Web 结构挖掘中 HITS 算法的改进*

郭 鸿，周 娅
（桂林电子科技大学 计算机与控制学院，广西 桂林 541004）

摘 要：HITS 算法是 Web 结构挖掘中一种经典的链接分析算法，其主要问题是容易发生主题漂移。针对这一问题，提出了一种基于文本内容和链接分析相结合的改进算法。实验证明改进后的算法提高了查询结果的相关度，降低了主题漂移的可能性。

关键词：HITS 算法；主题漂移；权威网页；中心网页
中图分类号：TP311 **文献标识码**：A

Improvement of HITS algorithm in web sructure mining

GUO Hong, ZHOU Ya
(College of Computer and Control, Guilin University of Electronic Technology, Guilin 542004, China)

Abstract: HITS is one of the classical link analysis algorithms in web structure mining, whose main problem of it is the topic drif. A new algorithm which is based on text and link analysis is proposed in this paper. Simulation experiments show that the improved algorithms increased the relevance of query results and decreased the probability of the topic drift.

Key words: HITS (Hyperlink-Induced Topic Search); topic drift ; text; authority web ; hub web

Internet 是一个巨大、分布广泛、全球性的信息服务中心，它提供了各种各样的信息服务。但如何从 Internet 浩如烟海的信息中获取所需信息或是从中提取有用知识，一直是相关专家探究的问题。将传统的数据挖掘技术和 Web 结合起来，对 Web 进行数据挖掘成为解决这一问题的重要途径。由于 Web 上的链接结构含有非常丰富和重要的信息，链接分析技术已经被成功地用于分析 Web 超链接数据来确定权威信息源。而在各种对网页进行链接分析并提取主题的算法中，HITS 算法是最典型的。

1 HITS 算法

1.1 HITS 算法的基本思想

HITS 算法[1]是一种 Web 结构挖掘算法[1]，该算法基于用户的查询，根据给定的查询通过分析 Web 的前向链接和后向链接来发现一组相关网页，从而找出 Web 集合中的 authority 网页(与给定查询主题的上下文最为相关并具有权威性的网页)和 hub 网页(提供指向权威网页链接集合的 Web 网页)。为每个网页定义两个度量值:权威权重(authority weight)和中心权重(Hub weight)，通过这两个权重来判定该网页对特定主题的重要性。

1.2 HITS 算法的具体过程

整个 HITS 算法主要可以分为以下几个步骤:

(1)在搜索引擎上输入给定的关键词，以此搜索到的最前面的 r 个等级最高的查询结果网页作为根集(root set) R，R 需满足如下 3 个条件：① R 中网页数量相对较小.② R 中网页大多数是与查询关键词 q 相关的网页。③ R 中网页包含较多的权威网页。

(2)通过向 R 中加入被 R 引用的网页和引用 R 的网页将 R 扩展成一个更大的基础集合(base set)B。扩展规则为：将根集中的全部网页加入进来，并加入最多 d 个链接到根集 R 中的 Web 网页。

(3)以 B 中的 Hub 网页为顶点集 V_1，以 authority 网页

基金项目：广西青年科学基金（桂科青 0832101）

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 $G=(V_1, V_2, E)$。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。假设 Web 链接结构子图 G 中包含 n 个节点(网页)，对这 n 个节点加以编号：1，2，…，n，这样就可以为 Web 链接结构子图 G 定义一个 $n \times n$ 的邻接矩阵 A，如果页面 i 指向页面 j，则矩阵中的项(i, j)为 1，否则为 0。同样把所有节点的 authority 和 hub 值定义为向量形式，即：$a=(a_1,a_2,...,a_n)$和 $h=(h_1,h_2,..., h_n)$。

算法如下：

a,h 初始化为 1, $a_0=1$, $h_0=1$

$t=1$

do $a_i = \sum_{j \in B(i)} h$

$h_i = \sum_{j \in F(j)} a$

for each v in V

do j

i

//$B(i)$和 $F(j)$分别代表网页集指向网页 i 和网页 j 指向网页集。

$a_t=a_t / \| a_t \|$

$h_t=h_t / \| h_t \|$

$t=t+1$

while $\| a_t - a_t{-}1 \| + \| h_t - h_t{-}1 \| < \xi$

return(a_t,h_t)

上面的迭代式链接分析算法相当于执行 $a=A^{\mathrm{T}}h$ 和 $h=Aa$，进一步展开可以得到：

$$a = A^{\mathrm{T}} h = A^{\mathrm{T}}Aa = (A^{\mathrm{T}}A)\, a$$

$$h = Aa = AA^{\mathrm{T}}h = (AA^{\mathrm{T}})h$$

根据线性代数的理论，向量 a 和 h 经过展开计算后，会收敛至对称矩阵 $A^{\mathrm{T}}A$ 和 AA^{T} 的主特征向量。$A^{\mathrm{T}}A$ 的主特征向量代表权威网页，而其主特征向量中数值越高代表网页的权威权重也越高；同样，AA^{T} 的主特征向量代表中心网页，而其主特征向量中数值越高代表网页的中心权重也越高。通过以上过程可以看出，经过若干次迭代计算后，即可得到每一页面的 authority 和 hub。基集 B 中网页的权威权重和中心权重从根本上说是由基集 B 中网页的链接关系所决定的，更具体地说，是由对称矩阵 $A^{\mathrm{T}}A$ 和 AA^{T} 所决定的。

2 HITS 算法中存在的问题

HITS 算法虽然在某些查询主题下能够较为准确地提取出权威网页，但在一些场合中仍会使得算法发生严重的“主题漂移”[2]的现象(authorities 集中到一些链接稠密的非相关网页的现象被称为“主题漂移”)。该现象的出现说明在传统 HITS 算法中仍存在一些缺点，这就要求对传统 HITS 算法进行改进，以使其具有更为广泛的适用性，提高权威页面搜索的效率。

3 HITS 算法的改进

HITS 算法遇到的问题，多是因为 HITS 是纯粹的基于链接分析的算法，没有考虑文本内容。继 KLIINBERG J 提出 HITS 算法以后，很多研究者对 HITS 进行了改进，提出了许多 HITS 的变种算法，主要有 IBM Almaden 研究中心 Clever 搜索引擎的 ARC(Automatic Resource Compilation)算法[4]和由 GEVREY J 和 RUGER S 于 2002 年提出来的两个基于超链接和内容的网页排序算法[5]：Average 算法和 Sim 算法等。

针对 HITS 算法发生的“主题漂移”的现象，本文在链接分析的基础上引入了网页内容信息[3]的判断，提出了一种改进的 HITS 算法。

3.1 改进思想

HITS 算法中，构造一个基本集 R 集，然后通过基本集扩展到 B 集，形成整个 Web 子图。这样做的原因是 R 集可能并不包含真正的用户需要的页面。例如搜索关键词“搜索引擎”时，文本搜索引擎返回的页面通常不会包含 Google、Yahoo 等搜索引擎的页面，因为它们的页面通常不会出现搜索引擎这样的字眼。这使得原本很重要的页面不能被包含在第一步得到的结果中。B 集可以解决这个问题，因为可以通过 R 集中网页的链接来得到需要的网页。但是也正是由于 HITS 算法的这种特性使得它在构造 B 集时，常常会引入过多与主题无关的页面，它们有些还由于拥有互相指向的链接而拥有较高的权威值。如果控制 B 集构造时的半径，可能得不到足够的页面，B 集半径足够大可能会找到真正的合适页面，但是这时也已经引入了过多的无关页面。

针对此，本文在链接分析的基础上引入网页内容信息[2]的判断，通过计算 B 集中每一网页与主题的相似度，设定阈值去掉相似度较低的页面，然后将网页的相似度用于最终的迭代计算，有效地去除“主题漂移”现象。

改进算法采用的模型和技术与当前 Web 检索系统大多采用的向量空间模型(VSM)和技术有最大的兼容性，以便算法的有效实现以及与当前检索系统的有效集成。改进后的算法主要包括 3 个过程：(1)有效地选取基集；(2)扩展基集时通过余弦公式对网页内容信息进行判断，使扩展后的网页与查询主题有最大的相关性，从而避免“主题漂移”；(3)迭代计算与返回结果[4-8]。

3.2 算法详细步骤

(1)合理地获取基集，构造链接结构子图 G，对于图 G 中的每一个节点 V(网页)有两个值，分别是 hub 值与

authority 值，用 $H(v),A(V)$表示，把所有节点的 authority 和 hub 值定义为向量形式，即：$a=(a_1,a_2,...,a_n)$和 $h=(h_1,h_2,...,h_n)$ $V=1,2,3..N;N$为 G中节点(网页)数量。

(2)对 $H(v),A(v)$进行初始化，使得 $H(v)=1$，$A(v)=1$。

(3)内容匹配：将 B集中扩展得到的网页看做一篇文档，把文档 d和查询式 q表示成向量形式($d=(d_1,d_2\cdots d_n)$ d_i代表第 i篇文档 $q=(q_1,q_2\cdots q_n)$ q_i代表查询主题中第 i个关键词)。文档 d(document)可看成是由相互独立的若干词条(term)($t_1,t_2...t_n$)组成，对于每一词条 t_i，根据词条在文档中隐含的语义及重要程度赋以一定的权值 W_{ti}，则文档的特征向量为($W_{t1},W_{t2}...W_{tn}$)，通过 Similarity(d_i,Q) 余弦公式来表示第 i篇文档与查询条件 Q的相关度。

$$Similarity(d_i,Q)=\cos\theta=\frac{\sum_{j=1}^{m}(W_{ij}\times q_j)}{\sqrt{\sum_{j=1}^{m}(W_{ij})^2\times\sum_{j=1}^{m}q_j{}^2}}$$

$$q_j=\begin{cases}1 \text{ 若 } t_j\in Q\\ 0 \text{ 若 } t_j\notin Q\end{cases}$$即特征向量 t_j出现在查询条件 W中，则 q_j为 1，否则为 0。

并以此作为权重赋予相应的节点(网页)，Web 节点的内容与查询主题相关度越大，对应的权值也越大。这样，链接结构图就成了节点带权的有向图，使用这样的权重来合理控制链接分析时节点对 authority/hub 值的影响，最终有效控制主题偏移现象。

(4)将相应的权值 weight(n)进行归一化。

(5)$\forall i\in V$，有 $a_i=\sum_{j\in B(i)}h_j W_j$，$h_i=\sum_{j\in F(j)}a_i W_j$。

(6)将计算出的 a和 h值进行规一化，使得 $\sum_{i=1}^{n}a(i)^2=1$及 $\sum_{i=1}^{n}h(i)^2=1$。

(7)若 a和 h值没有收敛时，转到(5)。

(8)设定阈值 y，将所有 a和 h值大于 T的网页挑选出来，输出查询结果。

4 实验结果与分析

在测试文档集的选择上，选用 BORODIN A 等人提供的 Web 文档集[9](包括“Abortion”、“Genetic”、“Movies”、“Harvard”等关键词依次对应的 2 849，2 613，5 613,1 583 个网页)对改进的 HITS 算法和原 HITS 算法进行了实验比较，实验数据如表 1 所示。

表1 实验数据统计

query	hubs	authorities	links
Abortion	949	933	2 849
Genetic	930	641	2 613
Movies	2 0b1	1 885	5 613
Harvard	537	462	1 583

通过实验数据，对搜索出来的前 30 位的网页进行相关率比较如图 1 所示。在前 30 位网页中发现原 HITS 算法将许多与查询主题无关的网页排了进来，使得网页相关率较低；而改进后的 HITS 算法排在前 30 内的网页相关率明显高于原 HITS 算法。

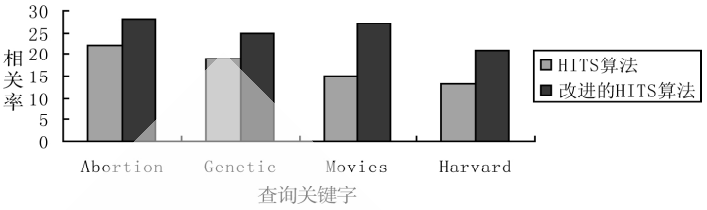

图 1 网页相关率示意图

再对获取网页的前 10 位进行权威度比较(这里网页权威度是根据大多数人的评价得来的)，发现原 HITS 算法由于获取相关网页的准确率不高，使得获取权威网页的总体效果也不佳，而改进后的 HITS 算法明显优于原 HITS 算法，如图 2 所示。

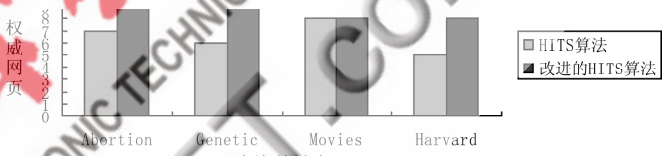

图 2 网页权威率示意图

以上结果说明，在原 HITS 算法中出现了 TKC 问题，排序较高的相关页面中存在与查询主题无关的网页，而改进的算法则有效地控制了 TKC 问题，通过加入对文本内容的分析使排序权值较高的页面与查询主题紧密相关。

文章在深入研究了 Web 挖掘和 Web 链接结构分析的基础上，重点分析了主题提取算法 HITS 的基本思想和算法步骤。针对 HITS 算法基于纯链接，容易发生“主题偏移”现象，本文从网页文本内容着手，提出一种将网页文本内容和链接结构相结合的改进 HITS 算法，并通过实验结果证明了改进后算法的有效性。

参考文献

[1] 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) : 1768-1780.

[2] 倪现军.结构挖掘中web有向图模型的改进算法[J].微计算机信息,2007,12-3:163-165.

[3] 黄丽雯, 钱微. 多文档文本摘要的一种改进HITS算法[J].计算机应用,2006,26(11):2625-2627.

[4] CHAKRABARTI S,DOM B,RAGHAVAN P,et al.Automatic resource compilation by analyzing hyperlink structure and associated text[J].Computer Networks and ISDN Systems,1998,30(4):1-7.

[5] GEVREY J,RUGER S.Link-based approaches for text retrieval.

(下转第 75 页)

(上接第72页)

Proceedings of TREC-10,NIST(Gaithersburg,MD,13-16Nov2001) [M].NIST Special Publication,2002.

[6] XINGW , GHORBANIA. Weighted pagerank algorithm[C].Proceedings of the Second Conference on Communication Networks and Services Research, 2004: 305- 314.

[7] KOSALA R, BLOCKEEL H. Web mining research: A Survey. ACMSIGKDD, 2000(07).

[8] MIZUUCHI Y. Finding Context Paths for web pages[J]. InProc. of 10th ACM Conference, 1999,2(2):13- 22.

[9] BORODIN A, ROBERTS G O, Rosenthal J S, etal.Finding authorities and hubs form link structures on the World Wide Web[C]. In Web,Hong Kong,China,May 2001.

(收稿日期：2009-03-18)