

# 一种更稳定的链接分析算法—— 子空间 HITS 算法

石 晶, 龚震宇, 裘杭萍, 张毓森

(解放军理工大学指挥自动化系, 上海 210016)

**摘要:** 在给定超链接情况下, Kleinberg 的 HITS 算法采用特征向量的办法来确定页面的权威性等级. 通过分析当链接模式发生小的变化时算法的稳定性, 提出一种新的算法——子空间 HITS 算法, 并通过 Web 超链接数据作为实验数据对其性能进行研究.

**关键词:** Web 挖掘; 链接分析; 稳定性

**中图分类号:** TP391   **文献标识码:** A   **文章编号:** 1671-5489(2003)01-0049-05

链接分析技术在信息检索中扮演着重要的角色, 已经被成功地用于分析 Web 超链接数据来确定权威的信息源, 或用于分析学术论文的引用数据来确定有影响力的论文<sup>[1,2]</sup>. 链接分析技术已经成为当前主流 Internet 搜索引擎的基础.

动态性是 WWW 信息的重要特性之一. 对一个搜索引擎而言, 页面链接可能因某些错误或某些变化而不可访问. 人们希望在链接结构发生变化时, 从超链接信息集中得到的权威页仍然能够具备一定的稳定性. 如果一个搜索引擎的返回结果每天都发生很大的变化, 那么其搜索结果是不可能得到用户信任的.

本文通过分析目前流行的链接分析算法——Kleinberg 的 HITS(Hyperlink-induced Topic Search, HITS)算法, 在此基础上提出子空间 HITS 算法, 并利用 Web 数据作为实验数据, 对两种算法的稳定性进行分析. 最后研究算法返回结果的多样性问题.

## 1 HITS 算法及稳定性分析

### 1.1 HITS 算法

一个好的搜索引擎算法不仅应该得到与搜索请求相关的 Web 页面, 而且检索到的页面还应该具有较高质量和权威性(Authority). 权威性可由 Web 页面链接来反映. Web 不仅由页面组成, 而且还包含了从一个页面指向另一个页面的超链接, 超链接有助于自动分析出权威性语义. 当一个 Web 页面的作者建立指向另一个页面的链接时, 可以看作是作者对另一个页面的注解. 把对一个页面来自不同作者的注解收集起来, 就可以用来反映该页面的重要性, 并可以很自然地用于权威 Web 页面的发现. Hub 页面提供了指向权威页面的链接集合, 是指一个或多个 Web 页面. Hub 页面本身可能并不突出, 或者说可能没有几个链接指向它们. 但是, Hub 页面却提供了指向就某个公共话题而言最为突出的站点链接. 此类页面可以是主页上的推荐链接列表, 如一门课程主页上的推荐参考文献站点. Hub 页面起到了隐含说明某话题权威页面的作用. 通常, 好的 Hub 页面是指向许多好的权威的页面; 好的权威页面是指由许多好的 Hub 所指向的页面. 这种 Hub 与 Authority 之间的相互作用, 可用于权威页面的挖掘和高质量 Web 结构和资源的自动发现, 这就是 Hub/Authority 方法的基本思想.

收稿日期: 2002-03-26.

作者简介: 石 晶(1976~), 女, 博士研究生, 从事数据挖掘、宽带网络技术与应用的研究, E-mail: doudouice@vip.sina.com.

基金项目: 国家“九七三”基金(批准号: G1998030414).

HITS 算法是利用 Hub/Authority 方法的搜索算法,其内容如下:

将查询提交给普通的基于相似度的搜索引擎,搜索引擎返回很多页面,从中取前  $m$  个页面作为根集,用  $s$  表示. HITS 算法通过向  $s$  中加入被  $s$  引用的页面和引用  $s$  的页面将  $s$  扩展成一个更大的集合  $T$ . 若以  $T$  中的 Hub 页面为顶点集  $V_1$ ,以 Authority 页面为顶点集  $V_2$ ,  $V_1$  中的页面到  $V_2$  中的页面的超链接为边集  $E$ ,则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  $h(v)$  表示页面  $v$  的 Hub 值,对  $V_2$  中的顶点  $u$ ,用  $a(u)$  表示页面  $u$  的 Authority 值. 设  $T$  中有  $n$  个页面,则可将  $SG$  表示为一个  $n \times n$  的矩阵  $A$ ,其中,若顶点(页面)  $v_i$  与  $v_j$  之间存在链接,则  $A$  中元素  $(i, j)$  为 1,否则为 0.

算法重复运算下列等式:

$$a_i^{(t+1)} = \sum_{\{j:i \rightarrow j\}} h_j^{(t)}; \quad h_i^{(t+1)} = \sum_{\{j:i \rightarrow j\}} a_j^{(t+1)}, \quad i \rightarrow j \text{ 表示页面 } i \text{ 链到页面 } j,$$

每次迭代后对  $a(u)$  和  $h(v)$  进行规范化处理:

$$a(u) = \frac{a(u)}{\sqrt{\sum_{q \in V_2} [a(q)]^2}}, \quad h(v) = \frac{h(v)}{\sqrt{\sum_{q \in V_1} [h(q)]^2}},$$

上面的等式也可以被写为

$$a^{(t+1)} = A^T h^{(t)} = (A^T A) a^{(t)}, \tag{1.1}$$

$$h^{(t+1)} = A a^{(t+1)} = (A A^T) h^{(t)}. \tag{1.2}$$

给定初始值  $[1, \dots, 1]^T$ , 算法计算出页面的 Hub 权重和 Authority 权重. 根据文献[1]的证明,迭代过程收敛,定义:

$$a^* = \lim_{t \rightarrow \infty} a^{(t)};$$

$$h^* = \lim_{t \rightarrow \infty} h^{(t)}.$$

在非严格的条件下,  $a^*$  和  $h^*$  分别为  $A^T A$  和  $A A^T$  的主特征向量<sup>[3]</sup>, 因此  $a_i^*$  是页面  $v_i$  的 Authority 权重,  $h_j^*$  是页面  $v_j$  的 Hub 权重.

### 1.2 例子

通过实验发现, Web 页面集合中一个很小的变化可能会导致矩阵特征向量的值发生很大的变化. 假设一个 Web 页面集合中含有 100 个链到页面 1 的页面和 103 个链到页面 2 的页面, 邻接矩阵  $A$  中除了与这两页相对应的两列元素之外, 其它的都为 0, 因此主特征向量  $a^*$  中只有与页面 1 和页面 2 对应的两项有非零值. 图 1(a) 给出了链接到这两个 Web 页面的散列图和对应的特征向量. 现在假设 5 个新的 Web 页面被加入到集合, 它们同时指向页面 1 和页面 2. 图 1(b) 给出了新的散列图, 可以发现特征向量发生了巨大的变化, 与原来的几乎成  $45^\circ$  角.

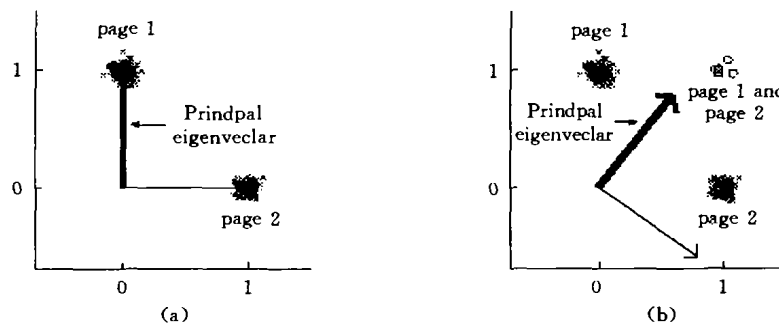


Fig. 1 Jittered scatter plot of hyperlink graph

HITS 采用矩阵  $S = A^T A$  的主特征向量来确定权威等级, 算法的稳定性由  $S$  的特征间隙决定. 第一大特征值(主特征值)和第二大特征值之间的数值差就是特征间隙  $\delta$ . 根据实验结果, 如果  $\delta$  很小, 那么

一个很小的扰动可能引起结果发生剧烈的变化. 特别是当  $\delta$  小到一定程度时, 一个微小的扰动就可能导致主特征向量与第二大特征值所对应的特征向量交换位置, 从而产生“翻转”现象.

## 2 子空间 HITS 算法

矩阵中独立的一个特征向量是不稳定的, 但是多个特征向量组成的子空间可能是稳定的<sup>[4]</sup>. 在图 1 中, 虽然两个特征向量可能自由旋转, 但是它们跨越的子空间却完全没有改变. 从而提出子空间 HITS 算法, 根据多个特征向量构成的子空间来确定页面的权威权重.

一般而言, 如果第  $k$  个和第  $k+1$  个特征值之间的特征间隙较大, 那么前  $k$  个特征向量跨越的子空间将保持稳定. 由此不仅要考虑主特征向量, 而且更一般地要考虑  $k=n$  时的情况, 即使用所有的特征向量. 当然要适当地分配权重, 使与大特征值对应的特征向量得到更多的重视. 考虑下列计算权威值的过程,  $f(\cdot)$  是一个非负单调增的函数:

(1) 找出  $S=A^T A$  (对中心权重  $S=AA^T$ ) 的前  $k$  个特征向量  $x_1, \dots, x_k$  和它们对应的特征值  $\lambda_1, \dots, \lambda_k$ .

(2) 设  $e_j$  是第  $j$  个基向量, 它除了第  $j$  个元素为 1, 其它元素都为 0, 则权威值

$$a_j = \sum_{i=1}^k f(\lambda_i) (e_j^T x_i)^2,$$

它是向量  $e_j$  在由  $x_1, \dots, x_k$  组成的子空间上投影的平方,  $x_i$  方向上投影的权重是  $f(\lambda_i)$ .

函数  $f$  有多种选择: 若取当  $\lambda \geq \lambda_{\max}$  时  $f(\lambda)=1$ , 否则  $f(\lambda)=0$ , 那么就是原始的 HITS 算法; 若取  $k=n$ ,  $f(\lambda)=\lambda$ , 则对应着简单的引用计数; 若取  $f(\lambda)=1$ , 则某页权威权重等于  $\sum_{i=1}^k x_{ij}^2$ . 在最后一种情况中, 权威权重依赖于跨越  $k$  个特征向量的子空间, 而并不仅仅是某个特征向量.

为计算方便, 忽略权重最小的特征向量, 使用  $k < n$  个特征向量来作为使用全集的一个近似. 在随后的试验中, 取  $k=20$ ,  $f(\lambda)=\lambda^2$ .

## 3 稳定性实验分析和多样性问题

为比较两种算法, 我们根据不同的主题做了 50 次 Web 查询试验. Kleinberg<sup>[5]</sup>描述了一种获取 Web 页面集合的办法, 我们也用它来构造试验数据, 并通过任意删除根集中 20% 页面的办法来修改 Web 页面集合. 在试验中, 删除一个根页面的同时也会去掉与它相关的链接结构, 则一次去掉 20% 的根页面可能会去掉原邻接图中的一大部分. 这种数据集的变换方式更准确地模拟了搜索引擎错过某些页面的情形.

表 1 和表 2 列出了部分查询结果, 其中第 1 列是数据集合变化前得到的页面排名, 第 3 列~第 7 列是数据集合分别被修改 5 次后得到的页面排名, “\*”号表示此页原来排在前 10 位但集合变化后却跌出了前 10 位. 由表 1 和表 2 可见, 绝大多数试验都出现了翻转现象. 由于主特征值的变化会引起翻转现象, 因此, 这实际上就是主特征向量被其它的特征向量所取代.

为了更好地分析特征向量的翻转情况, 计算每次试验中排名跌出前 10 位的页面个数. 由于是删减数据集而不是增加数据集, 两种算法的结果中页面排名大幅上升的情况很少出现, 所以只研究排名下降的情况. 这里共有 50 次 Web 查询, 每次查询做 5 次试验. 图 2 给出了由前 10 位下降到 20 位之后的页面数, 即如果在一次试验中, 原来前 10 位页面中有 8 个下降到 20 位之后, 那么直方图中对应 8 的矩形的高度加 1, 称这些图为翻转直方图. 一幅翻转直方图实际上是一次试验中排名下降的页面数的试验分布. 从图 2 中可以看出, HITS 更容易出现大量页面排名同时下降的情况. 在 250 次试验中, 8~10 个页面同时下降的情况在 HITS 中出现了 49 次. 当  $f(\lambda)=\lambda$  或  $\lambda^2$  时, 子空间 HITS 的实验结果比 HITS 稳定. 当  $f(\lambda)=\lambda^3$  时, 它的直方图与 HITS 的类似. 这是因为当  $f$  的次数增加时, 子空间中主特征向量的权重也同时增加了.

Table 1 HITS results on query neural networks

No	Address	$R_1$	$R_2$	$R_3$	$R_4$	$R_5$
1	www.neci.nec.com/	*	1	*	*	*
2	researchindex.org/	*	2	*	*	*
3	citeseer.nj.nec.com/cs/	*	3	*	*	*
4	citeseer.nj.nec.com/terms.html	*	4	*	*	*
5	citeseer.nj.nec.com/yao93review.html	*	5	*	*	*
6	citeseer.nj.nec.com/17901.html	*	6	*	*	*
7	citeseer.nj.nec.com/yao91optimizat	*	7	*	*	*
8	citeseer.nj.nec.com/yao91simulated	*	8	*	*	*
9	citeseer.nj.nec.com/yao93evolution	*	9	*	*	*
10	citeseer.nj.nec.com/yao99evolving	*	10	*	*	*
12	www.ieee.org/	1	—	1	1	—
13	www.cs.washington.edu/research/jai	8	—	5	—	—
14	ftp://ftp.sas.com/pub/neural/FAQ.html	4	—	4	4	—
35	www.ieee.org/nnc/	2	—	2	2	—
36	www.okstate.edu/elec-engr/faculty/	3	—	3	3	—
37	www.icsi.berkeley.edu/~jagota/NCS/	5	—	—	5	—
38	www.elsevier.nl	6	—	—	—	—
39	www.inns.org/	7	—	6	6	—
40	www.ai.univie.ac.at/oefai/nn/nngro	10	—	—	7	—
41	Synapse2.eng.wayne.edu/tpage3.html	9	—	7	—	—
44	www.emsl.pnl.gov:2080/docs/cie/neu	—	—	—	10	—
48	www.classify.org/safesurf/	—	—	8	8	—
49	www.weburbia.com/safe/ratings.htm	—	—	9	9	—
50	www.nd.com/	—	—	10	—	—
64	www.kcl.ac.uk/neuronet/	—	—	—	—	6
86	www.mitgmbh.de/	—	—	—	—	5
195	www.kcl.ac.uk/	—	—	—	—	7
230	www.kcl.ac.uk/neuronet/about/exec-	—	—	—	—	8
231	www.kcl.ac.uk/neuronet/about/map/	—	—	—	—	9
232	www.kcl.ac.uk/neuronet/about/roadm	—	—	—	—	10
381	www.ubcom.net/	—	—	—	—	1
382	www.brd.net/brd-cgi/sendmail/sent	—	—	—	—	2
383	www.amazon.de/exec/obidos/redirect	—	—	—	—	3
384	Amazon.de/exec/obidos/ASIN/3528064	—	—	—	—	4

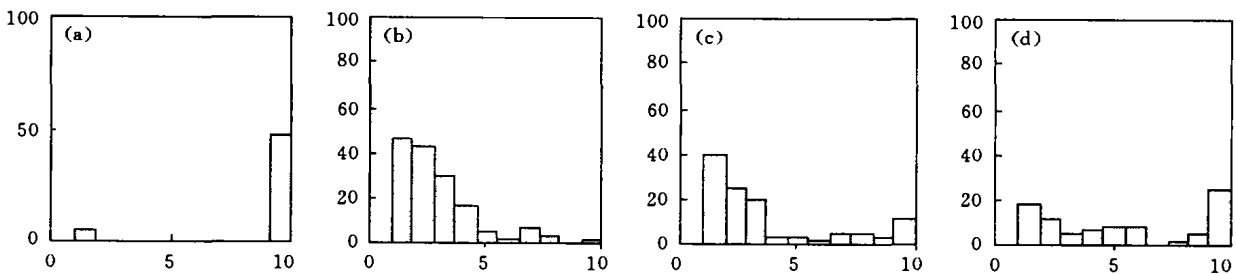


Fig. 2 Flip count histograms for the two algorithms

(a) HITS 1st Eig; (b) SP  $1(\lambda)=\lambda$ ; (c) SP  $1(\lambda)=\lambda^2$ ; (d) SP  $1(\lambda)=\lambda^3$ .

下面将进一步研究算法返回页面的范围,即返回页面的多样性问题.在所进行的 250 次实验中, HITS 算法和子空间 HITS 算法的前 10 位页面来自不同站点的平均值分别为 8.59 和 9.21. 由表 1 和表 2 可以看出, HITS 返回的前 10 个页面几乎来自同一个站点,因此即使排名顺序很稳定,它的用处

也不大. 相反, 子空间 HITS 返回页面的范围就更广一些. 在没有修改页面集合前两种算法得到的前两个页面都是 [www.neci.nec.com](http://www.neci.nec.com) 和 [reserachindex.org](http://reserachindex.org), 而在随后的第 1, 3, 4, 5 次试验结果中, 它们都退出了前 10 位. 而子空间 HITS 算法的结果还包括来自其它站点的页面, 因此当站点被删除时受的影响较小.

**Table 2 Subspace HITS results on query neural networks**

No	Address	$R_1$	$R_2$	$R_3$	$R_4$	$R_5$
1	<a href="http://www.neci.nec.com/">www.neci.nec.com/</a>	*	1	*	*	*
2	<a href="http://Researchindex.org/">Researchindex.org/</a>	*	2	*	*	*
3	<a href="http://www.ieee.org/">www.ieee.org/</a>	1	3	1	1	1
4	<a href="http://www.cum.edu/">www.cum.edu./</a>	2	4	2	2	5
5	<a href="http://www.ubcom.net">www.ubcom.net</a>	3	5	3	3	2
6	<a href="http://www.brd.net/brl-cgi/sendemial">www.brd.net/brl-cgi/sendemial</a>	4	6	4	4	3
7	<a href="http://Dmoz.org/about.html">Dmoz.org/about.html</a>	5	7	5	5	*
8	<a href="http://Ads.admonitor.net/clicktrack.cgi?">Ads.admonitor.net/clicktrack.cgi?</a>	6	8	6	6	*
9	<a href="http://www.ieee.org/nnc/">www.ieee.org/nnc/</a>	7	9	7	7	*
10	<a href="http://www.unibo.it/">www.unibo.it/</a>	8	10	8	*	4
11	<a href="http://www.deis.unibo.it/">www.deis.unibo.it/</a>	9	—	9	—	6
12	<a href="ftp://ftp.sas.com/pub/neural/FAQ.html">ftp://ftp.sas.com/pub/neural/FAQ.html</a>	—	—	—	9	—
13	<a href="http://Mathworld.wolfram.com/">Mathworld.wolfram.com/</a>	—	—	—	8	10
14	<a href="http://www.erudit.de/erudit/index.htm">www.erudit.de/erudit/index.htm</a>	10	—	10	—	8
15	<a href="http://www.iau.dtu.dk/~jj/address.html">www.iau.dtu.dk/~jj/address.html</a>	—	—	—	—	9
16	<a href="http://www.slac.stanford.edu/~rhatcher/">www.slac.stanford.edu/~rhatcher/</a>	—	—	—	10	—
20	<a href="http://www.cs.cmu.edu/">www.cs.cmu.edu/</a>	—	—	—	—	7

### 参 考 文 献

- [ 1 ] Amento B, Terveen L G, Hill W C. Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents [C]. Proc 23rd Annual Intl. ACM SIGIR, 1998. 121~126.
- [ 2 ] Brin S, Page L. The Anatomy of a Large-scale Hypertextual(Web) Search Engine [C]. The Seventh International World Wide Web Conference. Behav Genet. 1998. 23~33.
- [ 3 ] 王 奇, 宋国新, 邵志清. 信息检索中基于链接的网页排序算法 [J]. 华东理工大学学报, 2000, (10): 27~32.
- [ 4 ] Golub G H, VanLoan C F. Matrix Computations [M]. Jopt Soc Am; Johns Hopkins Univ Press, 1996. 44~52.
- [ 5 ] Kleinberg J. Authoritative Sources in a Hyperlinked Environment [C]. Proc 9th ACM-SIAM Symposium on Discrete Algorithms. Phys Rev, 1998. 133~156.

## A More Stable Link Analyze Algorithm-Subspace HITS

SHI Jing, GONG Zhen-yu, QIU Hang-ping, ZHANG Yu-sen

(Department of C<sup>3</sup>I, PLA University of Science and Technology, Shanghai 210016, China)

**Abstract:** The Kleinberg HITS algorithm is an eigenvector method for identifying authoritative or influential articles under given hyperlink information. That such an algorithm should give reliable or consistent answers is surely a desideratum. On the basis of the fact that the algorithm can give stable rankings under small perturbations to the linkage patterns, the paper presents a new algorithm: Subspace HITS method, and deals with their performance empirically with Web hyperlink data.

**Keywords:** Web mining; link analysis; stability

(责任编辑: 赵立芹)