基于Web链接分析的HITS算法研究与改进

喻金平[1],朱桂祥[2],梅宏标[3]
YU Jinping[1], ZHU Guixiang[2], MEI Hongbiao[3]

1.江西理工大学 工程研究院,江西 赣州 341000
2.江西理工大学 信息工程学院,江西 赣州 341000
3.江西理工大学 应用科学学院,江西 赣州 341000
1.Engineering Research Institute, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China
2.School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China
3.College of Applied Science, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China

YU Jinping, ZHU Guixiang, MEI Hongbiao. Research and improvement of HITS algorithm based on Web link analysis. Computer Engineering and Applications, 2013, 49(21):42-45.

Abstract: Vertical search engines have two kinds of subject search strategy, one is based on content evaluation, the other is based on Web link analysis, and HITS algorithm is a classical search strategy that is based on Web link analysis. Its significant drawback is easy to engender topic drift. In order to avoid engendering topic drift in the maximal degree, this paper puts forward a modified F-HITS algorithm that combines Web's text analysis with diffusion rate. Experiment's results show that those improvements not only can decrease system spending but also raise the accuracy of Web page searching.
Key words: vertical search; search strategy; diffusion rate; text analysis; Hyperlink-Induced Topic Search(HITS)

摘 要: 垂直搜索引擎的主题搜索策略有基于内容评价的搜索策略和基于Web链接分析的搜索策略,其中HITS算法是一种经典的基于Web链接分析的搜索策略,其主要的缺点是容易发生主题漂移。为了最大程度地避免主题漂移,提出了一种结合网页文本分析和扩散速率改进的F-HITS算法。实验结果表明,这些改进不仅节省了系统的开销,并且提高了页面搜索的准确率。
关键词: 垂直搜索;搜索策略;扩散速率;文本分析;超链接分析主题搜索(HITS)
文献标志码: A **中图分类号:** TP309 doi:10.3778/j.issn.1002-8331.1304-0301

1 引言

随着Internet技术的飞速发展,互联网对现代生活的影响越来越大,网页已经成为人们获取和发布信息的重要媒介。垂直搜索引擎具有“专”、“精”、“深”特点且具有行业色彩,相对于通用搜索引擎的信息量大、查询不准确、深度不够等局限性,它是针对某一特定的人群、某个特定的领域或某一特定的需求提供的有一定价值的信息及相关服务。在垂直搜索引擎的实现过程中,网络爬虫的主题搜索策略、主题相关度大小的计算算法是垂直搜索引擎的核心部分。所以,垂直搜索引擎的网络爬虫所采取的搜集网页资源的策略,已经成为近年来研究的焦点问题[1]。

本文通过研究HITS[2]及基于HITS改进的一些算法存在的问题,提出了一种基于HITS改进的F-HITS算法,通过信息管理学科中的扩散理论和网页内容评价对Web页面的Authority和Hub值进行了加权修改,同时对HITS算法中根集和基集进行了缩减,有效地避免了“主题漂移”现象发生。

2 Web链接分析的发展

早期的搜索引擎主要基于检索网页内容与用户查询的相似性或者通过查找搜索引擎中被索引过的页面为用户查找相关的网页。从1996年起,仅仅依靠分析内容相似性来进行搜索的方法变得不再有效,这主要由于下面两个原因造成的[3]:

(1)从20世纪90年代初期到20世纪90年代末,整个互联网上网页数目的增长十分迅速,用户进行一次查询后,

基金项目: 江西省教育厅自然科学基金项目(No.GJJ12346)。
作者简介: 喻金平(1964—),男,教授,研究领域为数据挖掘;朱桂祥(1988—),男,硕士研究生,研究领域为数据挖掘;梅宏标(1976—),男,博士学位,副教授,研究领域为大规模仿真系统工程。E-mail:yjp8761@163.com
收稿日期: 2013-04-22 **修回日期:** 2013-06-25 **文章编号:** 1002-8331(2013)21-0042-04
CNKI出版日期:2013-09-05 http://www.cnki.net/kcms/detail/11.2127.TP.20130905.1047.001.html

得到的相关网页数量往往非常巨大。

(2)基于内容相似性的检索方式容易被一些作弊手段所欺骗，网页所有者可以重复一些关键字，并且在他们的网页中加入大量的相关关键字来提高其在搜索结果中的排名，企图使其网页在更多的查询结果中出现。

大约从1996年开始，学术机构以及搜索引擎的公司中的众多研究者开始转向超链接的研究。随后不久两个最有影响力的PageRank和HITS被设计出来，为链接分析提供了两种不同的方法和思维，被广泛应用在搜索引擎的页面评价和页面排序中。

此后，针对HITS算法的不足，国外研究学者对此作了很多改进，IBM的Almaden研究中心的Clever工程组提出了ARC(Automatic Resource Compilation)算法[4]，对原始的HITS作了改进，赋予页面集对应的连结矩阵初值时结合了链接的锚(anchor)文本，适应了不同的链接具有不同的权值的情况。Lempel和Moran则利用马尔可夫链的概念，对HITS算法进行了改进，淡化了Authorities和Hubs页之间的关系，提出了一种分析超链接结构的随机算法SALSA[5]。在这两种算法的基础上又有一些新的变种算法。Saeko等提出了一种新的HITS改进算法——空间投影HITS算法[6]，算法通过对各个社区的考察，近一步保证了算法结果的合理性。

3　HITS算法

3.1　HITS算法介绍

HITS算法是一种Web结构挖掘，通过挖掘Web链接结构，分析Web间的链接关系找出Web集合中的Authorities和Hubs[7]。其中，Authority是指网络上那些非常著名的且被人们普遍尊重的权威页面。Hub页面是指中心网页，页面提供了指向那些权威页面的链接集合。也就是说，Authority与Hub有一种相互促进的关系，这种Hub与Authority页面的相互加强关系，可用于Authority页面的发现，这就是HITS算法的基本思想。

HITS首先根据查询的关键词确定一网络子图$G=(V,E)$(V为网络子图的结点集，E为边集)，然后通过迭代计算得出每一个网页的权威值和中心值，具体步骤主要可分为四步[8]：

(1)通过搜索引擎获得与主题最相关的K个网页(K=200)的集合，称之为root集。

(2)通过链接分析扩展root集，扩展后得到的集合称之为base集，扩展方法是对于root集中的任一网页p，加入最多d(d=50)个链入网页p和链出网页p的链接到root集中，经过扩展形成base集。

(3)计算base集中所有页面的中心值和权威值：有向边$<p,q>\in E$表示页面p有一条链接指向页面q。首先初始化$A_0=1$，$H_0=1$，然后进行中心值和权威值的计算操作：

$$A_p=\sum_{q\to p}H_q \tag{1}$$

$$H_p=\sum_{p\to q}A_q \tag{2}$$

(4)对A_p和H_p的值进行规范化处理，将所有网页的权威值A_p都除以最高权威值以将其标准化，将所有网页的中心值H_p都除以最高中心值以将其标准化，按上面的规范化操作经过一定次数迭代，直到A_p和H_p的值收敛。

3.2　HITS算法存在的问题

虽然HITS取得了巨大的成功，但是也存在一些问题，主要有：

(1)易发生主题漂移：由于HITS算法局限于Web页面之间的链接关系，忽略了页面的内容，在扩展网页集合里通常会包含部分与查询主题无关的页面，而且这些页面之间有较多的相互链接指向，那么使用HITS算法很可能会给予这些无关网页很高的排名，导致搜索结果发生主题漂移。例如在网页制作过程中加入商业广告、赞助商和友情交换的链接[9]。

(2)计算效率较低：因为HITS算法是与查询相关的算法，所以必须在接收到用户查询后实时进行计算，而HITS算法本身需要进行很多轮迭代计算才能获得最终结果，这导致其计算效率较低，耗时长，在时间方面开销较大[10]。

(3)无关链接的影响：通常一个页面上的链接并不是都与主题相关，例如一些开发者在页面中加入广告、赞助商、导航等链接，这些链接对Authority和Hub的值没有贡献，在一定程度上影响了HITS算法的效果[11]。

(4)忽视新页面：互联网上每时每刻都有大量的新的网页发布，新的网页发布初期，尽管有的网页很重要，但是新网页与其他网页之间的链接较少，导致这些新网页容易被HITS算法所忽视。

4　HITS算法的优化与改进

在针对HITS算法深入分析的基础上，为了避免主题漂移的发生，同时能及时发现互联网上新的网页且不增加额外的系统开销，本文提出了一种基于文本内容和扩散速率改进的F-HITS算法。

4.1　F-HITS算法的改进思想

HITS算法在由root扩展到base集过程中常常会引入过多与主题无关的页面，虽然这些页面链接之间紧密联系，但是均与主题不是相关的，这是由于HITS算法是纯粹基于链接分析的算法，并没有考虑网页的文本内容，所以HITS算法容易发生“主题漂移”现象。因此，本文在构造root集和扩展base集过程中引入网页文本内容判断，对root集和base集进行精简，将会降低发生“主题漂移”的概率。同时，还节省了系统的开销。

随着互联网的迅速发展，互联网信息量呈爆炸式增长，每时每刻都有大量的新的网页发布，在发布初期，这些页面很少被其他网页引用，导致这些新的网页不能够及时出现在搜索引擎中，但是这些刚发布的网页很有可能是想要爬取的与主题相关的重要网站。故本文认为，页面的Authority值和Hub值的计算不应该仅考虑链接数量的累计，还应该考虑网页链接的增幅。这样新发布的网页能够很快地进入用户的视野，提高查询的准确度。根据扩散模型[12]的理论分析可知，每一项创新在互联网上都有精确定义的传播速度，代表了该页面被用户接收并被介绍给别人

的可能性。研究表明,创新的传播速率呈幂率分布[13],即网页刚刚出现时传播速率会很快,接下来会逐渐变慢,直至稳定。本文综合考虑这两个方面的改进思想,结合了网页文本内容分析和扩散速率理论,提出了一种新的F-HITS算法。

4.2 F-HITS算法的介绍

在文本内容分析方面应用比较广泛的是向量空间模型(VSM)[14],该方法将网页文本所含有的基本语言单位:字、词表示为 $d_i(T_1, T_2, \cdots, T_m)$,其中 T_i 代表各个特征项,在一个文本中,每个特征项都被赋予一个权重 W,以表示这个特征项在该文本中的重要程度。权重一般都是以特征项出现的频率为基础进行计算,简记为特征向量 $\boldsymbol{W}_i(\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_m)$,用户查询用查询向量 $\boldsymbol{q}=(\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n)$ 来表示($\boldsymbol{q}_i$ 代表查询主题中第 i 个关键字),通过 $Sim(d_i, \boldsymbol{q})$ 余弦公式来表示网页 d_i 与用户查询 $\boldsymbol{q}$ 之间的相似度:

$$Sim(d_i, \boldsymbol{q})=\frac{\sum_{i=1}^{m}(\boldsymbol{W}_i*\boldsymbol{q}_i)}{\sqrt{\sum_{i=1}^{m}\boldsymbol{W}_i^2*\sum_{i=1}^{m}\boldsymbol{q}_i^2}} \tag{3}$$

若特征项 T_i 出现在查询向量 $\boldsymbol{q}_i$ 中,则 $\boldsymbol{q}_i=1$,否则以0作为权重赋给对应的网页结点,最终将相似度 Sim 值作为权值 W 赋给每个结点,Web结点对应网页文本内容与查询主题相关度越大,其对应的结点的加权值也就越大。

(1)构造root集并进行精简:在构造root集过程中,通过公式 $Sim(d_i, \boldsymbol{q})$ 计算root集中页面 d_i 与查询 $\boldsymbol{q}$ 的相似度。若相似度值小于事先设定的阈值 Q,则将页面 d_i 从root集中删除,否则将页面 d_i 设为root集中的一个结点。

(2)扩展base集并精简:根据前面得到的root集进行base扩展,在扩展过程中,不仅仅考虑页面之间的链接关系,还要考虑新引入的网页文本内容与查询主题的相似度,通过 $Sim(d_i, \boldsymbol{q})$ 计算其相似度,若相似度值小于事先设定的阈值 Q,则将页面 d_i 从base集中删除。

(3)重复(1)(2),直到满足阈值 Q 的扩展页面数量达到 K 为止。

(4)计算页面Authority和Hub值:对构造的网络子图 G 中的每一个节点 p 的Authority和Hub分别用 $A(p)$ 和 $H(p)$ 来表示,将所有节点的Authority和Hub值用向量形式表示,即:$\boldsymbol{a}(a_1, a_2, \cdots, a_p)$ 和 $\boldsymbol{h}(h_1, h_2, \cdots, h_p)$,$p=1, 2, \cdots, K$,将 $A(p)$ 和 $H(p)$ 进行初始化,使得 $A(p)=1, H(p)=1$。接着对页面的Authority和Hub值进行相似度和页面增幅的加权,得到最终的Authority和Hub值。

经过相似度加权后的Authority和Hub计算公式为:

$\forall p \in K, \forall \boldsymbol{q} \in K$ 则有:

$$A(p)=\sum_{\boldsymbol{q}\to p} H(\boldsymbol{q})*\boldsymbol{W}_{\boldsymbol{q}} \tag{4}$$

$$H(\boldsymbol{q})=\sum_{p\to \boldsymbol{q}} A(p)*\boldsymbol{W}_p \tag{5}$$

其中 $\boldsymbol{W}_p$ 是结点 p 的加权值。

通常搜索引擎会定期进行更新,例如每隔一个月Google会利用网络爬虫爬取网络上的网页形成新的网络蜘蛛结构并对网页的重要性和排序进行重新分析。结合扩散理论知识计算出第 t 个周期页面 p 的扩散速率,公式如下:

$$Speed_p^{\text{in}}=\frac{l_t^{\text{in}}-l_1^{\text{in}}}{t\times l_1^{\text{in}}} \tag{6}$$

$$Speed_p^{\text{out}}=\frac{l_t^{\text{out}}-l_1^{\text{out}}}{t\times l_1^{\text{out}}} \tag{7}$$

其中 $Speed_p^{\text{in}}$ 表示页面 p 的链入数量的扩散速率,$Speed_p^{\text{out}}$ 表示页面 p 链出数量的扩散速率,l_t^{in} 表示第 t 个周期页面链入的数量,l_t^{out} 表示第 t 个周期页面链出的数量。

最终标准化处理的计算Authority和Hub的公式为:

$$A(p)=d\times A(p)+(1-d)\times Speed_p^{\text{in}} \tag{8}$$

$$H(p)=d\times H(p)+(1-d)\times Speed_p^{\text{out}} \tag{9}$$

其中,d 为权重因子 $(0.5<d<1)$,表示网页结点的历史Authority和Hub值和页面扩散速率占最终的Authority和Hub值的比重。

(5)对Authority和Hub的值进行归一化处理,使得:

$$\sum_{p=1}^{K}A(p)=1, \sum_{p=1}^{K}H(p)=1$$

(6)如果 $A(p)$ 和 $H(p)$ 未收敛,则返回(3)。

(7)选择 $A(p)$ 和 $H(p)$ 值最大的前10个页面作为最后的返回结果。

5 实验结果及对比分析

5.1 实验设计

本实验采用Lucene作为全文搜索工具包,采用开源的Heritrix爬取网页并采集实验数据,对“Olympic”、“movie”、“football”三个主题进行查询实验,分别使用改进后的F-HITS算法和SALSA算法进行实验结果对比。

搜索引擎的性能评价指标主要有检索时间、查全率和查准率[15]。本文采用查准率来衡量算法改进的效果:

$$precision=\frac{page_{\text{topic \& relative}}}{page_{\text{all}}} \tag{10}$$

在实验中,优先考虑网页链接之间的相互增益关系对网页结点Authority和Hub值的影响,故设权重因子 d 为0.7。此外经过多次实验比较得出:阈值 Q 合理取值为0.6,取值如果过小就没有起到精简root集和base集的作用,主题漂移现象就不能得到有效抑制,取值如果过大就会在扩展base集过程中会消耗大量的时间从而增加系统的开销。

5.2 实验结果对比

本文实验结果将与SALSA[16]算法进行比较。SALSA算法是在深刻研究了HITS算法的基础上提出来的,取消了Authority和Hub之间的加强迭代关系,有效地抑制了主题漂移现象,同时计算量明显比HITS算法要小,所以SALSA算法是目前最好的链接算法之一。

以“Olympic”作为主题查询的实验数据为例,应用SALSA算法及F-HITS算法所得结果如表1及表2所示。

在表1前10个排名中,排名一,二,四,五,六,七的网站相对最有权威性,其他都发生了明显的主题漂移现象;而在表2前10个排名中,排名一,二,三,四,五,六,七,八

表1 SALSA算法的结果

Authority	Hub	网站URL
0.468 3	0.327 8	www.olympic.cn
0.438 6	0.553 1	www.olympic.org
0.396 1	0.424 1	www.olympic.edu
0.384 2	0.485 7	www.london2012.com
0.358 1	0.286 3	http://en.wikipedia.org/wiki/Olympic_Games
0.323 5	0.267 8	www.nbcolympics.com
0.296 7	0.256 7	www.specialolympics.org
0.268 7	0.157 9	www.telegraph.co.uk/sport/olympics/
0.289 1	0.326 4	www.olympicholidays.com
0.278 4	0.259 7	www.olygamefarm.com

表2 F-HITS算法的结果

Authority	Hub	网站URL
0.742 1	0.572 1	www.olympic.cn
0.724 8	0.310 6	www.olympic.org
0.697 5	0.364 7	www.nanjing2014.org
0.665 2	0.486 1	www.nbcolympics.com
0.603 1	0.345 2	www.olympic.org/sochi-2014-winter-olympics
0.574 5	0.432 7	www.specialolympics.org
0.563 9	0.474 2	www.london2012.com
0.535 9	0.413 7	http://en.wikipedia.org/wiki/Olympic_Games
0.497 8	0.375 1	http://sochi2014olympicstickets.com
0.456 7	0.373 2	www.olygamefarm.com

的网站相对最有权威性,其他都发生了明显的主题漂移现象,而且第三,四,五是2014年奥运赛事官网,页面较新。

5.3 实验结果分析

实验分别对三个主题返回的结果进行查准率分析,通过表1和表2比较分析可以看出,对同一主题的查准率F-HITS算法高于SALSA算法,有效地抑制了主题漂移现象,而且通过F-HITS算法得到的链接排序中,相比于SALSA算法会有更多的发布不久的页面的链接。

通过图1分析可以看出,在页面较少时F-HITS算法的查准率明显高于SALSA算法,随着页面逐渐增大,F-HITS算法查准率也在下降,逐渐和SALSA算法接近。这是因为更多的跟文本相关的刚发布不久的页面链接也被考虑其中,而且发布前期扩散速率较快的页面对F-HITS算法的结果影响较大,随着集基中页面数量的增大,不同时间发布的页面越来越多地被F-HITS算法考虑到,扩散速率方面的因素对F-HITS值影响也越来越小。

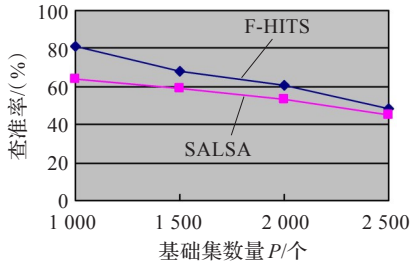

图1 查准率随着基础集数量的变化

此外,由于改进的算法通过网页文本内容和查询主题的相似度的计算,精简了根集合扩展中的基集,在一定程度上节省了系统的开销。

6 结束语

HITS算法对所有链接分配相等权重容易导致发生主题漂移现象。本文通过信息管理学科中的扩散理论和网页内容评价对Web页面的Authority和Hub值进行了加权修改,提出了一种结合网页文本分析和扩散速率改进的F-HITS算法。通过设计实验,对改进的算法进行对比分析。实验结果表明,F-HITS算法有效地降低了主题漂移现象发生的概率,提高了搜索的准确度,但是在节省系统开销方面跟SALSA算法比较并没有得到显著改善,因为在精简root集和base集的过程中同样占用了一定系统资源和开销。因此结合多线程网络爬虫技术可能会提高搜索的效率、节省系统的开销,这方面值得继续探讨和研究。

参考文献:

[1] 彭涛.面向专业搜索引擎的主题爬行技术研究[D].长春:吉林大学,2007:1-2.

[2] Kleinberg J M.Authoritativesources in a hyperlinked environment[C]//Proc 9th ACM-SIAM Symposium on Discrete Algorithms,1998:668-677.

[3] Liu Bing.Web数据挖掘[M].俞勇,薛贵荣,韩定一,译.北京:清华大学出版社,2009:174-175.

[4] Chakrabarti S,Dom B,Raghavan P.Automatic resource compilation by analyzing hyperlink structure and associated text[C]//Proc of the 7th International Conf on WWW,1998.

[5] 何晓阳,吴治水,连丽江,等.SALSA算法技术剖析[J].情报杂志,2004(7).

[6] Saeko N,Satoshi O,Toru I,et al.Analysis and improvement of HITS algorithm for detecting Web communities[C]//Proceedings of the 2002 Symposium on Applications and Internet(SAINT'02),2002.

[7] Madria S K.Research issues in Web data mining[J].Data Warehousing and Knowledge Discovery,1999,1676(99):303-312.

[8] 罗林波,陈绮,吴清秀.基于Shark-Search和Hits算法的主题爬虫研究[J].计算机技术与发展,2010,20(11):77-78.

[9] 张聪.基于HITS的链接分析算法的研究与改进[D].大连:大连理工大学,2007:22-23.

[10] 刘迪慧.一种基于相似度值的向量空间投影HITS算法[D].重庆:重庆交通大学,2010:25-26.

[11] 刘军.基于Web结构挖掘的HITS算法[D].长沙:中南大学,2008:18-19.

[12] Barabasi A L.Linked[M].徐彬,译.长沙:湖南科学技术出版社,2007.

[13] Bak P.How nature works[M].Oxford,England:Oxford University Press,1996.

[14] Salton G.Introduction to modem information retrieval[M].New York:McGraw-Hill,1983.

[15] 朱庆华,杜佳.搜索引擎评价指标体系的建立和应用[J].情报学报,2007,26(5):684-690.

[16] Lempel R,Moran S.The Stochastic Approach for Link-Structure Analysis(SALSA) and the TKC effect[J].Computer Networks,2000,33:387-401.