HITS算法在Web挖掘中的应用与改进

蔡 琼,罗雪松

(武汉工程大学 计算机科学与工程学院,湖北 武汉 430073)

摘 要:重点研究了HITS算法,并且在原有算法的基础上对其进行了改进。对搜索出的信息利用改进的HITS算法确定了权威Web页面,有效地去除了无效网页。

关键词:数据的预处理;HITS算法;权威Web页面

中图分类号:TP312 文献标识码:A 文章编号:1672-7800(2008)02-0150-02

1 Web数据挖掘过程中的数据预处理

数据的预处理是对Web上的数据检索后进行的数据预处理,为数据挖掘模块提供挖掘所需要的数据。确定挖掘主题后,可使用Google的Web API, 利用JBuilder实现对Google的巨大Web索引的搜索。但是,用户的一个查询请求往往会检索出庞大的结果集,而用户所需要的只是其中很小一部分,面对如此多的结果,用户仍然不知所措。所以必须用HITS算法来确定权威Web页面,这样就可以有效地去除无效网页。

2 HITS算法

HITS(hypertext induced topic search)[1]算法是一种有效的基于链接分析的主题提取方法,它所依赖的是对超链接环境下链接结构的分析。HITS算法为每个页面引入两个权值:authority权值和hub权值。最后分别输出一组具有最大authority权值的页面和一组具有最大hub权值的页面。authorities是那些与给定查询主题的上下文最为相关并具有权威性的网页;而hubs则是那些本身的内容虽然未必具有权威性, 但却包含了多个指向authorities的超链接的网页。整个HITS算法可以归纳为以下几个步骤[2]:

(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R_σ;

(2)将根集的规模扩展至n,形成基本集S_σ。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_σ中的Web网页;

(3)用$G(S_\sigma)$来表示从基本集S_σ中网页的链接关系所推导出的子图,则$G(S_\sigma)$中包含两种类型的链接——外部链接和内部链接。外部链接是在两个有着不同域名的网页之间的链接,而内部链接则是在两个有着相同域名的网页之间的链接。在实际使用过程中,所有S_σ中的内部链接都将被忽略,而仅仅只考虑外部链接;

(4)将根集S_σ构造为n×n的邻接矩阵A及其转置矩阵A^T,计算A^TA的最大特征值λ,并将λ所对应最大特征值的主特征向量e_1进行归一化;

(5)将经过归一化后的特征向量e_1中具有较大绝对值的元素作为authorities返回;

(6)计算AA^T的最大特征值λ_1',并将λ_1'所对应最大特征值的主特征向量e_1'进行归一化;

(7)将经过归一化后的特征向量e_1'中具有较大绝对值的元素作为hubs返回[3]。

3 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,但仍存在在一些场合中会使得算法发生严重的“主题漂移”的现象[4](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”问题)。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[6]。

4 HITS算法的改进

4.1 迭代过程中尽量向根集投影

仔细观察HITS算法的第(4)步和第(5)步,便会发现传统的HITS算法之所以会出现“主题漂移”现象,就是因为它从主特征向量中所提取的权威网页组, 虽然其内部链接关系最为稠密,但是该组网页与根集的关系却最小,即和原查询主题的关联程度非常小。也就是说传统的HITS算法是基于权威值迭代的算法,如果根集中存在着与主题不相关的一组稠密网页相连

作者简介:蔡琼(1961-),女,湖北武汉人,硕士,武汉工程大学计算机学院副教授,研究方向为数据库、数据挖掘;罗雪松(1982-),男,湖北武汉人,武汉工程大学计算机学院硕士研究生,研究方向为数据挖掘。

接的网页，那么，那组内部链接稠密的网页就会使得该组网页的权值在HITS算法的运行过程中不公平地迅速增加，使得结果不可避免地向该组网页集中。为了避免这种情形的发生，就应该考虑从其它非主特征向量中提取不同的、虽然内部链接并不是最为稠密、但与根集关系却更为密切的权威网页组。

所以要对原来HITS算法的第(4)、(5)步进行如下改动，以使其具有更为广泛的适用性，增强权威页面搜索的效果。

对于第(4)步的改进为：①根据S_σ构造$n\times n$的矩阵A及其转置矩阵A^T，计算其每个特征值$\lambda_1, \lambda_2, \cdots, \lambda_n$所对应的特征向量$e_1, e_2, \cdots, e_n$，并将它们归一化；②将每一特征向量的各项元素均取其绝对值，然后在根集子空间W上进行投影P，再进行比较，从中找到使$|P\lambda e_i|$的值达到最大的特征向量e^*。

对于第5步的改进为：将归一化后的特征向量e^*中具有较大绝对值的元素作为authorities返回。

通过修改原来算法的第(4)、(5)步，可以有效地抑止与查询主题无关的内部链接稠密网页组的提取，使提取结果更为向根集靠近，从而避免了“主题漂移”现象的产生。

4.2　基本集收缩

通过前面的分析，“主题漂移”的产生是由于基本集中包含了大量互不相关的网页。因此对基本集还可以进一步精简：一方面，剔除掉与根集关系不大的网页可以对“主题漂移”问题加以控制；另一方面，还可以大大减少运算量。所以可以对传统HITS算法的第(2)步进行如下改进：

(1)将根集R_σ的规模扩展至n，形成基本集S_σ。扩展规则为：将根集中网页所链接的全部网页加入进来，并加入最多d个链接到根集R_σ中的Web网页；

(2)设定一个参数k，对已经获取的基本集进行进一步筛选。即只选取那些链向多于k个根集网页的网页，以及被多于k个根集网页所链向的网页，得到缩减后的基本集。

随着基本集规模的缩减，邻接矩阵的阶数也大大减小，因此该改进不但可以有效地降低特征值的计算开销，也可以很大程度上抑止“主题漂移”现象的产生。

5　HITS算法实验结果及其分析

选择了“Artificial Intelligence”这一查询主题进行了实验，“Artificial Intelligence”主题下的基本集包含1206个网页。用未改进的HITS算法得到的结果产生了严重的“主题漂移”。用改进的HITS算法，其中authorities中前几项的结果包括American Association for Artificial Intelligence的主页，这是美国一个非营利性科学团体；《Journal of Artificial Intelligence Research》，这是人工智能方面的国际期刊网页；还有介绍人工智能方面科学的网页和全面介绍人工智能学科的综合网站等。显然，它们都符合权威网页的要求。在hubs中，包含了大量指向人工智能学科领域中团体、文献、成果等各方面的有用链接，也完全符合hubs包含丰富的有价值链接的要求。

从实验结果来看，应该说改进的HITS算法较为圆满地完成了提取authorities和hubs的任务，算法是可行的。

参考文献：

[1] 聂培尧.基于XML的半结构数据管理及数据集成问题研究[D].西安:西北工业大学，2002.
[2] 吴共庆，陈恩红. 一种基于XML的半结构化数据存储方法研究[J].计算机工程，2004(4).
[3] 王宁，徐宏炳，王能斌.基于带根连通有向图的对象集成模型及代数[J].软件学报，1998(5).
[4] 韩江洪，郑淑丽，魏振春等.面向XML的Web数据模型研究[J].小型微型计算机系统，2005.
[5] 刘洋.基于Web的内容挖掘技术研究[D].哈尔滨:哈尔滨工业大学，2003.
[6] 蒲秋菊.基于XML的Web数据挖掘技术的研究[D].武汉:武汉大学，2004.

（责任编辑：赵　峰）

Application and Improvement of HITS in Web Mining

CAI Qiong, LUO Xue-Song
(Institute of Computer Science and Technology, Wuhan Institute of Technology, Wuhan 430073,China)

Abstract:The paper mainly researches the algorithm HITS and improve it based on the origion algorithm. And on the base of pretreatment, using improved HITS algorithm can confirm the authority of Web page.
Key Words:pretreatment;HITS; authoritative Web page

勘　误

本刊 2008 年 1 月刊《DS18B20 数字传感器温度检测显示系统》作者李元斌简介为：李元斌(1962-)，男，湖北天门人，华中科技大学工程师，研究方向为单片机控制和应用。