

浅论 SEO 及其实现

张红宇

摘要: 搜索引擎优化 (SEO) 是近几年来发展起来的、为使网站达到良好的网络营销效果而进行的提高网站搜索排名的优化工作。本文从 SEO 的发展现状到 SEO 的优化实现, 以及当前伪 SEO 作弊方法进行了技术分析和阐述。

关键词: SEO; 搜索引擎优化; 实现

一、什么是 SEO

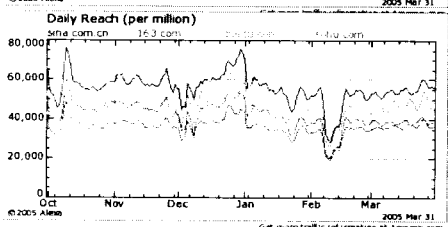
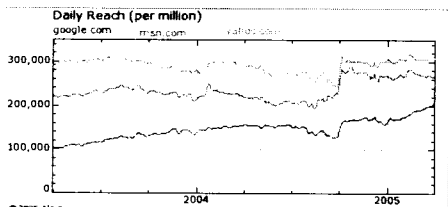
SEO 是英文 Search Engine Optimization 的缩写, 意思是搜索引擎优化。它主要研究的是搜索引擎友好 (Search Engine Friendly)。即通过优化网站结构、网页代码和内容, 使网站对全文搜索引擎友好, 从而帮助全文搜索引擎找到含有最好内容的网页, 提高网站在搜索结果中的自然排名, 获得良好的网络营销效果, 更准确的说是搜索引擎营销效果。

二、搜索引擎发展现状

随着互联网信息的成倍增长, 加上 blog (博客) 和垂直门户的夹击, 门户网站在网民的影响力日益下降。相反搜索引擎在“去中心化”的浪潮中却让网民更加的依赖搜索引擎。

以 Alexa (以发布世界网站排名而引人注目的一个网站) 的数据为例:

我们看下国内的情况, 拿三大门户和百度的 Alexa 数据进行分析



从图中分析看到, 在目前中文搜索引擎的竞争中, 百度暂时领先其他对手, 3721 靠着网络实名软件的推广以及捆绑上网助手目前在中文搜索领域排在第二位, 第三位是 Google。但在 iResearch (艾瑞) 发布的《2004 中国搜索引擎研究报告》中, 百度、雅虎、Google 分别以 36.29%、22.72%、21.22% 的用户占有率占据国内搜索引擎市场的前三位。所以在中文搜索领域需要特别关注这三个搜索引擎。

三、SEO 现状

搜索引擎优化的研究起源于美国, 美国的互联网从业人员于 1993 年开始了搜索引擎优化的相关研究, 并提出了 SEO 的概念。而中国在 2003 年才真正涉足 SEO 研究, 起步较晚。国内早期相关的 SEO 的业务主要集中在搜索引擎的注册, 在 2003 年随着个人网站的推动下, SEO 在国内取得了长足的发展。虽然经过了两年的发展, 但是 SEO 最关注的团体还是集中在个人主页, 商业网站大部分还是缺少 SEO 的意识。目前国内从事 SEO 研究的企业极少 (截止 2004 年 6 月底不超过 10 家),

而在利益的驱动下, 大量伪 SEO 公司纷纷涌入, 使得整个行业比较混乱。它们采用作弊等危害极大的手法欺骗客户 (通过作弊可以使网站排名迅速上升, 但一旦被发现就会被重罚, 直至从搜索引擎消失)。

四、SEO 的常用术语

在分析 SEO 开展前我们先了解几个 SEO 的常用术语:

PageRank: PageRank 取自 Google 的创始人 Larry Page, 它是 Google 排名运算法则 (排名公式) 的一部分, 用来标识网页的等级 / 重要性。级别从 1 到 10 级。PR 值越高说明该网页越受欢迎 (越重要)。

google 左侧排名: Google 搜索结果的排名, 因为 google 将 Adwords 的广告投放在网页右侧, 而左侧严格按照内部的机制, 让系统自动对搜索结果进行排序。

竞价排名: 竞价排名是搜索引擎关键词广告的一种形式, 按照付费最高者排名靠前的原则, 对购买了同一关键词的网站进行排名的一种方式。

外部链接 (Link IN): 也叫反向链接。从其他页面连接到你的网站的链接, 它会影响页面的 PageRank 值。

Link Farm: 直译是“链接农场”的意思, 是 Google 检索常用的作弊方法之一。亦称“大量链接机制”指由大量网页交叉链接而构成的一个网络系统。

Wiki (维客): 其实这个不应该属于 SEO 范畴, wiki 是一个任何人可以修改网站内容的网站 (网站有账号权限的分级, 同时可以随时恢复任意时间点的内容), 主要用来维护大百科知识的网站。

Spider: 是一种用于搜索引擎收集网页资料的专业的 Bot 程序, 又称“网络机器人”或“蜘蛛程序”。

五、影响网站搜索排名的因素

1. 页面重要性, 对于 Google 来说是 PageRank (PR) 值

PageRank 会将网站的外部链接数考虑进去。我们可以这样说: 一个网站的外部链接数越多、外部链接站点的级别越高其 PR 值就越高, PageRank 值的打分还考虑外部链接的质量, 你可以下载和安装 Google 的工具条来检查你的网站级别 (PR 值)。

计算 PR 值公式: $PR(A) = (1-d) + d(PR(t1)/C(t1) + \dots + PR(tn)/C(tn))$ 其中 d 为阻尼因数 (damping factor), 是当你投票或链接到另外一个站点时所获得的实际 PR 分值, 阻尼因数一般是 0.85; PR(A) 表示的是从一个外部链接站点 t1 上, 依据 Pagerank (网页级别技术) 系统给你的网站所增加的 PR 分值; PR(t1) 表示该外部链接网站本身的 PR 分值; C(t1) 则表示该外部链接站点所拥有的外部链接数量。一个网站的投票权值只

有该网站 PR 分值的 0.85, 而且这个 0.85 的权值平均分配给其链接的每个外部网站。

2. 关键字密度

除了 PR 值的影响以外, 影响最大的是关键字密度。很容易理解, 你页面涉及的某个关键字的密度越高, Google 认为你的页面跟某个关键字的关系越大。

不考虑作弊方式 (恶意拷贝关键字, 使用 CSS 隐藏) 增加关键字密度的原则: 适当重复关键字, 尽量符合 W3C 标准, 将页面的 CSS、Javascript 于页面分离, 降低页面的长度, 从而提高页面整体关键字密度。

3. 关键字位置

关键字所在的位置对于搜索引擎判断某个关键字在页面的份量起到很重要的作用。按照次序:

(1) 域名和 html 文件名

域名被搜索引擎认为网站最固定的因素, 例如: 域名里面含有 mp3 关键字的域名, 在用户检索关键字 mp3 的时候具有先天的优势。

(2) 标题

标题是网站的最宝贵的资源, 搜索引擎认为标题是在浏览器标题栏里面显示, 因为要显示给用户, 所以他是 html 文件最重要和最简洁的摘要。适当突出关键字在你标题的比重非常有利于排名的提高。

(3) <H1> 标记

搜索引擎认为这个是仅次于 title 的 html 标记。作弊网站也是最多利用的标记之一; 通过 css 将 h1 设为不显示, 或者颜色跟背景底色相同, 然后将热门关键字放置在 h1 标记中间, 这样可以对网站的用户透明, 而搜索引擎无法区分, 错误地认为页面跟一些热门关键字相关。

(4) 文件时间 (容易忽略的问题)

搜索引擎尽量给用户最新的页面。而页面的修改时间可根据 HTTP 协议里面的 HTTP HEADER 来取的。随着时间的推移页面在搜索引擎的排名会越来越低, 所以建立用 CMS 系统进行管理可根据搜索引擎蜘蛛程序的访问频率保持文件名相同的前提下进行适当的全部重新发布。

(5) 中文搜索引擎不得不提的话题

国内的搜索引擎为了能够让自己占据更大的市场份额, 百度和 Yisou 都可能与其他网站链接回搜索引擎的统计, 在链接回搜索引擎较多的网站在搜索结果的排名上要靠前。而对于页面有竞争对手链接的网站的排名尽量靠后或者不再收录。但是在做 Google 优化的时候无需考虑这方面问题。

六、SEO 的开展和实现

了解了排名原则以后, 开展和实现 SEO 优化, 可以通过以下方式:

1. 让搜索引擎收录尽可能多的页面

搜索引擎收录页面是开展 SEO 的前提, 可以通过搜索引擎里输入“site:站点域名”查询。如果

发现搜索引擎的收录页面为 0, 那么要先到搜索引擎注册你的页面, 对于搜索引擎未收录的网站, 需要进行手工注册 (到 DMOZ (<http://dmoz.org/>) 进行提交)。国外的搜索引擎基本使用 Open Directory Project (人工编辑目录索引类搜索引擎) 项目提交的数据, 搜索引擎一方面使用 ODP 的数据作为自身的目录服务, 也使用目录服务的数据作为蜘蛛爬行的起始点。

为了让浏览器收录更多的页面, 很多网站都构建自己的网站地图 SiteMap, 把尽可能多的页面放置在 SiteMap 的页面里面, 并把网站地图链接在网站的首页, 合理的 SiteMap 使得网站对于搜索引擎来说更加扁平, 使搜索引擎收录更加多的页面。同时页面收录影响着页面的 PageRank 值, 对提高网站的综合排名也有好处。

2、尽量让页面成为静态页面

搜索引擎认为通过动态页面生成的文件, 可变因素较高, 内容有着非常多的不确定性, 所以在收录页面和排名的时候对于静态页面有较高的优先级, 因此在开展 SEO 的时候尽量使用静态页面, 如果网站很难在短时间内进行调整, 可以采取以下策略:

1) Apache 服务器可以使用 mod_rewrite 模块进行页面重写规则来实现。

2) IIS 服务器可以通过技术手段把 news.asp?id=234 这样的链接映射成 news/234.html

这个技术非常简单, 你只需要在服务器上装一个 ISAPI REWRITE (Google888.com 有免费下载), 然后进行相应的参数设置就可以了。

3、尽量使用文字

图片对于搜索引擎来说很难判断图片里面的文字, 所以基本上没有搜索引擎去识别图片里面的意思, 而是引用图片边上文字的语义以及图片的文件名、图片的 Alt 标记来判断图片的。尤其网站的栏目导航条, 可以考虑使用文字为主。美工通过定义 css 和图片背景的结合使得二级栏目链接的美观程度和搜索引擎优化很好的结合起来。

4、搜索引擎来访的跟踪

网站设计不仅仅只是被动的迎合搜索引擎的索引, 更重要是充分利用搜索引擎带来的流量进行更深层次的用户行为分析。目前, 来自搜索引擎关键词统计几乎是各种 WEB 日志分析工具的标准功能。

以 Apache/webalizer 为例, 具体的做法如下: 记录访问来源:

在 Apache 配置文件中设置日志格式为 combined 格式, 这样的日志中会包含相应访问的转向来源: HTTP_REFERER, 如果用户是从某个搜索引擎的搜索结果中找到了你的网页并点击过来, 日志中记录的 HTTP_REFERER 就是用户在搜索引擎结果页面的 URL, 这个 URL 中包含了用户查询的关键词。

在 webalizer 中缺省有针对 yahoo, google 等国际流行搜索引擎的查询格式, 这里我增加了针对国内门户站点的搜索引擎参数设置:

```
SearchEngine yahoo.com p=
SearchEngine altavista.com q=
SearchEngine google.com q=
SearchEngine sina.com.cn word=
SearchEngine baidu.com word=
SearchEngine sohu.com word=
SearchEngine 163.com q=
```

通过这样设置 webalizer 统计时就会将

HTTP_REFERER 中来自搜索引擎的 URL 中的 keyword 提取出来, 比如: 所有来自 google.com 链接中, 参数 q 的值都将被作为关键词统计下来, 从汇总统计结果中, 就可以发现用户是根据什么关键词找到你及访问你的次数, 以及找到你的用户最感兴趣的是哪些关键词等, 进一步在 webalizer 中设置将统计结果倒出成 CSV 格式的日志, 便于以后导入数据库进行历史统计, 做更深层次的数据挖掘等。

5、关键字的运用和维护

一方面热门关键字可以让我们更加清楚地了解用户最关心的是什么内容, 可以通过查询热门关键字来分析用户的需求调整网站的内容。在网站的标题设计和关键字的设置的时候尽可能吸收热门关键字。下面是几个搜索引擎的关键字查询网址:

Google 年度关键字统计: <http://www.google.com/press/zeitgeist/archive2005.html>

百度搜索风云榜: <http://top.baidu.com>

6、定期全部发布网站所有页面

定期完整发布网站所有页面, 可以让页面的最后修改日期得到更新, 有利于排名的提高。

七、提高排名的常用作弊方法

1、隐藏文本 / 隐藏链接

一般指网页专为搜索引擎所设计, 但普通访问者无法看到的文本内容或链接。在形形色色的隐藏技术中, 最常见的就是把文本或链接文字的字体颜色设置为与背景色相同或十分接近。其实通过添加可视文本内容并保证一定的关键词密度可达到相同的优化效果。

隐藏文本内容 (Invisible/hidden text): 意欲在不影响网站美观的前提下通过包含大量关键词的网页提高关键词相关性得分, 从而达到改善搜索引擎排名的目的。

隐藏链接 (Invisible/hidden links): 意欲在不影响网站美观的前提下通过在其它页面添加指向目标优化页的隐形链接, 通过提升链接得分而改善搜索引擎排名。

2、网页与 Google 描述不符

一般做法是先向搜索引擎提交一个网站, 等该网站被收录后再以其它页面替换该网站。“诱饵行为 (Bait-&-Switch)” 就属于此类偷梁换柱之举——创建一个优化页和一个普通页, 然后把优化页提交给搜索引擎, 当优化页被搜索引擎收录后再以普通页取而代之。

3、误导或重复关键词

误导关键词 (Misleading Words): 在页面中使用与该网页毫不相干的误导性关键词来吸引查询该主题的访问者访问网站。

重复关键词 (Repeated Words): 这种作弊技术也被称为“关键词堆砌欺骗 (Keyword Stuffing)”, 它利用搜索引擎对网页正文和标题中出现的关键词的高度关注来对关键词进行不合理的 (过度) 重复。类似的其它做法还包括在 HTML 元标识中大量堆砌关键字或使用多个关键字元标识来提高关键词的相关性。

4、隐形页面 (Cloaked Page)

对实际访问者或搜索引擎一方隐藏真实网站内容, 并向搜索引擎提供非真实的搜索引擎友好的内容以提升排名。

5、欺骗性重定向 (Deceptive redirects)

指把用户访问的第一个页面 (着陆页) 迅速重

定向至一个内容完全不同的页面。

“鬼域 (Shadow Domain)”: 这是最常见的欺骗性重定向技术, 通过欺骗性重定向使用户访问另外一个网站或页面。一般利用 HTML 刷新标识 (Meta Refresh) 来实现。

6、门页 (Doorway Page)

也叫“Bridge/Portal/Jump/Entry Page”。是为某些关键字特别制作的页面, 专为搜索引擎设计, 目的是提高特定关键词在搜索引擎中的排名所设计的富含目标关键词的域名, 且重定向至另一域名的真实网站。

7、复制的站点或网页

最常见的当属镜像站点 (Mirror Sites)。通过复制网站或网页的内容并分配以不同域名和服务器的, 以此欺骗搜索引擎对同一站点或同一页面进行多次索引

8、作弊链接技术 / 恶意链接 (Link Spamming)

典型的作弊链接技术包括: 链接工厂 (link farms)、大宗链接交换程序 (bulk link exchange programs) 和交叉链接 (Cross Link)。

9、其它

日志欺骗行为: 通过对一些页面等级较高的站点进行大量的虚假点击以求名列在这些站点的最高引用者日志中, 从而获得它们的导入链接。

门域 (Doorway Domain): 专为提高特定关键词在搜索引擎中的排名所设计的富含目标关键词的域名, 然后重定向至其它域名的主页。

参考文献:

- 1、《Programming Spiders, Bots, and Aggregator in Java》[美] Jeff Heaton 著
- 2、《搜索引擎与信息获取技术》徐宝文、张卫丰著

(上接第 52 页)

- [2] A. Celentano, V. D. Lecce, A FFT based image signature generation. Proc. SPIE, Storage and retrieval for image and video database, 1997, 3022: 457-466
- [3] A. M. Eftekhari-Moghadam, et al. Image retrieval based on index compressed vector quantization. Pattern Recognition, Elsevier, 2003, 36: 2635-2647
- [4] 魏海, 沈兰荪. 基于分类矢量量化的图像压缩和检索算法. 电子学报, 2001, 29(7): 933-963
- [5] M. K. Mandal, F. Idris and S. Panchanatha. A critical evaluation of image and video indexing techniques in the compressed domain. Image and Vision Computing, 1999, 17: 513-529
- [6] F. Idris, S. Panchanathan. Image indexing using vector quantization. SPIE Proc. Storage and Retrieval for image and video databases. 1995, 2420: 373-380
- [7] A. E. Jacquin. Fractal Image Coding: A review. IEEE Proc. 1998, 81(10): 1451-1464
- [8] Wei Hai, Shen Lansun. Fractal-based image storage and indexing. SPIE Proc. Storage and Retrieval for Media databases, San Jose, CA, USA, 2000, 3972: 421-429