基于 PageRank 和 HITS 的 Web 搜索

常 庆,周明全,耿国华
(西北大学 可视化研究所,陕西 西安 710127)

摘 要:介绍了目前应用较为广泛的两种算法——PageRank 算法和 HITS 算法。PageRank 算法是基于用户随机的向前浏览网页的直觉知识,HITS 算法考虑的是 Authoritive 网页和 Hub 网页间的加强关系。PageRank 算法的基本思想是:如果一个页面被许多其他页面引用,则这个页面很可能是重要页面;一个页面尽管没有被多次引用,但被一个重要页面引用,那么这个页面很可能也是重要页面;一个页面的重要性被均分并传递到它所引用的页面。而 HITS 算法则专注于改善泛指主题检索的结果,通过一定的计算(迭代计算)方法以得到针对某个检索提问的最具价值的网页,即排名最高的 authority。

关键词:PageRank;HITS;特征向量;检索主题;链接分析
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)07-0077-03

PageRank and HITS-Based Web Search

CHANG Qing, ZHOU Ming-quan, GENG Guo-hua
(Institute of Visualization Technology, Northwest University, Xi'an 710127, China)

Abstract: Introduce the wider application of the present two algorithms: PageRank algorithm and HITS algorithm. PageRank algorithm is based on random users browse the website ahead of intuitive knowledge. HITS algorithm considered is Authoritive and Hub website homepage the strengthening of relations. PageRank algorithm's basic idea: if a page is used in many other pages, this page is likely to be important pages; although no one page was repeatedly quoted, but it was an important quote pages, this page may also be important page; the importance of a page are transfered to the pages which it cites. HITS algorithm focus on improving the generic theme of the search results, through some calculation (iterative) method in order to get a response to a search of the most valuable pages, the highest ranking authority.

Key words: PageRank; HITS; eigenvector; search theme; link analysis

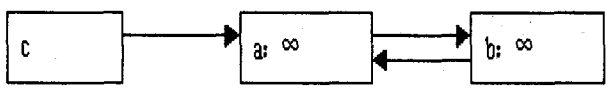
1 PageRank 算法

PageRank 算法描述如下:u 是一个网页,$F(u)$ 是 u 指向的网页集合,$B(u)$ 是指向 u 的网页集合,$N(u)$ 是 u 指向外的链接数,显然 $N(u) = |F(u)|$,c 是一个用于规范化的因子(Google通常取0.85,这种表示法也适用于以后介绍的算法),则 u 的 Rank 值计算如下:

$$R(u) = c\sum_{v\in B(u)} R(v)/N(v)$$

这就是算法的形式化描述,也可以用矩阵来描述此算法,设 $\boldsymbol{A}$ 为一个方阵,行和列对应网页集的网页。如果网页 i 有指向网页 j 的一个链接,则 $A_{ij} = 1/N_i$,否则 $A_{ij} = 0$。设 $\boldsymbol{V}$ 是对应网页集的一个向量,有 $\boldsymbol{V} = c\boldsymbol{AV}$,$\boldsymbol{V}$ 为 $\boldsymbol{A}$ 的特征根为 c 的特征向量。实际上,只需要求出最大特征根的特征向量,就是网页集对应的最终 PageRank 值,这可以用迭代方法计算[1]。

如果有两个相互指向的网页 a,b,它们不指向其它任何网页,另外有某个网页 c,指向 a,b 中的某一个,比如 a,那么在迭代计算中,a,b 的 rank 值因为不分布出去而不断地累计,如下图:



为了解决这个问题,Sergey Brin 和 Lawrence Page 改进了算法,引入了衰退因子 $E(u)$[2],$E(U)$ 是对应网页集的某一向量,对应 rank 的初始值,算法改进如下:

$$R'(u) = c\sum_{v\in B(u)} R(v)/N(v) + cE(u)$$

其中,$\|R'\|_1 = 1$,对应的矩阵形式为 $V' = c(AV' + E)$。

2 HITS 算法

HITS(Hyperlink-Induced Topic Search)算法是利

收稿日期:2007-10-11
基金项目:国家自然科学基金(F020503)
作者简介:常 庆(1981-),男,硕士研究生,研究方向为图形图像;周明全,教授,研究方向为虚拟现实与可视化技术。

用 Hub/Authority 方法的搜索方法[3,4]，算法如下：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(root set)，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 中的任一个顶点 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。

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

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

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)反映了若一个网页有很多好的 Hub 指向，则其权威值会相应增加(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[5](Hub 值增加为该网页链接的所有网页的权威值之和)。

3 PageRank 技术和 HITS 技术的比较

PageRank 和 HITS 的迭代算法都利用了特征向量作为理论基础和收敛性依据[6]。这也是超链接环境下此类算法的一个共同特征。但两种算法也有着明显的不同点，下面着重阐述两种算法的不同点。

从两者的权值传播类型来看，PageRank 算法基于随机冲浪(Random Surfer)模型[7]，将网页权值直接从 authority 网页传递到 authority 网页；而 HITS 算法则是将 authority 网页的权值经过 hub 网页的传递进行传播。

从算法思想上看，虽然均同为链接分析算法，但二者之间还是有一定的区别。HITS 的原理如前所述，其 authority 值只是相对于某个检索主题的权重，因此 HITS 算法也常被称为 query - dependent 算法。而 PageRank 算法独立于检索主题，因此也常被称为 query - independent[8] 算法。PageRank 的发明者(Page & Brin)把引文分析思想借鉴到网络文档重要性的计算中来，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的等级数。当然 PageRank 并不是引文分析的完全翻版，根据因特网自身的性质等，它不仅考虑网页引用数量，还特别考虑了网页本身的重要性[9]。

从处理的数据量及用户端等待时间来分析。表面上看，HITS 算法对需排序的网页数量需求较小，所计算的网页数量一般为 1000 至 5000 个[10]，但由于需要从基于内容分析的搜索引擎中提取根集并扩充基本集，这个过程需要耗费相当的时间，而 PageRank 算法表面上看，处理的数据数量上远远超过了 HITS 算法。

从两者的处理对象来看，都是针对整个万维网上的网页的一个子集进行排序、筛选[11,12]，没有一个搜索引擎能够将万维网上的网页全部搜索下来。但是，PageRank 算法的处理对象是一个搜索引擎上当前搜索下来的所有网页，一般在几千万个页面以上；而 HITS 的处理对象是搜索引擎针对具体查询主题所返回的结果，从几百个页面扩展到几千几万个页面。

从两者的具体应用来看，PageRank 算法应用于搜索引擎服务端[13]，可以直接用于标题查询并获得较好的结果，若要用于全文本查询，需要与其它相似度判定标准(向量模型等)进行复合，以针对具体查询形成最终排名，搜索机器人(Crawler)可以将 PageRank 作为搜索优先次序的标准，算法中 E 的取值可以用来定制个人搜索引擎；HITS 算法一般用于全文本搜索引擎的客户端[14]，对于宽主题的搜索相当有效，可以用于自动编撰万维网分类目录，通过找到指向某网页的 Hub 网页并以此为根集 R，查找该网页的相关网页，也可以用于元搜索引擎的网页排序。

4 结束语

基于链接分析的算法，目前的研究都还很不成熟，无论是 PageRank 算法，还是 HITS 算法等，有一些共同的问题影响着算法的精度。

(1)根集的质量。根集质量应该是很高的，否则，扩展后的网页集会增加很多无关的网页，产生主题漂移[15]、主题泛化等一系列的问题，计算量也增加很多。算法再好，也无法在低质量网页集找出很多高质量的网页。

(2)锚文本的利用。锚文本有很高的精度，对链接和目标网页的描述比较精确[16]。上述算法在具体的实现中利用了锚文本来优化算法。如何准确充分地利用锚文本，对算法的精度影响很大。

(3)噪音链接。Web 上不是每个链接都包含了有

用的信息,比如广告,站点导航,赞助商,用于友情交换的链接[17,18],对于链接分析不仅没有帮助,而且还影响结果。如何有效去除这些无关链接,也是算法的一个关键点。

(4)查询的分类。每种算法都有自身的适用情况,对于不同的查询,应该采用不同的算法,以求获得最好的结果。因此,对于查询的分类也显得非常重要。

文中对两种算法进行了深入探讨和比较,但在这几个方面需要继续做深入的研究,相信在不久的将来会有更多的有价值的成果出现。就目前的研究来看,值得关注的是,已有学者对这两种算法相结合的可能性作了理论上的探讨。

参考文献:

[1] 戚华春,黄德才,郑月峰. 具有时间反馈的PageRank改进算法[J]. 浙江工业大学学报, 2005,33(3):272-275.

[2] Lempel R, Moran S. The Stochastic Approach for Link-Structure Analysis(SALSA) and the TKC effect[J]. Computer Networks,2000,19(2):33-36.

[3] Brin S, Page L. Anatomy of a Large-Scale Hypertextual-Web Search Engine[C]//Proc. 7th International World Wide Web Conference. Stanford: Stanford University,1998.

[4] 赖茂生. 计算机情报检索[M]. 北京:北京大学出版社, 1993:89-112.

[5] 杨思洛. 搜索引擎的排序技术的研究[J]. 信息检索技术, 2005,119(1):43-45.

[6] 陈定权. Web信息检索技术最新进展[J]. 现代图书情报技术,2002(2):2-3.

[7] Henzinger R. Hyperlink Analysis for the Web[J/OL]. IEEE Internet Computing,2001:45-50. http://computer.org/internet/.

[8] 曹　军. Google的PageRank技术剖析[J]. 情报杂志,2002(10):10-12.

[9] 彭绪富,邹友宽,邓荣华. INTERNET搜索引擎探解[J]. 高等函授学报:自然科学版,2001(2):15-16.

[10] 苏宁新. 信息检索理论与技术[M]. 北京: 科学技术文献出版社,2004:231-233.

[11] Goldberg D E. Genetic Algorithms in Search, Optimization and Machine Learning[M]. New York: Addison Wesley, 1989.

[12] Mernik M, Crepinsek M, Zumer V. A metavolutionary approach in searching of the best combination of crossover operators for the TSP[C]// Proceeding of the IASTED ICNN. Pittsburgh, PennsyIvania: IASTED/ACTA Press, 2000:32-36.

[13] Kamvar S, Haveliwala T, Golub G. Adaptive methods for the computation of PageRank[C]// Linear Algebra and its Applications 386. Proc of International Conference. [s.l.]: American Pacific University,2004:51-65.

[14] Kleinberg J M. Authoritative Source in a hyperlinked Environment[J]. Journal of ACM, 1999,46(5):604-632.

[15] 肖　文,庞丽萍. 电子出版物的全文检索技术研究[J]. 计算机与数字工程,2002(4):45-48.

[16] 钟敏娟. 基于Web的文本信息检索算法研究[D]. 长沙: 湖南大学, 2004:9-19.

[17] Lee Min-Hyung, Kim Yeon-Seok, Lee Kyong-Ho. Logical structure analysis: From HTML to XML[J]. Computer Standards & Interfaces, 2007,29(1):109-124.

[18] Gupta S, Kaiser G E. Context-based content extraction of html documents[D]. Columbia: Columbia University,2006.

(上接第76页)

检测逐渐发展到动态随机端口号,近期涌现的新型P2P应用越来越具有反侦察的意识,采用一些加密的手法,伪装HTTP协议,传输分块等来逃避识别和检测,可见将来P2P软件必将走向加密通信的方向。针对现在P2P应用发展的趋势,P2P流量管理将逐步走向根据其不变传输特性建立相应的分析模型,应用更高级的机器学习和数据挖掘的方法去识别和管理流量,从而提高网络服务质量。

参考文献:

[1] CacheLogic[EB/OL]. 2006. http://www.cachelogic.com/.

[2] BitTorrent[EB/OL]. 2007. http://www.bittorrent.com/.

[3] Karagiann I T, Broidoa, Faloutsosm. Transport Layer Identification of P2P Traffic[C]// Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2004:121-134.

[4] Sen S, Wang Jia. Analyzing Peer-to-Peer Traffic across Large Networks[C]// In: IEEE/ACM Transactions on Networking. NJ: IEEE Press, 2004:219-232.

[5] Karagiann I T, Bro I I, Brownlee N, et al. Is P2P dying or just hiding[C]// Globecom. Dallas, TX, USA: [s.n.], 2004.

[6] Kim M S, Kang H J, Hong J W. Towards Peer-to-Peer Traffic Analysis Using Flows[C]// SelfManaging Distributed Systems, 14 th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management. Heidelberg, Germany:[s.n.], 2003:55-67.

[7] Kim M S, Won Y J, Hong J W K. Application-Level Traffic Monitoring and an Analysis on IP Networks[J]. ETRI Journal, 2005,27(11):22-42.

[8] P-Cube Inc. Approaches to Controlling Peer-to-Peer Traffic: A Technical Analysis[EB/OL]. 2004. http://www.pcube.com/doc_root/products/Engage/WP_Approaches_Controlling_P2P_Traffic_31403.pdf.