

基于 PageRank 算法和 HITS 算法的期刊评价研究

喻 依 甘若迅 樊锁海 刘 庆 邵 晴
(暨南大学信息科学技术学院 广州 510632)

摘 要 由 31 种计算机技术类核心期刊形成期刊引文网络,利用 PageRank 算法和 HITS 算法分别计算其 PageRank 值、权威值(authority)和中心值(hub)。通过与影响因子的对比研究,讨论了 PageRank 值、权威值、中心值和影响因子用于期刊评价的区别和优缺点。实例结果表明期刊影响因子的排名与 PageRank 算法、HITS 算法的结果排名略有差距,说明不同的评价算法反映出不同的期刊评价目的。

关键词 PageRank 算法, HITS 算法, 影响因子, 期刊评价
中图法分类号 TP311 文献标识码 A

Journal Evaluation Based on PageRank Algorithm and HITS Algorithm

YU Yi GAN Ruo-xun FAN Suo-hai LIU Qing SHAO Qing

(School of Information Science and Technology, Jinan University, Guangzhou 510632, China)

Abstract Journal citation network is formed with 31 kinds of computer technical core journal in this paper. The PageRank value, authoritative value(authority) and the central value(hub) of these journals are calculated using the algorithms of PageRank and HITS. The comparative study for advantages and disadvantages among impact factor with PageRank value, authoritative value and center value was discussed. The results show that journal impact factor ranking is slightly different from PageRank algorithm, HITS algorithm results ranking, indicating that different evaluation algorithms reflect different journal evaluation purposes.

Keywords PageRank algorithm, HITS algorithm, Impact factor, Journal evaluation

1 引言

评价期刊的方法有很多,其中引文分析是评价期刊质量的一种常用方法。而在引文分析及期刊评价中引入影响因子的概念有重要意义,通常影响因子定义为:该刊前两年发表论文在统计当年被引用的总次数除以该刊前两年发表论文总数^[1]。影响因子容易被理解且易获得,因此成为评价期刊重要性的最常用的指标。但是,影响因子存在固有的局限性^[2-4],特别是它把一篇论文的所有的引用视为等同,没有考虑到期刊类型的不同、学科之间的差别、自我引用的概率等诸多因素的影响。

随着网络的发展,网络信息量的不断增加,如何把用户所需求的信息按相关性、重要性的大小排序提供给用户是搜索引擎重点考虑的方向,目前提出的网页排序的算法很多,其中 Google 的 PageRank 和 Kleinberg 提出的 HITS 算法已成为搜索引擎领域中最经典的两个算法。PageRank 借鉴了文献计量学中引文分析的基本思想,即它认为一个超链接就是一个推荐,类似引文分析中的一次引用,一个拥有更多超链接的网页应该比一个更少超链接的网页更重要。但是,与影响因子不同的是,PageRank 不仅考虑入链网页的数量,还考虑

入链的网页的质量。PageRank 算法基于这样的假设:如果一个网页被一个重要的网页链接,那么它也应该是一个重要的网页^[5]。而 HITS 算法的基本思想是:每个网页的重要性有两种刻画指标即权威性(authority)和中心性(hub),它是建立在一个相互依存的循环的假设之上的,认为一个网页被多个具有高 hub 值的页面所指向,那么该页面就有较高的 authority 值,一个网页指向多个具有高 authority 值的页面,那么该页面就具有高的 hub 值。因此,PageRank 算法综合考虑入链网页的数量和质量,从而对网页排序,而 HITS 算法不仅考虑入链网页的多少,还考虑入链的网页的重要性。

期刊引用与网页链接之间具有相似性,期刊引用网络与网页链接网络均可视为有向图,图中一个节点代表一个期刊或网页,节点间的连线代表期刊之间的引用关系或者网页间的链接关系。因此,利用 PageRank 和 HITS 算法来评价期刊的质量从理论上是可行的。

2 期刊互引网络的构建

本文研究使用的数据取自“中国引文数据库”,期刊选自北大图书馆 2008 年公布的核心期刊目录中的 31 种计算机技术类种期刊,期刊的出版时间选取为 2007 年—2011 年。期

本文受国家自然科学基金(10671076),广东省自然科学基金(10151063201000005)资助。

喻 依(1990—),女,硕士生,主要研究方向为系统优化;甘若迅(1988—),男,硕士生,主要研究方向为智能优化;樊锁海(1967—),男,教授,主要研究方向为复杂网络、智能计算,E-mail:tfsh@jnu.edu.cn(通信作者);刘 庆(1989—),女,硕士生,主要研究方向为应用统计;邵 晴(1989—),女,硕士生,主要研究方向为应用统计。

刊互引网络与网页链接网络在本质上是相同的,但是两者还是存在一些差异:1)期刊互引网络中期刊之间可以相互引用多次,而网页链接网络中网页间只能相互链接一次;2)期刊互引网络中存在自引现象。为了使期刊互引网络的计算与网页链接网络尽可能相似,本文将排除₃₁种期刊自引的情况,于

是期刊引用网络矩阵 $L=(l_{ij})_{n \times n}$ 可以表示为:

$$l_{ij} = \begin{cases} m, & \text{期刊 } i \text{ 引用了期刊 } j \text{ 共 } m \text{ 次} \\ 0, & \text{期刊 } i \text{ 未引用期刊 } j \end{cases}$$

由于期刊排除了自引,因此 $l_{ii}=0$ 。由此,可以得到期刊互引网络矩阵,如表₁所列。期刊的名称与编号如表₂所列。

表₁ 期刊互引矩阵(排除自引)

| | | 引用期刊 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-------|----|------|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|------|----|----|-----|-----|-----|-----|-----|-----|-----|----|-----|------|----|-----|-----|-----|----|
| 编号 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
| 被引用期刊 | 1 | 0 | 299 | 475 | 197 | 64 | 134 | 1381 | 232 | 987 | 163 | 89 | 528 | 22 | 1013 | 9 | 45 | 733 | 37 | 960 | 79 | 92 | 178 | 274 | 5 | 95 | 944 | 8 | 153 | 15 | 228 | 14 |
| | 2 | 326 | 0 | 312 | 219 | 56 | 167 | 1239 | 223 | 775 | 109 | 128 | 475 | 35 | 805 | 8 | 51 | 602 | 31 | 781 | 36 | 103 | 155 | 215 | 4 | 26 | 669 | 8 | 35 | 15 | 168 | 6 |
| | 3 | 184 | 124 | 0 | 96 | 33 | 50 | 847 | 156 | 536 | 76 | 57 | 285 | 13 | 544 | 3 | 51 | 443 | 33 | 580 | 45 | 65 | 131 | 165 | 3 | 29 | 555 | 7 | 49 | 11 | 119 | 8 |
| | 4 | 46 | 31 | 56 | 0 | 20 | 216 | 487 | 174 | 172 | 212 | 4 | 74 | 2 | 243 | 1 | 0 | 210 | 7 | 151 | 19 | 21 | 89 | 74 | 2 | 4 | 305 | 4 | 9 | 7 | 57 | 0 |
| | 5 | 12 | 14 | 19 | 25 | 0 | 43 | 312 | 217 | 98 | 26 | 288 | 35 | 245 | 204 | 39 | 4 | 98 | 94 | 71 | 47 | 37 | 118 | 25 | 4 | 25 | 111 | 17 | 21 | 159 | 37 | 40 |
| | 6 | 9 | 13 | 22 | 116 | 32 | 0 | 687 | 111 | 291 | 11 | 7 | 67 | 3 | 320 | 6 | 1 | 302 | 13 | 155 | 51 | 34 | 180 | 82 | 65 | 31 | 318 | 14 | 18 | 7 | 76 | 1 |
| | 7 | 27 | 15 | 39 | 43 | 22 | 115 | 0 | 206 | 479 | 101 | 55 | 164 | 30 | 904 | 10 | 53 | 758 | 27 | 392 | 171 | 40 | 395 | 282 | 19 | 57 | 1699 | 51 | 72 | 45 | 263 | 23 |
| | 8 | 16 | 16 | 27 | 84 | 38 | 42 | 578 | 0 | 288 | 183 | 114 | 70 | 72 | 337 | 25 | 4 | 233 | 54 | 140 | 188 | 15 | 813 | 71 | 7 | 42 | 418 | 68 | 45 | 139 | 55 | 74 |
| | 9 | 24 | 21 | 33 | 35 | 20 | 53 | 948 | 149 | 0 | 72 | 22 | 155 | 9 | 767 | 5 | 29 | 706 | 20 | 341 | 185 | 23 | 262 | 264 | 12 | 46 | 1847 | 47 | 45 | 24 | 282 | 19 |
| | 10 | 2 | 10 | 4 | 81 | 12 | 2 | 213 | 124 | 119 | 0 | 81 | 31 | 6 | 208 | 3 | 0 | 71 | 20 | 46 | 8 | 3 | 48 | 27 | 0 | 2 | 149 | 3 | 0 | 9 | 12 | 3 |
| | 11 | 6 | 4 | 10 | 15 | 105 | 7 | 417 | 331 | 115 | 98 | 0 | 41 | 254 | 229 | 27 | 1 | 138 | 141 | 85 | 75 | 41 | 208 | 31 | 1 | 43 | 124 | 23 | 23 | 759 | 34 | 68 |
| | 12 | 17 | 24 | 30 | 16 | 9 | 21 | 422 | 65 | 290 | 36 | 35 | 0 | 5 | 249 | 3 | 15 | 206 | 15 | 203 | 32 | 14 | 82 | 99 | 3 | 7 | 448 | 8 | 15 | 11 | 95 | 7 |
| | 13 | 0 | 3 | 0 | 1 | 96 | 4 | 181 | 194 | 35 | 23 | 295 | 12 | 0 | 92 | 20 | 0 | 29 | 105 | 30 | 56 | 13 | 106 | 14 | 0 | 17 | 44 | 12 | 6 | 177 | 11 | 52 |
| | 14 | 15 | 5 | 15 | 16 | 6 | 27 | 509 | 88 | 371 | 36 | 19 | 68 | 3 | 0 | 5 | 15 | 348 | 13 | 181 | 94 | 8 | 125 | 136 | 16 | 35 | 911 | 23 | 35 | 10 | 114 | 6 |
| | 15 | 3 | 0 | 6 | 2 | 23 | 0 | 122 | 109 | 30 | 6 | 32 | 9 | 35 | 55 | 0 | 0 | 23 | 13 | 22 | 37 | 12 | 64 | 4 | 0 | 65 | 49 | 21 | 22 | 55 | 7 | 0 |
| | 16 | 40 | 8 | 25 | 3 | 16 | 2 | 247 | 0 | 109 | 0 | 3 | 24 | 0 | 135 | 0 | 0 | 100 | 2 | 106 | 0 | 22 | 13 | 48 | 0 | 0 | 110 | 0 | 0 | 1 | 18 | 0 |
| | 17 | 14 | 17 | 14 | 25 | 9 | 36 | 692 | 75 | 523 | 27 | 15 | 70 | 4 | 401 | 2 | 28 | 0 | 11 | 170 | 83 | 13 | 173 | 150 | 8 | 33 | 967 | 16 | 35 | 14 | 147 | 12 |
| | 18 | 4 | 1 | 3 | 5 | 25 | 5 | 138 | 86 | 58 | 32 | 85 | 18 | 51 | 63 | 3 | 0 | 42 | 0 | 28 | 34 | 4 | 66 | 11 | 1 | 17 | 56 | 12 | 22 | 84 | 14 | 44 |
| | 19 | 27 | 31 | 42 | 19 | 9 | 19 | 590 | 56 | 366 | 33 | 22 | 144 | 6 | 420 | 1 | 16 | 298 | 15 | 0 | 34 | 28 | 81 | 122 | 2 | 29 | 572 | 8 | 29 | 9 | 110 | 4 |
| | 20 | 2 | 0 | 3 | 0 | 1 | 3 | 39 | 33 | 35 | 3 | 7 | 3 | 3 | 24 | 1 | 0 | 33 | 6 | 6 | 0 | 0 | 31 | 10 | 2 | 47 | 97 | 92 | 33 | 24 | 19 | 22 |
| | 21 | 6 | 2 | 6 | 10 | 13 | 16 | 120 | 10 | 45 | 2 | 16 | 10 | 7 | 61 | 0 | 6 | 40 | 9 | 48 | 5 | 0 | 19 | 12 | 1 | 3 | 38 | 1 | 1 | 0 | 10 | 2 |
| | 22 | 4 | 4 | 7 | 13 | 5 | 19 | 253 | 493 | 98 | 15 | 11 | 18 | 8 | 110 | 8 | 1 | 138 | 13 | 47 | 104 | 1 | 0 | 35 | 6 | 27 | 264 | 49 | 17 | 42 | 53 | 25 |
| | 23 | 14 | 12 | 16 | 4 | 0 | 8 | 175 | 23 | 121 | 8 | 3 | 28 | 3 | 100 | 0 | 4 | 92 | 3 | 56 | 19 | 4 | 30 | 0 | 4 | 2 | 214 | 6 | 6 | 2 | 37 | 2 |
| | 24 | 0 | 0 | 0 | 1 | 3 | 11 | 33 | 5 | 14 | 0 | 0 | 1 | 0 | 13 | 0 | 0 | 8 | 1 | 6 | 2 | 0 | 5 | 5 | 0 | 0 | 6 | 2 | 2 | 0 | 0 | 0 |
| | 25 | 0 | 1 | 1 | 0 | 0 | 1 | 34 | 7 | 15 | 0 | 10 | 5 | 0 | 13 | 3 | 0 | 23 | 2 | 9 | 66 | 0 | 16 | 4 | 0 | 0 | 20 | 26 | 248 | 7 | 6 | 14 |
| | 26 | 7 | 3 | 7 | 11 | 4 | 16 | 370 | 43 | 294 | 18 | 9 | 44 | 0 | 217 | 4 | 11 | 260 | 6 | 100 | 100 | 3 | 129 | 93 | 4 | 26 | 0 | 27 | 16 | 18 | 94 | 15 |
| | 27 | 0 | 0 | 0 | 2 | 3 | 2 | 27 | 25 | 22 | 3 | 0 | 3 | 1 | 22 | 4 | 0 | 16 | 1 | 6 | 192 | 0 | 35 | 9 | 2 | 36 | 55 | 0 | 41 | 14 | 9 | 9 |
| | 28 | 3 | 2 | 4 | 0 | 7 | 1 | 74 | 15 | 46 | 3 | 12 | 12 | 1 | 47 | 9 | 0 | 34 | 5 | 8 | 80 | 4 | 50 | 11 | 0 | 311 | 45 | 36 | 0 | 11 | 15 | 11 |
| | 29 | 0 | 1 | 0 | 0 | 0 | 1 | 54 | 65 | 13 | 13 | 17 | 3 | 23 | 25 | 10 | 0 | 19 | 11 | 4 | 45 | 0 | 46 | 6 | 0 | 10 | 31 | 12 | 3 | 0 | 4 | 33 |
| | 30 | 0 | 2 | 8 | 13 | 2 | 0 | 148 | 24 | 131 | 9 | 6 | 24 | 2 | 108 | 0 | 0 | 78 | 4 | 46 | 50 | 4 | 47 | 39 | 1 | 12 | 285 | 21 | 11 | 9 | 0 | 4 |
| | 31 | 0 | 0 | 1 | 0 | 2 | 1 | 22 | 17 | 11 | 5 | 8 | 0 | 7 | 9 | 0 | 0 | 4 | 5 | 2 | 21 | 0 | 19 | 2 | 0 | 13 | 12 | 11 | 10 | 26 | 4 | 0 |

表₂ 期刊名称与编号

| 编号 | 期刊名称 | 编号 | 期刊名称 | 编号 | 期刊名称 | 编号 | 期刊名称 |
|----|---------------|----|-----------|----|-----------|----|----------|
| 1 | 软件学报 | 9 | 计算机工程 | 17 | 计算机应用 | 25 | 传感器与微系统 |
| 2 | 计算机学报 | 10 | 计算机集成制造系统 | 18 | 信息与控制 | 26 | 计算机工程与设计 |
| 3 | 计算机研究与发展 | 11 | 控制与决策 | 19 | 计算机科学 | 27 | 测控技术 |
| 4 | 计算机辅助设计与图形学学报 | 12 | 小型微型计算机系统 | 20 | 计算机测量与控制 | 28 | 传感技术学报 |
| 5 | 自动化学报 | 13 | 控制理论与应用 | 21 | 模式识别与人工智能 | 29 | 控制工程 |
| 6 | 中国图象图形学报 | 14 | 计算机应用研究 | 22 | 计算机仿真 | 30 | 微电子学与计算机 |
| 7 | 计算机工程与应用 | 15 | 机器人 | 23 | 计算机工程与科学 | 31 | 化工自动化及仪表 |
| 8 | 系统仿真学报 | 16 | 中文信息学报 | 24 | 遥感技术与应用 | | |

3 期刊评价的原理及方法

3.1 PageRank 算法

PageRank 算法是由斯坦福大学的博士研究生 Sergey Brin 和 Lawrence Page 于 1998 年提出的,并成功应用在 Google 搜索引擎中。其基本思想是:当从网页 A 链接到网页 B 时,就认为“网页 A 投了网页 B 一票”,根据网页的得票数评定其重要性,并根据投票来源和投票目标的等级来决定新的等级,一个高等级的页面可以提升其他低等级页面的等级。一个页面的“得票数”由所有链向它的页面的重要性决定。一个有很多链入的页面会有很高的等级,相反如果一个页面没有任何链入页面,那么它就没有等级。

基于 PageRank 的基本思想,计算期刊的 PageRank 值包括以下几个步骤:

(1)对期刊互引网络矩阵 L 进行规范化,也就是用矩阵中的每个元素除以该列的所有元素的总和,即:

$$h_{ij} = \frac{l_{ij}}{\sum_k l_{ik}}$$

其中, l_{ij} 表示互引矩阵 L 第 i 行第 j 列的元素, h_{ij} 表示期刊 i 引用期刊 j 的次数占期刊 i 总引用次数的比例,由此可以得

到规范化矩阵 $H=(h_{ij})_{n \times n}$ 。

(2)计算文章向量,即每个期刊发表文章数占总发表文章数的比例的列向量,即:

$$a_i = \frac{\sum_j l_{ij}}{\sum_{i,j} l_{ij}}$$

其中, a_i 表示期刊 i 发表文章数占总发表文章数的比例。

(3)构建概率转移矩阵 P 。

$$P = \alpha H' + (1 - \alpha) a e^T$$

其中, e^T 是一个全为 1 的行向量, α 是阻尼系数,一般取 0.85,它表示研究者以 α 的概率跟随链接进入下一期刊,以 $1 - \alpha$ 的概率跳转到另外的期刊中。

(4)计算构建的概率转移矩阵 P 的主特征向量,即概率转移矩阵 P 最大特征值 1 所对应的标准化特征向量 π 。但是,对于期刊互引网络来说, $\pi(i)$ 计算出的是某一期刊所有论文的 PageRank 值之和,然而不同的期刊载文量的差别很大,因此还需要考虑期刊的载文量大小,那么 PageRank 值可以定义为:

$$PR(i) = \frac{\pi(i)}{a_i}$$

其中, $\pi(i)$ 表示期刊 i 所对应的主特征向量的值, a_i 表示期刊

i 发表文章数占总发表文章数的比例。

3.2 HITS 算法

HITS 算法是由 Kleinberg 于 1999 年提出的, 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 利用网页的入链和出链(类似于论文间的被引用和引用)产生权威值(authority)和中心值(hub)这两个分数。

HITS 算法最初是用于信息检索, 在其后的研究中, HITS 算法逐步被推广到期刊或者文章的重要性评价中。处于某个主题文献群中的单篇文献与网页类似, 承载了大量的知识信息内容, 文献之间的引用关系也与网页之间的链接关系原理相同, 且链接可能出于多种原因, 而引用关系则更严格地反应了内容之间的相关, 因此 HITS 算法同样适用于文献群的分析。利用 HITS 算法计算各篇文献的中心度和权威度, 权威度和中心度排序均靠前的文献集中反映了该主题研究的主要内容。

基于 HITS 算法的基本思想, 计算期刊的中心度和权威度, 包括以下几个步骤:

(1) 计算文章向量, 亦即每个期刊发表文章数占总发表文章数的比例的列向量, 即:

$$a_i = \frac{\sum_j l_{ij}}{\sum_i \sum_j l_{ij}}$$

其中, a_i 表示期刊 i 发表文章数占总发表文章数的比例。

(2) 用 hub_i 表示期刊 i 的中心度, 用 $auth_i$ 表示期刊 i 的权威度, 假设 $hub_i^0 = auth_i^0 = 1$, 对于期刊 i 有:

$$auth_i^k = \sum_{j: l_{ij} \neq 0} hub_j^{k-1}$$

$$hub_i^k = \sum_{j: l_{ij} \neq 0} auth_j^{k-1}$$

$$k=1, 2, \dots$$

其中, 上式可以用矩阵的形式改写为:

$$auth^k = L^T L \cdot auth^{k-1}$$

$$hub^k = L L^T \cdot hub^{k-1}$$

(3) 计算权威值和中心值。由于矩阵 $L^T L$ 决定了 $auth$ 的值, 矩阵 $L L^T$ 决定了 hub 的值, 故 $auth$ 对应于矩阵 $L^T L$ 最大特征值的特征向量, hub 对应于矩阵 $L L^T$ 最大特征值的特征向量。但是实际上 $auth_i$ 或 hub_i 计算的却是某期刊所有权威值或中心值之和, 由于不同的期刊的载文量差别很大, 因此与 PageRank 算法相同, 要考虑期刊载文量的大小, 那么期刊的权威值和中心值可以通过下面的公式计算:

$$H_{auth}(i) = \frac{auth_i}{a_i}$$

$$H_{hub}(i) = \frac{hub_i}{a_i}$$

4 实证分析

我们利用 31 种计算机技术类种期刊的互引矩阵计算出了期刊的 PageRank 值、authority 值和 hub 值, 具体结果如表 3 所列。由表 3 可知, 期刊影响因子的排名与 PageRank 算法排名相差不大, 而与 HITS 算法算出的结果排名相差很大。这主要是由于 hub 值考虑的是期刊所引用期刊的质量, 即期刊引用的高质量期刊越多, 其 hub 值就越高, 但一篇期刊若引用了一篇质量高的期刊, 并不能代表其质量也高, 且 authority 值又与 hub 值相关。因此 HITS 算法的排名与影响因子的排

名相差很大。由此可见不同的评价体系对期刊评价结果截然不同。

表 3 期刊评价结果与影响因子值

| 期刊名称 | PageRank | HITS | | 影响因子 |
|---------------|----------|-----------|---------|-------|
| | | authority | hub | |
| 软件学报 | 1.4857 | 3.9307 | 0.3413 | 1.685 |
| 计算机学报 | 1.6125 | 3.8775 | 0.3989 | 1.528 |
| 计算机研究与发展 | 1.5892 | 4.0520 | 1.1170 | 1.103 |
| 计算机辅助设计与图形学学报 | 1.2859 | 3.9062 | 1.5688 | 0.649 |
| 自动化学报 | 1.1805 | 2.6161 | 0.6811 | 1.064 |
| 中国图象图形学报 | 0.7636 | 4.3447 | 1.2673 | 0.541 |
| 计算机工程与应用 | 0.6451 | 4.2496 | 5.8238 | 0.393 |
| 系统仿真学报 | 0.7810 | 3.4138 | 2.0063 | 0.405 |
| 计算机工程 | 0.5642 | 4.9454 | 3.5595 | 0.518 |
| 计算机集成制造系统 | 0.9709 | 4.0841 | 3.4514 | 0.758 |
| 控制与决策 | 0.8266 | 2.4201 | 0.9298 | 0.816 |
| 小型微型计算机系统 | 0.8652 | 4.5997 | 4.2853 | 0.312 |
| 控制理论与应用 | 1.1224 | 2.1348 | 0.9778 | 0.662 |
| 计算机应用研究 | 0.5911 | 4.8042 | 8.9922 | 0.422 |
| 机器人 | 0.8366 | 2.8188 | 0.5263 | 0.723 |
| 中文信息学报 | 1.4174 | 4.6786 | 1.4271 | 0.836 |
| 计算机应用 | 0.6233 | 4.9552 | 6.1105 | 0.539 |
| 信息与控制 | 0.9063 | 2.7823 | 1.5852 | 0.366 |
| 计算机科学 | 0.8755 | 4.7091 | 6.4445 | 0.444 |
| 计算机测量与控制 | 0.5411 | 2.9920 | 8.4274 | 0.466 |
| 模式识别与人工智能 | 1.1006 | 4.1595 | 4.3672 | 0.475 |
| 计算机仿真 | 0.7089 | 3.5782 | 5.8849 | 0.33 |
| 计算机工程与科学 | 0.9887 | 4.8528 | 9.0108 | 0.304 |
| 遥感技术与应用 | 0.7613 | 4.2472 | 4.4205 | 0.85 |
| 传感器与微系统 | 0.5398 | 1.7496 | 4.3533 | 0.306 |
| 计算机工程与设计 | 0.6321 | 3.7458 | 22.5047 | 0.449 |
| 测控技术 | 0.3976 | 2.3919 | 2.3887 | 0.354 |
| 传感技术学报 | 0.5763 | 2.2350 | 3.2158 | 0.677 |
| 控制工程 | 0.7416 | 2.6852 | 6.1182 | 0.309 |
| 微电子学与计算机 | 0.5062 | 4.8351 | 7.3835 | 0.446 |
| 化工自动化及仪表 | 0.6261 | 2.3563 | 4.5043 | 0.616 |

为了进一步说明 PageRank 值、authority 值、hub 值和影响因子之间的关系, 我们进行相关性分析。根据统计学中的知识, 利用 Pearson 线性相关系数来计算两个变量线性相关程度。利用 SPSS 统计分析软件计算 PageRank 值、authority 值、hub 值和影响因子之间的关系, 其相关系数矩阵如表 4 所列。

表 4 相关系数矩阵

| | | pagerank | authority | hub | 影响因子 |
|-----------|---------------------|----------|-----------|--------|--------|
| pagerank | Pearson Correlation | 1 | .154 | -.448* | .738** |
| | Sig. (2-tailed) | | .409 | .011 | .000 |
| authority | Pearson Correlation | .154 | 1 | .233 | .001 |
| | Sig. (2-tailed) | .409 | | .206 | .994 |
| hub | Pearson Correlation | -.448* | .233 | 1 | -.447* |
| | Sig. (2-tailed) | .011 | .206 | | .012 |
| 影响因子 | Pearson Correlation | .738** | .001 | -.447* | 1 |
| | Sig. (2-tailed) | .000 | .994 | .012 | |

*. Correlation is significant at the 0.05 level(2-tailed).

** . Correlation is significant at the 0.01 level(2-tailed).

从表 4 可以看出, PageRank 值与影响因子的相关系数为 0.738, 说明这两者之间的相关性较大, 也就是说影响因子的值高时 PageRank 值也可能较高。但是, 从图 1 可以看出, 并不是对所有的个案都成立, 因为如果大多数引用都是不重要的引用, 那么其影响因子会高, 但其 PageRank 值不会高; 而 PageRank 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, 其 PageRank 值也会很高,

而影响因子会很低。在影响因子算法中,每次被引用都是同等看待的,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,PageRank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了区分,因此可以认为其是对重要性的一种测度。

而另一方面,authority值与影响因子的相关系数为0.001,hub值与影响因子的相关系数为-0.447,说明authority值与影响因子之间几乎不存在相关性,而hub值与影响因子之间存在负相关性,说明期刊影响因子的高低不能反映期刊的权威性,由图3也可知影响因子越高,期刊的中心性可能会越低。这与前面的结论基本相同,也说明了选取的期刊中,大多数的引用都不是重要的引用,反而有的期刊的引用次数不多,但是其引用都比较重要。与影响因子不同的是authority值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了区分,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。图1—图3分别为影响因子与pageRank值、authority值和hub值的散点分布图。

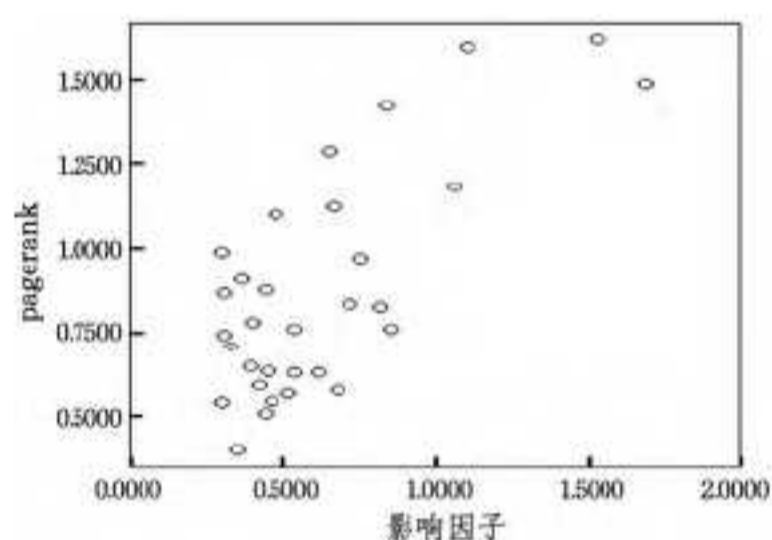


图1 影响因子与PageRank值的散点分图

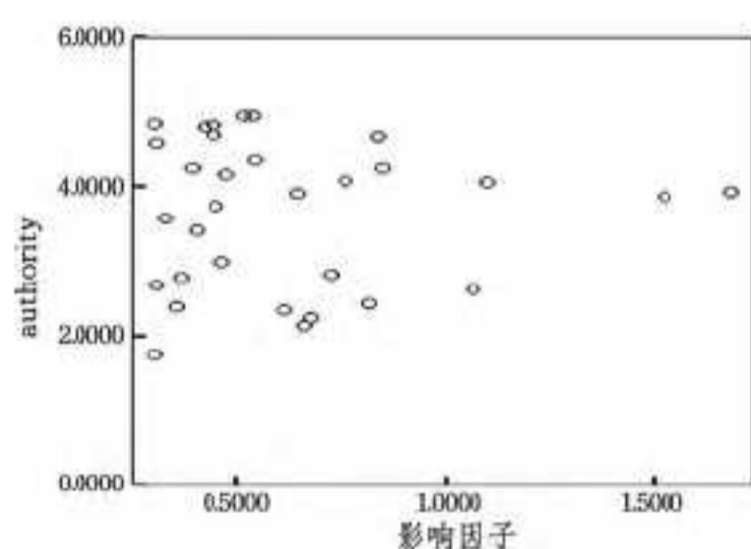


图2 影响因子与authority值的散点分图

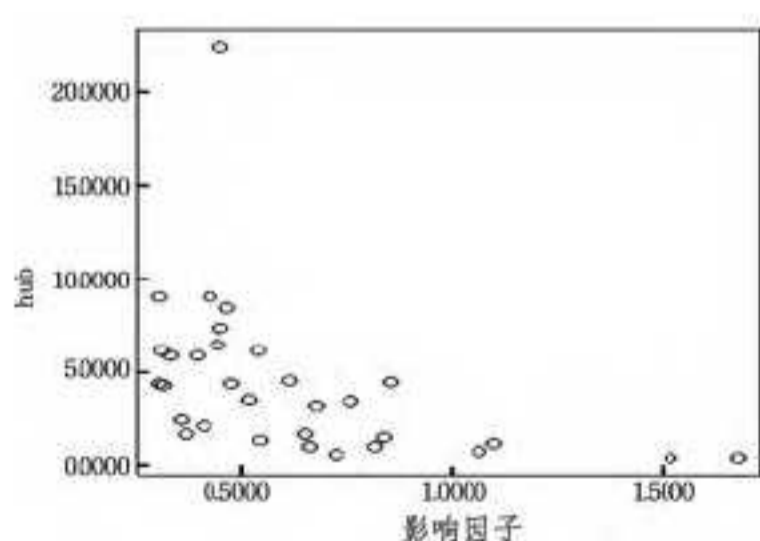


图3 影响因子与hub值的散点分布图

结束语 影响因子作为一个流行的期刊评价工具,容易

进行计算且易理解,但仍存在一些问题,特别是在计算过程中只考虑引用的数量而不考虑引用的质量是不科学的。它在很大程度上仅仅反映的是期刊的流行程度,还不能反映期刊的权威性。而PageRank算法和HITS算法正好弥补了影响因子评价的这一缺点,它们综合考虑了引用的数量和引用的质量,更能客观地反映期刊的权威性。通过研究期刊影响因子与PageRank值、authority值和hub值之间的不同关系,可以反映期刊的重要性、权威性和中心性。

本文受采样样本的限制,对PageRank算法和HITS算法的计算和讨论只限于31种计算机技术类核心期刊形成的引用网络,基于更学科、更大样本的相关研究还有待进一步深入展开。

参考文献

- [1] Garfield E. Indexing C. It's theory and Application in Science, Technology, and Humanities [M]. John Wiley and Sons, New York, 1979
- [2] Buena-Casal G. Assessing the Quality of Articles and Scientific Journals: Proposal for Weighted Impact Factor [J]. Psychology in Spain, 2004, 8(1): 60-76
- [3] Amin M, Mabe M. Impact factor use and abuse [J]. Perspectives in Publishing, 2000(1): 1-6
- [4] Bollen J, Rodriguez M A, de Sompel H V. Journal Status [OL]. <http://arxiv.org/abs/cs DL/0601030>, 2008-12-13
- [5] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117
- [6] Kleinberg J M. Authoritative sources in a hyper linked environment [C] // Proceedings of the 9th ACM SIAM Symposium on Discrete Algorithms. Baltimore, 1998: 668-677
- [7] 苏成, 潘云涛, 等. 基于PageRank的期刊评价研究 [J]. 中国科技期刊研究, 2009, 20(4): 614-617
- [8] 邱均平, 李江. 链接分析与引文分析的比较 [J]. 中国图书馆学报, 2008, 34(1): 60-64
- [9] 苏成, 潘云涛, 等. 基于HITS的期刊评价研究 [J]. 编辑学报, 2009, 21(4): 366-368
- [10] 李稚楹, 杨武, 谢治军. PageRank算法研究综述 [J]. 计算机科学, 2011, 38(10A): 185-188
- [11] 尹莉. 一种基于PageRank算法的期刊评价理论模型 [J]. 情报科学, 2012, 30(12): 1799-1803
- [12] 朱戈. 一种基于PageRank的文献相似性搜索算法 [J]. 计算机工程与应用, 2013, 49(8): 142-145
- [13] 许海云, 方曙. 基于引文分布和引文网络的学术期刊评价指标研究 [J]. 情报科学, 2013, 31(1): 130-134
- [14] 吴海芳. 几种学术期刊评价指标的相关性分析 [J]. 大学图书馆学报, 2013, 31(6): 86-89
- [15] 邱均平, 李爱群. 国内外期刊评价的比较研究 [J]. 重庆大学学报, 2007, 13(3): 60-65
- [16] 陶乃航. 社会网络分析方法在国内学术期评价中的应用研究 [D]. 重庆, 重庆大学, 2010