Clever Algorithm 的一种改进*)

何拥军[1] 骆嘉伟[2] 孙星明[2]

(湖南大学软件学院 长沙410082)[1] (湖南大学计算机与通信学院 长沙410082)[2]

摘 要 解决了Clever算法在Web超链接结构研究方面的一个问题:Clever算法在多重连续的超链接情况下忽略了用户的浏览行为,我们引入了可行性矩阵,提出了一种改进算法,同时应用Warshall算法解决了算法复杂度问题。

关键词 万维网,信息检索,超链接,矩阵理论

An Improvement to Clever Algorithm

HE Yong-Jun[1] LUO Jia-Wei[2] SUN Xing-Ming2

(Hunan University,Changsha 410082)

Abstract This article addresses a problem with the existing approach by Clever Algorithm on studies of Web hyperlink structure: It does not take into consideration of user behavior when following multiple consecutive hyperlinks we present a improved algorithm using Warshall Algorithm and show that any connected hyperlink graph has a unique hyper-weight distribution. Our formulation has been proved by some theorems and some examples are presented in order to compare its effectiveness.

Keywords World wide web,Information retrieval,Hyperlink,Matrix theory

1 引言

搜索引擎和信息检索是大家非常熟悉的Web服务,我们更多的是通过关键词检索我们所需要的信息,事实上,分析Web网页的超链接结果可以给我们很大的启示:当一个网页创建者将他的某一个页面向别的页面作链接(不是指那些用来导航或纯广告的链接)时,在某种意义上正是在表达对所链接页面重要性的一种认可,这种潜在的信息可以用来定位Web上的信息资源。

在超链接环境下,最先把链接分析应用到Web挖掘的是Dr. Jon. M. Kleinberg[1],他提出了HITS算法,该算法的直观思想非常简单:如果页面A有一个链接指向页面B,那么页面A的创建者便认为页面B一定包含了一些有价值的信息资源。该算法的实现过程:首先由某一基于关键词查询得到一初始结果集,然后对结果页面集构建一个包含该结果集合和它邻居页面的一个超链接子图,该子图通过HITS算法反复地进行迭代计算,最后收敛以得到每一节点的权威值。该思想后来在IBM Almadem实验室实现,并把它命名为Clever Algorithm[2]。

Clever Algorithm对每一页面计算两个非负权重值:权威权重值和hub权重值,结果输出具有较大权威权重值和hub权重值的页面,那些具有较高权威权值的页面很有可能是内容相关的页面,而那些具有较高权值的hub页面即可能是那些包含有指向内容相关页面超链接的页面。实现过程如下:

假定$G=(V,E)$是所构建的超链接子图,图中每个节点$P_i\in V$,a[i]和h[i]分别表示它的权威权值和hub权值,初始化a[i]和h[i]都为1,反复下面的计算直至向量a和h收敛:

{

对V中的每个节点有:$a[i]:=\sum_{(P_j,P_i)\in E} h[j]$;

$h[i]:=\sum_{(P_i,P_j)\in E} a[j]$;

最后规范化向量a和h。

Kleinberg[1]证明向量a和h最终收敛,并且它们和权重的初始值设置无关。最后的输出结果页面可以根据权值的不同进行分级。

但是Clever Algorithm存在这么一个问题,下面我们给出一个简单的例子用来说明,假定一个有四个节点的子图如下:

P1←P2→P3→P4

很明显,图中有三个链接:P2到P1,P2到P3,P3到P4,如果对图应用Clever Algorithm,P4的权威权重和P3的hub权重都将收敛为0,这是一个不正常的分值,因为从直观上来看,P4的权威权重分值

*)该课题得到湖南省自然科学基金资助(编号:03092),何拥军 硕士,主要研究方向:数据挖掘。

应该由P3来指向，同样P3指向P4，它的hub权重也不应该为0。

另外再仔细观察就会发现，如果把P4从图中去掉，改为图：P1←P2→P3，应用Clever Algorithm算法P1，P2，P3得到的结果将和上面情况的结果是一样的，因此P4的有无对Clever Algorithm算法也就没有什么意义，这种情况也是不能接受的。

Clever Algorithm算法之所以会出现以上问题的主要原因是：权威值和hub值的相互加强只是通过单一的有向链接产生的，而没有考虑到多重的连续链接情况，因此在多重连续的链接情况下，页面的权值无法得到更新，以上问题需要有新的方法来解决，本文我们就探讨如何有效地解决这个问题。

1.1 相关的研究工作

自Clever Algorith提出以来，该算法已成为计算Web上权威权值和hub权值的主要方法，也已有一些改进算法对它做出了新的应用，但都没有涉及到算法执行过程的本身[3]。

Bharat等人[3]考虑了文本内容的相似性，提出了一个相似链接分析算法，但该方法并没有对Clever算法本身的执行过程进行改进，其思想是基于文本的相似度比较。

D Herbach[4]提出了HITS-SW算法对它进行了改进，但该方法只是为了弥补HITS算法的纯链接分析的不足，把文本内容也考虑到该算法的权值计算当中去，对解决HITS算法的主题漂移问题有一定的帮助。

文[5]提出了一个对HITS算法很好的改进思想，该算法首次提出了HITS算法的多重超链接的问题，并提出了多重连续链接可行性矩阵概念。但是，该方法的时间复杂度有待进一步改善，本文主要是继承该算法的思想然后应用Warshall算法对该算法的时间复杂度进行改进。

2 新算法

在新算法中我们主要是要考虑如何能够解决链接图中多重连续链接的问题，我们引入了一个新的矩阵用来代替Clever Algorithm算法中的邻接矩阵。为了考虑到多重连续链接，也即从一个页面P_i到P_j，其中可能经过的所有路径的可能性，我们的方法是要能计算出这种可能性以更新计算的权威分值。因为我们的方法主要是用到Warshall算法，下面首先介绍图论中的Warshall算法。

2.1 Warshall算法

Warshall算法是用来判定有向图G中的两节点间是否存在道路，或者判定它是否连通。

设$A=(a_{ij})^{n\times n}$是G的邻接矩阵。由A的定义，$a_{ij}=1$表示$(v_i,v_j)\in E(G)$，即v_i可以通过某条边到达v_j，或者说G中有道路从v_i到达v_j，根据矩阵乘法，设$A2=(a_{ij}^{(2)})$，有

$$a_{ij}^{(2)}=\sum_{k=1}^{n}a_{ik}\cdot a_{kj}$$

$a_{ij}^{(2)}\neq 0$当且仅当存在k，使$a_{ik}=a_{kj}=1$，也就是说，如果G中存在节点v_k，满足$(v_i,v_k),(v_k,v_j)\in E(G)$，即经过两条边$(v_i,v_k),(v_k,v_j)$，$v_i$可以到达$v_j$时，$a_{ij}^{(2)}\neq 0$。同理，$A^l((l\leqslant n)$中的元素$a_{ij}^{l}\neq 0$表示了$v_i$可以经过l条边到达$v_j$。因此令

$$P=A+A^2+\cdots+A^n$$

如果$P_{ij}=t$，说明v_i有t条道路可以到达v_j。若$P_{ij}=0$，即n步之内v_i不能到达v_j则在G中不存在v_i到达v_j的道路。否则，若v_i经过$l(l\geqslant n)$步可达v_j，由抽屉原理，该道路上一定存在重复出现的节点v_k，而v_k之间的这段路一定是一个回路，去除这段回路，v_i仍然可以到达v_j。由于G中存在n个不同的节点，因此只要v_i有道路到v_j，则一定有$P_{ij}=0$。

在许多实际问题中，往往只要求了解v_i与v_j之间是否存在道路。对此可以采用逻辑运算的方法，即

$$a_{ij}^{(l)}=\vee_{k=1}^{n}(a_{ik}^{l-1}\wedge a_{kj}),\ l=2,3,\cdots,n$$

相应地

$$P=A\vee A^2\vee\cdots\vee A^n$$

就是图G的道路矩阵。

但是用上述方法求G的道路矩阵，计算复杂性为$O(n^4)$。以下介绍的Warshall算法是一个更好的方法，其计算复杂性是$O(n^3)$。

Warshall算法

```
Begin
P←A,
for i=1 to n do
    for j=1 to n do
        for k=1 to n do
            p_jk ← p_jk ∨ (p_ji ∧ p_ik)
End
```

2.2 定义

定义1 设图$G=(V,E)$是超链接图，$V=\{p_1,p_2,\cdots,p_n\}G$，这里的$n>1$，设P为超链接图的可行性矩阵；矩阵P中的每一项(i,j)要么取值1要么为0，如果从P_i到P_i存在有向链接，则为1。否则为0。

现在我们可以生成矩阵H，定义矩阵H为多重连续链接可行性矩阵，H中每一项(i,j)表示从节点P_i到P_j的可行性。

定义2 多重连续链接可行性矩阵：$H=\sum_{m=1}^{n}p^m$

这里H计算的复杂度非常重要。如果按照传统的矩阵相乘的方法，那么它的时间复杂度将是我们上面说的$O(n^4)$。但是，在这我们的主要目的是要获得两节点之间是否存在着有向路经，所以我们可以

应用上面的 Warshall 算法来计算获得多重连续可行性矩阵 H，复杂度将减少到 $O(n^3)$。

我们改进的整个算法发现权威分值的过程如下：

首先由一初始集构建一连通的超链接图 G=(V,E)，其中的 $V=\{p_1,p_2,\cdots,p_n\}$，这里的 n>1。

然后获得超链接可行性矩阵 P，然后应用 Warshall 算法计算多重连续链接可行性矩阵 H。对每一个 V 中的节点 P_i，设 $a[i]$为它的权威值，$h[i]$为它的 hub 分值，V 中所有节点的 $a[i]$和 $h[i]$初始值为 1。反复计算如下向量 a 和 h 直至收敛：

{ 对 V 中所有节点，$a[i]:=\sum_{j=1}^{n} H_{j,i}h[j]$；

对 V 中所有节点，$h[i]:=\sum_{j=1}^{n} H_{i,j}a[J]$；

最后规范化向量 a 和 h

}

3 主要的理论依据

我们下一步要说明我们新方法在计算权威分值和 hub 分值的理论可行性和结果唯一性。

定义3 超链接图 G 的权重分布定义为一有序向量对(a^*,h^*)，这里的 a^* 是一个 n 维的向量元素为非负的列向量(又叫做权威权值向量)，h^* 是另一个向量元素为非负的 n 维向量(又叫 hub 权值向量)，且有 $a^*=H^Th^*/\|H^Th^*\|$ 和 $h^*=Ha^*/\|Ha^*\|$。

提示：$\|\cdot\|$是向量范数，(a^*,h^*)对实际上是权威分值和 hub 分值所期望的平衡值，以上的两个等式是权威分值和 hub 分值得以相互加强的基本方法。

如果(a^*,h^*)是一个超链接图 G 的权重分布，那么 a^* 就是对应于矩阵 H^TH 的一个非负特征值的一个特征向量，因为对一个连通的超链接图都有一个确定的矩阵 H^TH，应用矩阵理论原理很容易得知矩阵 H^TH 的最大特征值也是很简单的，它的一个特征向量正是我们要找的 a^*。

主要理论：任意一个连通的超链接图都有一个唯一的超链接权重分布，我们算法的收敛性将以 $a\to a^*$ 和 $h\to h^*$ 得到保证。

4 例子和结论

我们用几个 Web 试验例子对我们新的方法和 Clever Algorithm 进行了比较，发现我们新的算法得出了令人信服的结果，这里我们仅仅给出两个简单的例子，这里的第二个例子是从一个实际的 Web 搜索结果获得的。

例1 以前面给出的四个节点图 P1←P2→P3→P4为例。

	权威分值	Hub 分值
Clever algorithm	(1/2,01/2,0)	(0,1,0,0)
新算法	(1/4,0,1/4,1/2)	(0,1/2,1/2,0)

例2 另外一个4节点的图

	权威分值	Hub 分值
Clever algorithm	(0,0.3820, 0.6180,0)	(0.6180,0.3820, 0,0)
新算法	(0.0.1706,0.4737, 0.3558)	(0.4317,0.3676, 0.2007,0)

我们算法的改进之处在于不仅考虑到了超链接多重连续链接，而且在计算权威分值的时候我们应用 Warshall 算法较好地解决了算法的执行速度问题。

参 考 文 献

1 Kleinberg J. Authoritative Sources in a Hyperlinked Environment. In: Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998, extended version in Journal of the ACM 46(1999), also appears as IBM Research Report RJ 10076, May 1997

2 Kleinberg J, Kumar S R, Raghavan P, Rajagopalan S, Tomkins A. The Web as a Graph: measurement, Models and Methods. Invited survey at the Intl. Conf. on Combinatorics and Computing, 1999

3 Bharat K, Henzinger M. Improved Algorithms for Topic Distillation in Hyperlinked Environments. In: Proc. 21st SIGIR, 1998

4 Herbach J D. Improving authoritative sources in a hyperlinked environment via similarity weighting. 2001

5 Wang Minhua. A Significant Improvement to Clever Algorithm. In: Hyperlinked Environment, 2003

6 Borges J, Levene M. Data Mining of User Navigation Patterns, in Web Usage Analysis and User Profiling. Published by Springer-Verlag as Lecture Notes in Computer Science, Vol. 1836, 2000. 92～111

7 Horn R, Johnson C. Matrix Analysis. Cambridge University Press, 1990

8 Levene M, Borges J, Loizou G. Zipf's Law for Web Surfers. Knowledge and Information Systems an Intl. Journal, 2001, 3