基于优化 PageRank、HITS 和 SALSA 算法的期刊评价研究*

苏 成 潘云涛 袁军鹏 马 峥
中国科学技术信息研究所情报方法研究中心,100038, 北京

摘 要 本研究构建了适合期刊引用网络的优化 PageRank、HITS 和 SALSA 算法,利用原始和优化后各算法计算了 2013 年的中国科技论文与引文数据库中 1 989 种期刊的权值,并与传统的影响因子和总被引频次进行了对比研究,分析讨论了各算法的特性、优缺点和适用范围。

关键词 PageRank;HITS;SALSA;影响因子;期刊评价

Evaluation of journals based on optimized PageRank, HITS and SALSA//SU Cheng, PAN Yuntao, YUAN Junpeng, MA Zheng

Abstract This research constructs optimized PageRank, HITS and SALSA for journal evaluation. We use the original and optimized algorithms to compute the weights of 1 989 journals of Chinese Science and Technology Paper Citation Database (CSTPCD), and compare the results with different algorithms. Finally, the features, merits, demerits and application scope of each algorithm are discussed, and some conclusions are given.

Keywords PageRank; HITS; SALSA; impact factor; journal evaluation

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

利用引文分析进行期刊评价的研究很多。其中 Garfield 提出的影响因子影响最大。但影响因子在计算时把所有引用视为等同是不合理的[1]。受网页链接分析算法如 PageRank[2]、HITS[3]等区分链接的重要性的做法的启发,国内外不少研究者借鉴网页排序算法 PageRank[4-9]和 HITS[10]进行了期刊评价研究。基于 PageRank 算法的期刊评价方法的优点是区分了不同引用的重要性差别,缺点是有利于创刊较早的期刊[11];基于 HITS 算法的期刊评价方法的优点是可以提供 2 个分数:Authority 值反映期刊的权威性,Hub 值反映期刊利用外部资源能力的大小,缺点是 HITS 算法因为其具有 TKC 效应[12],不太适合全学科的期刊排序[10]。为改善这 2 种算法的缺点,Lempel 和 Moran 于 2000 年提出了 Stochastic Approach to Link Structure Analysis (SALSA)算法[12],此算法综合了 HITS 和 PageRank 算法思想。与 HITS 算法类似,也产生 2 个分数,即权威(Authority)值和中心(Hub)值;它也借鉴了 PageRank 算法中的"随机冲浪"模型,Hub 网页根据其外链随机指向 Authority 网页,反之亦然。但是此算法得到的结果与网页的入链数(即论文的被引次数)的十分相似[13],这表明其在改善 PageRank 和 HITS 算法的同时,也失去了对网络链接结构的敏感度。

以上这些研究丰富了期刊评价研究工作,但是它们大多数没有像"影响因子"那样考虑期刊规模大小对被引次数多少的影响。受"影响因子"定义启发,本研究拟根据 CSTPCD2013 年的期刊引用网络,构建适用期刊引用网络并根据期刊规模大小标准化的优化 PageRank、HITS 和 SALSA 算法,对比分析各算法排序结果和影响因子、总被引频次等传统期刊评价指标。

1 数据

本研究使用的数据取自 2013 年中国科技论文与引文数据库(CSTPCD),包含 1 989 种中国出版的中英文科技期刊。所有这些期刊平均被引次数 1 180.21 次,表 1 显示大于平均数的期刊共有 624 种;大于2 000次的期刊共有 299 种,占总数的 15.03%。大于10 000次的期刊共有 7 种(表 1)。

表 1 期刊引用网络基本情况

总被引频次	期刊数	百分比/%
0 ~ 500	663	33.33
501 ~ 1 000	559	28.10
1 001 ~ 2 000	468	23.53
2 001 ~ 3 000	152	7.64
3 001 ~ 4 000	70	3.52
4 001 ~ 5 000	28	1.41
5 001 ~ 6 000	21	1.06
大于 6 000	28	1.41
合计	1 989	100.00

2 方法

CSTPCD 收录期刊中创刊时长差异很大,如果不选取特定的时间窗口,对于较新的期刊是不公平的。另外各期刊的年载文量差别也很大,如果不对载文量进行标准化,对于专注发表少而精论文的期刊是不合理的。为消除创刊时长、期刊载文量的影响,本研究借鉴"影响因子"选取 2 年时间窗口的成功经验,从"全时期刊引用网

* 国家科技支撑计划项目(2015BAH25F01)资助

络”中抽取 2 年时间窗口的引用关系,构建可对期刊规模大小标准化的优化 PageRank、HITS 和 SALSA 算法。具体步骤如下:

1)构建“全时期刊引用矩阵”。从 CSTPCD2013 年数据库中抽取 1 989 种期刊的引用关系。比如说 A 刊引用 B 刊 100 次,那么表示 A 刊在 2013 年发表的论文引用了在 B 刊自创刊以来发表论文的总次数,我们称之为“全时期刊引用矩阵”$\boldsymbol{L}$。$\boldsymbol{L}$ 可以用矩阵方式表示:

$$L_{ij}=\begin{cases} m,\text{如果期刊 } i \text{ 引用了期刊 } j \text{ 共 } m \text{ 次;} \\ 0,\text{否。} \end{cases} \tag{1}$$

2)构建“2 年期刊引用矩阵”。从“全时期刊引用矩阵”中抽取“2 年期刊引用矩阵”$\boldsymbol{M}$。比如在 2013 年“全时期刊引用矩阵”$\boldsymbol{L}$ 中 A 刊引用 B 刊 100 次,假设 A 刊引用了 B 刊在 2011 和 2012 年所发论文 50 次,其他年份所发论文 50 次,那么“2 年期刊引用矩阵”中 A 刊引用 B 刊为 50 次。那么 $\boldsymbol{M}$ 可表示如下:

$$\boldsymbol{M}_{ij}=\begin{cases} m,\text{如果期刊 } i \text{ 在统计年度引用了期刊 } j \text{ 在} \\ \quad \text{统计年度前 2 年发表的论文共 } m \text{ 次;} \\ 0,\text{否。} \end{cases} \tag{2}$$

2.1　优化 PageRank　针对期刊引用网络的原始 PageRank算法可以用如下矩阵形式[14]表示:

$$\boldsymbol{\pi}^{(K)\mathrm{T}}=\boldsymbol{\pi}^{(K-1)\mathrm{T}}\left(\alpha\boldsymbol{L}_r+(\alpha a+(1-\alpha)\boldsymbol{e})\frac{1}{n}\boldsymbol{e}^{\mathrm{T}}\right), \tag{3}$$

式中 $\boldsymbol{L}_r$ 为式(1)中矩阵 $\boldsymbol{L}$ 每行的元素除以该行之和得到的新矩阵,$\boldsymbol{\pi}^{(k)\mathrm{T}}$ 为第 k 次迭代计算的 PageRank 值,α 是介于 0 与 1 之间的阻尼系数,a 表示没有外链的孤立页面,$\boldsymbol{e}^{\mathrm{T}}$ 表示值均为 1 的行矩阵,n 为期刊总数。

那么能够消除创刊时长、期刊规模大小影响的优化 PageRank 算法可以表示如下:

$$\begin{aligned} &\boldsymbol{\pi}^{(K)\mathrm{T}}=\boldsymbol{\pi}^{(K-1)\mathrm{T}}\left(\alpha\boldsymbol{M}_r+(\alpha a+(1-\alpha)\boldsymbol{e})\frac{1}{n}\boldsymbol{e}^{\mathrm{T}}\right), \\ &\boldsymbol{M}_{PR}(J)=\boldsymbol{\pi}^{(K)\mathrm{T}}/(p_{y-1}+p_{y-2}), \end{aligned} \tag{4}$$

$\boldsymbol{M}_r$ 为式(2)中矩阵 $\boldsymbol{M}$ 每行的元素除以该行之和得到的新矩阵,$\boldsymbol{M}_{PR}(J)$ 为期刊 J 在 y 年的标准化 PageRank 值,p_{y-1}、p_{y-2}为期刊 J 在 $y-1$ 和 $y-2$ 年发表的论文数。

2.2　优化 HITS　原始 HITS 算法的矩阵形式如下:

$$\begin{aligned} &\boldsymbol{\pi}_a^{(k)}=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\pi}_a^{(k-1)}, \\ &\boldsymbol{\pi}_h^{(k)}=\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\pi}_h^{(k-1)}, \end{aligned} \tag{5}$$

$\boldsymbol{\pi}_a^{(k)}$ 和 $\boldsymbol{\pi}_h^{(k)}$ 表示在 k 次迭代计算得出的 Authority 值与 Hub 值,$\boldsymbol{L}^{\mathrm{T}}$ 表示矩阵 $\boldsymbol{L}$ 的转置矩阵。

那么能够消除创刊时长、期刊规模大小影响的优化 HITS 算法可以表示如下:

$$\begin{aligned} &\boldsymbol{\pi}_a^{(k)}=\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\pi}_a^{(k-1)}, \\ &\boldsymbol{\pi}_h^{(k)}=\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{\pi}_h^{(k-1)}, \\ &\boldsymbol{M}_{HA}(J)=\boldsymbol{\pi}_a^{(k)}/(p_{y-1}+p_{y-2}), \\ &\boldsymbol{M}_{HH}(J)=\boldsymbol{\pi}_h^{(k)}/(p_{y-1}+p_{y-2}), \end{aligned} \tag{6}$$

$\boldsymbol{M}_{HA}(J)$为期刊 J 在 y 年的标准化 Authority 值,$\boldsymbol{M}_{HH}(J)$为期刊 J 在 y 年的标准化 Hub 值。

2.3　优化 SALSA　原始 SALSA 算法的矩阵形式如下:

$$\begin{aligned} &\boldsymbol{\pi}_a^{(k)\mathrm{T}}=\boldsymbol{\pi}_a^{(k-1)\mathrm{T}}\boldsymbol{L}_c^{\mathrm{T}}\boldsymbol{L}_r, \\ &\boldsymbol{\pi}_h^{(k)\mathrm{T}}=\boldsymbol{\pi}_h^{(k-1)\mathrm{T}}\boldsymbol{L}_r\boldsymbol{L}_c^{\mathrm{T}}, \end{aligned} \tag{7}$$

$\boldsymbol{\pi}_a^{(k)\mathrm{T}}$表示第 k 次迭代计算的 Authority 值,$\boldsymbol{\pi}_h^{(k)\mathrm{T}}$表示第 k 次迭代计算的 Hub 值,$\boldsymbol{L}_r$ 为矩阵 $\boldsymbol{L}$ 每行的元素除以该行之和得到的新矩阵,$\boldsymbol{L}_c$ 为矩阵 $\boldsymbol{L}$ 每列的元素除以该列之和得到的新矩阵。

那么能够消除创刊时长、期刊规模大小影响的优化 SALSA 算法可以表示如下:

$$\begin{aligned} &\boldsymbol{\pi}_a^{(k)\mathrm{T}}=\boldsymbol{\pi}_a^{(k-1)\mathrm{T}}\boldsymbol{M}_c^{\mathrm{T}}\boldsymbol{M}_r, \\ &\boldsymbol{\pi}_h^{(k)\mathrm{T}}=\boldsymbol{\pi}_h^{(k-1)\mathrm{T}}\boldsymbol{M}_r\boldsymbol{M}_c^{\mathrm{T}}, \\ &\boldsymbol{M}_{SA}(J)=\boldsymbol{\pi}_a^{(k)}/(p_{y-1}+p_{y-2}), \\ &\boldsymbol{M}_{SH}(J)=\boldsymbol{\pi}_h^{(k)}/p_{y-1}+p_{y-2}), \end{aligned} \tag{8}$$

$\boldsymbol{M}_{SA}(J)$为期刊 J 在 y 年的标准化 Authority 值,$\boldsymbol{M}_{SH}(J)$为期刊 J 在 y 年的标准化 Hub 值。

2.4　算法计算　在本研究中,我们利用数据库软件 Foxpro 9.0 构建期刊引用网络矩阵,利用 Matlab2014a 进行 PageRank、HITS 和 SALSA 计算。具体步骤如下:

1)构建“全时期刊引用网络矩阵”和“2 年期刊引用网络矩阵”。选取 2013 年 CSTPCD 中 1 989 种的期刊,根据式(1)构建一个 1 989×1 989 的“全时期刊引用网络矩阵”,矩阵元素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。根据式(2)构建“2 年期刊引用网络矩阵”。

2)迭代计算。利用“全时期刊引用网络矩阵”,采用式(3)、(5)和(7),利用幂法运算分别求 PageRank、HITS 和 SALSA 相关矩阵的最大特征值。利用“2 年期刊引用网络矩阵”,采用式(4)、(6)和(8)计算优化 PageRank、HITS 和 SALSA 相关矩阵的最大特征值。计算矩阵最大特征值时采用的是 Matlab2014a,收敛值取 10^{-8}。

3　结果

一般来说,人们往往更关注排名靠前的期刊,也就是说位于排序列表前面的结果的重要性要大于后面的。所以我们重点考察了各排序列表的前 10 位。

3.1 各算法排名前10位期刊 从表2可以看出,CT与SA前10列表完全一致,这表示这2个算法具有高度一致性。IF与MSA前10列表中前9位都是一致的,只有第10不同,这表示这2个算法同样具有很高的一致性。HA算法中前10均是物理学领域期刊,而MHA算法均是医学期刊,这是因为HITS算法具有TKC效应。TKC效应是指HITS算法在Authority和Hub的相互迭代加强过程中,权重会越来越集中于紧密度最大的社区。各算法前10的权重值之和占总数之比可以反映权重的集中程度。HA算法前10期刊占所有期刊值的69.30%,TKC效应很强。而MHA算法前10占所有权值的25.77%,与HA相比,TKC效应有所减弱,但仍较强。其他算法前10占比都低于5%,SA为4.87%,CT为4.86%,MPR为3.56%,PR为3.06%,MSA为2.29%,IF为2.27%。PR与CT的前10列表中有8个共同元素,而PR与IF的前10列表中只有1个共同元素,所以可以认为PR与CT算法更加接近。MPR算法前10中有9种英文刊,这是因为这些刊的引用次数虽然不是最多的,但是它们的被引均是来自较好的期刊,这说明PR算法能更好地反映重要性。

表2 各算法排名前10期刊

排序	IF	CT	PR	MPR	HA	MHA	SA	MSA
1	石油勘探与开发	生态学报	中国电机工程学报	CELL RESEARCH	物理学报	中华医院感染学杂志	生态学报	石油勘探与开发
2	地理学报	中国电机工程学报	生态学报	JOURNAL OF GENETICS AND GENOMICS	CHINESE PHYSICS B	中华临床感染病杂志	中国电机工程学报	地理学报
3	石油与天然气地质	中华医院感染学杂志	中华医院感染学杂志	MOLECULAR PLANT	CHINESE PHYSICS LETTERS	中国消毒学杂志	中华医院感染学杂志	石油与天然气地质
4	草业学报	食品科学	农业工程学报	石油勘探与开发	光学学报	中国感染控制杂志	食品科学	草业学报
5	断块油气田	农业工程学报	电力系统自动化	COMMUNICATIONS IN MATHEMATICAL RESEARCH	强激光与粒子束	中国抗生素杂志	农业工程学报	断块油气田
6	中国感染与化疗杂志	物理学报	应用生态学报	NEUROSCIENCE BULLETIN	COMMUNICATIONS IN THEORETICAL PHYSICS	中国临床药理学杂志	物理学报	中国感染与化疗杂志
7	石油学报	应用生态学报	食品科学	JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY	中国电机工程学报	中华检验医学杂志	应用生态学报	石油学报
8	中国沙漠	中国组织工程研究	岩石力学与工程学报	ACTA MATHEMATICA SCIENTIA	科学通报	临床肺科杂志	中国组织工程研究	中国沙漠
9	冰川冻土	电力系统自动化	电网技术	VIROLOGICA SINICA	中国激光	检验医学	电力系统自动化	冰川冻土
10	电力系统自动化	科学通报	科学通报	PEDOSPHERE	大气科学	中华护理杂志	科学通报	石油实验地质

注:IF为影响因子,CT为总被引频次,PR为PageRank,MPR为优化PageRank,HA为HITS算法Authority,MHA为优化HITS算法Authority,SA为SALSA算法Authority,MSA为优化SALSA算法Authority。

一般来说数据的标准差等可以反映数据的离散趋势。但因为本研究中各算法得出数据的测量尺度相差太大,数据量纲也不同,直接利用标准差来进行比较不合适,必须消除测量尺度和量纲的影响,而变异系数就可以做到这一点。变异系数越大表示数据距离均值越远,数据分布越离散。表3显示HA和MHA算法的变异系数远远大于其他算法,这说明HITS算法因为TKC效应的存在,数据分布很离散。

表3 各算法的变异系数、偏度与峰度

参数	IF	CT	PR	MPR	HA	MHA	SA	MSA
变异系数	0.658 62	1.151 33	0.663 87	0.824 52	19.265 09	6.301 03	1.151 31	0.661 78
偏度	2.067	3.756	3.992	4.500	35.504	34.592	3.768	2.078
峰度	6.888	21.976	25.689	43.883	1 348.866	1 384.621	22.222	6.975

3.2 各算法相关性分析 通过表4所列各算法Spearman相关系数发现,PR vs CT > PR vs IF;HA vs CT > HA vs IF;SA vs CT > SA vs IF。这说明原始PageRank、HITS和SALSA算法因为没有根据期刊规模进行标准化,其排序结果与总被引频次更加相似,其本质偏向于期刊所载论文质量"总量"的测度。我们发现MPR vs CT < MPR vs IF;MHA vs CT < MHA vs IF;MSA vs CT < MSA vs IF。这说明优化PageRank、

HITS 和 SALSA 算法因为选定了特定时间窗口并对期刊规模进行标准化,其排序结果与影响因子更加相似,其本质偏向于期刊所载论文的“平均”质量的测度。这结果与我们的直观感觉相符合。

通过表 4 发现,MSA 与 IF 的 Spearman 相关系数等于 1.000,SA 与 CT 的相关系数也等于 1.000,这说明 MSA 与影响因子高度相似,SA 与总被引频次高度相似。另外,PR 与 SA 算法的 Spearman 相关系数比 PR 与 HA 的要大,这说明 PR 与 SA 算法比 PR 与 HA 算法更加相似。MPR 与 MSA 算法的 Spearman 相关系数比 MPR 与 MHA 的要大,这说明 MPR 与 MSA 算法比 MPR 与 MHA 算法更加相似。这也说明优化后的 PageRank、HITS 和 SALSA 算法并没有根本性改变其特性。值得指出的是 MPR 与 MHA 呈负相关关系,这可能与 HITS 算法的 TKC 效应关系很大。

表 4　各算法 Spearman 相关性

算法比	Spearman 相关系数	算法比	Spearman 相关系数	算法比	Spearman 相关系数
PR vs CT	0.991 **	MPR vs IF	0.618 **	PR vs IF	0.690 **
PR vs HA	0.307 **	MPR vs MHA	-0.142 **	HA vs IF	0.115 **
PR vs SA	0.991 **	MPR vs MSA	0.620 **	SA vs IF	0.682 **
HA vs CT	0.284 **	MHA vs IF	0.323 **	MPR vs CT	0.154 **
HA vs SA	0.284 **	MHA vs MSA	0.322 **	MHA vs CT	0.302 **
SA vs CT	1.000 **	MSA vs IF	1.000 **	MSA vs CT	0.679 **

** 在置信度(双测)为 0.01 时,相关性有统计学意义。

4　讨论与结论

用原始的 PageRank、HITS 和 SALSA 算法进行期刊排序的结果与总被引频次算法排序列表更加接近,而优化的 PageRank、HITS 和 SALSA 算法得到的结果与影响因子排序列表更加接近。PageRank 算法得出排序列表没有明显的学科偏向,能更好说明期刊的重要性,适合“全局”排序。但优化 PageRank 算法有过于偏向英文版期刊的缺陷。HITS 算法提供 Authority 和 Hub 两个分数,但是因为其使用的矩阵没有进行行或列的标准化,在 Authority 和 Hub 的相互迭代加强过程中,会产生 TKC 效果,往往排序结果前列的期刊均来自紧密度最大的社区,不适合“全局”排序。SALSA 算法结合了 PageRank 算法和 HITS 的优点,它采用了行和列均标准化的矩阵,消除了 TKC 效果。但是它的缺点是对引用结构不敏感,排序结果与被引次数结果太相似。很明显不存在一个理想算法适合所有情况。为了得到更好的排序结果,我们应该根据任务的不同目标选择合适的算法。比如,PageRank、SALSA 和被引次数等算法没有 TKC 效果,适合全领域排序,而 HITS 算法适合分领域排序。原始 PageRank、HITS 和 SALSA 适合质量“总量”测度,而优化 PageRank、HITS 和 SALSA 适合“平均质量”测度。

5　参考文献

[1] BUELA-CASAL G. Assessing the quality of articles and scientific journals: proposal for weighted impact factor [J]. Psychology in Spain, 2004, 8(1): 60

[2] BRIN S,PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks,2012,561(18):3825

[3] JON M. Authoritative sources in a hyperlinked environment [C]// Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms, Baltimore, MD, 1998:668-677

[4] BOLLEN J, RODRIGUEZ M A, VAN DE SOMPEL H. Journal status[J], Scientometrics, 2006, 69(3): 669

[5] SCImago Research Group. Description of SCImago Journal Rank Indicator[EB/OL]. [2015-04-15]. http://www.scimagojr.com/SCImago Journal Rank.pdf

[6] BERGSTROM C T,WEST J D,WISEMAN M A. The Eigenfactor™ metrics[J]. The Journal of Neuroscience,2008,28(45):11433

[7] 苏成, 潘云涛, 袁军鹏, 等. 基于 PageRank 的期刊评价研究[J]. 中国科技期刊研究, 2009, 20(4): 614

[8] SU Cheng, PAN Yuntao, ZHEN Yanning, et al. Prestige Rank: a new evaluation method for papers and journals[J]. Journal of Informetrics, 2011, 5(1): 1

[9] 苏成, 潘云涛, 马峥, 等. 权威因子:一个新的期刊评价指标[J]. 编辑学报, 2010, 22(4), 369

[10] 苏成, 潘云涛, 袁军鹏, 等. 基于 HITS 算法的期刊评价研究[J]. 编辑学报, 2009, 21(4): 366

[11] CHEN Y L, CHEN X H. An evolutionary PageRank approach for journal ranking with expert judgements[J]. Journal of Information Science, 2011, 37(3):254

[12] LEMPEL R, MORAN S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect[J]. Computer Networks, 2000, 33 (1):387

[13] BORODIN A, ROBERTS G O, ROSENTHAL J S, et al. Link analysis ranking: algorithms, theory, and experiments[J]. ACM Transactions on Internet Technology (TOIT), 2005, 5(1): 231

[14] LANGVILLE A, MEYER C. Google's pagerank and beyond: the science of search engine rankings[M]. Princeton: Princeton University Press, 2006

(2015-03-20 收稿;2015-04-28 修回)