

文章编号: 1006-2467 (2003) 03-0397-04

对网页 PageRank 算法的改进

宋聚平, 王永成, 尹中航, 滕伟

(上海交通大学 电子信息学院, 上海 200030)

摘要: 分析了著名搜索引擎 Google 采用的 PageRank 算法, 指出其偏重旧网页、忽视专业站点以及对网页中的超链接评估不恰当等不足之处。改进算法考虑了网页日期这一重要因素, 并重新计算网页中超链接对网页的影响。网页结构中蕴涵着丰富的信息, 在 href、title 等标记中文字对网页主题有重要作用, 利用结构标记可以辅助判断网页的主题内容。试验结果表明, 采用改进的算法可以提高判断网页重要性的准确度。

关键词: 搜索引擎; 网页; 超链分析; PageRank

中图分类号: TP 391 **文献标识码:** A

Improved PageRank Algorithm for Ordering Web Pages

SONG Ju-ping, WANG Yong-cheng, YIN Zhong-hang, TENG Wei

(School of Electronics & Information Technology, Shanghai Jiaotong Univ., Shanghai 200030, China)

Abstract: This paper analyzed Google's algorithm on PageRank and presented some disadvantages of this algorithm. Those disadvantages are preferring to old pages, ignoring special sites and inaccurate judge of hyperlinks pointed out from one page. Furthermore, the improved algorithm was described. The experiments show that the consideration on evaluating the importance of pages can make an improvement over the original algorithm.

Key words: search engine; authority of pages; hyperlink analysis; PageRank

目前的搜索引擎返回结果过多, 用户很难从中快速筛选出真正需要的信息。如果搜索引擎只返回相关度高的重要网页, 既可以很大程度地节省用户时间, 又可以减轻网络流量。

搜索引擎中网页的采集工作主要由 Spider 完成, 开发出性能良好的 Spider 是一个艰巨的工作。由于网络带宽窄、网页更新快, 搜索引擎的 Spider 搜集所有网页已经成为不可能的事情。优先获取重要网页逐渐成了网络信息搜索中重点研究的问题。

人们已经开发出多种基于特定搜索算法的 Spider, TueMosaic 和 WebCrawler 是早期基于 Best-

First-Search 算法的佼佼者, 后来 DeBra 根据 Fish-Search 算法对 TueMosaic 进行改进^[1], 这种搜索方式的最大缺点是不能搜索到潜在网页, 而且随着用户设置搜索深度的增加, Spider 的搜索量将急剧增大。随着网页数量的膨胀, 许多更加智能的 Spider 被开发, 如卡耐基梅隆大学开发的 WebAnts, 其研究重点是强调多个 agent 的协作。另外, 文献[2]中描述了一种“Shark”搜索算法, 该算法主要是面向主题的搜索。WTMS 是文献[3]中提到的另一种针对主题的搜索系统, 该系统中的 Spider 采取启发式搜索判断来自同一个网站的网页在主题上是否相关, 既加快 Spider 的访问速度, 又提高相关网页的判断精度。Grouper 则是文献[4]中介绍的另一个根据网页的主题对网页进行归类的系统。

在众多智能搜索系统所采用的算法中, 超链分

收稿日期: 2002-01-16

基金项目: 国家自然科学基金资助项目(60082003)

作者简介: 宋聚平(1974-), 男, 山东菏泽市人, 博士生, 主要研究网络信息检索

析是近期大家主要研究的问题,而搜索引擎 Google 所采用的 PageRank 算法尤其得到认可,从实际应用来看,这种算法也确实解决了一些问题 本文将对 PageRank 算法作详细研究并改进其不足之处

1 评测网页 PR 值的算法分析

文献[5,6]中指出:如果网页 A 存在一条指向网页 P 的超链,则认为 P 得到了 A 的认可;如果有许多网页指向网页 P ,则可以说 P 相对比较重要 从直觉上来说,这种判断是有道理的 计算 PageRank 的公式为^[5,6]

$$PR(P) = (1-d) + d \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)} \quad (1)$$

式中: $T_i (i=1,2, \dots, n)$ 为指向网页 P 的其他网页; d 为界于 $(0,1)$ 区间的衰减系数; $C(T_i)$ 为网页 T_i 向外指出的链接数目 通过简单的迭代算法可以计算出 $PR(P)$ 的值

1.1 PageRank 算法偏重旧网页

由式(1)可以看出,决定一个网页 $PR(P)$ 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,通过式(1)计算出的 PR 值也就很低 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,用户想首先看到最新的网页 因此,式(1)计算出的网页 PR 值并不能很好地反映网页的重要性,当然也就不能很好地满足用户需求

通过指向某个网页的超链来计算该网页的 PR 值时,应该考虑到网页的日期,即

$$w_i = CD/T \quad (2)$$

式中: w_i 为网页关于日期的权重; C 为搜索引擎对其库存 URL 列表索取一遍所需要的时间,即一个访问周期; T 为一个网页被 Spider 访问时的日期与网页的更新日期相距的天数; D 为常数,它的取值受到式(1)中 d 的影响,且也和 Spider 的访问周期相关 w_i 与 T 成反比关系,这是考虑到目前 Spider 的访问周期以及实际测试的结果,随着 Internet 的迅速发展,当前条件改变时,两者的关系也应随之改变 在此基础上做对比试验,利用 Spider 获取了 100 万个网页,分两种方式计算出这些网页的重要性,表 1 列出的是返回前 100 个结果中用户满意的网页数目(测试日期:2001.4.9)。

表 1 结果显示,不考虑网页的生成日期,用户对查询结果的满意度平均值为 20.9%,考虑网页的日

表 1 网页的日期对判断网页重要性的影响

Tab 1 Influence on evaluating the importance of pages resulted from date

查询关键词	满意网页数	
	不考虑网页日期	考虑网页日期
数据挖掘	13	28
撞机事件	12	70
内陆河	40	45
北京旅游	23	38
世乒赛	19	51
中文处理	8	11
人机界面	13	28
显示器	30	41
内存条	25	27
克隆人	26	33

期,用户对查询结果的满意度平均值为 37.2%,比原来提高了 16.3%。显然,考虑网页的生成日期,可以增加用户对返回结果的满意度,尤其是对于时效性强的查询,网页日期更是一个关键性因素

1.2 PageRank 算法忽视了专业站点

站点的权威性是一个重要因素,文献[5,6]中的依据仅仅是 URL 的表面现象,如以 com 结尾的比其他的重要 这个判断依据实际上并不恰当 因为许多以 com 结尾的网站一般是规模大涉及领域广的站点,但正是因为其规模大,所以很难对一些专业内容做深刻论述 在特定领域与其他专业站点相比,前者网页的权威性不如后者,尤其是目前关于某些专业的个人主页或者研究院所站点大量增加(这些站点的 URL 很多不是以 com 结尾),这些站点一般对所论述内容研究较深,显然应该比综合站点的泛泛论述更重要

考虑站点的权威性并不能仅仅根据该站点的 URL,而应该主要从该站点所有网页的 PR 值来评判,即在搜索引擎的 URL 列表库中,如果来自于同一个站点的网页都具有较高的 PR 值,则可以认为该站点具有较大的权威性 当然,这需要考虑到几个不同域名对应同一个站点的问题,而且还应优化 URL 列表的存储,把来自同一个站点的 URL 存放在一起,并且 URL 只存储根路径后增加的部分,这种存储方式,既方便计算同一站点内所有网页的平均 PR 值,又节省 URL 列表的存储空间

1.3 网页中的超链接对网页 PR 值的影响

网页中的超链接形式各异应该区别对待 式(1)计算网页 PR 值时对于一个网页中指出的超链接的

处理并不恰当 如: 图 1 中 A、B、C、D、E 表示 5 个任意的不同网页, 有向箭头表示网页间的超链关系, 图 1(b) 表示在图 1(a) 的基础上, 网页 A 又增加了两个指向权威网页的链接 从客观上来分析, 既然网页 A 增加了两个指向重要网页的超链, 它本身的 PR 值也应该有所增加, 可式 (1) 的推导结果却是: 网页 A 的 PR 值没有变化(由于指向 A 的所有网页都没有发生任何变化), 网页 B 的 PR 值减少了许多(因为网页 A 的指出超链接数目增加了), 这显然不符合常理 产生这种情况的主要原因是式 (1) 过于简单地判定一个网页指出的超链接对该网页都是负面影响, 其实网页所包含的指出超链对于该网页 PR 值的影响应该分为两大类: 如果网页 A 中的一个超链接所指向的网页 B 在内容上与本网页内容相关, 则该超链接增加网页 A 的 PR 值; 否则, 该超链将减少网页 A 的 PR 值

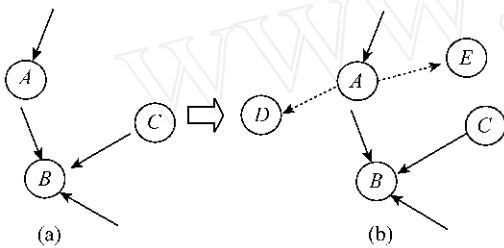


图 1 网页间超链接结构变化示意图

Fig.1 The change of the structure of hyperlinks

本文以 $f(T_{nk})$ 表示网页 T_n 的所有 m 个超链中第 k 个超链对网页 T_n 的影响因子, 当 A 中的一条超链接 H 所指的网页与 A 在内容上相关, 则 $f(T_{nk})$ 为正; 当 H 所指的网页与 P 内容无关, 则 $f(T_{nk})$ 显然为负值 对 PR 值的计算公式修正如下:

$$PR(P) = (1-d) + d[PR(T_1) \sum_{k=1}^m f(T_{1k}) + \dots + PR(T_n) \sum_{k=1}^m f(T_{nk})] \quad (3)$$

1.4 判断网页内容上的相似性

判断两个网页内容上是否相似, 主要是根据向量空间模型(Vector space model) 算法, 即文献关键词的权重主要与其在文献中的出现频率决定

由于网页采用了半结构化的 HTML 语言, 其包含有丰富的结构信息, 故在抽取网页的主题内容时应加以利用 位于 head、title、meta 以及 a href = ... 等标记之内的关键词无疑应该重视, 赋予较大的权重系数 本文在对大量网页的实际操作中, 在诸多网页标记中最能够反映网页内容的并不是通常认为的 title 或者 meta 间的文字, 而

是 a href = ... 与 /a 之间的超链文字 这主要是因为许多网页的 title 并未经过作者的仔细推敲, 有的是由网页制作工具自动生成(如 index 1、index 2 等), 有的是作者赋以与主题无关的 title (如欢迎你的到来), 还有的是为了提升在搜索引擎结果中的排名而故意造假欺骗 Spider, 尤其是在 meta 标记中, 对 Spider 的欺骗更为常见 对于关键词在网页中的权重修正如下:

$$W_{ij} = tf_{ij} \cdot \lg \left(\frac{N}{df_j} + 0.5 \right) \cdot \text{func}(t_j) \quad (4)$$

$$\text{func}(t_j) = \begin{cases} 3.0 & \text{在链接文字中} \\ 2.0 & \text{在 head/title/H}_1/\text{H}_2 \text{ 标记中} \\ 1.8 & \text{在 meta 标记中} \\ 1.0 & \text{其他} \end{cases}$$

2 试验结果

为了验证上述改进算法, 本文做了两次对比试验 在一个具有 10 万个网页的数据库中, 第 1 次是根据未曾改进的算法计算出所有网页的权威度, 然后对随机的关键词进行 20 次查询, 在返回的前 100 个结果中, 统计符合查询的网页篇数; 第 2 次是根据改进后的算法计算出所有网页的权威度, 用第 1 次查询的关键词同样进行 20 次查询 两次试验中符合查询的篇数如图 2 所示

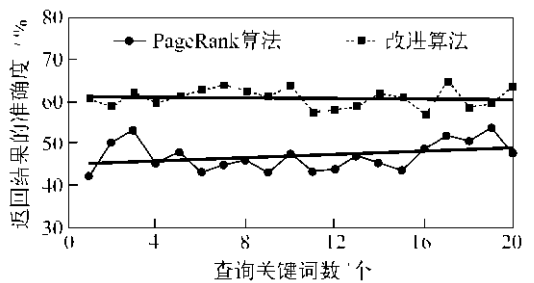


图 2 网页评估算法改进前后返回结果的比较

Fig.2 Comparison between the return lists result from two algorithms

由图 2 可见, 第 1 次查询结果满意度的平均值 45% 左右, 第 2 次是利用改进的算法可以把准确度提高到 62% 左右, 即改进后的算法可以更加准确地判断网页的权威性, 返回更加符合查询条件的结果

3 结论

(1) 在计算网页的 PR 值时应该充分考虑网页生成日期与 Spider 访问日期的间隔, 需要增加新近放到 Internet 上网页的重要度;

(2) 对于网页来源的站点评价, 不能仅仅局限

于其 URL 的表现形式, 而是应该根据该站点中所有网页 PR 值的平均值评测;

(3) 对于网页 A 中指出的超链接应该视不同情况具体分析, 如果所指网页与网页 A 内容相关, 则此超链增加网页的重要度, 否则减少;

(4) 计算网页内容与查询关键词的相关度时, 除了从网页文本内容上考虑之外, 还需要提高在重要标记中出现的内容的权重, 试验结果表明在这些标记中最重要的是出现在超链文本中的关键词

参考文献:

- [1] DeBra P, Post R. Information retrieval in the World Wide Web: making client-based searching feasible [A]. *Proc 1st International World Wide Web Conference* [C]. Geneva: CERN, 1994.45-55.
- [2] Hersovici M, Jacovi M, Maarek Y, et al. The shark-search algorithm - an application: tailored Web site mapping [J]. *Computer Networks and ISDN System*, 1998, 30: 256-264.
- [3] Sougata Mukherjea. WTMS: a system for collecting and analyzing topic-specific Web information [J]. *Computer Networks*, 2000, 33: 124-138.
- [4] Oren Zamir, Oren Etzioni. Grouper: a dynamic clustering interface to Web search results [J]. *Computer Networks*, 1999, 31: 58-63.
- [5] Brin S, Page L. The anatomy of a large-scale hyper-textual Web-search engine [A]. *Proc 7th International World Wide Web Conference* [C]. Brisbane: SIGIR, 1998.146-164.
- [6] Jughoo Cho, Hector G M, Lawrence P. Efficient crawling through URL ordering [A]. *Proc 7th International World Wide Web Conference* [C]. Brisbane: SIGIR, 1998.220-235.

下期发表论文摘要预报

双曲型线性方程三阶和四阶 TVD 格式的新构造

王嘉松¹, 倪汉根², 何友声³

(1. 上海交通大学 机械与动力工程学院, 上海 200030; 2. 大连理工大学 土木工程系, 大连 116024; 3. 上海交通大学 建筑工程与力学学院, 上海 200030)

摘要: 利用 Taylor 级数理论和总变差减小(TVD)格式的充分条件构造了时间二阶、空间五点三阶和四阶新 TVD 格式。给出了新 TVD 格式与传统 TVD 格式及近期建立的二阶新 TVD 格式用于线性双曲型方程的计算结果, 表明本文新格式特别是四阶 TVD 格式具有比二阶新 TVD 格式和传统 TVD 格式峰值衰减更慢、间断更陡, 而计算工作量具有与传统二阶 TVD 格式相当的良好数值性能

盾构隧道装配式管片接头三维有限元分析

张厚美^{1,2}, 张正林¹, 王建华¹

(1. 上海交通大学 建筑工程与力学学院, 上海 200030;
2. 广州市盾建地下工程有限公司, 广州 510030)

摘要: 根据 Saenz 公式将混凝土本构模型简化为双折线性强化弹塑性模型, 推导了混凝土弹塑性参数的计算公式。应用大型结构分析有限元软件 Algor 对装配式管片接头进行三维线弹性和弹塑性有限元分析, 得到了混凝土应变、接头位移、接缝转角、螺栓拉力等计算结果, 并将有限元计算结果与接头荷载试验测试结果进行了对比。研究表明, 有限元计算值与试验值两者的变化规律是一致的, 计算结果在反映结构应力/应变分布规律方面有重要参考作用