基于 HITS 算法的期刊评价研究*

苏 成 潘云涛 袁军鹏 马 峥 郭 红 张玉华 俞征鹿 胡志宇
(中国科学技术信息研究所情报方法研究中心,100038, 北京)

摘 要 构建了适用于期刊引用网络的 HITS(hypertext induced topic search)算法,利用 2006 年的中国科技论文与引文数据库(CSTPCD)分别计算所有及化学类统计源期刊的权威(authority)值与中心(hub)值,并与影响因子进行了对比研究,讨论了 HITS 算法用于期刊评价的优缺点以及适用范围。

关键词 HITS; 影响因子; 期刊评价

HITS for journal ranking//SU Cheng, PAN Yuntao, YUAN Junpeng, MA Zheng, GUO Hong, ZHANG Yuhua, YU Zhenglu, HU Zhiyu

Abstract We tried to adjust the popular webpage algorithm hypertext induced topic search (HITS) to rank journal, and with HITS compute respectively the authority score and hub score of all journals and chemistry journals. We also compared ISI impact factor and HITS for journal ranking, and discussed the merits and demerits of the ranking journal HITS algorithm and the scope of its applicability.

Key words HITS; impact factor; journal ranking

Author's address Institute of Scientific and Technical Information of China, 100038, Beijing, China

期刊评价方法很多,其中引文分析是被广泛使用的一种。1963 年 Garfield 提出影响因子的概念[1],并每年出版期刊引证报告[2]。因为影响因子容易理解并易于获得,所以成为了评价期刊最流行的工具之一。但是影响因子有其内在不足[3-5],我们姑且不论它选取 2 年是否合理,但它把所有的引用视为等同,不管这次引用是来自一篇高质量文章还是一篇低水平文章,其不合理是不言而喻的。

随着网络的兴起,网络信息量的不断膨胀,人们对于快捷准确地获取自己需要的信息提出了更高的要求。如何把用户需求的信息按重要性、相关性的大小排序提供给用户是搜索引擎重点考虑的方向,有不少研究人员提出了网页排序的很多算法,其中很有名的是 Kleinberg[6] 提出的 hypertext induced topic search (HITS) 算法。HITS 算法利用网页的入链和出链(类似论文间的被引与引用)产生 2 个分数,即权威(authority)值和中心(hub)值,它是建立在一个相互依存的循环的假设之上的,认为一个指向好的 authority 值的网页是个好的 hub,而一个被好的 hub 指向的网页是个好的 authority 的网页[6]。HITS 算法与影响因子最大不同之处在于它不仅是考虑入链网页的多少,还考虑入链的网页的重要性。

尽管说期刊引用网络与网页链接网络存在不少差异,但是期刊引用网络与网页链接网络均可看作是个有向图,有向图的一个节点代表一个期刊或网页,节点间的连线代表期刊的引用关系或者网页的链接关系。因此,利用 HITS 这种源自网页排序的算法来进行期刊排序从理论上讲是完全可行的。HITS 算法不但考虑了期刊被引用的次数,还区分了引用期刊的重要性,与影响因子单纯考虑期刊被引用次数相比无疑更合理些。

本研究使用的数据取自 2006 年中国科技论文与引文数据库(CSTPCD),包含 1723 种中国出版的中英文科技期刊。我们计算了全学科统计源期刊以及化学类期刊的 authority 值和 hub 值,并与影响因子进行了对比研究,讨论了 HITS 算法的优缺点以及适用范围。

1 研究方法

1.1 期刊引用网络与网页链接网络 期刊引用网络与网页链接网络本质上相同,但二者存在不少差异,最主要的有 2 个:一个是期刊引用网络中一个期刊可以引用另外一个期刊很多次,而一个网页只能链接另外一个网页 1 次;另一个是在期刊引用网络中期刊绝大多数是存在自引的,而且自引率还不低,在 2006 年 CSTPCD 的数据中,1723 种期刊的平均自引率是 19.89%[7]。因此,网页链接网络矩阵和期刊引用网络矩阵在表现形式上会有一些不同。如果用矩阵表示的话,HITS 算法的网页链接网络矩阵 $\boldsymbol{L}$[6] 可以表示为:

$$L_{ij}=\begin{cases}1, \text{如果网页 } i \text{ 链接了网页 } j,\\ 0, \text{网页 } i \text{ 未链接网页 } j。\end{cases} \quad (1)$$

而 HITS 算法期刊引用网络矩阵 $\boldsymbol{L}$ 可以表示为:

$$L_{ij}=\begin{cases}m, \text{如果期刊 } i \text{ 引用了期刊 } j \text{ 共 } m \text{ 次},\\ 0, \text{期刊 } i \text{ 未引用期刊 } j。\end{cases} \quad (2)$$

在式(1)中,因为网页自己不能链接自己,所以 $L_{ii}=0$,而在式(2)中,$L_{ii}=m$, 其中 m 为期刊 i 的自引次数。

1.2 适用于期刊引用网络的 HITS 算法 Kleinberg 认为,搜索开始于用户的检索提问,每个页面的重要性也依赖于用户的检索提问,而 HITS 算法利用网页链接结构来决定网页的重要性,它利用网页的入链和出链产生 authority 值和 hub 值这 2 个分数。HITS 算法

* 国家科技支撑计划资助项目(2006BAH03B05);国家自然科学基金资助项目(70673019);中国科技信息所科研项目预研基金资助项目

的目标就是通过一定的计算(迭代计算)方法以得到针对某个检索提问的最具价值的网页,即排名最高的 authority 值。它可以用以下公式[6]表示:

$$x_i^{(k)} = \sum_{j:e_{ji}\in E} y_j^{(k-1)},$$
$$y_i^{(k)} = \sum_{j:e_{ji}\in E} x_j^{(k)}, \quad k = 1, 2, \cdots。 \tag{3}$$

式(3)可以用矩阵形式改写[8]为:

$$x^{(k)} = \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}x^{(k-1)}, \quad y^{(k)} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}Ly^{(k-1)}。 \tag{4}$$

式中 $\boldsymbol{L}$ 为式(1)表示的矩阵,$x^{(k)}$ 和 $y^{(k)}$ 表示在 k 次迭代计算得出的 authority 值与 hub 值。式(4)是针对网页链接网络而言的,矩阵 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ 决定了网页的 authority 值,矩阵 $\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ 决定了网页的 hub 值,但对于期刊引用网络来说,式(4)中 $\boldsymbol{L}$ 不是式(1)所表示的矩阵,而是式(2)所表示的矩阵。根据式(4)算出的是某期刊的 authority 值和 hub 值实际上是这个期刊所有论文的这 2 个值之和,这类似于文献计量学中期刊总被引频次概念。在 2006 年的 CSTPCD 中,期刊载文量的差异很大,载文量最多的期刊 4 566 篇,最少的仅 16 篇[7],因此我们在考察期刊权威性时还应考虑期刊的载文量大小,借鉴影响因子思想,如果某期刊发表论文的平均的权威性较高,可以认为此期刊的权威性较高。那么期刊 J 在 n 年的 authority 值和 hub 值可以定义为:

$$H_{\mathrm{ITS}xn}(J) = \frac{x_n^{(k)}}{p_{n-1}+p_{n-2}}, H_{\mathrm{ITS}yn}(J) = \frac{y_n^{(k)}}{p_{n-1}+p_{n-2}}。 \tag{5}$$

式中:$x_n^{(k)}$ 与 是利用式(4)算出的值;$p_{n-1}+p_{n-2}$ 为期刊 J 在 $n-1$ 和 $n-2$ 年发表的论文总数。

1.3　期刊 authority 值和 hub 值算法计算步骤

1)构建期刊引用网络矩阵。为讨论 HITS 算法在不同环境下的应用,我们分别构建了所有统计源期刊和化学类期刊的引用网络矩阵。所有统计源期刊引用网络矩阵的构建过程为:选取 2006 年 CSTPCD 中 1 723种的期刊,根据式(2)构建一个 1723 × 1723 的矩阵,矩阵的行与列均表示期刊,其中的值代表这行的期刊引用这列的期刊的次数,主对角线代表期刊自引次数。化学类期刊引用网络的构建过程为:选取 2006 年 CSTPCD 中所有化学类期刊以及引用过这些期刊的其他相关领域的期刊,再根据式(2)构建矩阵,矩阵的行与列均表示期刊,其中的值代表这行的期刊引用这列的期刊的次数,主对角线代表期刊自引次数。

2)迭代计算。构建好期刊引用网络矩阵后,采用式(4),利用幂法运算求矩阵 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ 与 $\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ 的最大特征值,利用式(5)即可得期刊的 $H_{\mathrm{ITS}xn}(J)$ 和 $H_{\mathrm{ITS}yn}(J)$ 值。计算矩阵最大特征值时用 Matlab 软件,收敛值取 10^{-8}。

2　结果——HITS 与影响因子对比研究

我们利用得出的 1 723 种期刊的 $H_{\mathrm{ITS}xn}(J)$ 值与期刊的影响因子,考察了 $H_{\mathrm{ITS}xn}(J)$ 值与期刊的影响因子 Pearson 线性相关系数,经过 SPSS 统计分析软件分析,两者间的 Pearson 线性相关系数为 0.230 5,这说明这两者间相关性较小。也就是说,影响因子高并不代表其 $H_{\mathrm{ITS}xn}(J)$ 值就高,因为如果大多数引用都是不重要的引用的话,其影响因子可能会高,但其 $H_{\mathrm{ITS}xn}(J)$ 值不会高;而 $H_{\mathrm{ITS}xn}(J)$ 值高也不代表其影响因子就一定高,因为如果被引用次数不多但都是重要的引用的话其 $H_{\mathrm{ITS}xn}(J)$ 值也会很高,但影响因子会很低。

在影响因子算法中,每次被引用是同等看待的,被引用次数越多,其影响因子越高,并不考虑不同引用间的重要性差别,因此可以认为影响因子是一种流行性的测度。与影响因子不同的是 $H_{\mathrm{ITS}xn}(J)$ 不但考虑了被引用的次数,还对不同引用间的重要性差别进行区分,因此可以认为其是对权威性的一种测度。hub 值高表明这期刊引用的期刊具有较高的权威性,这在一定程度上反映了期刊利用外部资源能力的大小。

表 1 显示的是 2006 年 CSTPCD 中影响因子、HITS 算法得出的 $H_{\mathrm{ITS}xn}(J)$ 值与 $H_{\mathrm{ITS}yn}(J)$ 值的前 10 位期刊。从表 1 明显可见,采用不同算法得出的排序结果有很大变化。只有《电网技术》和《中国电机工程学报》出现在影响因子和 $H_{\mathrm{ITS}xn}(J)$ 这 2 个数据前 10 位中,而 $H_{\mathrm{ITS}xn}(J)$ 与 $H_{\mathrm{ITS}yn}(J)$ 数据的前 10 位中有 7 刊是 2 个数据共同的。从表 1 中的 $H_{\mathrm{ITS}xn}(J)$ 与 $H_{\mathrm{ITS}yn}(J)$ 数据可见,排前 10 位的基本属于动力与电力工程类期刊,这是因为 HITS 算法是基于查询主题的一种算法,如前所述,其计算为一个循环过程,存在一个主题不断加强的过程。在动力与电力工程类期刊中有《电网技术》和《中国电机工程学报》这 2 刊被引频次较高,而与之发生引用关系的大部分为同领域期刊,从而导致其权重大部分向同领域期刊传递,并在传递过程中得到不断加强,使得同领域中其他期刊的排位靠前。因此,HITS 算法用于期刊评价要特别注意学科差异,它不是很适合全学科领域期刊评价。

我们 CSTPCD 中 2006 年所有化学类期刊以及引用过这些化学类期刊的其他领域期刊,根据式(2)构建了化学类期刊矩阵,利用式(5)计算了化学类期刊的 $H_{ITSxn}(J)$ 值与 $H_{ITSyn}(J)$。我们利用得出的 34 种化学类期刊的 $H_{ITSxn}(J)$ 值与期刊的影响因子,考察 $H_{ITSxn}(J)$ 值与期刊的影响因子 Pearson 线性相关系数,经过 SPSS 统计分析软件分析,两者间的 Pearson 线性相关系数为 0.342 8,这说明这两者间相关性较小,但与所有期刊引用网络得出的结果相比,相关性要稍大些。其前 10 位的期刊见表 2。可以看出,化学类期刊中影响因子与 $H_{ITSxn}(J)$ 值的前 10 位列表中共有 7 个期刊是共有的,其排序有较大的不同。

表 1 2006 年 CSTPCD 影响因子、$H_{ITSxn}(J)$ 与 $H_{ITSyn}(J)$ 前 10 位的期刊

排名	影响因子		$H_{ITSyxn}(J)$		$H_{ITSyn}(J)$	
	数值	刊 名	数值 $\times 10^5$	刊 名	数值 $\times 10^5$	刊 名
1	2.857	电网技术	37.83	电网技术	38.25	电网技术
2	2.649	岩石学报	36.39	中国电机工程学报	28.78	中国电机工程学报
3	2.587	实验技术与管理	6.03	电力系统自动化	5.14	电力系统自动化
4	2.537	中国电机工程学报	2.69	高电压技术	4.42	高电压技术
5	2.455	中国沙漠	2.13	电工技术学报	3.84	电工技术学报
6	2.444	中国公路学报	2.05	电工电能新技术	3.52	电力系统及其自动化学报
7	2.331	PEDOSPHERE	1.47	动力工程	3.4	继电器
8	2.326	地质学报	1.36	中国电力	3.38	现代电力
9	2.302	地理学报	1.23	继电器	3.19	华北电力大学学报
10	2.212	地质科学	1.14	电力系统及其自动化学报	3.12	电力自动化设备

表 2 2006 年 CSTPCD 中化学类期刊影响因子、$H_{ITSxn}(J)$ 值与 $H_{ITSyn}(J)$ 值前 10 位的期刊

排名	影响因子		$H_{ITSxn}(J)$		$H_{ITSyn}(J)$	
	数值	刊 名	数值 $\times 10^6$	刊 名	数值 $\times 10^6$	刊 名
1	1.106	燃料化学学报	55.861	燃料化学学报	1.792	燃料化学学报
2	0.978	化学进展	0.817	高分子学报	1.023	高分子通报
3	0.977	催化学报	0.613	煤炭转化	0.241	化学学报
4	0.968	化学学报	0.583	化学学报	0.121	高等学校化学学报
5	0.96	分析化学	0.558	化学进展	0.061	分析试验室
6	0.957	物理化学学报	0.515	物理化学学报	0.052	煤炭转化
7	0.925	分子科学学报	0.501	高等学校化学学报	0.035	物理化学学报
8	0.865	高分子学报	0.497	中国科学 B	0.031	化学进展
9	0.851	无机化学学报	0.46	应用化学	0.021	催化学报
10	0.785	高等学校化学学报	0.424	催化学报	0.019	分子科学学报

图 1 中是化学类期刊影响因子与其 authority 值的散点图(因《燃料化学学报》的 Authority 值比其他期刊的值大太多,放在图中不利于观察其他期刊的情况,所以删除了该刊)。上方大椭圆中的期刊表示影响因子和 authority 值均较大,这表明这些期刊不但流行性高,权威性也高,它们的代表是《燃料化学学报》《高分子学报》等。下方小椭圆中的期刊表示影响因子较大而 authority 值较小,这表明这些期刊流行性高,权威性低,比较典型的有《分子科学学报》等。

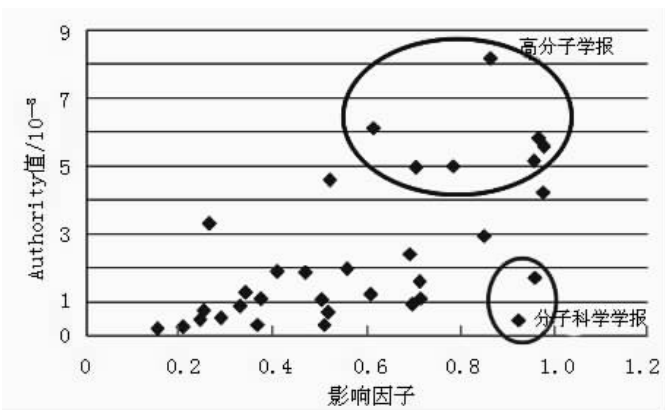

图 1 化学类期刊影响因子与 Authority 值散点图

3 讨论与结论

影响因子作为一个流行的期刊评价工具,有易理解和结果易获取的优点,但其只考虑被引数量而不区分不同被引重要性的算法不是很科学合理,它反映的是一种期刊的流行程度,还不能完全反映一种期刊的权威性。HITS 算法正好弥补了影响因子评价的弊端,它综合考虑了引用次数和引用质量,更能客观地反映期刊权威性。通过考察期刊影响因子与 authority 值的不同关系,可以反映期刊的流行度和权威度。

HITS 算法带来了 2 个值,其中 authority 值反映的是期刊的权威性,hub 值一定程度上反映的是期刊利用外部资源能力的大小。

HITS 算法最初设计是基于查询主题的网页排序用的,所以,在用于期刊引用网络时还要进行适当的调整,并注意其适用范。从我们的研究来看,用 HITS 算法进行期刊评价,要特别注意学科差异,它比较适合分学科的期刊排序,而并不很适合全学科的期刊排序。

4 参考文献

[1] Glanzel W. Bibliometrics as a reseatch field [EB/OL]. [2008-11-15]. http://www.norslis.net/2004/Bib_Module_KUL.pdf