

一种改进的搜索引擎网页排序算法

王秀平^{1,2} 李治柱¹

(上海交通大学信息安全工程学院 上海 200030)¹ (江苏信息职业技术学院 江苏无锡 214101)²

摘要 本文在分析目前流行的基于链接分析的搜索引擎网页排序算法的基础上,指出基于链接分析网页排序算法未充分考虑网站性能因素。针对这一缺点,本文提出了一种改进的网页排序算法,它考虑了网站的性能等因素,对提高搜索引擎的服务质量有很好的参考价值。

关键词 网站性能,搜索引擎,网页排序算法,数据挖掘

An Improved Page Ranking Algorithm in Search Engine

WANG Xiu-Ping^{1,2} LI Zhi-Zhu¹

(School of Information Security Engineering, Shanghai JiaoTong University, Shanghai 200030)¹

(Jiangsu College of Information Technology, Wuxi 214101)²

Abstract In this paper we point out that link-based Web page ranking algorithms do not adequately considerate the performance of Web sites, based on analyzing these algorithms which are popular at present. In order to cut this shortcoming, we propose an improved page ranking algorithm which can enhance QoS (quality of service) of search engine.

Keywords Website performance, Search engine, Web page ranking algorithm, Data mining

1 前言

随着互联网的迅猛发展与广泛应用,越来越多的人借助于搜索引擎进行网上信息的检索。这在给搜索引擎带来无尽的商业机遇的同时,也带来了巨大的挑战。挑战之一是 Web 信息资源呈指数级增长,搜索引擎无法索引所有页面;挑战之二是 Web 信息资源的动态变化,搜索引擎无法保证对信息的及时更新;挑战之三是传统的搜索引擎提供的信息检索服务,不能满足人们日益增长的对个性化服务的需要。这些挑战有一个共同的焦点,就是如何提高搜索引擎的服务质量(QoS)。

面对这些挑战,国内外不少学者和研究人员提出了多种解决方案。本文结合作者自己的研究,在分析目前流行的搜索引擎网页排序算法的基础上,指出网页排序应考虑网站性能因素,并提出了一种考虑网站性能因素的网页排序算法的改进方法。

2 基于链接分析的网页排序算法

目前主流的搜索引擎网页排序算法主要有 PageRank 算法和 HITS 算法等。

2.1 PageRank 算法

PageRank 算法是 Page L. 等人提出的,并在搜索引擎 Google 中应用且取得了巨大的成功。

PageRank 算法基于这样的认识:如果网页 A 存在一条指向网页 B 的超链,则认为 B 得到了 A 的认可;如果有许多网页指向网页 B,则可以说 B 相对比较重要;如果网页 A 比较重要并链接到网页 B,则说明网页 B 也比较重要。计算网页 A 的 PageRank 值的公式为:

$$PR(A) = (1-d) + d(PR(t_1)/C(t_1) + \dots + PR(t_n)/C(t_n))$$

式中: t_i ($i = 1, 2, \dots, n$) 为指向网页 A 的其他网页; d 为界于 (0, 1) 区间的衰减系数,表明一个网页的用于投票的权值,Google 通常取 0.85; $C(t_i)$ 为网页 t_i 向外指出的链接数目。 $PR(t_i)$ 表示该外部链接网站本身的 PR 分值。通过简单的迭代算法可以计算出 $PR(A)$ 的值。搜索引擎的返回结果按 PageRank 值由高到低排序。

2.2 HITS 算法

HITS(Hypertext Induced Topic Search)算法由 Kleinberg 提出。该算法将网页分成 hubs 网页和 authorities 网页, authorities 为具有较高价值的网页, hubs 为指向较多 authorities 的网页。通过迭代可以最终得到排名最高的 hub 网页和 authority 网页。

HITS 算法描述如下:将查询 q 提交给传统的基于关键字匹配的搜索引擎。搜索引擎返回很多网

页,从中取前 n 个网页作为根集,用 S 表示。 S 满足如下 3 个条件:a. S 中网页数量相对较小;b. S 中网页大多数是与查询 q 相关的网页;c. S 中网页包含较多的权威网页。通过向 S 中加入被 S 引用的网页和引用 S 的网页将 S 扩展成一个更大的集合 T 。以 T 中的 Hub 网页为顶点集 V_1 ,以权威网页为顶点集 V_2 , V_1 中的网页到 V_2 中的网页的超链接为边集 E ,形成一个二分有向图 $SG=(V_1, V_2, E)$ 。对 V_1 中的任一顶点 v ,用 $h(v)$ 表示网页 v 的 Hub 值,对 V_2 中的顶点 u ,用 $a(u)$ 表示网页的 Authority 值。开始时 $h(v)=a(u)=1$,对 u 执行 I 操作修改它的 $a(u)$,对 v 执行 O 操作修改它的 $h(v)$,然后规范化 $a(u)$, $h(v)$,如此不断地重复计算下面的操作 I,O,直到 $a(u)$, $h(v)$ 收敛。

$$I \text{ 操作: } a(u) = \sum_{v: (v,u) \in E} h(v) \quad (1)$$

$$O \text{ 操作: } h(v) = \sum_{u: (v,u) \in E} a(u) \quad (2)$$

每次迭代后需要对 $a(u)$, $h(v)$ 进行规范化处理:

$$a(u) = a(u) / \sqrt{\sum_{q \in V_2} [a(q)]^2}$$

$$h(v) = h(v) / \sqrt{\sum_{q \in V_1} [h(q)]^2}$$

式(1)反映了若一个网页由很多好的 Hub 指向,则其权威值会相应增加(即权威值增加为所有指向它的网页的现有 Hub 值之和)。式(2)反映了若一个网页指向许多好的权威页,则 Hub 值也会相应增加(即 Hub 值增加为该网页链接的所有网页的权威值之和)。HITS 算法输出一组具有较大 Hub 值的网页和具有较大权威值的网页。

3 基于链接分析的网页排序算法未考虑网站性能因素

从以上介绍可以看到,尽管 PageRank 算法和 HITS 算法有着明显的不同点,但两者均是对存储在搜索引擎数据库中的网页进行链接分析从而确定搜索返回结果排序的算法。这种算法未充分考虑返回结果集合中网页所属的网站的性能。

影响网站性能的主要因素有页面载入时间、有效访问次数、有效访问时间、网页的更新时间、与其它网站的链接数、其它网站与其链接数以及网站的内容与设计等。网页排序算法只考虑了网页间的链接信息,忽略了网站性能的其它因素。

3.1 未考虑页面载入时间

页面载入时间是指从用户发出指令到页面被显示所用的时间。这是影响网站性能最重要的因素之一。页面载入时间过长最容易招致访问者的不满,访问者要么中断操作要么离开网站。互联网行业有一

个比较流行的“八秒钟规则”,即如果 Web 站点在 8 秒钟后没有还反应,那么访问者一般会弃之而去。

我们在中国教育科研网宽带环境下,在网络非繁忙时间(23:00~8:00)对目前常用的中文搜索引擎 Google 和百度进行 15 个随机选取的关键词,每个关键词前 100 个检索结果进行了测试,发现网页不可达的情况达 5~13%,另有 1~2% 的网站的面载入时间超过了 8s。文[7]的研究表明,在较慢的网络传输速度(小于 3kb/s)下或上网高峰时间,载入时间超过 8s 的网页将会大大增加。

如果搜索引擎返回较多的载入时间过长的页面,会在一定程度上影响其用户服务质量(QoS)。

3.2 未考虑用户点击信息

在通用搜索引擎中,用户的参与也很重要。用户的选择,即搜索引擎用户的对返回结果的每一次点击,实际上就是对相应网页的一次选择。用户的这种选择应该作为评价网页重要性的一个因素。

针对这一问题,有研究人员提出了改进的算法,依照用户的选择来调整搜索返回结果集的排序。这一思想已经得到了一些搜索引擎(如 Google)的采纳。分析这些搜索引擎的返回页面的代码,我们发现其中包含了统计用户点击次数的代码(Google 中的代码是 `onmousedown=return clk(1,this)`)。我们在测试中也发现,在搜索引擎的一个更新周期(一般为 3~4 周左右,由 spider 程序的效率及搜索引擎的规模决定)内,对同一关键词检索返回结果的排名存在着变动(见表 1),这种排名变化不是因为搜索引擎数据库更新引起的,正是因用户的点数产生的。

表 1 Google 排名变化表(关键词:蘑菇云)

url	3 月 31 日	4 月 1 日
http://www.huaxia.com/zt/2002-29/104683.html	5	6
http://www.insm.org/war/nucleus.file/199919061739.shtml	6	5
http://www.szed.com/n1/ca731679.htm	7	8
http://www.peacehall.com/news/gb/pubvp/2004/02/200402091528.shtml	8	7
http://www.pep.com.cn/200309/ca289195.htm	9	11
http://mil.qianlong.com/5051/2003-7-30/54@976029.htm	10	9
http://www.pep.com.cn/200401/ca354832.htm	11	10

这种考虑用户点击信息的思想应该得到其他一些搜索引擎的借鉴。

3.3 未考虑网页的更新时间

基于链接分析的网页处理技术在计算页面重要性(PageRank 值或 Authority 值)时所使用的信息仅限于链接结构(或者还有文本内容),而决定一个网页的 PageRank 值或 Authority 值的主要因素是指向该网页的链接个数,如果一个网页被放到 Internet 上不久,由于时间短暂,许多其他网页还没有指向它,自然 PageRank 值或 Authority 值就很低。在搜索引擎返回的结果中往往就会把它排在较后的位置,这样,返回结果中新的网页反而被放在后面,可能正好与用户的需求恰恰相反,因为很多情况下,用户希望看到最新的信息。

目前,全球虽然有数以千万计的网站,但不少网站内容却很少更新,成为占用网上资源的“垃圾网站”,影响到人们在互联网上搜索信息的质量。表 2 列出了对 15 个关键词,每个关键词前 100 个检索结果的最后更新时间的统计,从中我们发现,有 40% 是半年以上未更新的信息,一年以上未更新的旧信息近 30%。

表 2 最后更新时间比率(%)

搜索引擎	一周内	一月内	三月内	半年内	一年内
Google	10.7	25.1	43.9	57.0	72.6
百度	13.1	27.7	38.1	54.1	70.3

忽视网页的更新时间除了返回给用户大量的旧信息外,还带来另一个方面的问题,容易返回过多的不可达链接(如前所述)。这主要是因为互联网上每时每刻都在变化,有的服务器可能被关闭,有的域名可能已经过期,有的网页可能已经被删除。

4 网页排序算法的改进

针对网页排序算法的以上缺陷,我们提出了如下改进方法:

为了让新页面在搜索引擎返回结果集中有比较好的排名,我们在用网页排序算法计算出所有页面的 PageRank 值或 Authority 值之后,再将其乘上一个更新时间因子 t 。对较新的页面 t 较大,对较旧的页面 t 较小。此外,还要考虑到有些情况下,如对新闻类或娱乐类内容,用户一般希望搜索到最新的内容,这时应选择合适的 t 值,以能够较好地地区分出新旧网页。

对于网站的页面载入时间因素,我们可以根据搜索引擎统计的结果或专业的网站性能评测机构的结果,对 PageRank 值或 Authority 值也乘以一个网

站性能因子 p 进行修正。对于性能好的网站 p 较大,对于性能差的网站 p 较小。对网站的平均页面载入时间(或响应时间),搜索引擎可以根据自己的数据库进行统计,也可以借助专门的网站性能评测机构的评测结果。

综合以上思想,对 PageRank 算法,我们得到改进的 PageRank 值计算公式:

$$PR(A)' = PR(A) * t(A) * p(A) + c(A)$$

其中 $PR(A)$ 为用 PageRank 算法计算的页面的 PageRank 值, $t(A)$ 为 A 页面的更新时间因子, $p(A)$ 为 A 页面的网站性能因子, $c(A)$ 为 A 页面的用户点击率调整量。

当用户用搜索引擎进行信息检索时,检索的结果将以 $PR(A)'$ 的值进行排序返回给用户。

对 HITS 算法的改进,与 PageRank 算法类似,对其输出一组网页,我们也将 Hub 值和 Authority 值乘上更新时间因子 t 和网站性能因子 p 进行修正。由于 HITS 算法的结果集是依据其它搜索引擎的搜索结果动态生成的,因而较难实现用户点击率修正。

结论 目前,搜索引擎已成为上网用户使用的主要功能之一,也成为用户在网上获取信息的主要手段之一,因此搜索引擎的服务质量的提高就具有很高的现实意义。本文针对搜索引擎所面临的挑战,分析了基于链接分析网页排序算法的缺陷,提出了考虑网站性能、页面更新时间和用户点击率的改进的网页排序算法,对通用搜索引擎服务质量的改进有很好的参考价值,在设计个性化的专业搜索引擎时也应加以考虑。

参 考 文 献

- 1 中国互联网络信息中心. 第十三次中国互联网络发展状况分析报告. 2004. 1
- 2 李学勇, 欧阳柳波, 等. 搜索引擎中网络蜘蛛搜索策略比较研究. 计算技术与自动化[J], 2003, 22(4): 63~67
- 3 Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[EB/OL]. <http://www-db.stanford.edu/~backrub/google.html>
- 4 Kleinberg J M. Authoritative Sources in a Hyperlinked Environment[J]. Journal of the ACM, 1999, 46(5)
- 5 王有为, 汪定伟. 基于网页可达性和平均载入时间的网站评估方法[J]. 东北大学学报(自然科学版), 2003, 23(10): 926~929
- 6 李凯, 赫枫龄, 等. PageRank2Pro--一种改进的网页排序算法[J]. 吉林大学学报(理学版), 2003, 41(2): 175~179