

PageRank 技术分析 with 搜索引擎检索效率研究

杨海东¹, 张莉²

(1. 淮阴师范学院 计算机科学系, 江苏 淮安 223001; 2. 淮阴师范学院 图书馆, 江苏 淮安 223001)

摘要: 通过对 Google 的 Pagerank 算法的分析, 对当前互联网搜索引擎的检索效率的改进与提高提出了自己的看法.

关键词: Google; PageRank; 搜索引擎; 检索效率

中图分类号: TP391 **文献标识码:** A **文章编号:** 1671-6876(2003)03-0230-04

0 引言

搜索引擎技术的发展还是近十年的事, 随着网络技术的迅速发展, 国际互联网已经成为越来越多的人获取信息的渠道, 然而如何检索数十亿的网页成为每个网民都关心的问题. 搜索引擎(Search Engine) 的出现暂时缓解了这一矛盾. 目前, 网上可供使用的搜索引擎比较多, 其中优秀的如 AltaVista、Infoseek、yahoo!、Google 等. 根据搜索引擎的技术原理, 我们可以将搜索引擎分成基于 Robot 的搜索引擎、基于分类目录的搜索引擎和元搜索引擎(Meta Search Engine). 这三类搜索引擎各有各的优点, 如基于 Robot 的搜索引擎可以相对比较全面地收集网页; 基于分类目录的搜索引擎由于人工参与可以得到更有价值的信息; 元搜索引擎没有自己的数据库, 但是它可以同时使用多个搜索引擎进行搜索, 再将排名靠前的结果经过技术处理合并后送交用户, 使搜索结果更加合理.

Google 搜索引擎属于基于 Robot 的搜索引擎, 诞生四年来它获得了巨大的成功, 著名分类目录搜索引擎 YAHOO! 也开始采用它的搜索引擎技术. Google 的成功在于它改变了过去的基于 Robot 的搜索引擎只注意网页数量而不重视质量的状况, 实现了网页搜索结果由量到质的转变, 而这一切都得益于它的网页重要性评价算法——PageRank(网页级别).

1 PageRank 算法及分析

PageRank 算法 1998 年由斯坦福大学(Stanford University) 的 Sergey Brin 和 Lawrence Page 提出, 它借鉴了传统情报检索理论中的引文分析方法: 当网页 1 有一个链接指向网页 2 时, 就认为网页 2 获得了一定的分数, 该分值的多少取决于网页 1 的重要程度, 即网页 1 的重要性越大, 网页 2 获得的分数就越高. 由于国际互联网上的链接相互指向的复杂程度, 该分值的计算过程是一个迭代过程, 最终网页将依照所得的分数进行排序并将检索结果送交用户, 这个量化的分数就是 PageRank 值, 其计算公式如下:

$$PR(A) = (1 - d) + d \left(\frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)} \right) \quad (1)$$

其中 $PR(A)$ 是网页的页面级别, d 为界于 $(0, 1)$ 区间的衰减系数, 一般取 0.85 左右, T_1, T_2, \dots, T_n 为指向网

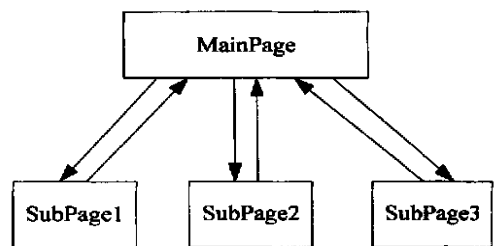


图 1 简单的网站结构图

页 A 的其它网页, $C(T_n)$ 是网页 T_n 中向外指出的链接数目. 如果有图 1 所示的网页结构, 每个网页的初始 PageRank 值增均为 1.00000, 应用上述公式我们对图中各网页进行迭代计算, 其运算次数和结果如下表所示:

表 1 简单结构网站的 PageRank 值的计算

计算次数	MainPage	SubPage1	SubPage2	SubPage3	总 值
1	2.70000	0.43333	0.43333	0.43333	4.00000
2	1.25500	0.91500	0.91500	0.91500	4.00000
3	2.48325	0.50558	0.50558	0.50558	4.00000
4	1.43924	0.85358	0.85358	0.85358	4.00000
5	2.32665	0.55778	0.55778	0.55778	4.00000
6	1.57235	0.80921	0.80921	0.80921	4.00000
7	2.21350	0.59550	0.59550	0.59550	4.00000
8	1.66852	0.77716	0.77716	0.77716	4.00000
9	2.13176	0.62275	0.62275	0.62275	4.00000
10	1.73801	0.75399	0.75399	0.75399	4.00000
20	1.88330	0.70557	0.70557	0.70557	4.00000
30	1.91191	0.69603	0.69603	0.69603	4.00000
50	1.91864	0.69378	0.69378	0.69378	4.00000
80	1.91892	0.69369	0.69369	0.69369	4.00000
90	1.91892	0.69369	0.69369	0.69369	4.00000

从 PageRank 的计算公式中我们可以看出, 由于各个网页之间存在相互链接, PageRank 值的计算是一个递归的计算, 随着递归次数的增加, 计算结果会越来越精确地接近某一确定值, 但实际的计算应该是有限次数的计算, 因此, 当结果充分接近某一值的时候就停止计算, 如表 1 所示, 当运算次数越来越多时(80 次左右时), 各个网页的 PageRank 也越来越接近最终结果. 现在的精确计算是 Google 的商业机密, 但根据 L Page(1999 年)发表的公开文献, 对 3.22 亿个链接进行的递归计算, 52 次即可得到稳定的 PageRank 值. 而且计算次数与网页的数量是呈对数增长的. 目前, Google 的能搜索到网页数量已经达到 30 多亿个页面, 上百次的 PageRank 运算次数是 Google 遍布全球的一万台服务器完全能够承受的.

2 PageRank 算法与检索效率研究

对于搜索引擎的检索效率, 我们主要从三个方面进行研究: 查全率、查准率和相关性排序.

2.1 查全率 (Recall Ratio)

搜索引擎查全率有赖于样本的大小及检索关键字的准确性. 经过近二十年的发展, 目前 Internet 上的信息资源浩如烟海, 很多信息甚至在其尚未发挥作用之前就已经湮没了. 同时, 也有相当数量的信息因为缺乏与外部网络的链接而无法被搜索引擎搜索到, 所以查全率在很大程度上具有相对性, 对于网络搜索引擎在查全率这一指标上我们首先倾向于它收集网页数量的多少. 搜索引擎从其诞生以来就一直致力于收集更多的网页并将它们罗列在一起, 迄今为止, 基于 PageRank 算法的 Google 搜索引擎已经收集了超过 30 亿张网页 (3 083 324 652, 2003 年 5 月), 如此多的网页自然给用户更多的选择.

为了更准确地检索到所需要的资源, 检索词的选取也是关键. 不管在何种语言中, 同一事物具有不同的指示词都是平常的事, 如计算机和电脑, 化学中化合物的俗名及学名、分子式等等, 在国际互联网的检索中, 还存在着不同语言对同一事物的称呼, 因此在使用搜索引擎时检索词的准确选取是提高查全率的关键, Google 共提供 35 种语言供用户选择, 默认情况下, 世界上任一地区的检索者都可以直接使用所在地区的语言进入搜索.

2.2 查准率 (Precision Ratio)

基于 Robot 的搜索引擎从初始网页出发, 顺着网页的链接进行查找, 当到达一个新网页时就在其 $\langle \text{head} \rangle$ 、 $\langle \text{title} \rangle$ 、 $\langle \text{meta} \rangle$ 、 $\langle \text{a href} = \dots \rangle$ 中查找关键字, 如果这些因素都明确反映了网页的内容, 我们就能得到一个适用的网页. 但是在实际的网络中存在着下面两种情况:

1) 主要元素中使用无意义的词汇

现在的网页已经很少用记事本逐行书写 HTML 代码,更多是使用网页编辑工具如 FrontPage、DreamWeaver 等,许多网页使用系统默认的名称 INDEX、HTML、DEFAULT、HTML 等等,或者使用一些欢迎词语。更多的网页在编写时不使用 Meta,致使搜索引擎无法识别网页的重要性,有些网页为了美观,重要的标题往往使用图片形式,并且不使用替代性文字,从而错过了被搜索引擎搜索到的机会,造成漏检。

2) 使用具有欺骗性的词汇

有些个人网站或商业网站,为了取得在搜索引擎中靠前的位置,提升自己网站的点击率,故意在 Meta 等重要标记中堆砌关键词,或者设置与其网页无关的但比较流行的词汇,用来欺骗搜索引擎。甚至有的网站只有框架,而无任何内容。

对于前者,网络管理员除了完备网页元素外,还要让自己被列入一个开放式的目录计划(如 www.dmoz.org),再经过一段时间的宣传、交换链接,就可能使其在 Google 中的排名向前靠拢;对于后者,Google 的网站的网络优化员会努力侦测并对违规网站作出处罚。然而,这还要更依赖于搜索引擎对网站内容的甄别。

2.3 相关性排序(Relevant sort)

这是 PageRank 算法的精华所在。对网络的使用情况调查表明:正常情况下,搜索引擎用户只会查看靠前的检索结果,下表是科研人员对北大天网搜索引擎 2001 年 4 月的统计:

页号	1	2	3	4	5
百分比(%)	47.0	12.1	7.8	5.0	3.7

数据表明:前 5 页的点击率就占到总点击率的 75%,这就要求搜索引擎要将最重要的结果尽量向前排列,以使网络浏览者在感到厌倦前就已经获得所需结果。Google 是笔者最常使用的搜索引擎,大多数情况下,总可以在前几页找到所需结果。但某些情况下,在开始的几页中难以找到合适的结果,这可能源于 PageRank 算法的几个出发点:

1) COM 网站比其它网站更加重要,但这一点并不是十分肯定的。许多 .com 网站规模大、涉及面广,对某一领域的研究并不是很深入,而该领域的专业网站对所论述的内容较深刻,因而要比综述性的商业网站更有价值。

2) 认为网页中向外指出的链接(出链)降低了网页的重要性(PageRank 值)。公式(1)不加区分地认为凡是向外指出的链接都有负值,而不具体考虑出链与本网站在内容上的相关程度。

为了减少这种负面影响,对公式(1)的修正公式如下:

$$PR(A) = (1 - d) + d \left[PR(T_1) \sum_{k=1}^m f(T_{1k}) + \dots + PR(T_n) \sum_{k=1}^m f(T_{nk}) \right] \quad (2)$$

公式(2)中当 A 的一条的超链接与 A 在内容上相关时 $f(T_{ik})$ ($i = 1, 2, \dots, n$) 为正,否则为负值。这样,专业网站就不会因为其指向类似的专业网站而降低自身的 PageRank 值。

相关性排序重要的一点还体现在用户点击率上。相关调查表明,与用户浏览搜索引擎的结果类似,用户对 URL 的点击也集中在部分 URL 上。确定用户行为对搜索引擎的生存与发展至关重要,一个好的搜索引擎系统应该考虑这些因素,将点击率最高的网页向前排,30%左右的的点击率最高的 URL 就可以满足 70%~80%用户的检索需求。

3 结束语

对一个优秀的搜索引擎的检索效率来说,除了上述指标而外,还包含响应时间(Response Time)、新颖率(Novelty ratio)等多项重要指标。尽管 Google 在技术方面领先于其它搜索引擎服务商,但应该看到,PageRank 技术本身也并不是十全十美的,Google 还存在着许多问题,搜索引擎的完善与发展还有很长的路要走,研究 PageRank 是为了能在此基础上提出更有创意的排序算法。同时,更多新技术的应用也将会给搜索引擎技术带来新的活力,搜索引擎将会有有一个美好的未来。

参考文献:

- [1] Brin S, Page L. The anatomy of a Large Scale Hypertextual Web Search Engine[EB/OL]. <http://www-db.stanford.edu/~backrub/google.html>. 2003-6-10/2003-6-25.
- [2] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank Citation Ranking:Bringing order to the Web[EB/OL]. <http://www-db.stanford.edu/~backrub/pageranksub.ps>, January, 1998-3-5/2003-2-10.
- [3] 曹军. Google 的 PageRank 技术分析 [J]. 情报杂志, 2002, 19(10): 15 - 18.
- [4] 阎放, 张海涛, 朱宏谊. Google 搜索引擎 PageRank 技术的优化 [J]. 情报科学, 2002, 22(12): 1333 - 1335.
- [5] 赖茂生, 王延飞, 赵丹群. 计算机情报检索 [M]. 北京: 北京大学出版社, 1993.

The Technical Analysis of Page Rank and the Efficiency Study of Search Engines

YANG Hai-dong¹, ZHANG li²

(1. Department of Computer Science, Huaiyin Teachers College, Huanan 223001, China)

(2. Library, Huaiyin Teachers College, Huanan 223001, China)

Abstract: Based on the analysis of PageRank technique employed by Google, ideas are put forward of improving search efficiency of current internet search engines.

Key words: Google; PageRank; search engine; search efficiency

[责任编辑:李晓薇]

(上接第 206 页)

参考文献:

- [1] 唐竞新. 模拟电子技术基础解题指南 [M]. 清华大学出版社, 1998.
- [2] 陈大钦. 模拟电子技术基础 [M]. 武汉理工大学出版社, 2001.
- [3] 康华光. 电子技术基础 [M]. 高等教育出版社, 1999.
- [4] 陈小虎. 电工电子技术 [M]. 高等教育出版社, 2000.
- [5] 龚淑秋, 李忠波. 电子技术试题题型精选汇编 [M]. 机械工业出版社, 2000.
- [6] 金长义, 李朝鲜, 束亦清. 计算机电路 [M]. 电子工业出版社, 1994.
- [7] 江晓安. 模拟电子技术 [M]. 西安电子科技大学出版社, 1993.
- [8] 哈益明, 刘连伟. 中国高等教育研究论丛 [M]. 中国社会出版社, 1999.

An Analysis of How to Distinguish Feedback Patterns in Amplified Circuits

ZHANG Qiu-yun

(Donggang College, Huaihai Institute of Technology, Lianyungang 222069, China)

Abstract: Briefly stated in this article are the concepts, the characteristics and effects of various feedback patterns. With the help of charts and diagrams, the methods to distinguish those patterns are introduced in detail.

Key words: amplified circuits; feedback; judgement of feedback patterns

[责任编辑:李晓薇]