具有时间反馈的 PageRank 改进算法

戚华春, 黄德才, 郑月锋
(浙江工业大学 信息工程学院, 浙江 杭州 310014)

摘要: 针对某一类网页(比如新闻网页)在互联网上发布时间越长, 其信息的重要性将随之下降这一事实, 在传统的 PageRank 算法中加入时间反馈因子, 实现网页因发布时间的长短, 其 PageRank 值也随之上下浮动. 并采用 Seidel 迭代算法加速迭代收敛过程. 实验结果表明, 改进后的算法在计算这类与发布时间相关的网页的 PageRank 值时, 符合人们的一般期望, 是有效的. Seidel 迭代算法有利于提高算法效率.

关键词: PageRank; Seidel 迭代; 时间反馈; 搜索引擎
中图分类号: G202 **文献标识码**: A **文章编号**: 1006-4303(2005)03-0272-04

An improved PageRank algorithm with time feedbacking

QI Hua-chun, HUANG De-cai, ZHENG Yue-feng
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310014, China)

Abstract: PageRank is a web page ranking algorithm proposed by Google, a well known search engine. The algorithm is an iterative process that determines web page ranking based on page link structure, or co-citation. PageRank is a successful, but not a perfect algorithm. For instance, an older page is always an important page because the more older it is, the more link-in pages it has. So a new page is usually not important. For this, we first integrated page time information with PageRank calculation, and then employed Seidel's method to speed up the convergence of the iteration process. Experimental results show that the new algorithm is good and reasonable.

Key words: PageRank; Seide literation; time feedbacking; search engine

0 引 言

随着互联网技术日益深入生活, 使我们面对的信息出现了爆炸式的增长. 1994 年, 最早的搜索引擎 World Wide Web Worm 标引了 11 万网页, 到 1997 年, 搜索引擎所标引的网页已达 2～100 M, 2000 年可标引的网页已超过 10 亿张. 著名的搜索引擎 Google 拥有 10 亿个网址, 30 亿个网页, 3.9 亿张图像, 而且, 今天仍然以每天超过 100 万张的速度在增长. 面对互联网如此巨大的信息量, 人们开始注意到如何对网络数据进行挖掘是一个重要的问题, 并对此展开了大量的研究. 本文的论述集中在如何对网络的组织结构和链接关系进行挖掘, 以产生我们需要的信息.

目前基于网络的组织结构和链接关系进行挖掘的主要算法有两种[1]:

(1) PageRank 算法[2]: 该算法提取网页的超链接信息, 进行离线计算, 得出网页的 PR 值, 并进行排序, 以发现网络中最主要的页面.

收稿日期: 2004-10-13
作者简介: 戚华春(1979-), 男, 浙江杭州人, 硕士研究生, 主要研究方向为网络应用、数据挖掘和遗传算法.

(2) HITS 算法[3]: 该算法将网页分为锚页(Hub)和权威页(Authority), 并通过这两种网页相互增强, 进行迭代, 以最终的网页权威值为依据对结果进行排序, 以发现网络中最主要的页面.

上述两种算法各有优缺点, 本文就主要针对 PageRank 算法的一些不足, 提出一个校正的算法 PageRank-Times 算法.

1 PageRank 算法与缺陷

1.1 PageRank 算法

传统情报检索理论中的引文分析方法是确定学术文献权威性的一个重要方法, 即根据引文的数量来确定文献的权威性. PageRank 算法的发明者对网络的超链接结构和文献引文机制的相似性进行了研究, 借鉴引文分析思想计算网络文档的重要性, 利用网络自身的超链接结构给所有的网页确定一个重要性等级数. 比如, 当从网页 A 链接到网页 B 时, 就认为"网页 A 投了网页 B 一票", 即增加了网页 B 的重要性. 最后根据各网页的得票数来评定其重要性, 以此来帮助实现排序算法的优化, 而这个重要性的量化指标就是网页的 PR 值.

1.2 PageRank 算法描述

PageRank 算法的数学表示为

$$PR(A) = (1-d) + d \cdot \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)} \right) \quad (1)$$

其中: $PR(A)$ 为网页 A 的 PageRank 值; $T_1, T_2, \cdots, T_n$ 为网页 A 的链入网页; $PR(T_i)$ 为网页 T_i 的 PR 值, $i = 1, 2, \cdots, n$; $C(T_i)$ 为网页 T_i 的链出链接数量, $i = 1, 2, \cdots, n$; d 为衰减因子, $0 < d < 1$, 通常取值为 0.85.

通过迭代算法可以计算出 $PR(A)$ 的最终值.

1.3 PageRank 算法的缺陷

PageRank 算法的主要缺陷是: 该算法偏重旧网页. 由式(1) 可以看出, 决定一个网页的 PR 值的主要因素是指向该网页的链接个数, 也就是说一个网页在网络上存在的时间越长, 就越有可能被其他更多的网页所链接, 按照算法则可以认为旧的网页比新的网页更可能具有较高的 PR 值, 旧网页比新网页更重要, 即 PageRank 算法偏重旧网页. 但这在实际的网络环境中是有出入的, 因为人们更希望看到的是新内容, 特别是有关新闻和商务信息. 比如, 一个很重要的网页被放到网络上不久, 由于时间短暂, 许多其他网页还没有指向它, 那么通过式(1) 计算出的网页 PR 值就很低, 在搜索引擎返回的结果中往往会把它排在较后的位置, 这正好与用户的要求相反. 所以, PageRank 算法需要改进.

2 算法改进

2.1 基本思路

通过式(1) 计算出的网页 PR 值并不能很好地反映网页的实际重要性, 当然也就不能很好地满足用户需求. 面对这个问题, 本文认为通过指向某个网页的超链接来计算该网页的 PR 值时, 应该考虑到其他因素, 比如网页的发布日期. 把网页的发布日期引入 PageRank 算法, 使在网络上存在比较长时间的网页慢慢沉下去, 新的网页能迅速浮上来. 但由于现在很多网页是由程序自动生成的, 并且大量的 HTML 网页的格式不规范, 很难从网页中提取到该网页的发布时间. 为此, 我们把网页的存在时间通过搜索引擎搜索的周期数来表征, 这一转换的核心思想是基于这样一个事实: 一般而言, 搜索引擎的搜索周期为半个月到一个月, 如果一个网页存在的时间较长, 那它将在每个搜索周期里都将被搜索到(在同一个搜索周期里不管搜索到该网页几次, 都算作 1 次), 即页面的存在时间正比于搜索引擎搜索到该页面的次数 T.

网页的时间反馈因子 W_t, 即

$$W_t = e/T \quad (2)$$

式中: W_t 为网页的时间反馈因子; T 为一个网页被搜索引擎访问的周期次数; e 为常数, 它的取值受到式(1) 中 d 的影响, 且也和搜索引擎的搜索周期相关, 是一个实验数据.

2.2 算法的改进

在新算法 PageRank-Times 中, 我们加入了时间反馈因子 W_t, 公式(1) 修正为

$$PR(A) = (1-d) + d \cdot \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)} \right) + W_A \quad (3)$$

显然, 式(3) 也可以写为

$$PR_{i+1} = (1-d) + d \cdot M \cdot PR_i + w \quad (4)$$

其中, 令 F_u 为页面 u 的链出链接集合, $N_u = |F_u|$, 则 $M_{u,v} = 1/N_u$, 如果有从网页 u 到 v 的链接, 反之 $M_{u,v} = 0$. 故根据线性代数的知识, 因为 $\|M\|_1 = 1$, 所以迭代是收敛的, 并且是有限次迭代.

2.3 加速收敛的 Seidel 技巧

对于迭代解出的过渡矩阵 M 的特征向量, 要用它对关键词匹配的搜索结果进行排序, 显然我们关心的不是它的大小, 而是它的方向(甚至只是其各分量的大小顺序). 由此, 可以在不影响 M 特征向量的方向的前提下对 M 进行适当的变换以加快迭代的收敛速度. 简单迭代法的分量形式为

$$x_{i+1}^{k} = \sum_{j=1}^{i-1} m_{kj} x_i^{i} + \sum_{j=1}^{n} m_{kj} x_i^{j}$$

其中, $k = 1, 2, 3, \cdots, n$. 显然, 在计算 X_{i+1}^{k} 时, $X_{i+1}^{1}, \cdots, X_{i+1}^{k-1}$ 都已经求出, 一般地, 后计算出的结果更接近最终结果. 因此可以用这些新值来计算 X_{i+1}^{k}, 于是迭代形式变为

$$x_{i1}^{k} = \sum_{j=1}^{i-1} m_{kj} x_{i+1}^{i} + \sum_{j=1}^{n} m_{kj} x_i^{j}$$

这就是 Seidel 技巧(计算过程中总是利用最新算出来的值, 也称 Seidel 迭代法).

2.4 新算法 PageRank-Times 描述

输入是: 通过搜索引擎搜索网络, 获得对应网络子图的邻接矩阵 A, 及对应的搜索次数即时间向量 T. 可以调整时间向量 T 的值, 用于表示网页存在时间的长短.

输出是: 页面排序的 PR 值.

算法的步骤:

(1) 获得网络子图.

(2) 确定每个 URL 的查询周期次数, 即时间向量 T.

(3) 用每个 URL 的查询周期次数的倒数, 计算该网页的时间反馈因子 W.

(4) 按式(3)计算每个分量的 PR 值.

(5) 迭代, 直至收敛.

(6) 结果规范化输出.

3 算法实验

图 1 是用于仿真的网络子图(为便于分析, 假设该子图中的页面是关于同一查询主题的).

3.1 仿真结果分析

图 2, 系列 1 为时间反馈因子 $W = (1, 1, 1, 1, 1, 1)$, 用于模拟传统的 PageRank 算法, 即无时间反馈的情况; 系列 2 为时间反馈因子 $W = (1, 1, 1, 1/10, 1, 1)$, 是改进的 PageRank-Times 算法.

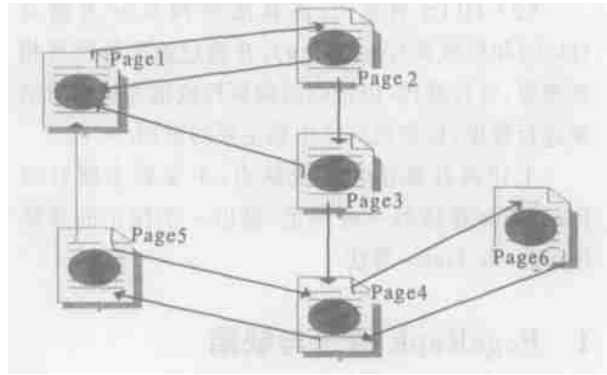

图 1 实验网络仿真子图

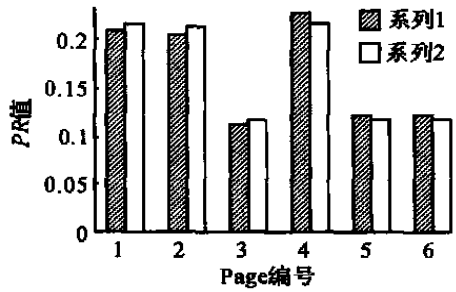

图 2 新旧算法实验结果比较图

由于系列 2 的时间反馈因子的在 Page4 上的为 1/10, 即认为该子图中的 Page4 已经被搜索引擎搜索到了 10 次, 是一个过期页面, 根据新算法, Page4 应该下沉. 所以比较系列 1 和 2, 发现 Page4 的 PR 值是下降的; 同时可以观察到, 在系列 2 中, Page4 和 Page1 间的差值基本上消失(实验数据表明, Page1 的 PR 值比 Page4 的略高), 原来系列 1 中明显在 Page4 下的 Page1 现在和 Page4 基本保持一致, 这有利于 Page1 的 PR 值排名.

见图 3, 系列 1 为 $e = 0.015/n$ (n 为总的页面数); 系列 2 为 $e = 0.15/n$; 系列 3 为 $e = 0.015$; 系列 4 为 $e = 0.15$; 系列 5 为 $e = 1$; 由图可见, e 的取值不影响最后的 PR 值排名. 但在计算的过程中, e 的大小影响迭代收敛的快慢.

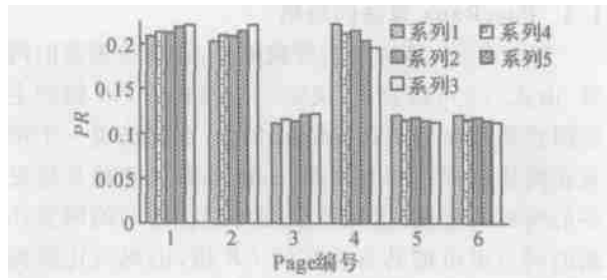

图 3 取不同 e 值的效果图

为验证新算法的实际有效性, 在一个月内, 我们还对新浪网爬行了 4 次, 每次获得 10 万张有效网

页, 应用传统 PageRank 算法和新算法分别进行排序运算, 然后进行查询分析. 在对比前后 4 次查询结果, 发现新算法提供的推荐网页质量要好于传统的 PageRank 算法, 特别一些近期的新闻能迅速出现在推荐结果中, 并且能在推荐结果中占据比较靠前的位置.

4 结　论

在新算法中, 引入时间反馈因子, 使网页的发布时间长短影响网页的 PR 值大小, 这是可行的. 并且, 实验显示, 新算法 PageRank-Times 有利于旧网页的下沉, 新网页的上浮, 这与人们的期望是一致的. 参数 e 的大小不影响最后的 PR 值分布, 但影响算法迭代的过程, 一般取 $e=0.15/n$ (n 为总的页面数) 较为合适.

参考文献:

[1] Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the World Wide Web [A]. 9th International Conference on Tools with Artificial Intelligence (ICTAI'97). IEEE Computer Society [C]. 1997. 558-567.

[2] Page L, Brin S, Motwani R, *et al*. The pagerank citation ranking: Bringing order to the WEB [EB/OL]. http://newdbpubs.stanford.edu/8090/pub/1999-66/1999-11-11.

[3] Jon M K. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 668-677.

[4] 王晓宇, 周傲英. 万维网的链接结构分析及其应用综述[J]. 软件学报, 2003, 14(10): 1 768-1 780.

[5] 宋聚平, 王永成, 尹中航, 等. 对网页 PageRank 算法的改进[J]. 上海交通大学学报, 2003, 37(3): 397-400.

[6] 张　岭, 马范援. 加速评估算法: 一种提高 Web 结构挖掘质量的新方法[J]. 计算机研究与发展, 2004, 41(1): 98-103.

(责任编辑: 刘　岩)

(上接第 267 页)

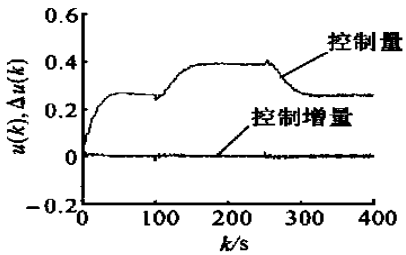

图 5　有 PID 反馈校正时的控制量及其增量

4 结　论

在建模过程中, 由于干扰、噪声和非线性等因素的存在, 不可避免的存在着模型失配问题. 通常广义预测控制通过不断的在线辨识克服建模误差, 它本质上是一种在线建模过程, 不仅计算耗时, 而且辨识结果与实际过程之间仍然存在着一定的偏差. 文中采用简单的 PID 算法作为反馈校正的手段, 能够有校克服建模误差, 减小计算量, 取得满意的控制效果. 提出的方法有三个特点: 体现了先进控制和经典控制的结合; 体现了基于过去误差的事后控制和基于预测信息的超前控制的结合; 体现了基于模型控制和基于非模型控制的结合. 基于这一思想的具体控制策略还有待进一步深入研究.

参考文献:

[1] Hapoglu H, Karacan S, Erten K Z S, *et al*. Parametric and nonparametric model based control of a packed distillation column [J]. Chemical Engineering and Processing, 2001, 40(6): 537-544.

[2] 张　峻, 席裕庚. 基于几何分析的约束预测控制直接算法[J]. 控制与决策, 1997, 12(2): 184-187.

[3] 毛志忠, 杨　林. 一种解决预测控制输入信号受约束问题的方法[J]. 控制与决策, 1994, 9(2): 230-233.

[4] 席裕庚. 复杂工业过程的满意控制[J]. 信息与控制, 1995, 24(1): 14-30.

[5] 陈增强, 袁著祉. PI型有约束后移限位预测控制器[J]. 应用科学学报, 1998, 16(2): 229-233.

[6] Lu J, Chen G, Ying H. Predictive fuzzy PID control: theory, design and simulation [J]. Information Sciences, 2001, 137(1-4): 157-187.

[7] Chen W H, Ballance D J, Gawthrop P J, *et al*. Nonlinear PID predictive controller [J]. IEE Proceedings: Control Theory and Applications, 1999, 146(6): 603-611.

[8] Tan K K, Huang S N, Lee T H. Development of a GPC-based PID controller for unstable systems with deadtime [J]. ISA Transactions, 2000, 39(1): 57-70.

[9] Miller R M, Shah S L, Wood R K. Predictive PID [J]. ISA Transactions, 1999, 38(1): 11-23.

[10] 王　伟. 广义预测控制理论及其应用[M]. 北京: 科学出版社, 1998.

[11] 余世明, 杜　维. 有约束多变量动态矩阵控制算法[J]. 控制与决策, 2001, 16(3): 299-302.

[12] 余世明, 杜　维. 目标规划法在预测控制滚动优化及在线辨识中的应用[J]. 自动化学报, 2002, 28(6): 955-1 000.

(责任编辑: 刘　岩)