

# 时间参数在 HITS 算法中的应用及改进

王学龙<sup>1</sup>, 张雪梅<sup>1</sup>, 李向伟<sup>2</sup>

(1. 甘肃省张掖中学, 张掖 734000; 2. 西北师范大学数学与信息科学学院, 兰州 730070)

**概要:** 针对 HITS 算法的不足及研究现状, 引入了时间参数, 并将其应用到 HITS 算法中, 对原有的 HITS 算法进行了改进。改进后的算法对提高 HITS 算法的有效性与准确性有较好的效果。

**关键词:** Web; 超链; HITS

## 引言

HITS 算法是 Web 结构挖掘中最具有权威性和使用最广泛的算法, 其基本思想是利用页面之间的引用链来挖掘隐含在其中的有用信息(如权威性), 具有计算简单且效率高的特点。然而 HITS 算法也有其明显的不足。首先, 它完全将网页的内容或文本排除在外, 仅考虑网页之间的链接结构来分析页面的权威性, 这与现实网络中的权威页面相比, 其不科学性显而易见。因为权威页面必须针对某一主题或关键词而言, 某一页面对一确定主题的具有较大权威性的页面, 并不意味着在其他与其无关的主题方面同样具有权威性。其次一个页面对另一页面的引用有多种情况, 其中包含了一页面对另一页面的认可, 但除此之外也有其他目的链接, 如为了导航或为了付费广告。而 HITS 算法在实现过程中均没有考虑以上情况, 导致了结果与目标的差距。文献[2,3,10,11]就 HITS 算法的思想与实现过程做了细致的研究与概括。针对前面第一种不足, 文献[2]提出了一种利用超链文字及其周围文字与关键字相匹配而计算超链权值的方法, 并引入系数对周围文字和超链文字进行权值的相对控制, 很好地将页面文本信息引入到 HITS 算法, 提高了算法的可靠性, 并在现实中取得了很好的效果。对 HITS 算法的第二个不足, 即非正常目的的引用, 在 HITS 算法看来, 也误认为是正常引用, 导致实际结果与目标的出入。本文引入时间参数来弥补 HITS 算法的这一不足, 即利用对一节点引用的时间长短来评价是否为正常引用。因为非正常引用其引用时间肯定不

会很长(如导航的引用), 相反, 如果一页面对另一页面的引用时间较长, 则必然反映此页面就是用户的寻找页面, 即目标页面或至少是正常引用, 如果设定时间阈值, 则可以将非正常引用的链接在 HITS 算法的实现过程中筛选出来, 如设定访问时间少于 1 分钟者为非正常引用。另外可构造时间访问函数, 控制权威页面的相对大小, 如随访问时间的增大而其权威性也逐渐非线性增大, 这样可为 HITS 算法的权威页面提供更合理、更科学的解释。

## 1 HITS 算法思想及研究现状

### (1) HITS 算法基本思想

HITS 算法是利用 Web 页面链接结构进行权威页面挖掘的一种最权威、最广泛的算法, 目前被许多高性能的搜索引擎广泛使用。其基本思想为:

①将页面分为两种类型, 一种为表达某一主题的权威页面, 称为 Authority 页面, 另一种为能把这些 Authority 页面联结在一起的页面, 称为 Hub 页面, 图 1 和图 2 表示了这两种类型的页面。而 Authority 和之间相互优化的关系构成了 HITS 算法的基础。

利用 Hub 页面找出权威页面的过程为: 首先, 由查询关键词借助传统的搜索引擎得到一初始结果集, 作为根集(root set), 也称为开始集(start set)。由于这些页面中的许多页面是假定与搜索内容相关的, 因此它们中应包含指向最权威页面的指针。故此, 根集可进一步扩展为基本集(base set), 它包含了所有由根集中的页所指向的页, 以及所有指向根集页的页。

②开始权重传播。这一过程是递归过程, 用于决

定 Hub 与权威权重的值。先为基本集中的每一个页面设定一个非负的权威权重  $ap$  和非负的 Hub 权重  $hp$ , 并将其初始化为同一常数。权重可按如下公式计算:

$$ap = \sum_{q:q \rightarrow p} hq \quad (1)$$

$$hp = \sum_{q:q \leftarrow p} aq \quad (2)$$

公式(1)反映了若一个页面由很多好的 Hub 所指, 则其权威权重会相应增加; 公式(2)反映了若一个页面指向好的权威页, 则 Hub 权重也会相应增加。

最后, HITS 算法输出一组具有较大 Hub 权重的页面和具有较大权威权重的页面。

如果有向图来描述 Web 的链接结构, 则其包含了一个节点集合和有向图的边的集合, 而此节点集合中的子集  $S$  和  $S$  中的所有节点和节点之间的边构成了 Web 的子图。而 HITS 算法就是为每个页面引入两个权重: Authority 权值和 Hub 权值, 最后输出具有最大的页面。

可以合页面标号  $\{1, 2, \dots, n\}$  并且定义它们的  $nn$  阶邻接矩阵, 如果页面  $i$  指向页面  $j$ , 则矩阵中的项  $(L_{ij})$  为 1, 否则为 0。同样把所有的 Authority 权值和 Hub 权值定义为向量,  $x=(x_1, x_2, \dots, x_n), y=(y_1, y_2, \dots, y_n)$ , 则式(1)(2)的矩阵形式为:

$$x \leftarrow A^T y \quad (3)$$

$$y \leftarrow Ax \quad (4)$$

将式(3)(4)进一步展开, 可以得到:

$$x \leftarrow A^T y \leftarrow A^T Ax \leftarrow (A^T A) x \quad (5)$$

$$y \leftarrow Ax \leftarrow A^T A y \leftarrow (A^T A) y \quad (6)$$

因此向量  $x, y$  均可由式(3)(4)经过多次迭代而得到。根据线性代数理论, 迭代序开经过标准化最终将收敛于矩阵的特征向量, 即计算机的 Authority 权值和 Hub 权值是页面集合的固有属性, 并不是由初始向量和参数的选择决定的<sup>[2]</sup>。

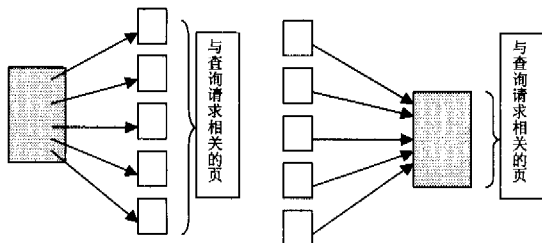


图 1 Hub 页

图 2 Authority 页

## (2) 目前对 HITS 算法的改进

虽然 HITS 算法将超链结构创造性地应用到实际并产生了很好的效果, 然而其不足之处也是显然的。首先 HITS 算法完全不考虑页面的文本内容, 在实际应用中产生了所谓的“主题漂移”, 其主要原因是算法在实现过程中将所有超链设为具有相同权值, 即只要两个页面之间存在链接, 则其邻接矩阵中对应的值设为 1, 而这种设定在某种意义上不尽合理; 其次, 对于 Authority 页面的认定也存在片面性, 算法中认为只要 Hub 页面所引用的页面其 Authority 的值就增大, 而在实际引用中, 用户的偶然引用或非正常目的的引用频繁发生, 必然导致 HITS 算法的偏移或出错。其原因是没有分清对页面的真实引用和虚假引用。

针对 HITS 算法的第一个不足, 许多研究者提出了改进算法; Kleinberg 提出了改进算法来评定网页内容的重要性, 并以此为核心技术解决了搜索引擎 Clever 的检索结果相关度排序问题, 他认为网页重要性依赖于用户提出的查询请求。而且对每一个网页应该将其和 Hub 权重分开来考虑, 通过分析页面之间的超链接结构来发现两种类型的页面, 即 Authority 页面和 Hub 页面。Krishna Bharat 和 Monika R. Henzinger 通过对超链引入相关权值 (Relevance Weight) 的方法来修改 Authority 权值和 Hub 权值, 如果相关权值小于一定阈值, 则认为该超链对页面权值的影响可以忽略不计, 该超链将从子图中删除。在一定意义上提高了权威页面产生的质量。而 HITS 算法的应用实例 Clever 系统中, 作者利用超链的周围文字中匹配查询关键字并计算词频的方法来计算超链权值, 用计算出的权值来代替邻接矩阵中相应的值, 从而达到引入语义信息来减少“主题漂移”的问题, 收到了一定的效果。文献[2]在 Clever 系统的基础上引入控制系数, 通过适当的控制系数控制超链周围文字在超链权值中的占有比例来更精确地控制“主题漂移”现象。

以上算法均是在 HITS 算法的基础上为克服“主题漂移”现象而对原有算法进行的改进, 而对于 HITS 算法的第二个缺陷, 目前没有较好的方法来解决, 本文通过引入时间参数的办法来进行改进。

## 2 本文对 HITS 算法的改进

### (1) 时间参数的引入

对于某一确定网页引用(如: 节点  $P$  引用节点  $Q$ ) 其应用时间的长短在很大程度上反映了被引用节点的权威与否, 在现实中, 真正的用户对其想要访问的

## 实践与经验

权威页面的访问时间应该比较长,而对于其偶然、作为导航或出于其他目的访问应该一扫而过,即访问时间比较短。而反过来,如果一个用户对某特定页面的访问时间较长,则我们可认为此页面就是用户想要访问的页面,即目标页面,将这一信息应用到 HITS 算法的 Authority 权重的计算,则可极大地提高 HITS 算法的准确性。

### (2) 时间参数控制模型

定义从页面 P 指向页面 Q 关于查询关键字 K 的超链权值为  $W(P, Q, K, T)$ , 此数值由三个因素决定最后的 W 的值: ① P 指向 Q 的链接; ② 查询关键字在超链文字中出现次数的多少 (K); ③ P 访问 Q 的访问时间的多少 (T)。

为了更精确地控制结果, 可引入系数来控制 K 中周围文字的语义信息在超链权值中的比例, 而引入参数  $T = \sqrt{t}$  来控制访问时间对权值的影响。带有时间参数的权值控制模型可由下式表示:

$$W(P, Q, K, T) = 1 + \Phi(k) + \alpha * \Phi(k) + \sqrt{t}$$

上式的  $\alpha$  可以根据不同页面集进行调整, 而 W 的值会在 Authority 权重计算迭代过程中不断增大, 然而本文只关心他们之间的相对大小, 而不是绝对数值。其中  $\sqrt{t}$  反映了随访问时间的增大而其 Authority 权重非线性增大。也可以构造其他函数来控制访问时间在权值中的比例, 以上只是最简单的一种形式。

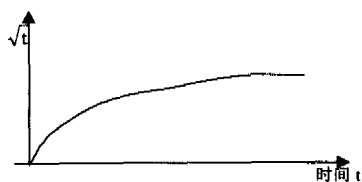


图3  $\sqrt{t}$  的函数图像

### (3) 算法描述

输入: 子集节点, 访问序列与时间及时间的最小阈值  $t_{min}$ 。

输出: 具有较大 Authority 权重和 Hub 权重的页面。

步骤:

① if  $t > t_{min}$

    计算节点权值  $W(P, Q, K, T) = 1 + \Phi(k) + \alpha * \Phi(k) + \sqrt{t}$

    else 从节点集中删除

② 依 W 迭代计算 Authority 权重 ( $ap$ ) 和 Hub ( $hp$ )

权重:

$$ap = \sum_{q:q \rightarrow p} hq$$

$$hp = \sum_{q:q \rightarrow p} aq$$

③ 输出具有最大 Authority 权重和 Hub 权重的页面。

## 结语

本文在分析传统 HITS 算法思想及实现原理的基础上, 对其不足进行了剖析, 并根据具体情况列举了不同的改进方案。仅依据页面链接结构推测权威页面, 虽然有其合理的方面, 但有失全面性与客观性, 后来许多学者在其基础上进行了改进, 将文本信息引入到 HITS 算法中, 较理想地克服了传统 HITS 算法不全面推理的不足。然而对于非正常引用对结论的影响却力不从心, 本文引入时间参数这一访问页面的重要指标, 很好地解决了这一问题, 在现实中有较好的应用价值。

## 参考文献

- [1] 丁国栋, 王斌, 白硕. Web 超链挖掘: 中国境内 Web 图结构研究, 2005, 31(14): 24-26
- [2] 李昕, 朱永胜, 武港山. Web 结构分析算法 HITS 的改进及应用, 2005, 31(6): 40-42
- [3] 杨炳儒, 李岩, 陈新中等. Web 结构挖掘. 2003, 29(20): 28-30
- [4] 宋建康, 张礼平. Web 结构挖掘算法探讨. 2003, 29(5): 537-540
- [5] 张岭, 马范援. 一种提高 Web 结构挖掘质量的新方法. 2004, 41(1): 98-102
- [6] 刘琨, 郑有才. 搜索引擎剖析. 2004, 14(3): 18-22
- [7] 王晓宇, 周傲英. 万维网的链接结构分析及其应用综述. 2003, 14(10): 1768-1778
- [8] 刘丽珍, 宋瀚涛, 陆玉昌. 网络结构挖掘的关键分析. 2003
- [9] 石晶, 龚震宇, 袁抗萍. 一种更稳定的链接分析算法-子空间 HITS 算法. 2003, 41(1): 49-53
- [10] 韩家炜, 孟小峰, 王静等. Web 挖掘研究. 计算机研究与发展, 2001, 38(5): 405-414
- [11] 范明, 孟小峰. 数据挖掘概念与技术. 北京: 机械工业出版社, 2003, 91

(收稿日期: 2006-03-15)

# The Application of Time Parameter in HITS Algrithom

WANG Xue-long<sup>1</sup> , ZHANG Xue-mei<sup>1</sup> , LI Xiang-wei<sup>2</sup>

(1. The Zhangye Middle School of Gansu Province, Zhangye 734000 China;

2. College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070 China)

**Abstract:** In view of the shortage and the research status of the HITS algrithom, we import time parameter to it, and improve the primitive HITS algrithom. The improved algrithom has better effectiveness and veracity compared with the primitive algrithom.

**Key words:** Web; Hyperlink; HITS

(上接第 87 页)

# Design of Intelligent Home Center Control Unit on Embedded Web Server

ZHANG Heng , YE Wei-qiong , LIN Wei

**Abstract:** In term of the characteristic of Intelligent Home application, it provides a new kind of design scheme of Intelligent Home Center Control Unit. The Center Control unit uses Embedded Web Technology, and is based on ARM9 & Linux. This paper describes software and hardware design for the system, and introduces how to perform the main functions of the system such as HTTP protocol, Dynamic Web Page and MiniGUI Manage program etc.

**Key words:** Intelligent Home; ARM; Embedded Web Server; MiniGUI