

文章编号:1001-9081(2006)05-1171-03

基于超链接和内容相关度的检索算法

张娜,张化祥

(山东师范大学 信息科学与工程学院, 山东 济南 250014)

(cocobaby1900@yahoo.com.cn)

摘要:在网络环境下,经典的链接分析方法(HITS 算法)过多的关注网页的权威性,忽视了其主题相关度,易产生主题漂移现象。在简要介绍 HITS 算法的基础上,分析了其产生主题漂移的原因,并结合内容相关度评价方法,提出了一种新的搜索算法——WHITS 算法。实验表明,该算法挖掘了超链接间的潜在语义关系,能有效的引导主题挖掘。

关键词:HITS 算法;超链接;向量空间模型

中图分类号:TP391 **文献标识码:**A

Retrieval algorithm based on hyperlinks and content similarity

ZHANG Na, ZHANG Hua-xiang

(School of Information Science & Engineering, Shandong Normal University, Jinan Shandong 250014, China)

Abstract: Under the circumstances of web, classical hyperlink analysis algorithms (such as HITS algorithm) mainly focused on the authority of a web page rather than its topic, so it was easy to drift away from the mining topic when traversing the hyperlinks. The cause of topic drifting away in HITS algorithm was analyzed. By combining the topic analysis method with the content relevance evaluation, a novel web information retrieval algorithm - WHITS was presented. Experiment results show that WHITS focuses on mining the potentially semantic relationship between hyperlinks and performs quite well in the topic-specific crawling.

Key words: HITS; hyperlink; VSM

0 引言

万维网信息的爆炸性增长使 Web 已经成为世界上最大的信息库。面对这个分散无序的海量信息库,Web 用户经常发现难以找到能满足他们需要的信息,造成“信息过载,知识匮乏”的问题^[1]。为了解决 Web 上的信息过载的问题,人们提出了 Web 信息检索理论(即 Web IR)^[2]。目前许多 Web IR 系统已经成为 Web 用户访问 Web 的必备工具,其中最广泛使用的工具是通用搜索引擎。

通用搜索引擎是将用户提交的查询词作为查询条件来获取与用户需求相关的网页。目前在 Internet Explorer 和 Netscape Browser 等浏览器以及像 Google、Alexa 等搜索引擎都提供了查找相关网页的功能。网页与普通文本之间的最大区别是,网页内有超链接而普通文本没有。超链接不仅可以作为用户浏览 Web 的导航信息,还可以表达网页作者的某些意图,例如:网页 A 和网页 B 之间有超链接,则网页 A 和网页 B 可能属于同一个主题。对 Web 上的超链接进行分析,可以增强检索结果的准确率,所以大部分主要的搜索引擎都声称使用了某种超链接结构分析技术。现在已形成的描述网络超链接拓扑结构的算法主要有 PageRank^[3]、HITS^[4]等。HITS (Hypertext Induced Topic Search)算法是利用页面的被引用次数及其链接数目来决定不同网页的价值。这种方法可以获得比较高的查全率,但是忽略了文本内容,容易产生主题漂移现象。本文在简要介绍 HITS 算法的基础上,分析了产生主题漂移的原因,并进一步提出了改进的检索算法。

1 HITS 算法

HITS 算法^[4]是由 Kleinberg 提出的基于超链接的检索算法。Authority 页面和 Hub 页面是 HITS 算法中两个重要的概念。Authority 页面是指与某个查询关键词和组合最相近的页面。Hub 页面是指它的出链中包含了很多的 Authority 页面的页面。它的主要功能就是把这些 Authority 页面联合在一起。

HITS 算法的基本思想是:先用基于文本的搜索引擎得到一个 Web 子集,称为“根集”(Root Set) R ,根集内的页面和用户查询都有较大的相关性;然后将 R 所有指向的网页集合以及其他指向 R 的网页集合包含进来形成“基集”(Base Set) S 。然后利用 Authority 页面和 Hub 页面的互增强属性,对基集进行链接分析,通过迭代的计算方法为基集中的每个页面计算一个 Authority 值和一个 Hub 值,作为结果页面排名的依据。

HITS 算法如下:假定基集 S 中的页面分别为:1, 2, ..., n 。每个页面 i 有一个 Authority 值 a_i 和 Hub 值 h_i ; 页面 i 的入链页面集表示为 $B(i)$, 出链页面集表示为 $F(i)$ 。

(1) 设定收敛的最大误差为 T , 初始化 a_i, h_i ;

(2) 执行迭代过程:

$$a_i = \sum_{j \in B(i)} h_j \quad h_i = \sum_{j \in F(i)} a_j$$

(3) 归一化 a_i 和 h_i 的值,使:

$$\sum_i a_i^2 = 1 \quad \sum_i h_i^2 = 1$$

(4) 当 a_i 和 h_i 没有收敛时,转步骤(3);

(5) 将所有 a_i 和 h_i 值大于 T 的网页挑选出来,排序输出

收稿日期:2005-11-07;修订日期:2006-01-20

作者简介:张娜(1980-),女,山东淄博人,硕士研究生,主要研究方向:Web 挖掘、信息检索、数据挖掘;张化祥(1966-),男,教授,博士,主要研究方向:多代理协同技术、机器学习。

查询结果。

网络拓扑用有向图 $G = (V, E)$ 表示,其中 V 是页面集合, E 是页面之间的超链接集合。页面抽象为图中的顶点,而页面之间的超链接抽象为图中的有向边。定义它们的 $n \times n$ 阶邻接矩阵,如果页面 i 指向页面 j ,则矩阵中的项 (i, j) 为 1,否则为 0。同样把所有的 authority 值和 hub 值分别定义为向量, $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n)$,则算法第三步可转化为:

$$x \leftarrow A^T y; y \leftarrow Ax$$

进一步展开,可以得到

$$x \leftarrow A^T y \leftarrow A^T Ax \leftarrow (A^T A)x;$$

$$y \leftarrow Ax \leftarrow AA^T y \leftarrow (AA^T)y;$$

因此向量 x, y 均可由式(3)经过多次迭代而得。根据线性代数的理论,迭代序列经过标准化最终将收敛于矩阵的特征向量,即上文计算的 hub 权值和 authority 权值是页面集合的固有特征,不是由初始向量和参数的选择决定的。

HITS 这种检索算法可以获得比较好的查全率,但也存在一些缺点^[6]:第一,由于算法认为页面中的所有超链接具有同等价值,只要两个页面之间有超链接,则邻接矩阵中对应的值即是 1,这样完全不考虑页面文本的内容,容易出现主题漂移^[5]。第二,Web 页面中的许多链接都是为了其他目的而创建的,例如广告标语、导航等等,因此单凭链接数目来判断页面的 authority 值和 hub 值,是不合理的,容易受到一些 Web 页面创建者的欺骗。

针对这些问题,本文将基于内容的检索方法与超链接分析技术结合起来,提出一种新的改进算法 WHITS 算法。

2 WHITS 算法

2.1 算法描述

(1) 将查询条件提供给两个或多个通用搜索引擎,取各自查询结果的前 c 名 url,删除重复项后,合并得“根集”,然后扩展“根集”得到“基集”。

(2) 把“基集”看作一个有向图 G ,每个顶点代表一个网页。对每个顶点 i 赋两个值: $a[i], h[i]$,分别代表 Authority 值和 Hub 值。

(3) 赋初值: $a[i] = 1, h[i] = 1 (i = 1, 2, \dots, n)$ 。

(4) 相关度分析:将查询条件与网页的超链接和超链接上下文进行匹配,如果匹配则计算出相应的权值 $w(i)$;如果不匹配,则引入向量空间模型,把网页、查询条件用向量的形式表现出来,用查询向量和网页向量进行匹配,计算出相应的权值 $w(i)$ 。

(5) 将权值 $w(i)$ 标准化。

$$(6) a[i] = \sum_{j \in B(i)} w(j) * h_j, h[i] = \sum_{j \in F(i)} w(j) * a_j;$$

(7) 将计算出的 $a[i]$ 和 $h[i]$ 进行标准化,使得:

$$\sum_i a_i^2 = 1 \quad \sum_i h_i^2 = 1$$

(8) 如果 $a[i]$ 和 $h[i]$ 没有收敛,则转向(4)。

(9) 设定阈值 M ,将所有 $a[i]$ 和 $h[i]$ 大于 M 的网页挑选出来,排序输出查询结果。

2.2 构造链接扩展邻接图

当我们给不同的搜索引擎提供相同的查询条件时,由于各个搜索引擎所建立的索引库的内容不完全相同,所运用的检索方法也各有千秋,因此所得到的页面集也不完全一样^[7]。本文采用多个搜索引擎进行检索,把得到的多个页面集进行合并,使根集内容更全面。

首先,将查询条件提交给两个通用搜索引擎,例如 www.sohu.com, www.yahoo.com,得到两个与查询内容相关的 url 集合。取各自排序在前 c 名(设 $c = 300$)的网页合并,得到一个含有大量重复页面和近似页面的集合。如果两个页面的 url 地址完全相同,则是重复页面,删除其中的一个;如果两个页面之间有超过 10 条链接,或是它们的链接指向有 90% 相同,则可能是由于镜像网页或镜像站点产生,我们认为这样的页面是近似页面,删除其中的任意一个。把新得到的集合作为根集,收集所有指向根集的网页和根集指向的网页,作为扩展的邻接图。

2.3 计算网页的相关度权值

我们分两种情况来计算网页的权值:第一种情况是查询条件与网页的超链接和超链接上下文进行匹配;第二种情况是查询向量和网页向量进行匹配。

2.3.1 情况 1

超链接文字是超链接的载体,能够很好的反映链接的语义信息。因此用查询条件与超链接文字相匹配可以取得好的效果。而超链接上下文也能够一定程度上反映链接的语义,为检索带来有用的信息,但有时会带来一些噪声。因此超链接文字和超链接上下文在评价网页相关度时,应具有不同的影响度^[8]。本文通过引入加权系数 a 来控制超链接上下文在网页相关度所占的比例。

$$w_n = \sum_{k=1}^n \frac{tf_k}{L} + a * \frac{tf'_k}{L'} \quad (1)$$

其中, n 为特征值的个数; tf_k, tf'_k 分别为特征值 t_k 在超链接文本和超链接上下文中出现的次数; L 和 L' 分别为超链接文本和超链接上下文的长度。

$$w(i) = \sum_{n=1}^l w_n \quad (2)$$

其中, $w(i)$ 为网页 i 的相关度值, l 为匹配的超链接条数。

2.3.2 情况 2

如果查询条件与超链接及其上下文不匹配,则引入向量空间模型,把查询条件和网页内容用向量的形式表现出来,比较两者的相似度。

$$w_{ij} = D_j * tf_{ij} * idf_j;$$

其中 tf_{ij} 是查询关键词 t_j 在文档 d_i 中出现的频率; idf_j 为 t_j 在整个文档集中出现的倒频率; D_j 为位置因子, t_j 在文档的不同位置出现应当赋予不同的权值。

$$idf_j = \log(N/n_j);$$

其中, N 表示文档集中的文档总数, n_j 表示文档集中包含词 t_j 的文档总数。

查询条件表示为向量空间 $Q, Q = \{q_1, q_2, \dots, q_n\}$ 。其中 q_j 为查询条件中的第 j 个关键词。根据 VSM 模型,文档 d_i 与查询条件的相似度就是这两个向量的内积。如下:

$$W(i) = \text{similarity}(d_i, Q) = \frac{\sum_{j=1}^n w_{ij} * q_j}{\sqrt{\sum_{j=1}^n w_{ij}^2 * \sum_{j=1}^n q_j^2}} \quad (3)$$

3 实验分析

常用搜索引擎评价标准是查全率和查准率^[9]。查全率是指检索到的相关文档与所有满足条件的文档数目的比例;查准率是指检索到的相关文档与检索到的全部文档的比率。这两者是相互制约的,相互影响的。但由于 Internet 资源海量

性的特点,要想统计 Internet 上所有主题相关的网页是比较困难的,因此查准率难以估计。实验是在我们自己开发的基于旅游信息的智能搜索和挖掘平台上进行的。对于用户以关键字形式提供的同一查询条件,我们对 HITS 和 WHITS 这两种算法的查准率进行对比测试。

实验一:由于一般 Web 用户最关心查询结果集中排在前 20 名的网页,因此我们只对查询结果子集中的前 20 个网页进行相关性评价。我们选定的查询关键词有“长城”、“泰山”、“九寨沟”、“布达拉宫”。图 1 给出了两种算法在执行了查询后所获得的查准率。

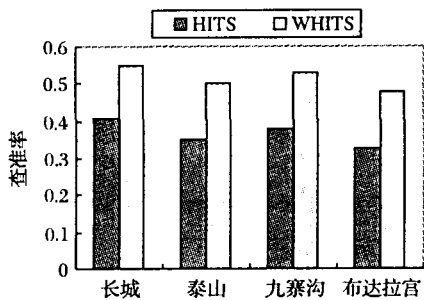


图 1 HITS 算法和 WHITS 算法的性能比较

从图 1 可以看出,WHITS 算法的查准率比 HITS 算法高,说明加入了内容相关度的评价对提高页面评价算法的性能是很有效的。

实验二:随着搜索网页数量的增多,主题漂移现象带来的后果不容忽视。本试验的目的是检测 HITS 算法和 WHITS 算法的主题保持度。 x 轴表示已经搜索过的网页数, y 轴表示主题保持度。

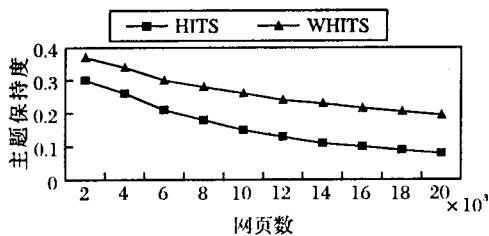


图 2 HITS 与 WHITS 性能比较

从图 2 中可以看出,WHITS 算法的主题保持度要比 HITS 算法高。随着搜索页面的增多,基于 HITS 算法的索引数据库中出现大量与主题无关的页面,影响了搜索的准确度。WHITS 算法加入了内容相关度评价策略,可以及时的判断页面的主题相关度,并裁减掉那些与主题无关的噪声页面,使搜索的精度得以提高。

4 结语

在网络环境下,经典的链接分析方法(HITS 算法)过多的关注网页的权威性,忽视其主题相关度,易产生主题漂移现象。本文在简要介绍 HITS 算法的基础上,分析了其产生主题漂移的原因,并结合内容相关度评价方法,提出了一种新的搜索算法——WHITS 算法。实验表明,该算法挖掘了超链接间的潜在语义关系,能有效地引导主题挖掘。下一步的工作是继续分析网络中超链接之间的关系,把超链接分析与内容分析有机地结合起来,提高检索的查全率和查准率。

参考文献:

[1] KOSALA R, BLOCKEEL H. Web Mining Research: A Survey[A]. SKGDD Explorations[C], 2000.
 [2] KOBAYASHI M, TAKEDA K. Information retrieval on the web[J].

ACM Computing Surveys, 2000, 32(2): 144 - 173.

[3] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation rank2ing: Bring order to the Web! Stanford University[R]. Tech Rep: 199720072, 1997.
 [4] KLEINBERG J. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604 - 632.
 [5] HENZINGER MR, BHARAT K. Improved algorithms for topic distillation in a hyperlinked environment[A]. Proceedings of the 21'st International ACM SIGIR Conference on Research and Development in IR[C], 1998.
 [6] LEMPEL R, MORAN S. SALSA: stochastic approach for link-structure analysis and the TKC effect[J]. ACM Trans. Information Systems, 2001, 19: 131 - 160.
 [7] KOSALA R, BLOCKEEL H. Web mining research: a survey[J]. SIGKDD Explorations, 2000, 2(1): 1 - 15.
 [8] 张敏,高剑峰,马少平. 基于链接描述文本及其上下文的 Web 信息检索[J]. 计算机研究与发展, 2004, (1): 221 - 226.
 [9] YATES RB, NETO BR. Modern Information Retrieval[A]. New York: ACM Press Series/Addison Wesley, 1999.

关于征集中国计算机事业 五十周年大事记的通知

为纪念中国计算机事业创建五十周年,中国计算机学会决定编辑出版“中国计算机事业五十周年大事记”。在“大事记”的基础上,由学会选评中国计算机事业发展历程中的五十件大事。编辑出版完成后,在今年 10 月下旬举行的“纪念中国计算机事业五十周年”活动上发布。

各相关单位或个人均可书面提供“大事记”的内容。“大事记”将反映对中国计算机事业发展有重要影响的事件、项目、发明或成果,“大事记”记录中还包括与该事件相关的主要单位和/或主要人士。

提供者须完整填写下表,电子邮件发至:ccf@ict.ac.cn; 原件同时寄至:北京 2704 信箱中国计算机学会(100080)。须注明“大事记”字样,也可传真至:010-62527485。

事件征集时间范围:1956 年~2005 年

征集截止日期:2006 年 6 月 30 日

中国计算机学会
2006 年 3 月 14 日

“大事记”推荐表

1 事件名称			
2 起始日期		3 结束日期	
4 参与该事件 主要单位	单位 1:		
	单位 2:		
	单位 3:		
5 主要人士			
6 提供者资料 (个人)	姓名:		现工作单位:
	任职:		电话:
	Email:		签字:
	单位名称:		
7 提供者资料 (单位)	联系人:		单位盖章
	电话:		
	Email:		

注: 4: 限 3 个单位; 5: 可选, 限 5 位; 6 和 7 选其一