基于更新网页排名算法的研究

高 臣[1],高斐斐[2],张家健[3]
(1.河海大学 商学院,江苏 南京 210000; 2.中国科学院 力学研究所,北京 100190;
3.江苏省邮电规划设计院有限公司 江苏 南京 210000)

摘要: 页面内容的内容评分与 PageRank 评分都需要频繁更新,以保证提供最新的结果。基于如何使得更新 PageRank 向量过程更为容易,并使得更为频繁的更新成为可能这一问题,本文通过对更新算法的数学内容分析,研究更新 PageRank 向量的问题,通过提出假设矩阵 $Q_{m\times m}$ 的 PageRank 向量 $\phi^{T}=(\phi_1,\phi_2,\cdots,\phi_m)$,文中立足于通过 3 种聚合更新算法来利用 ϕ^{T} 中的值计算 G 的更新后的 π^{T},文中分析了近似聚合更新、精确聚合更新、迭代聚合更新的算法,并对 3 种更新算法各自的使用条件进行分析。

关键词: PageRank; 近似聚合更新; 精确聚合更新; 迭代聚合更新
中图分类号: TN0 **文献标识码**: A **文章编号**:1674-6236(2017)01-0006-03

Research on updating PageRank vector

GAO Chen[1],GAO Fei-fei[2],ZHANG Jia-jian[3]
(1.*Business School of Hohai University*, *Nanjing* 210000,*China*; 2.*Institute of Mechanics*, *Chinese Academy of Sciences*,*Beijing* 100190,*China*;3.*Jiangsu Posts & Telecommunication Planning and Designing Institute Co. Ltd*, *Nanjing* 210000,*China*)

Abstract: The score of the page content and the PageRank will require frequent updates, to provide the latest results. How to make it easier to update the PageRank vector in order to make it possible to update more frequently In this paper, the problem of updating the PageRank vector is studied through the analysis of the mathematical content of the update algorithm. Based on the PageRank vector of the hypothesis matrix ($Q_{m\times m}$), we propose a new algorithm based on the three algorithms ($\phi^{T}=(\phi_1,\phi_2,\cdots,\phi_m)$).Based on the three polymerization update algorithm, we use ϕ^{T} value in calculation of G into π^{T}, this paper analyzes the approximate polymerization update, precise polymerization update, iterative aggregation update algorithm, and carries on the analysis to the three update algorithm respective application conditions.

Key words: pageRank; approximate polymerization update; precise polymerization update; iterative aggregation update

DOI:10.14022/j.cnki.dzsjgc.2017.01.002

网页变化可以是网页内容的改变或是页面出链的改变,研究表明,一半以上的网页在一周内发生了变化, 而近三分之一的.com 网页每天都在发生变化[1]。相比于较小的网页,大型网页中的变化则更为频繁[2]。对于新增的网页, 内容和链接的更新可能发生在以小时计算的时间尺度上[3]。因此,反映页面内容的内容评分与 PageRank 评分都需要频繁更新,以保证提供最新的结果。如何使得更新过程更为容易,得到研究者越来越多的重视。

PageRank 向量可能发生两类更新:1)当超链接被加入到万维网中或从万维网中被删除时, 超链接矩阵 H 的元素发生改变,而矩阵的大小未变。该类只有这一类型的更新,那么更新 PageRank 向量的问题就是链接更新问题;2) 网页本身可能被加入到万维网中或从万维网中被删除, 那么对于页面更新问

收稿日期:2016-03-29 稿件编号:201603379
基金项目:江苏省社科联研究基金(201035)
作者简介:高 臣(1991—),男,山东泰安人,硕士。研究方向:企业管理、技术经济。

题而言，发生的状态将被加入到马尔科夫链中或者从链中被删除,此时矩阵大小也会发生改变,此类型更新问题也更加复杂。将早期精确更新的理论结果用于 PageRank 问题[4],计算结果对链接更新问题给出了理论上的答案，对于仅有一两行发生改变而且没有页面被加入或删除的情况而言，已知的精确链接更新公式是有用的,但是从计算角度分析,由于万维网的动态性，该方法对更为一般的更新而言实际价值较小[5-6]。由旧的 PageRank 向量开始重启幂法对于链接更新问题而言作用也较小[7],因为不能简单地对幂法进行调整以处理更加复杂的页面更新问题,因此仅靠幂法本身来由旧的 PageRank 向量重启算法实际价值也较小。

1　近似聚合更新

状态聚合作为近似聚合技术方法的一部分,可以用作估计近解耦链的稳态分布。同理,可以利用一个机遇状态聚合的近似方法估计 PageRank[8-9]。虽然近似聚合只能给出 π^k 的估计值,但是近似聚合的计算量小并且可以同时处理链接更新和页面更新。利用已知分布($\phi^{\mathrm{T}}=(\phi_1,\phi_2,\cdots,\phi_m)$)以及 G 中的更新后的转移概率构建一个聚合马尔可夫链，其转移概率矩阵 C 比 G 更小,利用 C 的稳态分布 ξ^{T} 估计更新分布 π^{T},具体算法包括:

将更新的马尔可夫链链的状态空间 S 划分为两组,即 $S=L\cup\bar{L}$,其中,补集 $\bar{L}$ 包括所有其他状态,L 是由稳态概率可能受更新影响最大的状态构成的子集,如果新加入的状态被自动包含在 L 中,则将受影响的转移概率设为 0 以处理删除的状态。需要注意,如果一个扰动只涉及 PageRank 大的稀疏链中的少数状态，那么它对于稳态向量的影响将主要是局部性的,因此,大多数稳态概率不会受到显著影响。根据 $S=L\cup\bar{L}$ 导出更新后的转移矩阵及其对应的稳态分布的一个划分:

$$G_{n\times n}=\begin{array}{c} \\ L \\ \bar{L} \end{array}\begin{array}{c} \begin{array}{cc} L & \bar{L} \end{array} \\ \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \end{array}\quad \text{且 } \pi^{\mathrm{T}}=(\pi_1,\cdots,\pi_l|\pi_{l+1},\cdots,\pi_n)$$

该式中 G_{11}，的大小为 $l\times l$,$l=|L|$ 为 I 的势,G_{22} 的大小为 $(n-l)\times(n-l)$,原有分布 ϕ^{T} 中对应于 $\bar{L}$ 中的状态的稳态概率被存入一个行向量 ω^{T} 中,而 $\bar{L}$ 中的状态被聚合为一个超级状态，进而得出一个更小的聚合马尔可夫链,其转移矩阵大小为 $(l+1)\times(l+1)$,给出式子 $\tilde{C}=\begin{pmatrix} G_{11} & G_{12}e \\ \tilde{S}^{\mathrm{T}}G_{21} & 1-\tilde{S}^{\mathrm{T}}G_{21}e \end{pmatrix}$ 中 $\tilde{S}^{\mathrm{T}}=\dfrac{\omega^{\mathrm{T}}}{\omega^{\mathrm{T}}e}$（e 是全 1 列),进而利用近似程序计算出 $\tilde{C}$ 的稳态分布 $\tilde{\xi}^{\mathrm{T}}=(\tilde{\xi}_1,\tilde{\xi}_2,\cdots,\tilde{\xi}_l,\tilde{\xi}_{l+1})$,利用 $\tilde{\xi}^{\mathrm{T}}$ 中的前 I 个元素以及 ω^{T} 中的元素,得出产生精确的更新分布 π^{T} 的一个近似 $\tilde{\pi}^{\mathrm{T}}$,该近似值为 $\tilde{\pi}^{\mathrm{T}}=(\tilde{\xi}_1,\tilde{\xi}_2,\cdots,\tilde{\xi}_l|\omega^{\mathrm{T}})$,由此可知,与获得完整的更新 PageRank 向量 π^{T} 的精确值，该方法使用一个较小的稳态向量 $\tilde{\xi}^{\mathrm{T}}$ 以构建 π^{T} 的近似 $\tilde{\pi}^{\mathrm{T}}$。

马尔科夫链可能会表现出对微小扰动的敏感性,对于扰动对马尔可夫链的影响,目前可以衡量稳态概率对于转移概率中变化的敏感程度的度量包括:转移矩阵次主特征的绝对值接近于 1 的程度、不同种类条件数的微小程度、平均首达时间的微小程度等。结合上述近似聚合更新的算法,需要适当地构造出划分 $S=L\cup\bar{L}$，并保证 δ^{T} 的量级处于较小规模,那么 $\tilde{C}$ 将接近于 C,因此其各自的稳态分布 $\tilde{\xi}^{\mathrm{T}}$ 和 ξ^{T} 也会相互接近,进而保证对于 $i\leqslant l$,$\tilde{\pi}_i$ 相 π_i 互接近。如果 C 所定义的链在以上任何一个度量下都是良态的,那么 ξ^{T} 对于微小扰动将相对不敏感,即 $S=L\cup\bar{L}$ 的恰当程度能够更加直接地体现 $i\leqslant l$，$\tilde{\pi}_i\approx\pi_i$ 的程度,因此计算的关键在于确定的良态程度。

2　精确聚合更新

对于一个不可约的 n 状态马尔可夫链，假设其状态空间已被划分为 k 个互不相交的部分 $S=L_1\cup L_2\cup\cdots\cup L_k$,同时假设与之对应的转移概率矩阵具有分块矩阵的形式:

$$G_{n\times n}=\begin{array}{c} \\ L_1 \\ L_2 \\ \vdots \\ L_k \end{array}\begin{array}{c} \begin{array}{cccc} L_1 & L_2 & \cdots & L_k \end{array} \\ \begin{pmatrix} G_{11} & G_{12} & \cdots & G_{1k} \\ G_{21} & G_{22} & \cdots & G_{2k} \\ \vdots & & & \vdots \\ G_{k1} & G_{k2} & \cdots & G_{kk} \end{pmatrix} \end{array}$$

（对角线上的分块均为方阵）

该条由 G 所定义的父马尔可夫链可诱导出 k 条更短的马尔可夫链[10],具体的诱导方法为:与 L_i 这组状态相对应的受限马尔可夫链定义为一个马尔可夫过程,仅当父链对 L_i 中的状态进行访问时,该过程才会记录父链的位置，并忽略所有对 L_i 之外的状态的访问。已知第 i 条受限链的转移概率矩阵为第 i 个随机补,由 $S_i=G_{ii}+G_{i*}(I-G_i^*)^{-1}G_{*i}$ 给出,其中,G_{i*} 和 G_{*i} 分

别为 G_{ii} 被移除后的第 i 行和第 i 列的分块，通过去除第 i 行和第 i 列的分块可得出 G_i^* 为 G 的主子矩阵。

为了获得较小的 k 状态聚合链，可以将每个组 L_i 压缩为一个单一的状态，将父转移矩阵 G 压缩为聚合转移矩阵 $C_{k\times k}=\begin{bmatrix} S_1^{\mathrm{T}}G_{11}e & \cdots & S_1^{\mathrm{T}}G_{1k}e \\ \vdots & & \vdots \\ S_k^{\mathrm{T}}G_{k1}e & \cdots & S_k^{\mathrm{T}}G_{kk}e \end{bmatrix}$（该矩阵为随机且不可约）

对于正则链，在由 C 所定义的聚合链中的状态转移，对应当未聚合的父链达到平衡时，在父链中的 L_i 组之间的转移，其中，允许父链被分解为 k 个小的受限链且可以独立求解，由此解得的受限分布 S_i^{T} 可以通过 C 的稳态分布加以组合构造出父链的稳态分布 π^{T}。

对于计算 π^{T} 而言，其数值求解过程并非高效，原因在于要获得受限分布 S_i^{T} 需要计算随机补，但是随机补 $S_i=G_{ii}+G_{i*}(I-G_i^*)^{-1}G_{*i}$ 中包含了计算成本较高的求逆运算。解决这一问题可以对受限分布进行一定程度的近似，具体包括：1）估计出随机补 S_i，计算这些估计的分布以得到近似设限分布，得到近似聚合转移矩阵，利用精确聚合定理计算 π^{T} 的近似值；2）忽略随机补，直接对设限分布 S_i^{T} 进行估计。

3 迭代聚合更新

迭代聚合是一种求解近解耦马尔可夫链的算法[11-12]，假设某个不可约马尔科夫链 C 的稳态分布 $\phi^{\mathrm{T}}=(\phi_1,\phi_2,\cdots,\phi_m)$，对 C 进行更新，令更新后的链的转移概率矩阵和稳态分布分别为 G 和 $\pi^{\mathrm{T}}=(\pi_1,\pi_2,\cdots,\pi_n)$，其中，更新后的 G 不可约，并且由于更新过程可能会新增或删除状态以及改变转移概率，m 不一定等于 n。具体算法包括：将更新后的链的状态划分为 $S=L\cup\bar{L}$，对 G 进行重排：

$$G_{n\times n}=\begin{array}{c} \\ L \\ \bar{L} \end{array}\begin{array}{c} \begin{array}{cc} L & \bar{L} \end{array} \\ \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \end{array}$$

且 $\pi^{\mathrm{T}}=(\pi_1,\cdots,\pi_l|\pi_{l+1},\cdots,\pi_n)$

ω^{T} 对应于 $\bar{L}$ 状态的 ϕ^{T} 中的元素，给出式子 $C=\begin{pmatrix} G_{11} & G_{12}e \\ S^{\mathrm{T}}G_{21} & 1-S^{\mathrm{T}}G_{21}e \end{pmatrix}$ 中 $S^{\mathrm{T}}=\dfrac{\omega^{\mathrm{T}}}{\omega^{\mathrm{T}}e}$，进而利用近似程序计算出 C 的稳态分布 $\xi^{\mathrm{T}}=(\xi_1,\xi_2,\cdots,\xi_l,\xi_{l+1})$，进而得出 $X^{\mathrm{T}}=(\tilde{\xi}_1,\tilde{\xi}_2,\cdots,\tilde{\xi}_l|\tilde{\xi}_{l+1}S^{\mathrm{T}})$，最后令 $\psi^{\mathrm{T}}=\chi^{\mathrm{T}}G=(\psi_1^{\mathrm{T}}|\psi_2^{\mathrm{T}})$，以将循环移出不动点 χ^{T}。

迭代聚合更新的优点为：当使用一个良好的 L-集是，相比于幂法，迭代聚合算法可以带来明显的改善，时间成本有效减少，并且随着数据集规模的扩大而越明显[13-14]。其次，迭代聚合被用于更新方法时，在更新使得问题规模发生改变的同时不会带来不利后果，因此，迭代聚合算法可以用于同时处理链接和页面更新两类更新的算法。迭代聚合更新的缺点为：首先，迭代聚合更新并不是一种普遍用途的方法，对于并非近解耦的链，迭代聚合更新一般不能获得良好的运行效果[15]。其次，向量 χ^{T} 是一个不动点，如果直接利用 χ^{T} 重启算法，在后续的迭代中将在该计算环节复制出相同的 χ^{T}。最后，迭代聚合算法的收敛率直接依赖于主随机补 $S=G_{22}+G_{21}(I-G_{11})^{-1}G_{12}$，收敛率完全由 S 最大的次主特征值所决定。相比于幂法，迭代聚合算法的每次迭代都需要进行更多的计算。

4 结 论

更新 PageRank 向量的研究已经展现其有效性，研究方法和思路更加追求创新和效率，但是无论近似聚合更新、精确聚合更新、迭代聚合更新，目前的研究都还不尽完善。由于不同算法给出的矩阵彼此之间存在明显差别，因此未来的研究工作可以将多个相互独立的算法的结果加以融合。

参考文献：

[1] Junghoo Cho, Hector Garcia-Molina. The evolution of the Web and implications for an incremental crawler [C]// In Pro-ceedings of the Twenty-sixth International Conference on Very Large Databases, New York, 2000:198-210.

[2] Dennis Fetterly, Mark Manasse. A large-scale study of the evolution of web pages [C]// In The Twelfth International World Wide Web Conference, 2003.

[3] Konstantin Avrachenkov and Nelly Litvak [R]. The effect of new links on Google PageRank. Technical report, INRIA, 2014.

[4] Meyer C D, Shoaf J M. Updating finite Markov chains by using techniques of group matrix inversion [J]. Journal of Sta-tistical Computation and Simulation, 1980:161-179.

[5] Cho G E, Meyer C D. Comparison of perturbation bounds for the stationary distribution of a Markov chain [J]. Linear Algebra and Its Applications,

（下转第 12 页）

3 结束语

与已有的基站信息采集系统相比，本系统利用android 智能手机为开发平台，无需额外的硬件设计，节省了大量开发成本。同时，相较于以往采集基站信息时，工作人员需要携带繁重的专用采集设备来实现基站信息的采集存储，本系统将信息采集与信息处理进行分离，即有助于基站信息的精细化管理，也优化了采集的工作流程，提高了工作效率，为工作人员提供了便利。

参考文献：

[1] 刘长征，李伟. 多种定位技术融合构建LBS体系[J]. 地理信息世界，2013，1(3):24-27.

[2] 赵鸣翔. 蜂窝移动通信系统单基站定位技术研究[D].成都:西南交通大学，2012.

[3] 王婷. 多模基站信息采集系统的设计与实现[D]. 北京:北京邮电大学，2013.

[4] 公磊，周聪. 基于Android的移动终端应用程序开发与研究[J].计算机与现代化，2014，3(8):86-89.

[5] 张宇，王映辉，张翔南. 基于Spring的MVC框架设计与实现[J]. 计算机工程，2012，36(4):59-62.

[6] 刘引涛. 基于Spring的MVC模式网上银行系统的设计与实现[J]. 电子设计工程，2013，21(7):169-171.

[7] 查修齐，吴荣泉，高元钧.C/S到B/S模式转换的技术研究[J]. 计算机工程，2014，40(1):263-267.

[8] 张怡. 基于J2EE的基站管理信息系统的设计与实现[D]. 武汉：华中科技大学，2013.

[9] 姚嘉健. 基于Android的LBS定位系统的设计[D]. 南京:南京邮电大学，2013.

[10]韦峥. 第三代移动通信系统网络规划(桂林)[D]. 北京:北京邮电大学，2012.

[11]刘峰. 移动通信基站设备信息管理系统的分析与设计[D]. 吉林:吉林大学，2015.

[12]肖宇翔. GPS定位与干扰技术研究[D]. 成都:电子科技大学，2013.

[13]王敏，吴中博，徐德刚. 基于预测模型的传感器网络近似数据采集算法 [J]. 计算机工程与科学，2014，36(11):2148-2152.

[14]刘伟江，李振汉，唐余亮，等. 基于Android的嵌入式Web服务器设计 [J]. 电子设计工程，2013，21(9):4-6.

[15]刘靖桐. 面向Web2.0的web应用前端开发框架的设计与实现[D]. 北京:北京邮电大学，2014.

(上接第 8 页)

2010：135-155.

[6] Eugene Seneta. Sensivity analysis, ergodicity coefficients and rank-one updates for finite Markov chains [J]. In William J. Stewart, Editor, Numerical Solution of Markov chains, 1991：121-130.

[7] Meyer C D. Matrix Analysis and Applied Linear Algebra[M]. SIAM, Philadelphia, 2009.

[8] Steve Chien, Cynthia Dwork. Towards exploiting link evolu-tion[M]. In Workshop on Algorithms and Models for the Web Graph，2001.

[9] James H. Aggregation of variables in dynamic systems [J]. Infor-mation Processing and Management, 2013：111-139.

[10]Meyer C D. Stochastic complementation, uncoupling Markov chains and the theory of nearly reducible systems[J]. SIAM Review，1989：240-270.

[11]Stewart W J. Introduction to the Numerical of Markov Chains [M]. Princeton University Press, -12-

2004.

[12]杨博，陈贺昌，朱冠宇，等. 基于超链接多样性分析的新型网页排名算法 [J]. 计算机学报，2014(4)：833-847.

[13]Sergey Brin, Lawrence Page. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN Systems, 1998(33)：105-118.

[14]Ayman Farahat, Thomas Lofaro. Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization [J]. SIAM Journal on Scientific Com-puting, 2006(27):1181-1213.

[15]Matthew Richardson ,Petro Domingos. The intelligent surfer: Probabilistic combination of link and content inform-ation in PageRank [J]. Advances in Neural Information Proc-essing Systems，2002(14)：1398-1406.