一种基于PageRank和时间衰退因子的作者影响力评价方法

崔景洋
(河北地质大学,河北 石家庄 050031)

【摘 要】为便于找到某个领域的优秀作者,将PageRank算法应用于学术网络之中,通过加权的方式将与作者有关的属性进行整合,提出一个作者影响力评价方法。与之前的评价方法相比,不仅考虑了作者文章、工作机构等对于作者影响力的影响,还考虑了时间因素在计算过程中的衰减作用。

【关键词】PageRank; 引用半衰期; 作者影响力评价

A Method of Authors´ Influence Evaluation Based on PageRank and Time Factor

CUI Jing-yang
(Hebei GEO University, Shijiazhuang 050031, China)

【Abstract】 For the convenience of finding good authors in a certain field, we apply PageRank algorithm to academic networks through a weighted way to integrate the author's related attributes. Compared with the previous method of calculating PageRank, this algorithm not only considers the impact of author's paper and author's work organization for the author's influence, but also considers the decay effect of time factor in the process of calculating PageRank.

【Key words】 PageRank; cited half-life; author's influence evaluation

〔中图分类号〕TP301.6 〔文献标识码〕A 〔文章编号〕1674-3229(2018)02-0007-05

0 引言

在相关领域内根据发文等情况对作者的影响力进行评价,可以客观地反映这个作者实际的学术能力和文献生产力水平,为我们了解该领域的研究情况和发展趋势提供一个清晰明确的指导方向。从2008年开始,我国授予博士学位的人数就超过了美国,成为世界上每年授予博士学位人数最多的国家。同时,我国学者仅在2015年一年发表的论文数量就高达百万篇[1]。研究人员数量的庞大,势必导致学术文献数量的增长。所以,如何在海量文献中根据作者的发文情况找出各领域影响力较大的作者,成为一个热门问题。

对于作者的影响力评价方法主要包含:

(1)传统文献计量学评价方法。例如文献统计分析法、数学模型法、引文分析法等[2]。

(2)H指数及其扩展指数。例如H指数[3]、V指数[4]、X指数[5]等。

(3)基于网络的评价方法。合作网络[6]、引用网络[6]、社会网络[7]等。

上述方法除了以统计的方式计算作者影响力之外,有的还考虑了文献与文献之间的引用关系,但忽略了时间和其他因素在计算过程中对于评价指标的影响。李玉鑑[8]等考虑了时间因素对于PageRank值

[收稿日期] 2018-01-20
[基金项目] 国家自然科学基金项目"大规模网络半监督广义社区发现研究"(61503260);河北省研究生创新资助项目"基于Hadoop的学术网络作者影响力及文献推荐方法研究"(CXZZSS2017132)
[作者简介] 崔景洋(1992-),男,河北地质大学信息工程学院硕士研究生,研究方向:数据挖掘与机器学习。

的影响,但是计算过程过于复杂,使得原本就需要迭代很多次的PageRank算法的时间复杂度进一步提高,不符合当前环境下对于大数据处理速度的要求。本文根据作者影响力的含义,考虑了时间、机构等因素在作者影响力计算过程中的重要程度,提出了一种基于PageRank和时间衰退因子的作者影响力评价方法。

1 PageRank算法介绍

PageRank算法来源于谷歌公司网页排名方法,是一种将网页之间相互链接关系的计算结果作为网页排名的一种算法[9],由谷歌创始人Larry Page和Sergey Brin于1998年在斯坦福大学提出。在PageRank算法提出之前,搜索引擎还是采用的关键字匹配技术,于是在结果中经常出现名字与内容不相符的网站。PageRank的基本思想是一个投票机制,A页面链接B页面,就是A在给B投票。一个页面的PageRank值比较高,有两种可能:第一是很多页面都在链接这个页面;第二是有一个PageRank值比较高的页面链接了这个页面。无论哪一种情况,根据PageRank值对网页进行排序都比利用关键字更有意义。将其运用到学术排名领域时,页面与页面之间的链接关系,也就变成了文献与文献之间的引用关系。

$$PR(p)=\frac{(1-d)}{C_{Total}-1}+d\sum_{i=1}^{N}\frac{PR(X_i)}{C(X_i)}。$$

这是一个经典的PageRank算法公式,其中p代表某个待评价的学术文献,d是一个阻尼系数,通常取0.85。C_{Total}是文献总量。N表示N个引用了p的文献,X_i表示第i个引用了p的文献,$C(X_i)$表示X_i这篇文献总的参考文献数目。在计算PageRank值时,首先将每一个需要计算对象的PageRank值设置为1,然后开始迭代,每篇文章的权值在全部样本内部流动。迭代n次或者算法收敛后,停止算法。再根据PageRank值由大到小进行排名,PageRank值较高的文献就是影响力较大的文献。

2 基本概念

2.1 引用半衰期及时间衰减函数

文献半衰期(Half-life of Literature)最早是贝尔纳(J.D.Bernal)提出的用以表征文献老化速度的一种概念。放射性半衰期定义的是经过T时间后,只有一半原子有放射性。文献半衰期是T时间后,一半的文献将不再被引用,与放射性物质半衰期的定义基本相同,故可以使用同一半衰期公式。王丽雅[10]利用公式$HL(t)=\frac{C_{Total}*\left(\frac{1}{2}\right)^{t/T}}{C_{Total}}$,计算出的计算机学科的引用半衰期为4.76年,即$T$=4.76。

上式中,t是文献的年龄(以年为单位),C_{Total}是该学科中所有文献的数量。半衰期时间衰减函数$HL(t)$的含义是经过t时间后该篇文献的影响力降为之前影响力的$HL(t)$倍。这里为了便于理解,没有将上下的C_{Total}进行约分。上式中的分子部分是物理学半衰期公式,目的是求解经过t时间后还有多少原子仍有放射性。在除以原子总数后可以得到有放射性原子所占百分比。在文献排名中,得到的就是有效文献所占百分比,可以理解为影响力剩余百分比。

2.2 文章PageRank值

显然,作者的影响力应该跟他所发表的文章最大程度地有关联。如果一个作者所发表的论文PageRank值比较高,那么这个作者的影响力也就相应会很高,但需要注意的是不同作者在发表文献的过程中对于该文献所做的贡献是不同的,所以应该考虑作者在该文章中署名名次。利用文献-文献之间的引用关系构建引文网络,作者所发表的某一篇文章的PageRank值为:

$$PR(paper)=\left(\frac{(1-d)}{C_{Total}-1}+d\sum_{i=1}^{N}\frac{PR(paper_i)}{C(paper_i)}\right)\times weight(x)。$$

其中,C_{Total}是文献总量,N表示一共有N篇文献引用了该文献,$paper_i$表示第i篇引用$paper$的文献,$C(paper_i)$表示$paper_i$参考文献的数量,$weight(x)$中的x是该作者在写作过程中的顺序,$weight(x)$对应的多人权值如表1所示。

2.3 作者PageRank值

其次要考虑的是作者PageRank值,这个PageRank值不同于文献的PageRank值,文献的PageRank值代表的是每一篇文献的重要程度,而作者的PageRank值是利用每个作者之间相互引用的关系来计

表1　多人参与业绩计算权重对照表

参加人数	参加人名次									
	1	2	3	4	5	6	7	8	9	10
1	100%									
2	60%	40%								
3	50%	35%	15%							
4	48%	32%	20%	10%						
5	47%	26%	20%	10%	7%					
6	46%	25%	18%	10%	7%	6%				
7	45%	23%	17%	10%	7%	5%	4%			
8	40%	22%	17%	9%	7%	5%	4%	3%		
9	37%	21%	17%	9%	7%	5%	4%	3%	2%	
10	36%	21%	17%	9%	7%	5%	4%	3%	2%	1%

算出从整体来看每个作者的影响力程度，故同样应该作为计算作者影响力时的一个重要参考。在计算作者的PageRank值时，利用文献间的引用关系，统计作者全部的发文情况，构建作者–作者引用网络，用以计算作者的PageRank值。

$$PR(author)=\frac{(1-d)}{C_{Total}-1}+d\sum_{i=1}^{N}\frac{PR(author_i)}{C(author_i)},$$

其中，C_{Total}是某个领域内的作者总量，N表示曾经引用过该作者任何一篇学术文献的作者的数量，$author_i$代表N中的第i个作者，$C(author_i)$表示$author_i$引用过的作者的数量。

2.4　机构PageRank值

除了要考虑上述评价指标在作者影响力计算中的贡献率，还应该考虑作者所在单位的影响力。通常来说，名校的教师一般都比普通高校的教师水平要高。机构与机构之间的关系可以用二者作为发文单位所发表的所有学术文献之间的引用关系来代替。如果一个机构被其他机构或者个人引用的次数比较多，那么这个机构的影响力也就很大。利用机构–机构之间的引用关系构建网络，机构PageRank值的计算公式如下。

$$PR(organization)=\frac{(1-d)}{C_{Total}-1}+d\sum_{i=1}^{N}\frac{PR(organization_i)}{C(organization_i)},$$

其中，C_{Total}是机构总量，N代表引用过署名为该机构文献的机构数量，$organization_i$表示的是第i个引用该机构文献的机构，$C(organization_i)$表示$organization_i$引用的其他机构的机构数。

3　作者影响力计算公式

根据以上介绍的概念以及计算方法进行整合，得出基于时间因子（Time Factor）和PageRank的作者影响力计算公式为：

$$TF-PR(author)=\alpha\frac{PR(paper)HL(t)}{Sum(author's\ \ paper)}+\beta PR(author)+\gamma PR(organization)\ ,$$

权重限制为：$1=\alpha+\beta+\gamma$，

其中$Sum(author's\ paper)$代表该作者所发表的文献总数。当一个作者发表很多学术文献时，如果不加以归一化，那么在文献PageRank值的总量上对于其他发表量少质优的作者来说是非常不公平的，所以需要将其利用平均值的方式来衡量。

另外，各权重的取值也可以根据实际情况进行划定，推荐α、β、γ分别取0.5、0.3、0.2。

4　数据获取

本文研究中所使用的数据取自“维普网中文期刊服务平台”，以“题目中或关键词含有PageRank”为检索条件，检索时间为2018年1月3日，共获得了1989年至2018年的865条文献信息、引用关系6541条，其中涉及期刊249种、作者1242名、机构727个。数据集结构如表2、3、4所示。

表2　Pager信息表

文献	作者	发表年份	发表刊物
P1	A1，A2	2011	J1
P2	A1，A3	2016	J2
P3	A4，A5，A6	2017	J3
…	…	…	…

表3　Author信息表

作者	机构
A1	O1
A2	O2
A3	O1
…	…

表4　引用关系表

文献	被引文献
P1	P5、P6、P7
P2	P1、P7
P3	P4、P5
…	…

5　实验分析

与期刊影响因子类似，在实际计算过程中，作者本身的PageRank值与作者所在机构的PageRank值可

以不进行实时计算以保证计算速度。但由于作者发表文献的实时性与不确定性,所以对于文献的PageRank值,需要在每次研究作者排名时都进行计算。

5.1 作者PageRank值的计算

根据表4中文献与文献之间的引用关系,可以将每篇文献替换为该文献的作者,以构建作者-作者引用网络。但二者还是有些许差异的,因为在PageRank的计算过程中,需要避免自引现象。自引会导致PageRank值的流动陷入自我循环,也就削弱了该算法应有的价值。所以需要通过去除自引后的作者与作者引用网络来计算作者的PageRank值。前10位作者的PageRank值如表5所示。

表5 作者PageRank值前10位

作者姓名	PageRank值
刘大有	153.4564
程学旗	105.4561
孙建军	98.9834
李国杰	98.3354
黄丽华	85.9723
苏小英	73.4620
杨炳儒	60.3842
汪卫	51.1862
周明全	34.7486
陈志刚	23.2856

5.2 机构PageRank值的计算

与上一步骤类似,同样通过表4进行机构-机构引用关系的转化,前10位机构的PageRank值如表6所示。

表6 机构PageRank值前10位

机构名称	PageRank值
武汉大学	50.3546
复旦大学	48.7865
上海交通大学	39.6478
清华大学	28.5871
大连理工大学	20.1568
浙江大学	15.5670
中科院计算所	10.1956
北京大学	8.4562
安徽工业大学	8.7896
重庆大学	5.3482

5.3 作者影响力的计算

将作者之前所发表文章的PageRank值均考虑在内,并乘以时间衰退因子,可以对于当前计算时间后作者影响力水平有一个实时了解。另外由于机构PageRank值和作者本身的PageRank值都已经计算好了,也就省去了一部分计算时间,降低了计算强度。需要注意的是,作者的PageRank值并不能代替现在我们所计算的作者影响力,因为随着时间的推移,作者的发文数量和质量肯定有一个波动趋势,或升高或下降。所以应该将考虑了时间衰退影响的作者发表文章的PageRank值纳入对于作者影响力的评价当中。

通过计算后的作者影响力排名前10的作者如表7所示。

表7 作者影响力计算结果表

作者姓名	TF-PR值	H-index	总被引量	H-index排名	总被引量排名	PageRank值排名
程学旗	98.7357	18	2397	2	1	2
耿国华	94.3354	16	1750	5	4	11
刘大有	89.6178	19	1964	1	2	1
孙建军	76.9665	15	1053	8	9	3
周明全	60.8637	16	1628	5	5	9
黄丽华	53.2537	15	963	8	10	5
杨炳儒	51.4573	16	1393	5	7	7
吴朝晖	50.1822	18	1147	2	8	15
李红	38.4531	18	1796	2	3	16
李国杰	33.5210	9	1447	10	6	4

H指数和总被引量的排名是根据当前计算结果在10位学者当中的排名,PageRank值排名是表5中的排名。可以看出程学旗在本次的作者影响力计算过程中排名第一位,且总被引量和H指数的排名中也有不错的成绩。这说明程学旗发表文献质量较高,且近几年学术产出量也不错。

刘大有虽然在总被引量和H指数的排名中有着不俗的表现,但是因为其他相关性因子例如发文时间、单位因素的影响,所以在最后的排名中,没有得到较高的位次。

6　总结

作者在文献发表的过程中,与文献、机构等周围因素的关系是错综复杂的。而H-index等计算方法,单纯考虑文献与文献的关系、考虑被引次数等,这些计算方法在实际生产生活中的使用上存在不足。学术网络是一种复杂网络,对于作者影响力的研究应该纳入整个学术网络来进行。本文通过对于之前作者影响力评价指标的分析,考虑了时间衰退等因素,提出了一种用于作者影响力评价的新方法。

以CNKI的检索结果为数据集,取得更加权威的作者影响力的计算结果以及对于该评价方法的并行化设计,可以作为今后的研究方向,来完善实验结果。

[参考文献]

[1] 马丙超.基于引文网络的文献在线推荐系统研究和实现[D].大连:大连理工大学,2016.

[2] 安源,张玲.文献计量学在我国图书情报领域的应用研究进展综述[J].图书馆,2014,(5):63-68.

[3] Hirsch J E. An index to quantify an individual's scientific research output[J]. PageRankoceedings of the National Academy of Sciences of the United States of America, 2005,102(46):16569-16572.

[4] 李海英,许强,李恩科.评价领域作者影响力的新指标--v指数[J].情报杂志,2015,(12):38-43.

[5] 肖学斌.x指数:描述研究人员论文水平的文献计量新指数[J].图书情报知识,2015,(2):93-99.

[6] 贺焕振.基于合著与引用网络的专家知识地图构建研究[D].杭州:浙江大学,2015.

[7] Abbasi A, Altmann J, Hossain L. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures[J]. Journal of Informetrics,2011,5(4):594-607.

[8] 李玉鑑,张甫.基于PageRank和时间衰减的科技文献重要度评价方法,CN 105740452 A[P]. 2016.

[9] Page L. The PageRank citation ranking : Bringing order to the web[J]. Stanford Digital Libraries Working Paper, 1998,9(1):1-14.

[10] 王丽雅.基于CNKI的计算机科学学科半衰期分析[J].图书与情报,2015,(1):100-105.

(上接第6页)

射模。对任意右 R-模 M，由(1)知 M 是 n-X-余纯平坦模，故 $Tor_1^R(M,N)=0$，由此得 N 是平坦模。

(2) $\Rightarrow$ (1)。M 是右 R-模，对任意 n-X-内射左 R-模 N，由(2)知 N 是平坦模，故有 $Tor_1^R(M,N)=0$，由此得 M 是 n-X-余纯平坦模。

(2) $\Rightarrow$ (3)。由内射模是 n-X-内射模，(2)即为平坦模可得。

[参考文献]

[1] Anderson F W, Fuller K R. Rings and categories of modules[M]. New York: Springer-Verlag, 1974.

[2] Enochs E E, Jenda O M G. Relative homological algebra[M]. Berlin-New York: Walter de Gruyter, 2000.

[3] D. Bennis. n-X-Coherent Rings[J]. International Electronic Journal of Algebra, 2010, 7: 128-139.

[4] Y. Y. Liu. n-X-Projective Modules, n-X-Injective Modules, Cotorsion Theories[J]. Int. J. Contemp. Math.Sciences, 2010, 56(5):2753-2762.

[5] 陈翔,戴立辉.n-X-余挠模与n-X-余挠维数[J].闽江学院学报, 2012, 33(2):14-18.