

文章编号:1001-9081(2005)07-1711-02

基于 PageRank 算法的搜索引擎优化策略

张 巍,李志蜀

(四川大学 计算机学院,四川 成都 610065)

(zhangwei_scu@hotmail.com)

摘 要: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,详细阐述了各种网页链接结构对基于 PageRank 算法的网站搜索引擎排名结果可能产生的影响,并分析了实际应用中网站针对 PageRank 算法的各种优化策略,讨论了各自的优点。

关键词:PageRank; Google; 链接

中图分类号: TP391.3 **文献标识码:** A

Optimizing strategies for search engines based on PageRank algorithm

ZHANG Wei, LI Zhi-shu

(School of Computer Science, Sichuan University, Chengdu Sichuan 610065, China)

Abstract: PageRank algorithm used by Google and many other search engines was introduced. The possible influence that various kinds of Web linking structures exerted on the ranking results of search engines based on PageRank were analysed. The advantages of various optimizing strategies of PageRank were discussed.

Key words: PageRank; Google; link

最近几年来,Google 已经成为世界范围内使用最为广泛的搜索引擎之一。Google 的优点不仅仅在于去除无用的(广告)标语构成单一页面的功能、独自の Cache 系统、动态制成摘要信息、为实现高速检索而设置的分散系统(数千台规模的 Linux 群集器)等^[1],而最大的优点正是其检索结果的正确性,与其他搜索引擎相比更优的搜索结果排名,有利于用户尽可能快地找到所需的信息。而研究目前国内互联网网站,可以发现我们在处理网页间的链接结构时,往往存在很大的随意性。殊不知网页间的链接结构正是决定网站在搜索结果中排名的重要因素之一。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,分析各类网络链接结构对搜索结果(PageRank 值)的影响,以及由此产生的搜索引擎优化策略。

1 PageRank 算法

简单的说,PageRank 是代表互联网上某个页面重要性的一个数值。

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。就像后边即将阐述的一样,检索语句不会呈现在 PageRank 自己的计算式上。不管得到多少检索语句,PageRank 也是一定的、文件固有的评分量,该值仅仅依赖于网络的链接结构。

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。

计算某个网页 PageRank 值时所有的人链接都要考虑在内,页面 A 的 PageRank 值计算公式如下:

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

公式中的 PR 代表页面的 PageRank 数值,t1 ~ tn 代表有链接指向页面 A 的网页,C 是网页出链接的数量,d 是阻尼系数(常数,Google 通常取值 0.85)。由于用户在互联网浏览时可能不按当前页面中的链接前进,而随机跳跃到完全无关页面,所以 d 实际上代表的是用户跟随网页链接浏览,而不产生随机跳跃的概率值。

(1)式是在计算网页 PageRank 值的最初公式。由于至今 Google 都没有对外公布它所采用的算法,所以可能 Google 在使用时采取了该公式的一些变形。但这也几乎不影响下面的分析。

由(1)式可知,计算某个网页的 PageRank 值总是依赖于其他的相关页面,所以计算 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^[2]。

另外需要说明的是 PR 值与 PageRank 值的区别。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浏览器当前访问的网页在 Google 中的 PageRank 值的标记,该值在 0 至 10 范围内变化。之所以称之为“标记”是因为它并非网页的真实 PageRank 值,而是真实值的一个对数指标,对数基应该是 5 ~ 6 范围内的某个数值。

网页对它的所有出链接的页面进行 PageRank “投票”。由于随机跳转的可能性,投出的 PageRank 总值比网页本身的 PageRank 值稍小(本身值 * d)。这个值在所有出链接中平均分配。因而指向你的网页的页面的 PageRank 值固然重要,

收稿日期:2004-12-28;修订日期:2005-03-16

作者简介:张巍(1979-),男,天津人,硕士研究生,主要研究方向:计算机网络、信息系统;李志蜀(1946-),男,重庆人,博士生导师,主要研究方向:计算机网络、智能控制。

但该页面的出链接数量也同样不可忽略:出链接越多,你的网页得到的 PageRank 值就越少。另外由于 PR 值是 PageRank 真实值的对数指标,这意味着一个网页从某个较高的 PR 值得到提升要比从较低的 PR 值上升过来需要更多的 PageRank 值。在这种情况下,一个较多出度的 PR8 网页和另一个只有较少出链接的 PR4 页面相比,哪个更有效些?这可能依赖于 PR 值的对数基和具体的链接情况了。

需要注意的是当一个网页以“投票”方式影响其他页面的 PageRank 值时,不会减少自身的 PageRank 值。这并非 PageRank 的转移过程。

2 基于 PageRank 的优化策略

假设我们有一个网站,很显然将网站的 PageRank 平均分配到各个网页(如果可能的话)是非常不明智的,因为我们不可能、也没有必要使网站的所有网页搜索排名都做到很高。如果能将网站的大部分 PageRank 值以某种方式导向其中的一个或少数网页,使得它(们)的排名大大提高,其效果当然要好于平均分配的结果。所以下边讨论的重点不在单个网页的权值,而是考虑整个网站或者说网站中的重要页面的 PageRank 值,这些页面可能是索引页、中心页或是专门为某些搜索术语优化的页面。

2.1 考虑内部链接的影响

网站 PageRank 值即为网站内部所有页面 PageRank 值的和。一个网站的 PageRank 最大值等于它的页面数量。入站链接可以增加这个最大值,而出站链接则能减少之。网站内部链接组织不好,网站可能会达不到最大的 PageRank 值,但是却不可能超过该值。需要注意,虽然增加页面可以增加网站的 PageRank 值,但不是随便增加什么页面都能达到效果。那些完全相同或几乎相同的页面称为 cookie-cutter, Google 认为是垃圾信息并会引发相应的警报机制使得页面甚至是整个网站受到处罚。所以从根本上来说网页要有一定的质量。

下面分析一下网站内部链接如何影响 PageRank 值。在这里我们考虑的是一个相对独立的网站,入站链接和出站链接的影响暂不考虑在内。

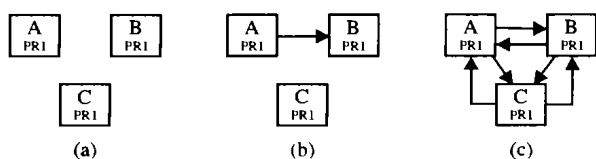


图1 内部链接对 PageRank 的影响

假设一个包含三个网页的网站,没有外部链接(图1)。在(a),(b),(c)的情况下,我们为每个网页分配初值1,阻尼系数保持与 Google 一致(0.85),经过迭代收敛后,得到三种情况的 PageRank 值如下:

(a): PageRank A = 0.15, PageRank B = 0.15, PageRank C = 0.15;

(b): PageRank A = 0.15, PageRank B = 0.2775, PageRank C = 0.15;

(c): PageRank A = 1, PageRank B = 1, PageRank C = 1;

网站(a)的 PageRank 值为 0.45,严重浪费了潜在的 PageRank 值。(b)的情况稍好一些,总值 0.5775 比上一个例子的摇摆页情况,在这里不作讨论)。在(c)的链接结构下,网站达到了 PageRank 最大值,也可以通过环形结果获得:A 到 B, B 到 C, C 再到 A。同样的情况也可以将页面增加到 3 个以上。

可见链接的不好,完全可能浪费潜在的 PageRank 值。根据试验的规律,得出针对内部链接结构的第一个优化策略:一般来说,环形链接或者任意两个页面都有相互链接时能达到网站 PageRank 大值。

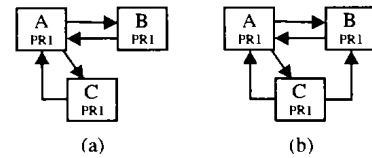


图2 索引页的 PageRank

假设把 A 作为索引页,有(a),(b)两种链接结构,省略迭代计算过程后迭代结果如下:

(a): Page A = 1.459459, Page B = 0.7702703, Page C = 0.7702703;

(b): Page A = 1.298245, Page B = 0.9999999, Page C = 0.7017543;

两种结构的总值仍然是 3(最大值),所以没有浪费。但(b)的情况下,A 明显损失了 PageRank,页面 C 也损失了一部分 PageRank,因为 A、B 分享方式替代了 A 独享,A 通过 A→C 链接反馈回 C 的值也就减小。

所以得出第二条优化策略:为了获得索引页的最大 PageRank 值,其他页面应该尽可能减少相互链接。如果某个页面链接到的页面有回路链接,那么在这个页面上增加一个新的出链接会导致间接损失一部分的 PageRank 值。如果没有这样的回路,则不会减少 PageRank 值。这在内部链接中并不十分重要,但对发生到网站外的链接时就不一样了。

可见,通过组织内部链接,可以将网站的 PageRank 值导向某个选定的页面。内部链接可以根据网站的 PageRank 需要来组织,但必须是 Google 认可的页面。

2.2 入站链接和出站链接

入站链接(由网站外部进入的链接)是增加网站 PageRank 值的方式之一。入站链接来自何处并不重要,Google 认为只要网站管理员没有对链入网站的其他网站产生控制,就不会因为这种链接做出处罚。

链接页的 PageRank 值很重要,但同时出链接的数量也相当关键。例如:如果是一个 PageRank 值为 2 的网页的唯一出链接,将得到 $0.15 + 0.85(2/1) = 1.85$ 的值;而一个 100 个出链接的 PageRank 8 网页,得到的是 $0.15 + 0.85(7/100) = 0.2095$ 。很明显,PR2 链接更有效。一旦有 PageRank 值注入到网站中,计算将需要重新进行。某些页面的值增加,某些保持不变,这依赖于内部链接结构,但肯定不会有页面会损失 PageRank 值。

入站链接指向你试图导向的重要页面更有好处,如果 PageRank 注入到其他页,则会因为内部链接分散到网站中。索引页也会得到提升,但不如直接链接提升得多。直接得到入站链接的页面得到的值最大。

第三条优化策略:将网站索引页作为引入入站链接的最佳目标。

出站链接会导致网站 PageRank 值的消耗。为了抵消这种消耗,需要确保链接是互给的。互惠链接可能得到也可能损失 PageRank 值,所以交换链接时需要特别小心。

当 PageRank 值随着指向另一个网站的链接而引出时,内部链接的所有页面都将受到影响。虽然具体的 PageRank 值变化情况依赖于链接结构,但一般来说给出链接的网页往往是损失 PageRank 值最大的,因此得出第四条优化策略:出站

(下转第 1718 页)

均仅有两个字段,即 HZ 和 SYM 字段,字段 HZ 存放单个汉字,字段 SYM 存放对应汉字的首音码。在表 GBKSYM1. DB 中,字段 SYM 宽度为 1 个字符即可,在表 GBKSYM2. DB 中,该字段宽度为 3 个字符。表 GBKSYM2. DB 的存储记录如表 4 所示,字段 SYM 中两首音码中间带“*”,表示对应汉字的首音码不能自动确定,需要人工选择。

由于 GBK 中 GB 汉字使用频度较高,为了提高查询速度,在表 GBKSYM1. DB 中 GB 汉字排序在前,扩充汉字排序在后。

4 汉字型姓名智能转换为首音码的实现

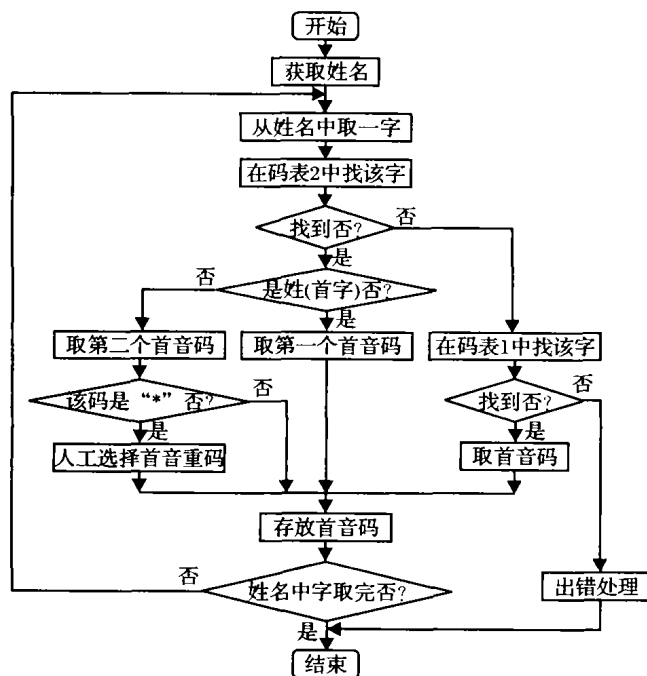


图1 汉字型姓名智能转换为首音码的算法流程

使用上述分析生成的首音码表,并应用如图 1 所示的转换算法就可以方便地将输入的姓名自动转换成其对应的首音码。图 1 中的码表 1 和码表 2 分别指表 GBKSYM1. DB 和表 GBKSYM2. DB。由于表 GBKSYM2. DB 中汉字很少,为了提高查询速度,先查询该表,若找不到再查询表 GBKSYM1. DB。

5 结语

通过对 64 000 多个无重复姓名进行测试,直接使用 GBK 拼音表首音重码选择率为 20%,去掉首音相同汉字后首音重码选择率降为 14.4%,而采用本文设计的码表及转换算法首音重码选择率降至 3.7%。可见,使用本码表及转换算法可以大幅度降低首音重码选择率,显著提高转换效率。该码表及算法已经在温州职业技术学院分院图书馆、温州菜篮子集团蔬菜种子批发公司中得到应用,效果很好。若对偶尔遇到的重码选择感到不便,也可以去掉图 1 中的“人工选择首音重码”功能,由系统自动组合各首音重码,就可以实现全自动转换。但这将增加一定的数据冗余,若想要去掉冗余数据,可以在转换后再进行人工编辑。当然,该码表可能忽略了一些汉字的首音重码,在使用时只要按上述规则添加即可。虽然本文仅对用作姓名的汉字的音码重码进行分析,但其分析方法也可以为解决汉字的其他方面(如药名、书名、歌曲名等)的首音重码问题提供参考。

参考文献:

- [1] 李琦. 基于汉字拼音首字母的信息查询法的分析与实现[J]. 四川轻化工学院学报, 2003, 4(3): 71-74.
- [2] 张春生, 廉洁, 包图雅. 拼音缩写码在医药行业管理中的应用[J]. 内蒙古民族大学学报, 2003, 18(2): 121-122.
- [3] GBK, 汉字内码扩展规范[S], 1995.
- [4] (宋)佚名, 王应麟, [梁]周兴嗣, (清)李毓秀, 木子. 百家姓三字经千字文弟子规[M]. 乌鲁木齐: 新疆青少年出版社, 1996.
- [5] 金山词霸 2005 专业版[CP/DK]. 金山公司, 2004.

(上接第 1712 页)

链接放在 PageRank 较低的页面,造成的 PageRank 损失较小。

任何一个网站都几乎不可能没有出站链接,但不巧的是,所有的“正常”链接都会泄漏 PageRank 值。但还有些“特别”的链接方式不用泄漏。PageRank 泄漏与否依赖于 Google 能否识别出链接,这样可以使用 Google 不能识别或不考虑的连接,包括表单处理(form action)和包含 JavaScript 代码的连接^[3]。

表单的 action 属性不一定是处理表单脚本的 url,它可以指向任何网站的任何一个页面。

例子:

```
<form name = "myform" action =
"http://cs. scu. edu. cn/somepage. html" >
<a href = "javascript: document. myform. submit()" >
四川大学计算机学院 </a >
```

此外, action 属性甚至可以不必位于 form 表单而在 JavaScript 代码中,而 JavaScript 代码可以位于存储路径的 js 目录下,而该目录一般 Google 的 spider 程序都不访问。

3 总结及 PageRank 改进

PageRank 值由网络链接结构决定,与具体的检索内容无关,因而检索期间消耗很小,优于早期的 HITS 算法。在不考虑网页内容具体需要的情况下,提出的优化策略有利于提高网站在基于 PageRank 算法排名的搜索引擎搜索结果中的排名。这种效应也许短时间内尚不明显,但随着页面的增加和网站间链接的逐渐增多,最终的效果还是可观的。

同时,由于 PageRank 算法的检索无关性,也可能导致一些不利的结果,例如对一些词汇在特定的上下文中有特定的含义,或是一些专业词汇,仅仅依靠 PageRank 排名的结果可能不太令人满意,比如同样是查找“结构”这个词,在建筑学的上下文中,和芯片制造的上下文中,用户希望得到的检索结果必然不尽相同。但由于 PageRank 是网页的固定属性,可能就达不到期望的效果了。如果将整个互联网看成一个维度,那么 PageRank 则是该维度上的一个矢量,针对以上的缺陷,可以考虑建立这类矢量的一个矢量集。换句话说,可以针对某些指定的主题词计算出多个 PageRank 值,然后根据检索内容匹配相应主题词的网页 PageRank 值^[4]。当然,在结果排序时用到的 PageRank 值仍然是唯一的。这种改进在检索期间的消耗上有所增加,但在结果排序上却有大大提高。

参考文献:

- [1] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[A]. Proceedings of the Seventh International World Wide Web Conference[C], 1998.
- [2] BABA H, 馬場肇. Google の秘密 - PageRank 徹底解説[EB/OL]. <http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html>, 2003.
- [3] JEH G, WIDOM J. Scaling personalized web search[R]. Stanford University, 2002.
- [4] HAVELIWALA TH. Topic-Sensitive PageRank[A]. Proceedings of the Eleventh International World Wide Web Conference[C], 2002.