

基于 PageRank 算法的一种搜索引擎优化方法及实现

张光年 李茂青

(厦门大学系统与控制中心 福建 厦门 361005)

摘要: 本文在介绍 Google 等搜索引擎最常用的 PageRank 搜索结果排名算法的基础上, 提出了一种针对 PageRank 算法的搜索引擎优化方法, 设计并用 Java 技术实现了一个采用此方法的搜索引擎优化工具。

关键词: PageRank; Google; 搜索引擎优化; Java

An SEO Method and Implement Based on PageRank Algorithm

Zhang Guangnian Li Maoqing

(Center for Systems and Control, Xiamen University Xiamen 361005)

Abstract: This article is on the basis of PageRank algorithm, which is most frequently used by search engine such as Google, etc. An SEO method based on the algorithm is proposed. Finally, the author designed and implemented an SEO tool based on the new method by Java language.

Keyword: PageRank; Google; SEO; Java

1. 引言

随着互联网信息的成倍增长, 搜索引擎的地位在网民心中日益重要。而同时大量的企业建立网站将产品营销推广出去。如何让网民可以通过搜索引擎更容易的找到自己的网站, 成为了企业网站经营的一个重要问题。Webmaster 们针对搜索引擎数据采集和标引算法频繁设计, 优化自己的网页, 以使其在搜索引擎相关关键词检索结果中排列在前。事实上近年来国外的 SEO (Search Engine Optimizing, 针对搜索引擎的网站优化) 研究风起云涌, 甚至已形成了一个新的业态。

搜索结果排序算法和组织技术的细节作为搜索引擎的商业机密是秘不示人的, 但综合迄今为止这方面的研究实践, 主要有关键字的词频、位置, 网页间的链接流行度这样几种思路。对关键字的词频、位置所进行的优化属于页面上 (onpage) 优化, 通过将关键字放于页面 title 中, 在页面正文中提高关键字词频等等手段, 来提高页面的关键字相似度。这类的页面上优化已经被广大的 Webmaster 所熟知, 并且是可以很容易实现的。而对网页间的链接流行度的优化属于页面外 (offpage) 优化。网页间的链接流行度也是决定页面在搜索结果中排名的重要因素。这种搜索结果排名技术建立在一种针对 Web 文档的复杂算法上, 称之为 PageRank 算法。

本文的目的是在对 PageRank 算法分析的基础上, 提出了一种提高网站主页的链接流行度的方法, 及其基于 Java 技术的实现。

2. PageRank 算法

PageRank^[1]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和重要性。级别从 1 到 10 级。PR 值越高说明该网页越受欢迎 (越重要)。一般搜索引擎将 PageRank 值与网页搜索结果相似度共同作为搜索结果的排序依据。

PageRank 算法的具体思路是, 将某个页面的 PageRank 除以存在于这个页面的正向链接, 由此得到的值分别和正向链接所指向的页面的 PageRank 相加, 即得到了被链接的页面的 PageRank。

算法基于“从许多优质的网页链接过来的网页, 必定还是优质网页”的回归关系, 来判定所有网页的重要性。一个网页的得票越多, 则认为它的重要性也就越高。进一步说, 投票网页的重要性也决定着票本身的重要程度。PageRank 的算法如下:

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)) \quad (1)$$

公式中的 PR 代表页面的 PageRank 数值, T_1, \dots, T_n 代表有链接指向页面 A 的网页, C 是网页出链接的数量, d^2 是阻尼系数 (常数, Google 通常取值 0.85)。由 (1) 式可知, 计算某个页面的 PageRank 值实际上是一个迭代的过程, 计算结果的精确程度依赖于初值的选取和迭代的次数。对于初值一般取 1, 而为了保证实际应用中这个结果总是收敛的, 则加入了阻尼系数 d。

另外需要说明的是, 在 IE 上安装了 Google 工具栏的用户也许看到工具栏上的 PageRank 显示条, 这个工具可以即时地反映出 IE 当前访问的网页在 Google 中的 PageRank 值, 该值在 0 至 10 的范围内变化。这个值并非该页面的真实 PR 值, 而是真实 PR 值的一个对数指标, 对数基应该是 5-6 范围内的某个数值。

3. 提高网站主页 PageRank 值的一般方法

通过对公式 (1) 的分析, 我们知道: 页面的 PageRank 是影响搜索

引擎排名的重要因素。拥有越多的入站链接, 将会提高网页的 PageRank 值。因此, 如何拥有比较多的入站链接就是提高网站主页 PageRank 值的关键所在。下面提出几种提高主页 PageRank 值的一般方法:

(1) 首先要将网站主页提交到各大搜索引擎, 这样搜索引擎才会知道你的网站的存在。

(2) 提供有趣、有价值的网站内容, 这样其他网站的 Webmaster 们会主动和你进行友情链接, 从而提高你的外部链接值。

(3) 可将网站主页添加到行业门户网站、网上论坛、留言板等等各种允许添加网址链接的地方。

(4) 尽量与其他 PageRank 值高的网站交换链接来提高链接权重。需要注意的是, 与其他网站交换链接时首先要查看对方站点是否被 Google 删除, 或是否被 Google 收录, 没有被 Google 收录的站点最好不要做链接。

可以看到上面的几种方法, 其实也是一个新网站进行网络宣传的基本方法。需要注意的是, 在网上论坛和留言板上进行网站宣传时, 一定要有指向网站主页的链接, 而不仅仅是写出网址。

4. 基于 PageRank 的一种搜索引擎优化方法

我们知道, 无论是通过交换链接, 还是在论坛和留言簿上添加网址链接, 都是人为的制造了一些拥有入站链接的页面。通过公式 (1) 可以得到, 拥有入站链接的页面越多, 我们网站主页的 PageRank 值就越高。但是, 这一切都建立在一个前提下, 就是这些拥有入站链接的页面必须要被搜索引擎索引, 即存在于搜索引擎的数据库中。而不同的搜索引擎爬虫的运行机制不尽相同, 相同的页面却不一定能够被不同的搜索引擎索引。

例如, 对于新浪主页: www.sina.com.cn, 通过查询 link:www.sina.com.cn, 我们可以得到在 Google, MSN 和 AlltheWeb 中, 拥有到 www.sina.com.cn 链接的页面数量分别是 423,000, 11,786,540 和 446,000。

因此, 我们可以做出这样的一些页面, 它们包含了各大搜索引擎中索引的所有拥有到我们主页入站链接的页面。将这些页面放在我们的网站上, 让搜索引擎能够发现。这些页面将帮助搜索引擎爬虫发现其他搜索引擎所找到的到我们主页的入站链接。我称这些页面为 LinkMap。

在文章的下半部分, 我将提出一种自动生成 LinkMap 的工具, 我称之为 LinkMap Producer, 及其基于 java 技术的实现。

5. LinkMap Producer 设计思路

在下面的叙述中, 我假设我们要优化的网站主页为 www.xyz.com。

第一步: 链接获取。首先, 我们尽可能从各个搜索引擎中获得拥有到 www.xyz.com 链接的页面的 URL。基本上各大搜索引擎都提供了这样的高级搜索。例如, 对于 Google, MSN, AlltheWeb, 使用 link:www.xyz.com 进行搜索, 就可以得到链接到 www.xyz.com 的页面的搜索结果; 而对于 Yahoo, ASK 就要用 links:www.xyz.com。我们可以通过 Web Service 来获得搜索结果, 像 Google, MSN 就提供了 Web Service 接口供程序调用。对于那些没有提供 Web Service 接口的搜索引擎, 我们就可以直接从搜索结果页面中提取我们所需要的信息, 如页面的 URL。一般情况下, 搜索引擎并不会将其数据库中的数据都提供出来, 但其所提供的页面也是对所要优化页面的 PR 值最有贡献的页面。

第二步: 链接清洗。在从搜索引擎得到链接页面的 URL 后, 我们还要对那些链接页面进行二次验证, 得到其更详细的信息, 并剔除那些对于 Google 搜索引擎不友好的页面。例如, 我们还要得到页面的 Description 信息, 页面的 Keyword 信息, 链接的 Anchor Text 信息; 我们要剔除那些已经是返回 404 错误的页面, 剔除那些含有超过 100 个链接的页面 (这种页面有可能会被 Google 搜索引擎认为是 SPAM 页面), 剔除重复的链接页面 URL。

第三步: 生成 LinkMap 页面。最后, 根据最后剩下的链接页面的 URL, 及其他信息, 生成类似搜索引擎结果页面的 LinkMap 页面。页面格式如下:

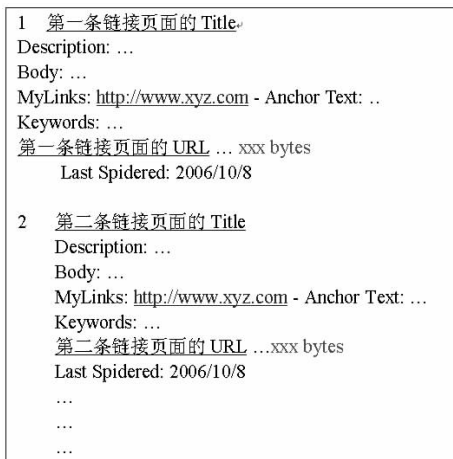


图 1 LinkMap 页面格式

注意在 LinkMap 页面的 <Head> 信息中必须包含 <meta name= "ROBOTS" content= "FOLLOW, NOINDEX ">。这条 meta 信息将告诉 Google 的爬虫, 此页面不要被 Google 收录, 并且允许 Google 的爬虫顺着页面上面的链接继续爬行。这样的 Google 的爬虫就会顺利的到达那些链接页面。

6.LinkMap Producer 的 Java 实现

最后, 我运用 Java 技术对上述的设计思路进行了编程实现。在第一步中, 采用了 Axis³ 与搜索引擎的 Web Service 进行通讯, 对于没有提供 Web Service 接口的搜索引擎, 利用了 HtmlParser 工具包直接对搜索结果页面进行信息提取。在第二步中, 运用了 Java 语言的多线程技术和 HtmlParser⁴ 工具包进行链接清洗。整个工具采用了 Mysql 数据库对数据进行持久化存储。最后, 采用了 SWT/Swing 技术实现了工具的界面, 见下图:

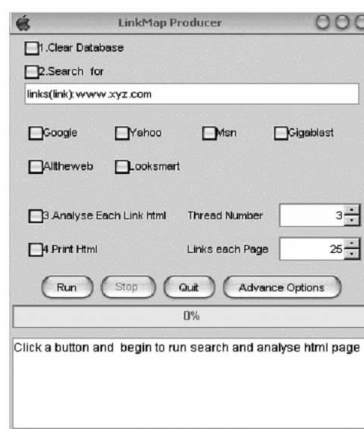


图 2 LinkMap Producer 界面

7. 结论

在此工具完成后, 在某国外保险公司的网站上进行了实施。在工具生成的 LinkMap 放在其网站上一个多月后, 公司网站主页的 PageRank 值由原来的 5 上升到了现在的 8, 证明了本文提出的搜索引擎优化方法的正确性与可操作性。

参考文献

- [1] Sergey Brin, Lawrence Page, The anatomy of a large - scale hypertextual Web Search Engine, [J. www7/Computer Networks, 1998, 30 (1- 7): 107- 117.
- [2] Chris Ridings, PageRank Explained, [EB. http://search.engine - submission.co.uk/ 2001.
- [3] http://ws.apache.org/axis/.
- [4] http://htmlparser.sourceforge.net/.

作者简介: 张光年(1982-), 男, 硕士研究生, 主要研究方向: 管理信息系统, 搜索引擎优化。

李茂青(1953-), 男, 教授, 博士生导师, 主要研究方向: 系统工程, 计算机系统集成。

注释:

以上数据采集于 2006 年 10 月 8 日。

基金项目: 由厦门大学 985 二期信息创新平台项目资助。

(上接第 57 页) 按键采用微动开关。

电源电路: 主要采用专用变压器 (静态电流小), 带有 12V 备用电池的具有自动充电功能电路的不间断电源供电, 以确保在停电时系统也能正常工作。

3.3.2 软件设计 软件功能框图如下:

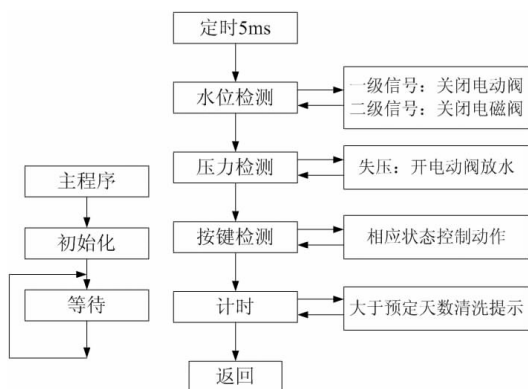


图 5

控制器的工作程序用汇编语言编写, 采用了模块化结构的程序设计方法, 便于使用维护与扩展。控制器软件主要分为主程序模块, 时间中断管理模块, 水和水压信号检测处理模块, 手动控制处理模块, 电磁阀和电动阀开关等模块。各模块主要以中断方式调用。系统设计在充分利用 CPU 功能的前提下, 尽量减少硬件数量。除合理选择硬件外, 在软件上还采取抗干扰陷阱与冗余处理, 提高了系统的稳定性和可靠性。

4. 结束语

本家居自动蓄水供水装置采用了先进可靠的单片机、多传感器和先进的电磁阀等实现了家居的不间断供水功能; 具有硬件电路简单、可靠性高、工作稳定、使用方便和性价比较高等特点。

参考文献

- [1] 李鲁强. 浅观智能化住宅[J]. 工程建设与档案. 2004 (02).
- [2] 李涛. 水传感器的技术分析[J]. 仪表技术与传感器, 2006 (04).

作者简介: 贾书洪, 男, 空军航空大学飞行基础训练基地计算机室教员, 高级实验师, 多年来主要从事电子和单片计算机等嵌入系统的开发研制。