但是却可能有外部链接，通过链接信息，搜索引擎就可以了解图片和视频的内容从而排名。

不同文字的页面排名也成为可能。比如在百度或 google.cn 搜索“SEO”，都可以看到英文和其他文字的 SEO 网站。甚至搜索“搜索引擎优化”，也可以看到非中文页面，原因就在于有的链接可能使用“搜索引擎优化”为锚文字指向英文页面。

链接因素现在已经超过页面内容的重要性。不过理解链接关系比较抽象。页面上的因素对排名的影响能看得到，容易直观理解。举个简单例子，搜索一个特定关键词，SEO 人员只要观察前几页结果，就能看到关键词在标题标签中出现有什么影响？出现在最前面又有什么影响？有技术资源的还可以大规模的统计，计算出关键词出现在标题标签中不同位置与排名之间的关系。虽然这种关系不一定是因果关系，但至少是统计上的联系，使 SEO 人员大致了解如何优化。

链接对排名的影响就无法直观了解，也很难进行统计，因为没有人能获得搜索引擎的链接数据库。我们能做的最多只是定性观察和分析。

下面介绍的一些关于链接的专利，多少透露了链接在搜索引擎排名中的使用方法和地位。

5-1 李彦宏超链分析专利

百度创始人李彦宏在回国创建百度之前就是美国最顶级的搜索引擎工程师之一。据说李彦宏在寻找风险投资时，投资人询问其他三个搜索引擎业界的技术高人一个问题：要了解搜索引擎技术应该问谁。这三个被问到的高人中有两个回答：搜索引擎的事就问李彦宏。由此投资人断定李彦宏是最了解搜索引擎的人之一。

这其实就是现实生活中类似于链接关系的应用。要判断哪个页面最有权威性，不能光看页面自己怎么说，而要看其他页面怎么评价。

李彦宏 1997 年就提交了一份名为“超链文件检索系统和方法”的专利申请，这比 Google 创始人发明 PR 要早得多，不得不说这是非常具有前瞻性的研究工作。在这份专利中，李彦宏提出了与传统信息检索系统不同的基于链接的排名方法。

这个系统除了索引页面之外，还建立一个链接词库，记录链接锚文字的一些相关信息，如锚文字中包含哪些关键词，发出链接的页面索引，包含特定锚文字的链接总数，包含特定关键词的链接都指向哪些页面。词库不仅包含关键词原型，也包含同一个词干的其他衍生关键词。

根据这些链接数据，尤其是锚文字，计算出基于链接的文件相关性。在用户搜索时，将得到的基于链接的相关性与基于关键词匹配的传统相关性综合使用，得到更准确的排名。

在今天看来，这种基于链接的相关性计算是搜索引擎的常态，每个 SEO 人员都知道。但是在十三四年前，这无疑是非常创新的概念。当然现在的搜索引擎算法对链接的考虑，已经不仅仅是锚文字，而要复杂得多。

这份专利所有人是李彦宏当时所在的公司，发明人是李彦宏本人。感兴趣的读者可以在这个地址查看美国专利局发布的“超链文件检索系统和方法”专利详情：

<http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=5,920,859>

5-2 HITS 算法

HITS 是英文 Hyperlink-Induced Topic Search 的缩写，意译为超链诱导主题搜索。HITS 算法由 Jon Kleinberg 于 1997 年提出，并申请了专利：

<http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=6,112,202>

按照 HITS 算法，用户输入关键词后，算法对返回的匹配页面计算两种值，一是枢纽值（Hub Scores），二是权威值（Authority Scores），这两个值是互相依存、互相影响的。所谓枢纽值指的是页面上所有导出链接指向页面的权威值之和。权威值指的是所有导入链接所在页面的枢纽值之和。

上面的定义比较拗口，我们可以简单的说，HITS 算法会提炼出两种比较重要的页面，也就是枢纽页面和权威页面。枢纽页面本身可能没有多少导入链接，但是有很多导出链接指向权威页面。权威页面本身可能导出链接不多，但是有很多来自枢纽页面的导入链接。

典型的枢纽页面就是如雅虎目录、开放目录或好 123 这样的网站目录。这种高质量的网站目录作用就在于指向其他权威网站，所以称为枢纽。而权威页面有很多导入链接，其中包含很多来自枢纽页面的链接。权威页面通常是提供真正相关内容的页面。

HITS 算法是针对特定查询词的，所以称为主题搜索。

HITS 算法的最大缺点是，它是在查询阶段进行计算，而不是在抓取或预处理阶段。所以 HITS

算法是以牺牲查询排名响应时间为代价的。也正因为如此，原始 HITS 算法在搜索引擎中并不常用。不过 HITS 算法的思想很可能融入到搜索引擎的索引阶段，也就是根据链接关系找出具有枢纽特征或权威特征的页面。

成为权威页面是第一优先，不过难度比较大，唯一的方法就是获得高质量链接。当你的网站不能成为权威页面时，就让它成为枢纽页面。所以导出链接也是当前搜索引擎排名因素之一。绝不链接到其他网站的做法，并不是好的 SEO 方法。

5-3 TrustRank 算法

TrustRank 是近年来比较受关注的基于链接关系的排名算法。TrustRank 中文可以翻译为信任指数。

TrustRank 算法最初来自于 2004 年斯坦福大学和雅虎的一项联合研究，用来检测垃圾网站，并且于 2006 年申请专利。TrustRank 算法发明人还发表了一份专门的 PDF 文件，说明 TrustRank 算法的应用。感兴趣的读者可以在这个网址下载 PDF 文件：

<http://www.vldb.org/conf/2004/RS15P3.PDF>

TrustRank 算法并不是由 Google 提出，不过由于 Google 所占市场份额最大，而且 TrustRank 在 Google 排名中也是一个非常重要的因素，所以有些人误以为 TrustRank 是 Google 提出的。更让人糊涂的是，Google 曾经把 TrustRank 申请为商标，但是 TrustRank 商标中的 TrustRank 指的是 Google 检测含有恶意代码网站的方法，而不是指排名算法中的信任指数。

TrustRank 算法基于一个基本假设：好的网站很少会链接到坏的网站。反之则不成立，也就是说，坏的网站很少链接到好网站这句话并不成立。正相反，很多垃圾网站会链接到高权威、高信任指数的网站，意图提高自己的信任指数。

基于这个假设，如果能挑选出可以百分之百信任的网站，这些网站的 TrustRank 评为最高，这些 TrustRank 最高的网站所链接到的网站信任指数稍微降低，但也会很高。与此类似，第二层被信任的网站链接出去的第三层网站，信任度继续下降。由于种种原因，好的网站也不可避免地会链接到一些垃圾网站，不过离第一层网站点击距离越近，所传递的信任指数越高，离第一级网站点击距离越远，信任指数将依次下降。这样，通过 TrustRank 算法，就能给所有网站计算出相应的信任指数，离第一层网站越远，成为垃圾网站的可能性就越大。

计算 TrustRank 值首先要选择一批种子网站，然后人工查看网站，设定一个初始 TrustRank 值。挑选种子网站有两种方式，一是选择导出链接最多的网站，因为 TrustRank 算法就是计算指数随着导出链接的衰减。导出链接多的网站，在某种意义上可以理解为“逆向 PR 值”比较高。

另一种挑选种子网站的方法是选 PR 值高的网站，因为 PR 值越高，在搜索结果页面出现的概率就越大。这些网站才正是 TrustRank 算法最关注的、需要调整排名的网站。那些 PR 值很低的页面，在没有 TrustRank 算法时排名也很靠后，计算 TrustRank 意义就不大了。

根据测算，挑选出两百个左右网站作为种子，就可以比较精确地计算出所有网站的 TrustRank 值。

计算 TrustRank 随链接关系减少的公式有两种方式。一是随链接次数衰减，也就是说第一层页面 TrustRank 指数是一百的话，第二层页面衰减为 90，第三层衰减为 80。第二种计算方法是按导出链接数目分配 TrustRank 值，也就是说一个页面的 TrustRank 值是一百，页面上有 5 个导出链接的话，每个链接将传递 20% 的 TrustRank 值。衰减和分配两种计算方法通常综合使用，整体效果都是随着链接层次的增加，TrustRank 值逐步降低。

得出网站和页面的 TrustRank 值后，可以通过两种方式影响排名。一是把传统排名算法挑选出的多个页面，根据 TrustRank 值比较，重新做排名调整。二是设定一个最低 TrustRank 值门槛，只有超过这个门槛 TrustRank 值的页面，才被认为有足够的质量进入排名，低于门槛的页面将被认为是垃圾页面，从搜索结果中过滤出去。

虽然 TrustRank 算法最初是作为检测垃圾的方法，但在现在的搜索引擎排名算法中，TrustRank 概念使用更为广泛，常常影响大部分网站的整体排名。TrustRank 算法最初是针对页面级别，现在在搜索引擎算法中，TrustRank 值也通常表现在域名级别，整个域名的信任指数越高，整体排名能力就越强。

5-4 Google PR

PR 是 PageRank 的缩写。Google PR 理论是所有基于链接的搜索引擎理论中最有名的。SEO 人员可能不清楚本节介绍的其他链接理论，但不可能不知道 PR。

PR 是 Google 创始人之一拉里佩奇发明的，用于表示页面重要性的概念。用最简单的话说就是，反向链接越多的页面就是最越重要的页面，因此 PR 值也越高。

Google Pr 有点类似于科技文献中互相引用的概念，被其他文献引用最多的文献，很可能是比较重要的文献。

PR 的概念和计算

我们可以把互联网理解为由节点及链接组成的有向图，页面就是一个节点，页面之间的有向链接传递着页面的重要性。一个链接传递的 PR 值决定于导入链接所在页面的 PR 值，发出链接的页面本身 PR 值越高，所能传递出去的 PR 也越高。传递的 PR 数值也取决于页面上的导出链

接数目。对于给定 PR 值的页面来说，假设能传递到下级页面 100 份 PR，页面上有 10 个导出链接，每个链接能传递 10 份 PR，页面上有 20 个导出链接的话，每个链接只能传递 5 份 PR。所以一个页面的 PR 值取决于导入链接总数，发出链接页面的 PR 值，以及发出链接页面上的导出链接数目。

PR 值计算公式是：

$$PR(A) = (1-d) + d(PR(t1)/C(t1) + \dots + PR(tn)/C(tn))$$

- A 代表页面 A
- PR(A) 则代表页面 A 的 PR 值
- d 为阻尼指数。通常认为 $d=0.85$
- $t1...tn$ 代表链接向页面 A 的页面 $t1$ 到 tn
- C 代表页面上的导出链接数目。C(t1) 即为页面 $t1$ 上的导出链接数目。

从概念及计算公式都可以看到，计算 PR 值必须使用迭代计算。页面 A 的 PR 值取决于链接向 A 的页面 $t1$ 至 tn 页面的 PR 值，而 $t1$ 至 tn 页面的 PR 值又取决于其他页面的 PR 值，其中很可能还包含页面 A。所以 PR 需要多次迭代才能得到。计算时先给所有页面设定一个初始值，经过一定次数的迭代计算后，各个页面的 PR 值将趋于稳定。研究证明，无论初始值怎么选取，经过迭代计算的最终 PR 值不会受到影响。

对阻尼系数做个简要说明。考虑如图这样一个循环（实际网络上是一定存在这种循环的）：

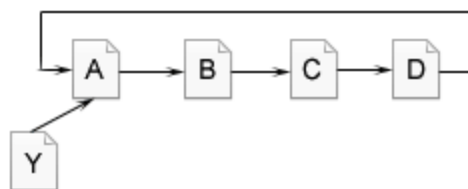


图 29 链接构成的循环

外部页面 Y 向循环注入 PR 值，循环中的页面不停迭代传递 PR，没有阻尼系数的话，循环中的页面 PR 将达到无穷大。引入阻尼系数，使 PR 在传递时自然衰减，才能将 PR 计算稳定在一个值上。

PR 的两个比喻模型

关于 PR 有两个著名的比喻。一个比喻是投票。链接就像民主投票一样，A 页面链接到 B 页面，

就意味着 A 页面对 B 页面投了一票，使得 B 页面的重要性提高。同时，A 页面本身的 PR 值决定了 A 所能投出去的投票力，PR 值越高的页面，投出的票也更重要。在这个意义上，传统基于关键词匹配的算法是看页面自己说页面内容是什么，基于链接的 PR 则是看别人怎么评价一个页面。

第二个比喻是随机冲浪比喻。假设一个访问者从一个页面开始，不停地随机点击链接，访问下一个页面。有时候这个用户感到无聊了，不再点击链接，就随机跳到了另外一个网址，再次开始不停地向下点击。所谓 PR 值也就是一个页面在这种随机冲浪访问中被访问到的概率。一个页面导入链接越多，被访问到的概率也越高，因此 PR 值也越高。

阻尼系数也与随机冲浪模型有关。 $(1-d)=0.15$ 实际上就是用户感到无聊，停止点击，随机跳到新 URL 的概率。

工具条 PR

真正的用于排名计算的 Google PR 值我们是无法知道的，我们所能看到的只是 Google 工具条 PR 值。需要清楚的是，工具条 PR 值并不是真实 PR 值的精确反应。真实 PR 值是一个准确的、大于 0.15、没有上限的数字，工具条上显示的 PR 值已经简化为 0-10 十一个数字，是一个整数，也就是说 PR 值最小的近似为 0，最大的近似为 10。实际上每一个工具条 PR 值代表的是很大一个范围，工具条 PR5 代表的页面真实 PR 值可能相差很多倍。

真正的 PR 值是不间断计算更新中的，工具条 PR 值只是某一个时间点上真实 PR 值的快照输出。工具条 PR 几个月才更新一次，过去一年工具条 PR 值更新的日期如下表所示。

2010 年 4 月 1 号
2009 年 12 月 31 号
2009 年 10 月 29 号
2009 年 6 月 23 号
2009 年 5 月 26 号
2009 年 4 月 1 号
2008 年 12 月 31 号

工具条 PR 与反向链接数目呈对数关系，而不是线性关系。也就是说从 PR1 到 PR2 需要的外部链接是 100 个的话，从 PR2 到 PR3 则需要大致 1000 个，PR5 到 PR6 需要的外部链接则更多。所以 PR 值越高的网站想提升一级所要付出的时间和努力比 PR 值比较低的网站提升一级要多得

多。

关于 PR 的几个误解

PR 的英文全称是 PageRank。这个名称来源于发明人佩奇 (Page) 的名字, 巧合的是 Page 在英文中也是页面的意思。所以准确地说 PageRank 这个名字应该翻译为佩奇级别, 而不是页面级别。不过约定俗成, 再加上形成巧妙的一语双关, 大家都把 PR 称为页面级别。

PR 值只与链接有关。经常有站长询问, 他的网站做了挺长时间, 内容也全是原创, 怎么 PR 还是零呢? 其实 PR 与站长是否认真、建站多少时间、内容是否原创都没有直接关系。有反向链接就有 PR, 没有反向链接就没有 PR。一个高质量的原创网站, 一般来说自然会吸引到比较多的外部链接, 所以会间接提高 PR 值, 但这并不是必然的。

工具条 PR 值更新与页面排名变化在时间上没有对应关系。在工具条 PR 值更新过程中, 经常有站长说 PR 值提高了, 难怪网站排名也提高了。肯定的说这只是时间上的巧合而已。前面说过, 真实的用于排名计算的 PR 是连续计算更新的, 随时计入排名算法。我们看到的工具条 PR 几个月才更新一次, 当我们看到有 PR 更新时, 真实的 PR 早在几个月之前就更新和计入排名里了。所以, 通过工具条 PR 变化, 研究 PR 值与排名变化之间的关系是没有意义的。

PR 的意义

Google 工程师说过很多次, Google PR 现在已经是一个被过度宣传的概念, 其实 PR 只是 Google 排名算法 200 多个因素之一, 而且重要性已经下降很多, SEO 人员完全不必太执着于 PR 值的提高。

当然, PR 还是 Google 排名算法中的重要因素之一。除了直接影响排名, PR 的重要性还体现在下面几点。

网站收录深度和总页面数。搜索引擎蜘蛛爬行时间以及数据库的空间都是有限的。Google 希望尽量优先收录重要性高的页面, 所以 PR 值越高的网站就能被收录更多页面, 蜘蛛爬行内页的深度也更高。对大中型网站来说, 首页 PR 值是带动网站收录的重要因素之一。

更新频率。PR 值越高的网站, 搜索引擎蜘蛛访问得就越频繁, 网站上出现新页面或旧页面上内容更新时, 都能更快速被收录。由于网站新页面通常都会在现有页面上出现链接, 更新频率高也就意味着被发现的速度快。

重复内容判定。当 Google 在不同网站上发现完全相同的内容时, 会选出一个作为原创, 其他作为转载或抄袭。用户搜索相关关键词时, 被判断为原创的那个版本会排在前面。而判断哪个版本为原创时, PR 值也是重要因素之一。这也就是为什么那些权重高、PR 值高的大网站, 转载

小网站内容却经常被当作原创的原因。

排名初始子集的选择。前面介绍排名过程时提到，搜索引擎挑选出所有与关键词匹配的文件后，不可能对所有文件进行相关性计算，因为返回的文件可能有几百万几千万，搜索引擎需要从中挑选出一个初始子集再做相关性计算。初始子集的选择显然与关键词相关度无关，而只能从页面的重要程度着手，PR 值就是与关键词无关的重要度指标。

现在的 PR 算法比当初拉里佩奇专利中的描述肯定有了改进和变化。一个可以观察到的现象是，PR 算法应该已经排除了一部分 Google 认为可疑或者无效的链接，比如付费链接，博客和论坛中的垃圾链接等。所以有时候我们会看到一个页面有 PR6 甚至 PR7 的导入链接，经过几次工具条 PR 更新后，却还维持在 PR3 甚至 PR2。按说一个 PR6 或 7 的链接，应该把被链接的页面带到 PR5 或 PR4，所以很可能 Google 已经把一部分它认为可疑的链接排除在 PR 计算之外。

PR 专利发明人是拉里佩奇，专利所有人是斯坦福大学，Google 公司拥有永久性排他使用权。虽然 PR 是 Google 拥有专利使用权的算法，但其他所有主流搜索引擎也都有类似算法，只不过不称为 PR 而已。

5-5 Hilltop 算法

Hilltop 算法由 Krishna Baharat 在 1999 年到 2000 年左右所研究，于 2001 年申请了专利，并且把专利授权给 Google 使用，后来 Krishna Baharat 本人也加入了 Google。

Hilltop 算法可以简单理解为与主题相关的 PR 值。传统 PR 值与特定关键词或主题没有关联，只计算链接关系。这就有可能出现某种漏洞。比如一个 PR 值极高的关于环保内容的大学页面，上面有一个链接连向一个儿童用品网站，这个链接出现的原因可能仅仅是因为这个大学页面维护人是个教授，他太太在那个卖儿童用品的公司工作。这种与主题无关，却有着极高 PR 值的链接，有可能使一些网站获得很好排名，但其实相关性并不高。

Hilltop 算法就尝试矫正这种可能出现的疏漏。Hilltop 算法同样是计算链接关系，不过它更关注来自主题相关页面的链接权重。在 Hilltop 算法中把这种主题相关页面称为专家文件。显然，针对不同主题或搜索词有不同的专家文件。

根据 Hilltop 算法，用户搜索关键词后，Google 先按正常排名算法找到一系列相关页面并排名，然后计算这些页面有多少来自专家文件的、与主题相关的链接，来自专家文件的链接越多，页面的排名分值越高。按 Hilltop 算法的最初构想，一个页面至少要有两个来自专家文件的链接，才能返回一定的 Hilltop 值，不然返回的 Hilltop 值将为零。

根据专家文件链接计算的分值被称为 LocalRank。排名程序根据 LocalRank 值，对原本传统排

名算法计算的排名做重新调整，给出最后排名。这就是前面讨论的搜索引擎排名阶段最后的过滤和调整步骤。

Hilltop 算法最初论文和申请专利时对专家文件的选择有不同描述。在最初的研究中，Krishna Baharat 把专家文件定义为包含特定主题内容，并且有比较多导出链接到第三方网站的页面，这有点类似于 HITS 算法中的枢纽页面。专家文件链接指向的页面与专家文件本身应该没有关联，这种关联指的是来自同一个主域名下的子域名，来自相同或相似 IP 地址的页面等。最常见的专家文件经常来自于学校、政府以及行业组织网站。

在最初的 Hilltop 算法中，专家文件是预先挑选的。搜索引擎可以根据最常见的搜索词，预先计算出一套专家文件，用户搜索时，排名算法从事先计算的专家文件集合中选出与搜索词相关的专家文件子集，再从这个子集中的链接计算 LocalRank 值。

不过在 2001 年所申请的专利中，Krishna Baharat 描述了另外一个挑选专家文件的方法，专家文件并不预先选择，用户搜索特定查询词后，搜索引擎按传统算法挑出一系列初始相关页面，这些页面就是专家文件。Hilltop 算法在这个页面集合中再次计算哪些网页有来自于集合中其他页面的链接，赋予比较高的 LocalRank 值。由于传统算法得到的页面集合已经具备了相关性，这些页面再提供链接给某一个特定页面，这些链接的权重自然应该很高。这种挑选专家文件的方法是实时进行的。

通常认为 Hilltop 算法对 2003 年底的佛罗里达更新有重大影响，不过 Hilltop 算法是否真的已经被融入进 Google 排名算法中，没有人能够确定。Google 从来没有承认，也没有否认自己的排名算法中是否使用了某项专利。不过从排名结果观察以及招揽 Krishna Baharat 至麾下等现象看，Hilltop 算法的思想得到了 Google 的极大重视。

Hilltop 算法提示 SEO，建设外部链接时更应该关注主题相关的网站。最简单的方法是搜索某个关键词，目前排在前面的页面就是最好的链接来源，甚至可能一个来自竞争对手网站的链接效果是最好的。当然，获得这样的链接难度最大。

6 用户怎样浏览和点击搜索结果

用户搜索关键词后，搜索引擎通常返回 10 个结果。用户对这 10 个结果列表的浏览和点击有很大差别。这一节介绍用户在搜索结果页面上的浏览方式，包括目光关注度及点击的一些研究。

6-1 英文搜索结果页面

页面浏览最主要的研究方法是视线跟踪（eye-tracking），使用特殊的设备跟踪用户目光在结果页面上的浏览及点击数据。enquiro.com 就是专门做这方面实验及统计的公司。2005 年初，enquiro.com 联合 eyetools.com 和 did-it.com 两家公司进行了一次很著名的视线跟踪实验，实验数据于 2005 年 6 月发表，提出在 SEO 业界很有名的金三角图像，也有人称为 F 型浏览图像。

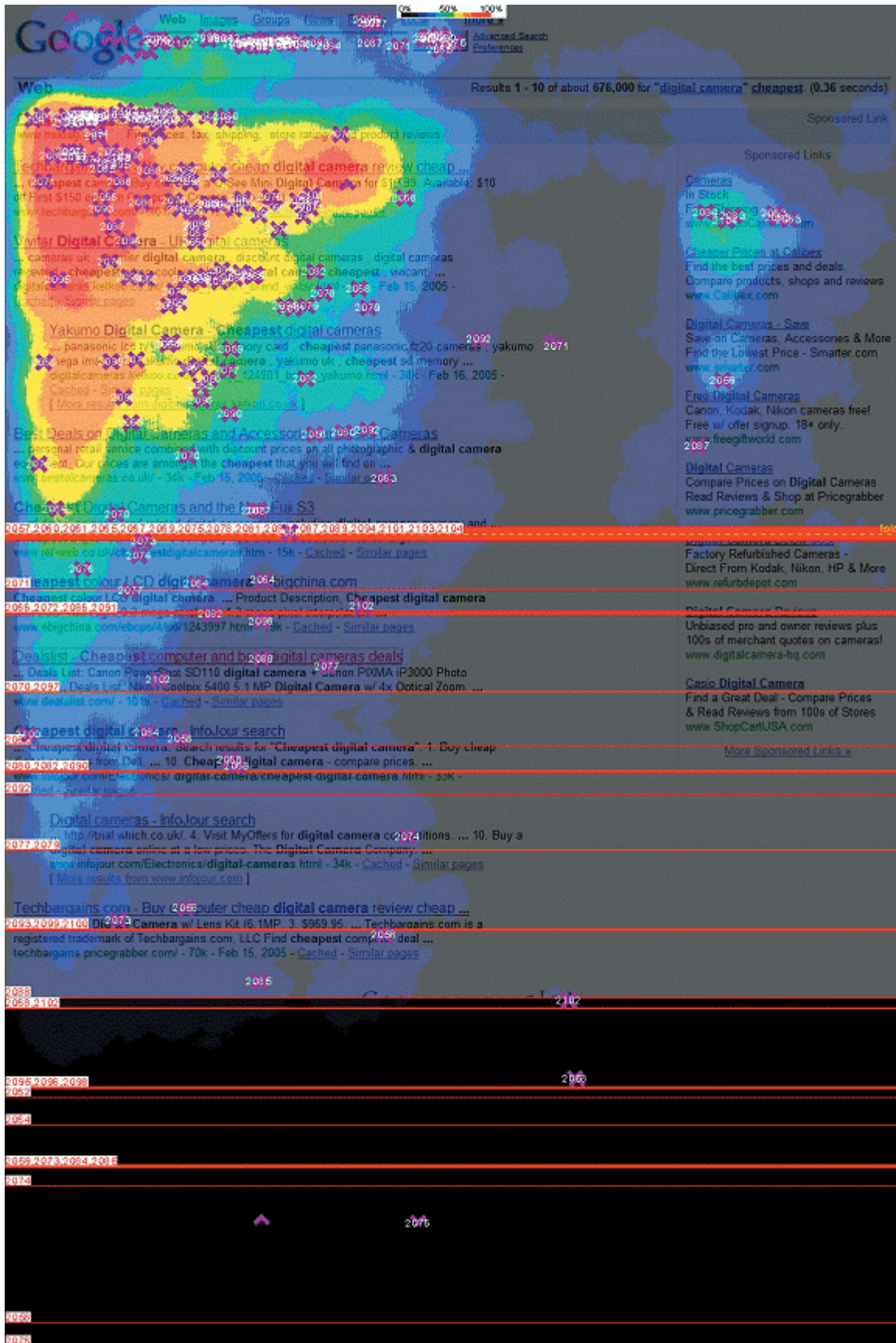


图 30 著名的用户视线分布金三角

如上图所示，颜色区块代表用户目光的停留位置以及关注时间，图像中的 X 号代表点击。从图中我们可以看到，典型搜索用户打开搜索结果页面后，目光会首先放在最左上角，然后向正下方移动挨个浏览搜索结果，当看到感兴趣的页面时，横向向右阅读页面标题。排在最上面的结果得到的目光关注度最多，越往下越少，因此形成一个所谓的金三角。金三角中的搜索结果都有比较高的目光关注度。这个金三角结束于第一屏最底部的排名结果，用户向下拉页面查看第二屏结果的概率大为降低。

这个浏览统计是针对 Google 搜索结果页面做的。后来 enquiro.com 针对雅虎及 MSN 搜索结果页面做的实验也得到大致相同的结果，如下图所示。

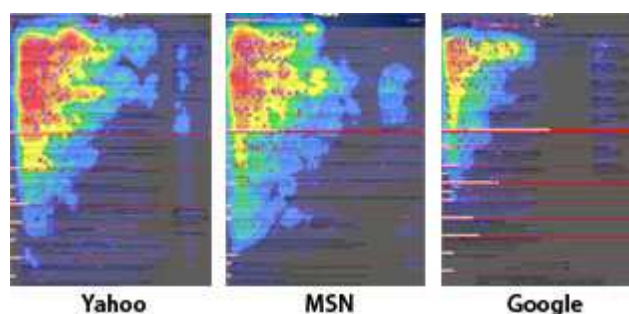


图 31 主流搜索引擎都存在视线分布金三角

2009 年 Google 官方博客也发布了一个类似的目光跟踪实验结果，确认了 enquiro.com 的金三角图像。Google 的实验结果如下图所示。



图 32 Google 官方发布的视线分布金三角

2006 年 10 月，康奈尔大学做了更进一步的实验和统计，记录 397 次实验对象（搜索用户）对搜索结果的关注时间及点击分布，实验数据如下图所示。

	% of Clicks	% Time Spent
Something	56.36	28.43
Something	13.45	25.08
Something	9.82	14.72
Something	4.00	8.70
Something	4.73	6.02
Something	3.27	4.01
Something	0.36	3.01
Something	2.91	3.68
Something	1.45	3.01
Something	2.55	2.34

图 33 康奈尔大学实验显示的搜索结果关注时间及点击分布

我们可以看到，排名前三位的页面得到的关注时间相差不大，尤其是前两位差距很小，但是点击次数却有很大差异。排名第一的结果占据了 56.36% 的点击，排名第二的结果只有不到第一位四分之一的点击量，从第四位以后点击率更是急剧下降。唯一的特例是排名第十的点击结果，比第九位稍微多了一点。原因可能是用户浏览到最后一个结果时没有更多结果可看，也没有其他选择，也就点击了最后一个页面。

中间还有一个值得注意的结果，排名第七位的页面点击率非常低，只占 0.36%。这是因为前六位结果都处在第一屏，用户在第一屏没有找到满意结果的话，就会拉动右侧滑动条看第二屏内容。不过大多数用户不会刚刚好把屏幕下拉到第七位结果排在最上面，而是直接拉动到页面最下面，这样第七位结果反倒已经跑到第二屏之外，很多用户根本没看到第七位排名页面。

下图显示的是把关注时间及点击次数按曲线显示：

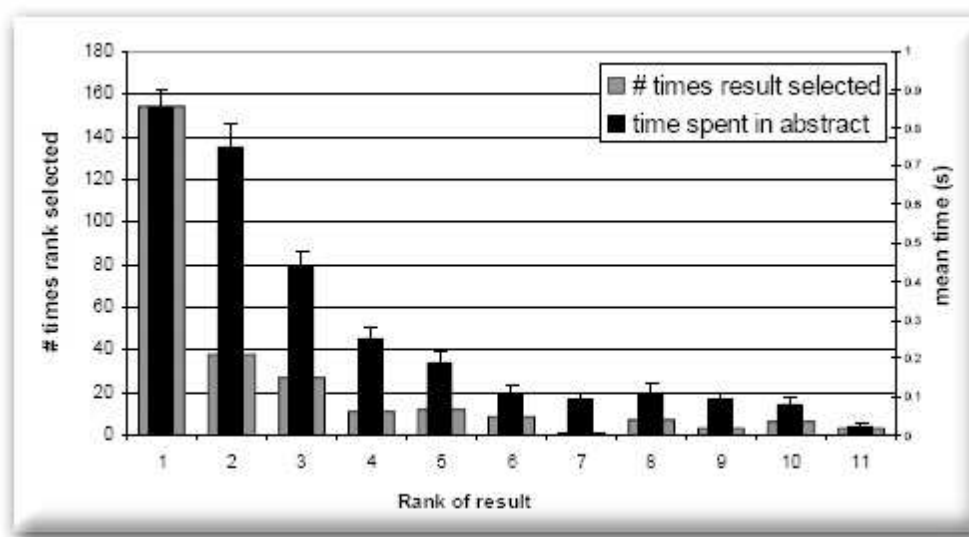


图 34 关注时间及点击次数按曲线显示

我们可以更清楚地看到，关注时间是按比较连续的曲线下下降，而点击次数在第一和第二位结果上有了巨大的差别，从第二位之后才形成比较平滑的曲线。

上面的实验数据来自于对用户搜索的观察记录，还不是来自搜索引擎的真正点击数据，而且样本数量有限。

2006年8月，美国在线（AOL）因为疏忽公布了三个月的真实搜索记录，包括2006年3月1号到5月31号1900万次搜索，1080多万不同的搜索词，还包括65万8000个用户ID。这份资料公布后引起轩然大波。虽然用户ID都是匿名的，但是搜索词本身就可能泄露个人隐私，使得有心人士可以从这份资料中挖掘出不少能与具体个人相关联的数据。

也有SEO人士对这些搜索记录做了大量统计，得出搜索结果页面的真实点击数据。有人从9038794个搜索中统计到4926623次点击，这些点击在前十个结果中的分布如下表所示。

页面排名	点击次数	占点击总数比例
1	2,075,765	42.1 %
2	586100	11.90%
3	418643	8.50%
4	298532	6.10%
5	242169	4.90%

6	199541	4.10%
7	168080	3.40%
8	148489	3.00%
9	140356	2.80%
10	147551	3.00%

把表格中的点击比例换算成相对于第一位结果点击减少的百分比，又可以得到下面表格。

页面排名	点击次数	比第一位点击减少倍数
1	2,075,765	
2	586100	3.5
3	418643	4.9
4	298532	6.9
5	242169	8.5
6	199541	10.4
7	168080	12.3
8	148489	14.0
9	140356	14.8
10	147551	14.1

从这个数据中可以看到，第一页点击分布与康奈尔大学的数据大体相当。排名第一的结果获得了 42.1% 的点击，排名第二的结果点击次数大幅下降，不到第一位的四分之一。排名第一页的 10 个结果，总共获得所有点击流量的 89.71%。第二页排名第 11 到 20 的结果，得到 4.37% 的点击。第三页只得到 2.42%。前 5 页占据了 99% 以上的点击。

这是目前为止我们所能看到的唯一一份来自搜索引擎的真实点击数据，对 SEO 有很大的参考价值。

比如，同样是提高一位排名，从第十提高到第九，与从第二提高到第一位获得的流量差距有天壤之别。很多公司和 SEO 人士把排名进入前十或前五当作目标，但实际上第十名或第五名与第一名流量上的差距非常大。这就给我们一个启示，有的时候我们可以找到网站有哪些关键词排名是在第二位，想办法把它提高到第一位，能使流量翻好几倍。

这两个搜索结果点击数据，对 SEO 人员预估流量也有重要意义。

6-2 中文搜索结果页面

上面介绍的目光跟踪及点击数据，都是针对英文网站及美国用户。那么中文搜索引擎情况如何呢？

2007年4月，enquiro.com做了 google.cn 及百度搜索结果页面实验。参加实验的是50个18-25岁的中国留学生，这些留学生来到美国不超过几个星期，正在就读语言培训班，所以其浏览习惯大体上还与主流中文用户相同，没有受英文用户浏览习惯太大的影响。这次试验的结果如下图所示。



图 35 中文用户视线分布与英文用户对比

上图是英文 google.com 与中文 google.cn 的对比。可以明显看到，相对于英文 Google 上比较规则的 F 型分布，中文用户在 google.cn 上的浏览更具有随机性。虽然大体上还是呈现最上面的页

面关注时间比较多，越往下越少，但是中文用户并不像英文用户那样垂直向下浏览结果，看到感兴趣的结果则向右方移动目光，阅读页面标题或说明。中文用户的眼光更多的像是横向随机跳动，点击也是比较随机的，目光及点击分布都更广。



图 36 百度与 Google 中文的视线分布对比

上图是百度和 google.cn 的搜索页面对比。如果说用户在 google.cn 上还大致符合越上面的页面关注越多，在百度上则连在垂直方向也呈现更多随机特性，用户目光从上向下并没有显现出关注时间的急剧下降，百度用户不仅浏览页面上部结果，也在页面下部的结果上花了不少时间。在页面底部的相关搜索上，更是呈现出聚集目光和点击的情况。

按照英文用户搜索引擎结果浏览习惯分析，中文用户无论在 Google 还是百度上，似乎都花了更长时间才能找到自己想要的结果。英文用户在 Google 上平均 8-10 秒就找到想要的结果，而中文用户在 Google.cn 上则需要花 30 秒，在百度上要花 55 秒。这一方面说明中文搜索比英文搜索结果准确度要低，另外也很可能有语言方面的差异。中文句子中的词都是连在一起的，用户必须花多一点时间真正阅读标题，才能了解列出的结果是否符合自己的要求。而英文单词之间有空格分隔，更利于浏览，用户很容易在一瞥之下就能看到自己搜索的关键词。

在百度上满天星似的浏览也可能与百度广告和自然结果都放在左侧又没有背景颜色区分有关，一些用户会很自然地跳过广告，去查看排在后面的结果。

目前还没有见到中文搜索结果页面的点击数据统计。显然，上面介绍的点击数据不适用于中文搜索结果页面，尤其不适用于百度。可以想象，中文搜索结果点击率没有英文那样急剧下降的趋势，排在第五六位比排在第一位不会相差 10 倍之多。预估中文关键词流量时，不能照搬英文点击数据，而要更多依靠自己网站的点击数据。

6-3 整合搜索及个人化搜索

上面的视线跟踪及点击实验，都是基于传统 10 个文字列表的搜索结果页面。近几年随着整合搜索和个人化搜索的流行，搜索结果页面中出现图片、视频、新闻等结果，整个页面排版方式的变化必然影响用户浏览方式。

2007 年 9 月，enquiro.com 又做了整合搜索的目光跟踪实验，结果如下图所示。

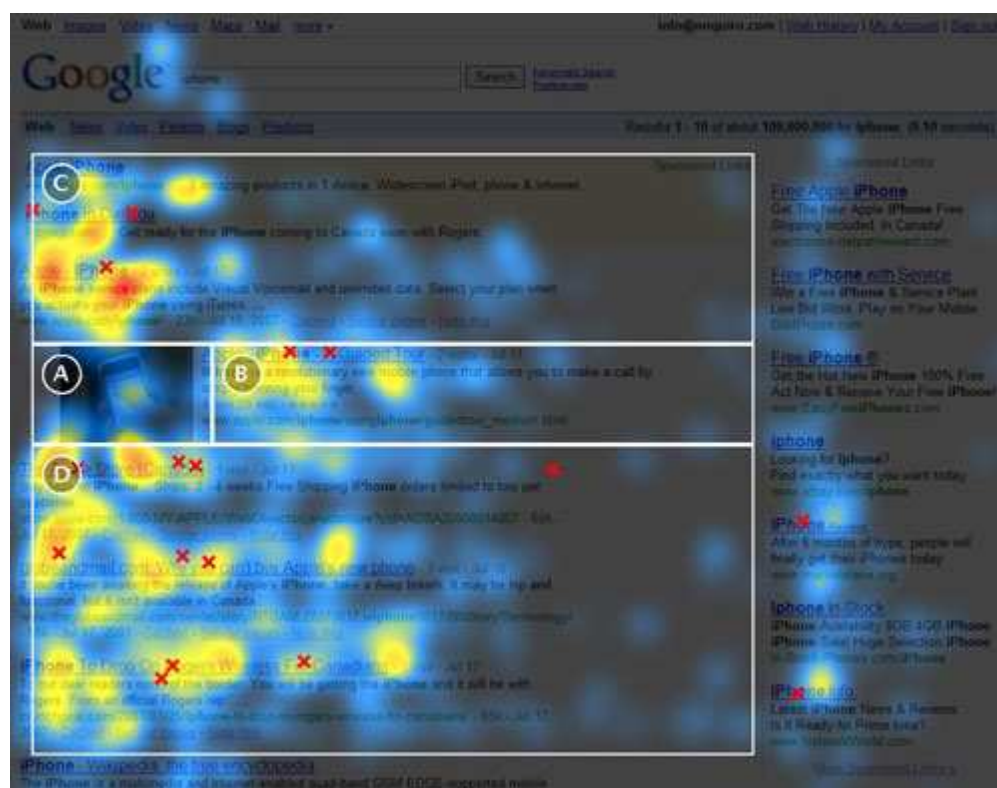


图 37 整合搜索结果对视线分布的影响

可以看到，当有图片出现在结果页面上部时，用户目光不再是从页面最左上角开始，而是首先把目光放在了图片上，接着向右移动看图片对应的结果是否符合自己的要求。然后用户视线再回到左上角重新向下浏览，看到合适的页面时再向右侧移动目光阅读页面标题和说明。

很明显，图片的出现完全改变了用户浏览方式，极大地吸引了用户目光。因此整合搜索结果不仅获得排名比普通页面要容易，竞争小，而一旦出现排名也更能吸引用户眼球和点击。

enquiro.com 对整合结果的实验也发现，带有图片的列表常常起到一种分隔作用。

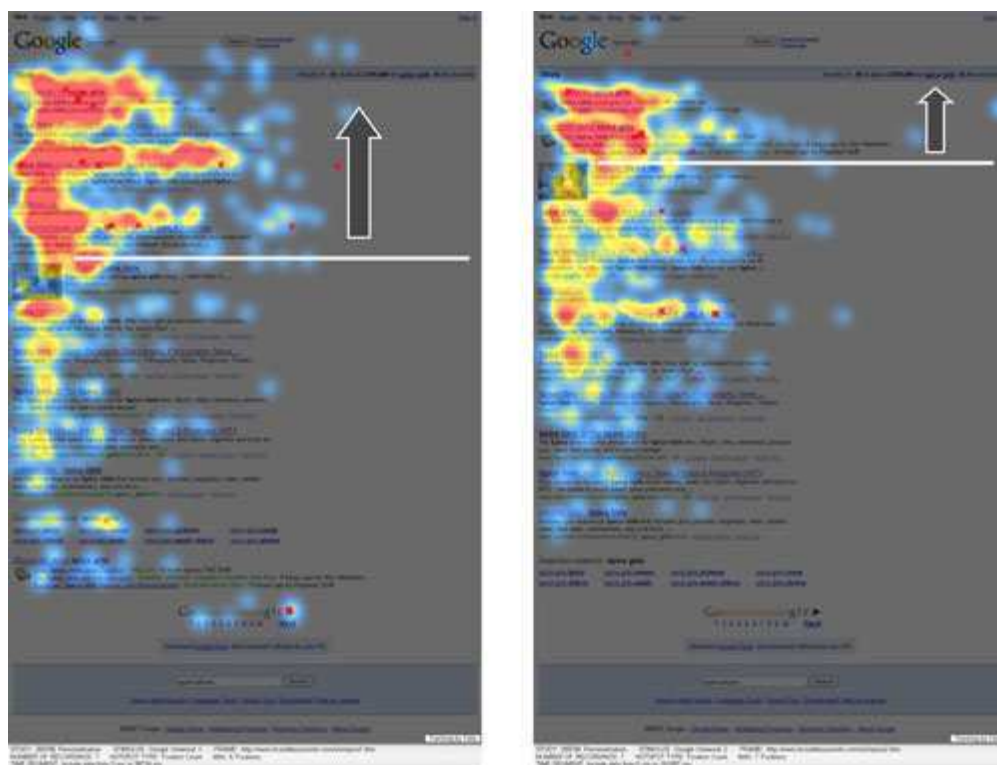


图 38 搜索结果中出现的图片起到分隔作用

如上图所示，以图片为横向分隔线，分隔线之上的结果获得很大的目光关注。用户把图片当成了一个阻隔，视线不再向下浏览图片之下的内容。

这对传统关注于页面排名的 SEO 来说是个挑战，而且是自己无法克服的挑战。好在这个实验是 2007 年所做，当时整合搜索结果还是个新鲜事物，用户不太习惯，因此会吸引更多的不成比例的视线。当用户对带有图片、视频的结果习以为常后，很可能浏览方式会向传统金三角模式靠拢。

enquiro.com 同时做了个人化搜索页面的目光跟踪实验。



图 39 个人化结果对视线分布的影响

如上图所示，左侧是非个人化结果页面，右侧是个人化结果页面，白色框中出现的三个结果是个人化结果，也就是用户以前曾经访问过的网站。明显可以看到，用户对自己访问过的熟悉的网站投入的目光关注度和时间要比陌生的网址高得多。在个人化搜索一节我们会提到，虽然目前看到的个人化结果还很少，但这是一个确定的搜索行业趋势，对 SEO 的影响不仅仅在于排名，还在于高得多的点击率。

下面图表显示了非个人化搜索及个人化搜索的具体点击比例：

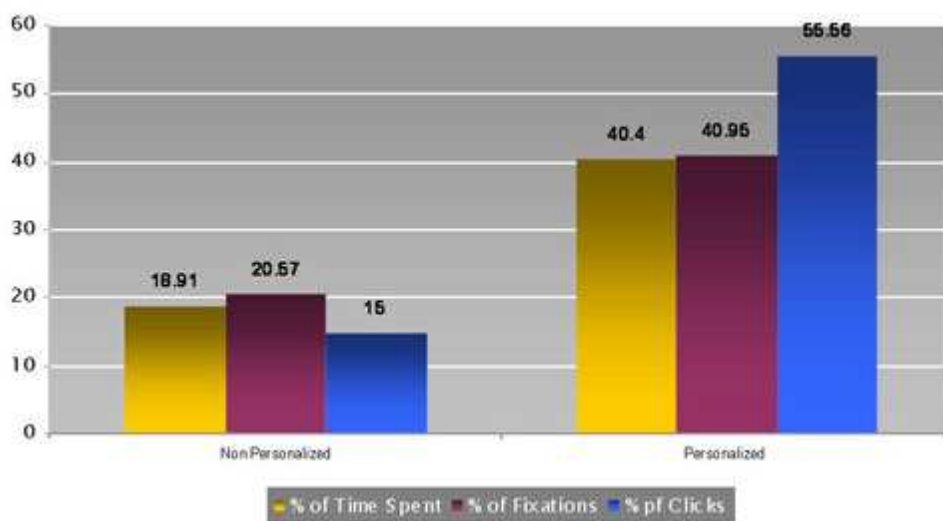


图 40 非个人化搜索及个人化搜索的点击比例

左侧是非个人化搜索，右侧是个人化搜索，三个数据分别为：

1. 所花时间比例
2. 目光注视比例
3. 点击比例

我们可以看到，用户对熟悉的网站点击可能性要比不熟悉的网站高出四倍多。

2010年3月，OneUpWeb 公司还做过实时搜索的视线跟踪实验，如下图所示，页面底部框中是实时搜索结果。

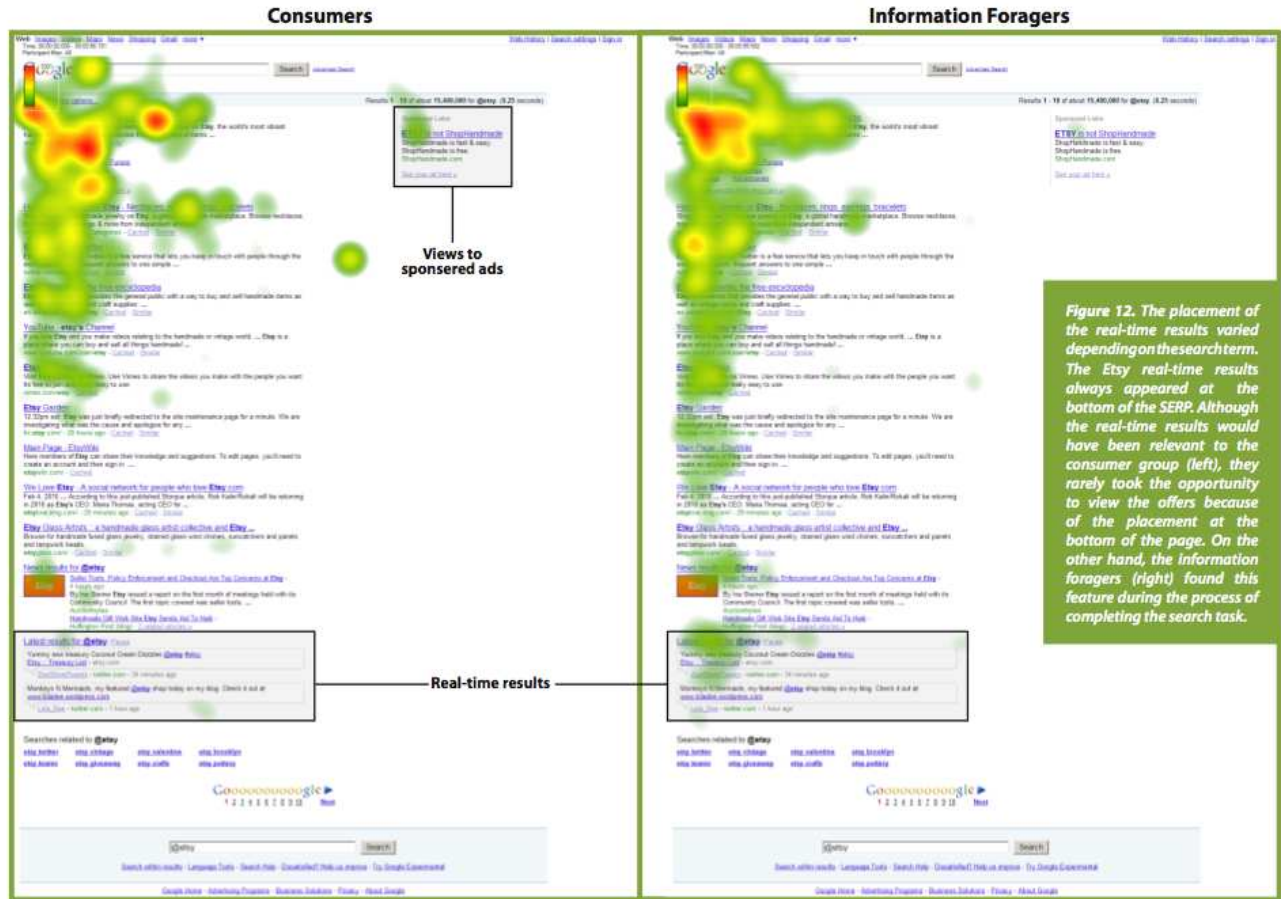


Figure 12. The placement of the real-time results varied depending on the search term. The Etsy real-time results always appeared at the bottom of the SERP. Although the real-time results would have been relevant to the consumer group (left), they rarely took the opportunity to view the offers because of the placement at the bottom of the page. On the other hand, the information foragers (right) found this feature during the process of completing the search task.

图 41 实时搜索对视线分布的影响

可以发现用户不太关注实时搜索，与整合搜索有较大差别。

7 高级搜索指令

用户除了可以在搜索引擎搜索普通关键词外，还可以使用一些特殊的高级搜索指令。这些搜索指令普通用户很少会用到，对 SEO 人员进行竞争对手研究和寻找外部链接资源却非常有用。这一节简单介绍常用的高级搜索指令。

7-1 双引号

把搜索词放在双引号中，代表完全匹配搜索，也就是说搜索结果返回的页面包含双引号中出现的所有的词，连顺序也必须完全匹配。百度和 Google 都支持这个指令。

比如搜索:

SEO 方法图片

新闻 网页 贴吧 知道 MP3 图片 视频 地图

Baidu 百度 百度一下 设置 | 高级搜索

把百度设为首页 百度一下, 找到相关网页约6,010,000篇, 用时0.022秒

[图片站SEO方法 标签的优化 - 搜索优化 - 站长之家-技术教程-中国...](#)
 图片站SEO方法 标签的优化 2008-9-20 11:19:00 大家对SEO了解的比我多, 我承认我不是老
 鸟, 今天不得不提的是百度图片的收录大乱了, 我们正常的需求, 我对...
www.suca.com/Tech/List1/323.htm 2008-12-3 - 百度快照

[几个简单好用的SEO优化方法\(图\) 勤加缘网社区](#)
 几个简单好用的SEO优化方法(图) 标签: 几个 简单 好用 优化 方法 上一篇: 我国建材市场状
 况(图)下一篇: 做业务员不能说的九种话(图)...
www.qjy168.com/forum/d_95658.html?page=1 2010-1-31 - 百度快照

[旗锦南昌SEO谈图片优化在搜索引擎中的方法 网易科技论坛](#)
 旗锦南昌SEO谈图片优化在搜索引擎中的方法 旗锦南昌SEO知道, 就目前的技术水平, 搜索
 引擎还无法判断图片的信息, 那我们就要帮助搜索引擎“读懂”图片的信息。一...
bbs.tech.163.com/bbs/tech_000e/169127242.html 2010-3-13 - 百度快照

[百度图片SEO方法 标签的优化 站长中国](#)
 百度图片SEO方法 标签的优化_站长中国 站长中国 版权所有 站长中国法律顾问: ITlaw-庄毅雄
 鲁ICP备08000081号 All rights reserved.按钮样式 360推出免费杀毒软件 ...
www.cq-web.com/Acquisition_title.asp?id=1851 2010-2-11 - 百度快照

[企业SEO方法之一:图片的SEO 唐兴通 新浪博客](#)
 scoail media marketing,,企业SEO做的好是,当用户检索你公司的名字时候,在第一页,出现的是
 有关你企业的相关情况:1,企业的网站2,企业的照片/产品的照片3,企业的...
blog.sina.com.cn/s/blog_4e6c9fcb0100dzc6.html 2009-6-8 - 百度快照

图 42 不带双引号的搜索结果

从图中可以看到, 返回的结果中不少页面出现的关键词并不是完整的“SEO 方法图片”, 有的页面中“SEO”, “方法”, “图片”这三个词出现在不同地方, 中间有间隔, 顺序也不相同。把“SEO 方法图片”放在双引号中再搜索:

"SEO 方法图片"



图 43 搜索时带双引号的结果

可以看到，返回结果只剩下四个，都是完整而且按顺序出现“SEO 方法图片”这个搜索字符串的页面。

使用双引号搜索可以更准确地找到特定关键词的竞争对手。

7-2 减号 -

减号代表搜索不包含减号后面的词的页面。使用这个指令时减号前面必须是空格，减号后面没有空格，紧跟着需要排除的词。Google 和百度都支持这个指令。

比如搜索“搜索”这个词时，返回结果如下：

网页 图片 视频 地图 资讯 问答 来吧 购物 更多 ▾



网页 [+ 打开百宝箱...](#) 搜索 搜索 获得约 598,000,000

相关搜索: [搜索引擎](#) [搜索 mp3](#) [国外搜索](#) [狗狗搜索](#) [狗狗搜索影视](#)

[百度中文搜索引擎](#)

百度网页[搜索](#)、MP3歌曲[搜索](#)、百度Flash[搜索](#)，分类信息[搜索](#)及百度大富翁游戏。

[+ 显示"BIDU"的股票报价](#)

[MP3 - 视频 - 图片 - 贴吧](#)

www.baidu.com/ - [网页快照](#) - [类似结果](#)

[搜狐搜索引擎](#)

123.sogou.com网址导航——最专业权威的上网导航。包含音乐、视频、小说、游戏、财经等上百个分类的优秀站点，提供最简单便捷的网上导航服务，是最受网民欢迎的上网 ...

[电影 - 小说 - 视频 - 娱乐](#)

123.sogou.com/ - [网页快照](#) - [类似结果](#)

[雅虎搜索](#)

[_全球领先的中英文搜索引擎.](#)

www.yahoo.cn/ - [网页快照](#) - [类似结果](#)

图 44 普通不使用减号的搜索

排在前面的都是关于搜索引擎的页面。如果我们搜索

搜索 -引擎

返回的则是包含“搜索”这个词，却不包含“引擎”这个词的结果，如下图所示：



图 45 搜索时使用减号

使用减号也可以更准确地找到需要的文件，尤其是某些词有多种意义时。比如搜索“苹果 -电影”，返回结果页面就排除了苹果这部电影的结果，而不会影响苹果电脑和苹果作为水果的内容。

7-3 星号 *

星号*是常用的通配符，也可以用在搜索中。百度不支持*号搜索指令。

比如在 Google 中搜索

搜索*擎

其中的*号代表任何文字。返回的结果就不仅包含“搜索引擎”，还包含了“搜索引擎”，“搜索引擎”等内容。

网页 图片 视频 地图 资讯 问答 贴吧 购物 更多 ▾

Google

搜索*擎

Google 搜索 高级

网页 [+ 打开百宝箱...](#)

搜索 搜索*擎 获得约 10,300,000

[免费搜索引擎 我行我素 湛江,东海岛,湛江钢铁厂,北斗卫星,维和警察 ...](#)

网站的推广登录 [搜索引擎](#) 免费网站登录. http://www.google.com/intl/zh-CN/add_url.html Google 登录/更新网站 http://www.baidu.com/search/url_submit.html 百度搜索 ... hi.baidu.com/fjeg/blog/item/6662f936fd19fd310b55a909.html - [网页快照](#)

[免费搜索引擎网站登录地址 铁蛋的地盘 百度空间](#)

[论坛] [免费搜索引擎网站登录地址](#). 2007-06-03 19:16:27. 1、Yahoo 网站登录地址: http://misc.yahoo.com.cn/search_submit.html 2、Google网站登录入口: ... hi.baidu.com/zjl9848/blog/.../15a783017a14f306738da5cb.html - [网页快照](#)

[搜索引擎的未来:谷歌正在寻求新发展-企业战略-世界经理人网站](#)

[搜索引擎的未来:谷歌正在寻求新发展,谷歌必须努力认清一个事实,即它已不再是敢打敢拼、斗志旺盛的商界新手,而是举足轻重的市场竞争者。](#) www.ceconline.com/strategy/ma/8800050377/01/?pa_art_8 - [网页快照](#)

[外电最新分析:Google百度同为搜索引擎的差异- baidu相关- 网站运营](#)

外电最新分析:Google百度同为 [搜索引擎](#) 的差异, 据外国媒体的最新报道称, 华尔街曾感叹道, www.xmsc.com.cn/InfoView/Article_30445.html - [网页快照](#)

图 46 搜索时使用星号

7-4 inurl:

inurl: 指令用于搜索查询词出现在 url 中的页面。百度和 Google 都支持 inurl 指令。inurl 指令支持中文和英文。

比如搜索

inurl:搜索引擎优化



图 47 inurl:指令

从图中可以看到，返回的结果都是网址 url 中包含“搜索引擎优化”的页面。由于关键词出现在 url 中对排名有一定影响，使用 inurl:搜索可以更准确地找到竞争对手。

7-5 inanchor:

inanchor:指令返回的结果是导入链接锚文字中包含搜索词的页面。百度不支持 inanchor。

比如在 Google 搜索

inanchor:点击这里



网页 [+ 打开百宝箱...](#) 搜索 inanchor:点击这里 获得约 86,300,000

[FLASH音画学习音画教程动画QQ音画素材音画欣赏视频教程 - 音画驿站学习...](#)

您好欢迎您来到本站本站的永久域名为www.6000y.com 如果你想一天学会FLASHMTV的制作, 请您仔细阅读本网页后再和我们联系, 提醒大家 有少数朋友设置了 拒绝任何人加 ...

[www.6000y.com/index.asp?xAction=xreadnews... - 7 小时前 - 网页快照](#)

[Adobe Reader 9.3 简体中文版下载- 华军软件园- 系统程序- 电子阅读](#)

2010年1月15日 ... Adobe Reader(也称为Acrobat Reader)是美国Adobe公司开发的一款优秀的PDF文档阅读软件。文档的撰写者可以向任何人分发自己制作(通过Adobe Acrobat制作) ...

[www.onlinedown.net > 下载分类 - 网页快照 - 类似结果](#)

[Adobe - Adobe Reader](#)

Download Adobe Reader to view, print and collaborate on PDF files.

[get.adobe.com/cn/reader/ - 美国 - 网页快照 - 类似结果](#)

[下载详细信息: Internet Explorer 6 Service Pack 1](#)

Internet Explorer 6 SP1, the latest version of Internet Explorer for users not running Windows XP, provides a flexible and reliable browsing experience with ...

[www.microsoft.com/downloads/details.aspx?...zh... - 网页快照 - 类似结果](#)

[注册 - 工商银行网上银行](#)

工商银行牡丹灵通卡、理财金账户卡、信用卡、贷记卡、国际卡、商务卡客户, 点击这里。开通个人网上银行, 可获得账户查询、网上购物支付等服务。友情提示: ...

图 48 inanchor:指令

从图中可以看到, 返回的结果页面本身并不一定包含“点击这里”这四个字, 而是指向这些页面的链接锚文字中出现了“点击这里”这四个字。

在后面章节我们会讨论, 链接锚文字是现在关键词排名最重要因素之一, 有经验的 SEO 会尽量使外部链接锚文字中出现目标关键词。因此, 使用 inanchor:指令可以找到某个关键词的竞争对

手，而且这些竞争对手往往是做过 SEO 的。研究竞争对手页面有哪些外部链接，就可以找到很多链接资源。

7-6 intitle:

intitle: 指令返回的是页面 title 中包含关键词的页面。Google 和百度都支持 intitle 指令。

Title 是目前页面优化的最重要因素。做 SEO 的人无论要做哪个词的排名，都会把关键词放进 title 中。所以使用 intitle 指令找到的文件才是更准确的竞争页面。如果关键词只出现在页面可见文字中，而没有出现在 title 中，大部分情况是并没有针对关键词进行优化，所以也不是有力的竞争对手。

比如搜索“搜索引擎优化”，我的博客在百度排在自然排名的第二位：



[在中国做搜索引擎优化,首选百度推广!](#)

百度占据中国搜索引擎市场80%份额,了解中国企业推广需求,按效果付费,针对性强,每个企业都适用,立即电话咨询:400-800-8888!

e.baidu.com 2010-03 - [推广](#)

[搜索引擎优化 百度百科](#)

搜索引擎优化(Search Engine Optimization, 简称SEO)是一种利用搜索引擎的搜索规则来提高目的网站在有关搜索引擎内的排名的方式。由于不少研究发现,搜索引擎的用户往往只会留意搜索结果最前面的几个条目,所以不少网站都希望通过...共87次编辑

baike.baidu.com/view/7147.htm 2010-3-11

[SEO每天一贴 - Zac的SEO博客,正在写书,暂时每月两三贴](#)

SEO每天一贴研究搜索引擎优化SEO技术,网络营销及电子商务思考。正在写一本SEO方面的书,所以不能每天一贴了,会尽快恢复。

www.chinamyhosting.com/seoblog/ 2010-3-15 - [百度快照](#)

[搜索引擎优化 SEO165.COM](#)

SEO165是专业的搜索引擎优化服务商,致力于搜索引擎优化(SEO)和Google优化排名研究,为企业提供搜索引擎优化,Google排名,搜索引擎排名,搜索引擎营销顾问服务。

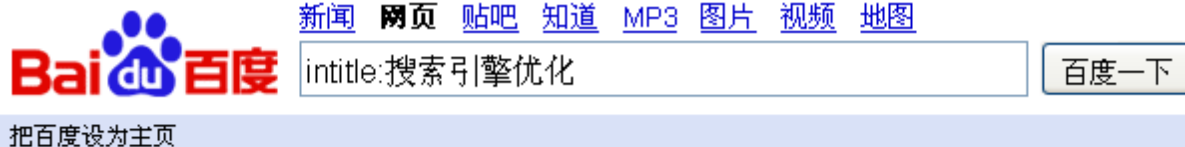
www.seo165.com/ 2010-3-11 - [百度快照](#)

图 49 普通搜索结果

但是搜索

intitle:搜索引擎优化

我的博客就不会被返回,因为博客标题中并没有“搜索引擎优化”这六个字。



[在中国做搜索引擎优化,首选百度推广!](#)

百度占据中国搜索引擎市场80%份额,了解中国企业推广需求,按效果付费,针对性强. e.baidu.com

[搜索引擎优化 百度百科](#)

搜索引擎优化即Search Engine Optimization, 用英文描述是to use some technics to make your website in the top places in Search Engine when somebody is using ...

baike.baidu.com/view/7147.htm 2010-3-15 - [百度快照](#)

[搜索引擎优化|优化|SEO优化-飞度时代: 400-600-3630](#)

搜索引擎优化 (SEO优化,优化), 就是针对各种搜索引擎的检索特点, 让网站建设和网页设计的基本要素适合搜索引擎的检索原则 (即搜索引擎友好), 从而获得搜索引擎收录并在...

www.youhuaseo.com/ssyqseo/ 2010-3-14 - [百度快照](#)

[搜索引擎优化 SEO165.COM](#)

SEO165是专业的搜索引擎优化服务商,致力于搜索引擎优化(SEO)和Google优化排名研究,为企业提供搜索引擎优化,Google排名,搜索引擎排名,搜索引擎营销顾问服务。

www.seo165.com/ 2010-3-11 - [百度快照](#)

[网站优化|搜索引擎优化|seo](#)

网站优化: 建于2003年, 收集搜索引擎优化、网站优化资料及SEO工具。致力搜索引擎优化研究, 与SEO爱好者分享搜索引擎优化技巧与经验, 最新推出网站优化及SEO培训!

www.seochat.org/ 2010-3-13 - [百度快照](#)

图 50 intitle:指令搜索结果

7-7 allintitle:

allintitle:搜索返回的是页面标题中包含多组关键词的文件。例如

allintitle:SEO 搜索引擎优化

就相当于

intitle:SEO intitle:搜索引擎优化

返回的是标题中既包含“SEO”，也包含“搜索引擎优化”的页面。

7-8 allinurl:

与 allintitle: 类似。

allinurl:SEO 搜索引擎优化

就相当于

inurl:SEO inurl:搜索引擎优化

7-9 filetype:

用于搜索特定文件格式。Google 和百度都支持 filetype 指令。

比如搜索

filetype:pdf SEO

返回的就是包含 SEO 这个关键词的所有 pdf 文件。



图 51 filetype:指令

百度只支持下面几种文件格式: pdf, doc, xls, ppt, rtf, all。其中的“all”表示搜索百度所有支持的文件类型。Google 则支持所有能索引的文件格式, 包括 HTML, PHP 等。

filetype:指令用来搜索特定的资源, 比如 PDF 电子书, word 文件等非常有用。

7-10 site:

site:是 SEO 最熟悉的高级搜索指令, 用来搜索某个域名下的所有文件。比如搜索

site:chinamyhosting.com

返回的就是 chinamyhosting.com 这个域名下的所有页面。



The screenshot shows a Baidu search interface. At the top, there are navigation links for '新闻', '网页', '贴吧', '知道', 'MP3', '图片', '视频', and '地图'. The search bar contains the query 'site:chinamyhosting.com' and a '百度一下' button. Below the search bar, a status bar indicates '把百度设为首页' and '百度一下, 找到相关网页约1,200篇, 用时0.028秒'. The main content area displays several search results, each with a title, a brief description, and a link to the source page. The results include: 1. '虚拟主机-域名注册-中新国外虚拟主机无需备案-域名注册转出自由' with a link to 'www.chinamyhosting.com/ 2010-3-14 - 百度快照'; 2. 'SEO每天一贴 - Zac的SEO博客, 正在写书, 暂时每月两三贴' with a link to 'www.chinamyhosting.com/seoblog/ 2010-3-11 - 百度快照'; 3. '电子商务-Zac的电子商务思考' with a link to 'www.chinamyhosting.com/seoblog/category/e ... 2010-3-14 - 百度快照'; 4. '网络营销实战密码 - 最实用的网络营销书' with a link to 'www.chinamyhosting.com/seoblog/book/ 2010-3-13 - 百度快照'.

图 52 site:指令

所以这个指令是查询网站收录页面数的最直接方法。

site:指令也可以用于子域名, 比如

site:blog.sina.com.cn

搜索的就是 blog.sina.com.cn 子域名下的所有收录页面。而

site:sina.com.cn

则包含 sina.com.cn 本身以及 sina.com.cn 下面所有子域名 (包括如 blog.sina.com.cn) 的页面。

不过 site:指令并不准确, 尤其是 Google, 返回的收录页面经常有大幅度波动, 只能作为参考。

7-11 link:

link:也是SEO常用的指令,用来搜索某个url的反向链接,既包括内部链接,也包括外部链接。比如搜索

link:chinamyhosting.com

返回的就是 chinamyhosting.com 的反向链接。



图 53 link:指令

不过可惜的是,Google的link:指令返回的链接只是Google索引库中的一部分,而且是近乎随机的一部分,所以用link:指令查反向链接几乎没有用。百度则不支持link:指令。

7-12 linkdomain:

linkdomain:指令只适用于雅虎,返回的是某个域名的反向链接。雅虎的反向链接数据还比较准确,是SEO人员研究竞争对手外部链接情况的重要工具之一。比如搜索

linkdomain:dunsh.org -site:dunsh.org

得到的就是点石网站的外部链接,因为-site:dunsh.org已经排除了点石本身的页面,也就是内部链接,剩下的就都是外部链接了。



图 54 雅虎的 linkdomain:指令

7-13 related:

related:指令只适用于 Google,返回的结果是与某个网站有关联的页面。比如搜索

related:dunsh.org

我们就可以得到 Google 所认为的与点石网站有关联的其他页面。



图 55 related:指令

这种关联到底指的是什么, Google 并没有明确说明, 一般认为指的是有共同外部链接的网站。

7-14 综合使用高级搜索指令

上面介绍的这几个高级搜索指令, 单独使用可以找到不少资源, 或者可以更精确地定位竞争对手。把这些指令混合起来使用则更强大。

比如下面这个指令

`inurl:gov 减肥`

返回的就是 url 中包含 gov，页面中有“减肥”这个词的页面。很多 SEO 人员认为政府和学校网站有比较高的权重，找到相关的政府和学校网站，就找到了最好的链接资源。

下面这个指令返回的是来自.edu.cn，也就是学校域名上的包含“交换链接”这个词的页面：

`inurl:.edu.cn 交换链接`

从中 SEO 人员可以找到愿意交换链接的学校网站。

或者使用一个更精确的搜索：

`inurl:.edu.cn intitle:交换链接`

返回的则是来自 edu.cn 域名，标题中包含“交换链接”这四个字的页面，返回的结果大部分应该是愿意交换链接的学校网站。

再比如下面这个指令：

`inurl:.edu.cn/forum/*register`

返回的结果是在.edu.cn 域名上，url 中包含“forum”以及“register”这两个单词的页面，也就是学校论坛的注册页面。找到这些论坛，也就找到了能在高权重域名上留下签名的很多机会。

下面这个指令返回的是页面与减肥有关，url 中包含 links 这个单词的页面：

`减肥 inurl:links`

很多站长把交换链接页面命名为 links.html 等，所以这个指令返回的就是与减肥主题相关的交换链接页面。

下面这个指令返回的是 url 中包含 gov.cn 以及 links 的页面，也就是政府域名上的交换链接页面：

`allinurl:gov.cn+links`

最后一个例子，在雅虎搜索这个指令：

`linkdomain:dunsh.org -linkdomain:chinamyhosting.com`

返回的是链接到点石网站，却没有链接到我的博客的网站。使用这个指令可以找到很多连向你的竞争对手或其他同行业网站，却没连向你的网站的页面，这些网站是最好的链接资源。

高级搜索指令组合使用变化多端，功能强大。一个合格的 SEO 必须熟练掌握这几个常用指令的意义及组合方法，才能更有效率地找到更多竞争对手和链接资源。

附录：《SEO 实战密码》目录

第 1 章 为什么要做 SEO	1
1.1 什么是 SEO	1
1.2 为什么要做 SEO	2
1.3 搜索引擎简史	6
第 2 章 了解搜索引擎	14
2.1 搜索引擎与目录	15
2.2 搜索引擎面临的挑战	15
2.3 搜索结果显示格式	17
2.3.1 搜索结果页面	17
2.3.2 经典搜索结果列表	20
2.3.3 整合搜索结果	21
2.3.4 缩进列表	21
2.3.5 全站链接	22
2.3.6 迷你全站链接	22
2.3.7 One-box	22
2.3.8 富摘要	23
2.3.9 面包屑导航	23
2.3.10 说明文字中的链接	23
2.4 搜索引擎工作原理简介	24
2.4.1 爬行和抓取	24
2.4.2 预处理	27
2.4.3 排名	31
2.5 链接原理	35
2.5.1 李彦宏超链分析专利	36
2.5.2 HITS 算法	36

2.5.3 TrustRank 算法	37
2.5.4 Google PR	38
2.5.5 Hilltop 算法	41
2.6 用户怎样浏览和点击搜索结果	42
2.6.1 英文搜索结果页面	43
2.6.2 中文搜索结果页面	46
2.6.3 整合搜索及个人化搜索	48
2.7 高级搜索指令	51
2.7.1 双引号	51
2.7.2 减号	51
2.7.3 星号	52
2.7.4 inurl:	53
2.7.5 inanchor:	54
2.7.6 intitle:	54
2.7.7 allintitle:	55
2.7.8 allinurl:	55
2.7.9 filetype:	56
2.7.10 site:	56
2.7.11 link:	57
2.7.12 linkdomain:	58
2.7.13 related:	58
2.7.14 综合使用高级搜索指令	59
第3章 竞争研究	60
3.1 为什么研究关键词	60
3.1.1 确保目标关键词有人搜索	60
3.1.2 降低优化难度	61
3.1.3 寻找有效流量	61
3.1.4 搜索多样性	61
3.1.5 发现新机会	62
3.2 关键词的选择	62
3.2.1 内容相关	62
3.2.2 搜索次数多, 竞争小	63
3.2.3 主关键词不可太宽泛	63
3.2.4 主关键词也不可太特殊	63
3.2.5 商业价值	63
3.3 关键词竞争程度判断	64
3.3.1 搜索结果数	64
3.3.2 intitle 结果数	65

3.3.3 竞价结果数	65
3.3.4 竞价价格	65
3.3.5 竞争对手情况	66
3.3.6 内页排名数量	66
3.4 核心关键词	67
3.4.1 头脑风暴	67
3.4.2 同事、朋友	68
3.4.3 竞争对手	68
3.4.4 查询搜索次数	69
3.4.5 确定核心关键词	70
3.5 关键词扩展	71
3.5.1 关键词工具	71
3.5.2 搜索建议	72
3.5.3 相关搜索	72
3.5.4 其他关键词扩展工具	72
3.5.5 各种形式的变体	73
3.5.6 补充说明文字	73
3.5.7 网站流量分析	74
3.5.8 单词交叉组合	74
3.6 关键词分布	75
3.6.1 金字塔形结构	75
3.6.2 关键词分组	75
3.6.3 关键词布局	76
3.6.4 关键词-URL 对应表	77
3.7 长尾关键词	77
3.7.1 长尾理论	77
3.7.2 搜索长尾	78
3.7.3 怎样做长尾关键词	79
3.8 三类关键词	80
3.8.1 导航类关键词	80
3.8.2 交易类关键词	81
3.8.3 信息类关键词	81
3.9 预估流量及价值	81
3.9.1 确定目标排名	82
3.9.2 预估流量	82
3.9.3 预估搜索流量价值	85
3.10 关键词趋势波动和预测	86
3.10.1 长期趋势	86

3.10.2 季节性波动	86
3.10.3 社会热点预测	87
3.11 竞争对手研究	89
3.11.1 域名权重相关数据	89
3.11.2 网站优化情况	91
3.11.3 网站流量	92
3.12 快速网站诊断	93
3.12.1 robots 文件检查	93
3.12.2 首选域设置	94
3.12.3 关键词排名	95
3.12.4 外部链接	97
3.12.5 网站内容	98
3.12.6 内部链接	99
3.12.7 抓取错误及统计	99
3.12.8 HTML 建议	100
3.12.9 模拟蜘蛛抓取	101
3.12.10 网站性能	102
第 4 章 网站结构优化	103
4.1 搜索引擎友好的网站设计	104
4.2 避免蜘蛛陷阱	109
4.2.1 Flash	109
4.2.2 Session ID	110
4.2.3 各种跳转	110
4.2.4 框架结构	110
4.2.5 动态 URL	111
4.2.6 JavaScript 链接	111
4.2.7 要求登录	111
4.2.8 强制使用 Cookies	111
4.3 物理及链接结构	112
4.3.1 物理结构	112
4.3.2 链接结构	113
4.4 清晰导航	114
4.5 子域名和目录	115
4.6 禁止收录机制	116
4.6.1 robots 文件	117
4.6.2 meta robots 标签	118
4.7 nofollow 的使用	119

4.8 URL 静态化	121
4.8.1 为什么静态化	121
4.8.2 怎样静态化 URL	122
4.8.3 URL 不需要静态化吗	122
4.9 URL 设计	123
4.10 网址规范化	125
4.10.1 为什么出现不规范网址	125
4.10.2 网址规范化问题	126
4.10.3 解决网址规范化问题	127
4.10.4 301 转向	127
4.10.5 Canonical 标签	129
4.11 复制内容	130
4.11.1 产生复制内容的原因	130
4.11.2 复制内容的害处	132
4.11.3 消除复制内容	132
4.12 绝对路径和相对路径	133
4.12.1 绝对路径	134
4.12.2 相对路径	134
4.13 网站地图	135
4.13.1 HTML 网站地图	135
4.13.2 XML 网站地图	135
4.14 内部链接及权重分配	137
4.14.1 重点内页	137
4.14.2 非必要页面	137
4.14.3 大二级分类	138
4.14.4 翻页过多	138
4.14.5 单一入口还是多入口	139
4.14.6 相关产品链接	140
4.14.7 锚文字分布及变化	141
4.14.8 首页链接 NoFollow	142
4.14.9 深层链接	142
4.14.10 分类隔离	142
4.15 CMS 系统	143
4.16 404 页面	145
4.16.1 404 错误代码	145
4.16.2 404 页面设计	146
4.16.3 404 错误与外链	146

第 5 章 页面优化	148
5.1 页面标题	148
5.1.1 独特不重复	148
5.1.2 准确相关	151
5.1.3 字数限制	151
5.1.4 简练通顺, 不要堆砌	152
5.1.5 关键词出现在最前面	153
5.1.6 吸引点击	153
5.1.7 组合两三个关键词	153
5.1.8 公司或品牌名称	154
5.1.9 连词符使用	154
5.1.10 不要用没有意义的句子	155
5.1.11 noodp 标签	155
5.2 描述标签	155
5.3 关键词标签	156
5.4 正文中的关键词	157
5.4.1 词频和密度	157
5.4.2 前 50~100 个词	157
5.4.3 关键词变化形式	158
5.4.4 关键词组临近度	158
5.4.5 词组的拆分出现	158
5.4.6 语义分析	158
5.4.7 分类页面说明文字	159
5.5 H 标签	160
5.6 ALT 文字	160
5.7 精简代码	161
5.8 内部链接及锚文字	162
5.9 导出链接及锚文字	162
5.10 W3C 验证	162
5.11 黑体及斜体	163
5.12 页面更新	163
5.13 Google 沙盒效应	163
第 6 章 外部链接建设	165
6.1 外部链接意义	165
6.1.1 相关性及锚文字	165
6.1.2 权重及信任度	166
6.1.3 收录	166

6.2	Google 炸弹	167
6.3	链接分析技术	169
6.4	什么样的链接是好链接	170
6.5	外部链接查询	173
6.5.1	链接查询指令	173
6.5.2	工具查询外链	174
6.5.3	影响排名的链接	174
6.6	外部链接原则	175
6.6.1	难度越大, 价值越高	176
6.6.2	内容是根本	176
6.6.3	内容相关性	176
6.6.4	链接来源广泛	176
6.6.5	深度链接	177
6.6.6	锚文字分散自然	177
6.6.7	平稳持续增加	177
6.6.8	质量高于数量	177
6.7	网站目录提交	178
6.7.1	提交前的准备	178
6.7.2	寻找网站目录	179
6.7.3	网站提交	180
6.8	友情链接	180
6.8.1	友情链接页面	181
6.8.2	软件使用	181
6.8.3	寻找交换链接目标	182
6.8.4	交换链接步骤	182
6.8.5	内页正文链接交换	183
6.8.6	交换链接中的小花招	184
6.9	链接诱饵	185
6.9.1	链接诱饵的制作	186
6.9.2	链接诱饵种类和方法	187
6.9.3	链接诱饵之度	198
6.10	其他常规外链建设方法	199
6.11	非链接形式的链接	207
6.12	竞争对手能否通过垃圾外链陷害你	209
6.13	链接工作表	210
第 7 章	SEO 效果监测及策略修改	212
7.1	为什么要监测	212

7.1.1	检验工作成效	212
7.1.2	发现问题, 修改策略	212
7.1.3	SEO 完整过程	213
7.2	网站目标设定及测量	213
7.2.1	网站目标	213
7.2.2	网站目标实例	214
7.2.3	网站目标确定原则	215
7.2.4	网站目标影响 SEO 策略	216
7.3	非流量数据监测	216
7.3.1	收录数据	216
7.3.2	排名监测	219
7.3.3	外部链接数据	220
7.3.4	转化和销售	220
7.4	流量数据监测	220
7.4.1	怎样读日志文件	221
7.4.2	常用流量分析工具	223
7.4.3	流量统计分析基础	224
7.5	策略改进	230
7.5.1	收录是否充分	230
7.5.2	哪些页面带来搜索流量	231
7.5.3	目标 URL 排名如何	232
7.5.4	挖掘关键词	233
7.5.5	其他搜索引擎流量	233
7.5.6	长尾效果	234
7.5.7	关键词排名下降	235
7.5.8	链接诱饵成效	235
7.5.9	发现链接伙伴	235
7.5.10	寻找有潜力关键词	236
第 8 章 SEO 作弊及惩罚		237
8.1	白帽、黑帽、灰帽	237
8.1.1	白帽黑帽是风险度判断	237
8.1.2	道德及法律底线	238
8.1.3	SEO 服务商的底线	238
8.1.4	黑帽 SEO 的贡献	239
8.1.5	承担风险, 不要抱怨	239
8.1.6	了解黑帽, 做好白帽	240
8.2	主要 SEO 作弊方法	241

8.2.1 隐藏文字 (Hidden Text)	241
8.2.2 隐藏链接 (Hidden Links)	241
8.2.3 垃圾链接 (Link Spam)	242
8.2.4 买卖链接 (Paid Links)	243
8.2.5 链接农场 (Link Farm)	244
8.2.6 链接向坏邻居 (Bad Neighborhood)	245
8.2.7 隐藏页面 (Cloaking, Cloaked Page)	245
8.2.8 PR 劫持 (PR Hijacking)	246
8.2.9 桥页 (Doorway Pages, Bridge Pages)	247
8.2.10 跳转	247
8.2.11 诱饵替换 (Bait and Switch)	248
8.2.12 关键词堆积 (Keyword Stuffing)	248
8.2.13 大规模站群	249
8.2.14 利用高权重网站	249
8.3 搜索引擎惩罚	250
8.3.1 作弊的积分制	250
8.3.2 不要学大网站	252
8.3.3 不要存侥幸心理	253
8.3.4 搜索引擎惩罚的种类	253
8.3.5 搜索引擎惩罚的检测	254
8.4 被惩罚了怎么办	255

第9章 SEO 专题 258

9.1 整合搜索优化	258
9.1.1 什么是整合搜索	258
9.1.2 机会和挑战	260
9.1.3 新闻搜索	260
9.1.4 图片搜索	261
9.1.5 视频搜索	261
9.1.6 地图搜索	262
9.2 更改域名	263
9.3 多个域名的处理	264
9.4 更换服务器	265
9.5 用户行为影响排名	265
9.5.1 用户行为信息收集	266
9.5.2 影响排名的用户行为	266
9.5.3 回归用户体验	267
9.6 域名与 SEO	267

9.6.1	域名后缀	268
9.6.2	域名年龄	268
9.6.3	域名第一次被收录时间	268
9.6.4	域名续费时间	268
9.6.5	域名包含关键词	269
9.6.6	连词符使用	269
9.6.7	品牌优先	269
9.6.8	域名长短	269
9.6.9	域名买卖历史	270
9.6.10	匿名注册信息	270
9.6.11	域名权重	270
9.7	主机与 SEO	270
9.7.1	IP 及整个服务器惩罚	270
9.7.2	服务器设置	271
9.7.3	稳定性	271
9.7.4	主机速度	271
9.7.5	URL 重写支持	271
9.8	多语种内容	272
9.8.1	多语种页面处理	272
9.8.2	当地语言习惯与 SEO	272
9.9	地理定位	273
9.9.1	什么是地理定位	273
9.9.2	地理定位的表现形式	274
9.9.3	地理定位的影响因素	274
9.10	社会化媒体的影响	275
9.10.1	带来链接	276
9.10.2	互动及口碑传播	276
9.10.3	新形式的链接流动成为排名信号	277
9.10.4	网络名誉管理	277
9.11	避免过度优化	277
9.12	SEO 与品牌	278
9.12.1	排名第一就是品牌	278
9.12.2	传统品牌建设与 SEO 结合	279
9.12.3	网上危机公关	279
9.13	针对不同搜索引擎的优化	280
9.13.1	SEO 原则不变	281
9.13.2	百度和 Google 的区别	281
9.13.3	英文网站优化	282

9.14 网站改版	282
9.14.1 设计还是 CMS 系统改变	282
9.14.2 不要改 URL	283
9.14.3 分步更改	283
9.15 Google Dance	283
9.15.1 什么是 Google Dance	283
9.15.2 Google 已不再 Dance	284
9.15.3 近年 Google 更新	284
9.16 Google 全站链接	285
9.16.1 全站链接的出现	285
9.16.2 屏蔽全站链接	286
9.16.3 迷你全站链接	286
9.17 个人化搜索	287
9.17.1 什么是个人化搜索	287
9.17.2 个人化搜索对 SEO 的影响	288
第 10 章 SEO 观念及原则	289
10.1 搜索引擎的目标	289
10.1.1 搜索引擎的目标是满足搜索用户	289
10.1.2 搜索引擎不在乎我们	290
10.1.3 搜索引擎在乎垃圾	290
10.2 相关性、权威性、实用性	291
10.2.1 网站内容的相关性	291
10.2.2 网站及网页的权威性	291
10.2.3 网站的实用性	292
10.3 SEO 与赚钱	292
10.3.1 给别人做 SEO	292
10.3.2 给自己做 SEO	294
10.4 SEO 不是免费的	295
10.4.1 人力成本	295
10.4.2 机会成本	295
10.4.3 失败风险	295
10.4.4 SEO 成功风险	296
10.5 不要做奇怪的事	296
10.6 解决基本问题就解决了 95% 的问题	297
10.7 自然和平衡的艺术	297
10.8 SEO 是长期策略	298
10.8.1 实施 SEO 需要时间	298

10.8.2	不进则退	299
10.9	没有 SEO 秘籍	300
10.9.1	为什么没有 SEO 秘籍	300
10.9.2	搜索引擎排名算法的秘密	300
10.9.3	SEO 绝招	300
10.10	SEO 不仅是排名	301
10.11	SEO 不是作弊	302
10.12	SEO 与网站运营	302
10.13	内容为王	303
10.13.1	原创内容是 SEO 的根本	303
10.13.2	内容策划是 SEO 策略	304
10.13.3	内容推广	304
10.14	具体问题具体分析	305
第 11 章	SEO 工具	307
11.1	Xenu	308
11.2	Alexa	310
11.3	谷歌趋势	313
11.4	百度指数	316
11.5	百度搜索风云榜	318
11.6	Google Adwords 关键词工具	319
11.7	微软广告工具	322
11.8	Google 百宝箱	324
11.9	Google 快讯	326
11.10	服务器头信息检测器	327
11.11	W3C 验证	328
11.12	雅虎外链检查工具	329
11.13	链接分析插件 SEO Link Analysis	331
11.14	外链概况工具	332
11.15	IP 地址检查工具	332
11.16	Google 基于搜索的关键字工具	333
11.17	关键词问答	335
11.18	Google 搜索解析	336
11.19	Google Sets	337
11.20	网站索引指数查询	338
11.21	SEO for Firefox	338
11.22	SEO 工具条	340
11.23	火狐浏览器 Search Status 插件	342

11.24	火狐 SeoQuake 插件	343
11.25	站长帮手	346
11.26	关键词排名批量查询工具	348
11.27	SEOmox 工具	349
11.28	Backlink Watch	353
11.29	TouchGraph	354
11.30	Quintura	354
11.31	Google Ad Planner	355
11.32	Majestic SEO	355
11.33	追词	362

第 12 章 SEO 项目管理 366

12.1	内部团队还是 SEO 服务	366
12.2	寻找 SEO 服务商	368
12.3	SEO 团队建设	370
12.4	流程及计划	372
12.5	绩效考核	374
12.6	获得高层支持	375
12.7	沟通、培训及规范	377
12.8	应急计划	378

第 13 章 搜索引擎排名因素调查 380

13.1	Google 排名因素 2009	380
13.1.1	与关键词有关的页面排名因素	382
13.1.2	与关键词无关的页面排名因素	383
13.1.3	特定页面链接流行度排名因素	384
13.1.4	全站链接排名因素	385
13.1.5	全站非链接排名因素	386
13.1.6	社会化媒体排名因素	388
13.1.7	用户数据排名因素	388
13.1.8	负面排名因素	389
13.1.9	影响外部链接价值的负面因素	391
13.1.10	地理定位因素	391
13.1.11	附加 SEO 数据	392
13.1.12	链接建设调查	393
13.2	百度排名因素调查 2010	396
13.2.1	与关键词有关的页面排名因素	396
13.2.2	与关键词无关的页面排名因素	398

13.2.3 特定页面链接流行度排名因素	398
13.2.4 全站链接有关排名因素	399
13.2.5 全站非链接相关排名因素	399
13.2.6 社会化媒体排名因素	400
13.2.7 用户数据排名因素	401
13.2.8 负面排名因素	401
13.2.9 地理位置定位因素	403

第 14 章 SEO 案例分析 407

14.1 竞争对手分析	408
14.1.1 了解网站基本数据	408
14.1.2 外部链接	409
14.1.3 Alexa 数据	410
14.1.4 Google 趋势流量	411
14.1.5 网站品牌名称热度	412
14.1.6 英文比较购物网站情况	413
14.2 竞争对手网站研究	414
14.2.1 域名注册信息	414
14.2.2 基本信息	415
14.2.3 外部链接	415
14.2.4 收录	416
14.2.5 QQ 书签	418
14.2.6 外链锚文字	418
14.2.7 网站首页优化	420
14.2.8 其他页面优化	425
14.3 亿赐客网站分析	429
14.3.1 域名注册	430
14.3.2 Google PR 值	430
14.3.3 收录	430
14.3.4 外部链接	432
14.3.5 QQ 书签	433
14.3.6 基本流量数据	433
14.3.7 Google 网管工具数据和分析	443
14.4 关键词研究	450
14.4.1 首页	450
14.4.2 分类页面	455
14.4.3 商家页面	460
14.4.4 品牌页面	460

14.4.5 产品页面	465
14.4.6 搜索页面	466
14.5 亿赐客网站优化建议	467
14.5.1 涉及全站的调整	471
14.5.2 首页修改	482
14.5.3 一级分类页面	484
14.5.4 二级分类页面	486
14.5.5 三级分类页面（产品列表页面）	487
14.5.6 产品页面	490
14.5.7 产品按属性过滤页面	491
14.5.8 搜索页面	491
14.6 执行、效果及后续	492
附录 A SEO 术语	495
结术语	511