

SARank: 一种学术社交网络用户影响力分析模型

顾瑞春, 王静宇

(内蒙古科技大学, 内蒙古 包头 014010)

摘要: 社交网络作为联系现实社会与网络社会的重要途径, 已经成为人们日常使用最为广泛的网络应用。目前, 国内外各大社交网络均拥有大量的活跃用户, 并且用户数量仍在不断上升。由众多用户不断产生的大量数据, 已经成为数据分析领域新的研究热点。而基于社交网络的科研信息共享协作平台, 也十分受科研工作者的青睐。同样, 在学术社交网络中不断产生的大量数据中也隐含着十分重要的信息, 有着极大的研究价值。因此, 以学术社交网络中海量的用户数据为研究对象, 以大数据分析处理技术为基础, 结合科研领域相关参数以及社交网络中信息传播的相关指标, 提出一种多元化的科研用户影响力计算模型-SARank。通过学术社交网络中的数据对科研用户影响力进行量化分析, 形成一种新的科研评价参考指标。经过实验分析, 得出了有益结论。

关键词: 学术社交网络; 影响力分析; 多元度量; 数据科学; 用户产生内容

中图分类号: TP399

文献标识码: A

文章编号: 1673-629X(2018)03-0032-05

doi: 10.3969/j.issn.1673-629X.2018.03.007

SARank: An Influence Analysis Model on Academic Social Network

GU Rui-chun, WANG Jing-yu

(Inner Mongolia University of Science and Technology, Baotou 014010, China)

Abstract: As an important connection of the real life and the virtual network space, the SNS (social network services) is becoming the most widely used network applications in daily life. At present, there are lots of active users with rising number in the most famous social networks. The huge volume contents generated by many users all the time have become a new research hotspot of the data analysis field. The scientific information sharing and collaborating platform based on social network has being favored very much by the scientific researchers. There are many potential precious information in the UGC (user generated contents) of the academic social network, with great research values of them. Therefore, taking the massive user data as an object in the scholar social network, based on the big data analyzing and processing technology, combined with the parameters of the scientific research field and the information communication indexes of the academic social network, we propose an altmetrics (alternative metric) scientific research user influence analysis model named SARank. It could be used in the quantitative computing of the scientific user influence of the scholar social network, forming a new reference indicator of the scientific research evaluation system. Some useful conclusions are obtained through the analysis and explanation of the experiment.

Key words: scholar social network; influence analysis; altmetrics; data science; user generated content

1 概述

学术社交网络, 是一种基于社交网络 (social network services) 的科研信息共享与协作平台, 用户可以通过各种网络终端参与其中, 进行在线交流、信息交互、技术协作等线上活动, 社交网络用户可通过某种网络联系进一步将线下关系迁移到线上, 形成在线虚拟社区。随着 Web2.0 技术和移动互联网的迅猛发展, 社交网络正极大地改变着人们获取信息和使用互联网

的方式, 并已经逐渐成为连接现实信息社会与虚拟网络社会的重要桥梁。

社交网络发展迅猛, 国内外有不少社交网络已经具有较大规模, 著名的社交网站 Facebook 目前用户数大约为 22 亿, 活跃用户数超过 13 亿, 并且 Facebook 旗下的移动端社交平台 WhatsApp 的月活跃用户数已经突破 10 亿大关, Twitter 的用户量也已经突破了 6 亿, 国内的腾讯网也已经有近 8 亿用户, 新浪微博用户量

收稿日期: 2017-04-22

修回日期: 2017-08-24

网络出版时间: 2017-12-05

基金项目: 国家自然科学基金 (61462069, 61662056); 内蒙古自然科学基金 (2014MS0622, 2015MS0622, 2016MS0609)

作者简介: 顾瑞春 (1982-), 男, 硕士, 讲师, CCF 会员 (35424M), 研究方向为网络通信、数据库技术。

网络出版地址: <http://jns.cnki.net/kcms/detail/61.1450.TP.20171205.0906.054.html>

约为4亿,新晋社交网络 Instagram 用户量也已经超过4亿。

如此多的用户在进行信息交流、转发、评论的同时,也会产生大量的数据。调查显示,国内平均每人每天花在社交网络上的时间,大约为60分钟。由于社交网络具有强大的交互性和实时性,大量用户不断地创建、转发、评论相关信息时,大数据(big data)便随之产生了。全球各大社交网络中每天生成新的数据量约为2.5 EB。深度挖掘与分析这些大数据中潜在的有用信息,成为数据挖掘领域新的研究方向,社交网络中社会关系识别、社会影响力挖掘已经成为数据挖掘研究中新的热点。

社交网络的用户总会受到其他用户的影响,同时也会影响到其他用户。在传统社交网络中,利用高影响力的用户的强大的号召力来进行相应的商业推广或品牌推荐,已经成为商业广告、企划营销的重要手段。高影响力用户的在线言论、行为等网络活动,能够形成社交网络中的主流舆论导向,并可引导其粉丝推动一轮新的舆论热点。社交网络中的用户影响力分析,已经成为目前数据挖掘与大数据研究领域的新方向。

近年来,学术社交网络的兴起,逐渐吸引了众多科研与学术人员的加入,进行科研成果的在线共享、学术问题的线上研讨以及科研项目的线上协作等。如 A-miner. org, SoScholar. com, Academia. edu, Research-Gate. net, ResearcherID. com 等。与其他社交网络相同,学术社交网络同样能够产生大量的数据,这些数据具有更加重要的研究价值和实际意义。

学术社交网络中的用户影响力分析,是以学术社交网络中的海量数据为依据,结合用户的科研领域的相关指数,如科研人员的 H 指数,其发表论文的他引数量、影响因子总和,以及项目经费,获奖级别,成果转化等数据。文中结合用户的各类科研贡献在社交网络中的传播情况,如文章的引用数、转发数、好评数,以及用户的粉丝数量与粉丝级别等多元化数据,对科研人员进行综合的影响力分析,提出一种多元化的学术社交网络用户影响力分析模型—SARank,为现有的科研评价体系建立一种新的参考指标,还能够为科研管理人员提供一套可靠的评判依据。

2 用户影响力分析方法

著名的 Google PageRank 算法^[1]是一种根据网页之间相互的超链接数量来进行网页排名的技术,该算法使用一种基于马尔可夫的随机游走思想来模拟用户浏览网页的行为。其核心思想是:某个网页被越多的优质网页所指向,则该网页的排名就越靠前。其具体计算公式如式(1)所示。

$$R(i) = c \sum_{j \in B(i)} \frac{R(j)}{N(j)}$$

(1)

其中, R 表示要计算的网页的 PageRank 值; B 表示所有指向即将计算排名页面的网页集合,即该页面的链入页面集合; N 表示该页面的链出网页数; c 为常数。

PageRank 算法最初仅是应用在搜索引擎中,用来计算网页排名,但随后,部分学者将 PageRank 算法引用到社交网络中,认为 PageRank 算法是社交网络用户个体影响力分析的基础算法。2009 年, Tunkelang^[2] 将 PageRank 算法应用到著名社交网络 Twitter 中的用户影响力计算中,使用粉丝的影响力来衡量个体用户的影响力,拥有高影响力粉丝的用户越多,且该粉丝关注的其他用户越少,则粉丝对该用户的影响力贡献越大。

与搜索引擎不同,社交网络中的影响力考虑的是某位用户个体,而不是一个静态页面。当然,PageRank 算法并没有考虑到具体个体用户特征参数, Haveliwala 等^[3] 在 PageRank 的基础上,结合社交用户个体特征因素,提出了 Personalized PageRank 算法。在该算法中,作者使用了用户个性化特征向量,如个体对社交网络话题的偏好程度、个体发布信息的新颖程度与敏感程度等^[4]。

针对社交网络用户个性化的问题,不少学者均提出了面向不同属性的影响力分析方法,如 Weng 等^[5] 提出的 TwitterRank 算法,针对知名社交网站 Twitter,根据账户连接结构和用户话题相似性等参数来计算个体在不同领域的影响力。

在研究 PageRank 算法时,研究人员发现某些网页仅仅因为存在时间较长,才获得了较多的指向入链接,反而使其 PageRank 值高于某些较新的页面的问题,通过分析新浪微博中用户转发行为时间间隔分布,通过转发时间间隔来确认粉丝对用户的关注度,认为关注度越高的粉丝对用户的影响力贡献越大。并认为,在同一时刻或同一事件中,粉丝将不同的关注度分配给不同的被关注用户。代表性研究有陈少钦等^[6-7] 提出的基于新浪微博的用户影响力分析模型 WURank 等。

3 学术社交网络用户排名算法

在学术领域,传统的科研人员的学术排名主要根据如下两种方式进行计算:

- (1) 根据科研人员成果质量来计算。如发表科研文章的数量,文章被引数量,以及由被引数量而产生的 H-指数和 G 指数等。
- (2) 根据科研成果所在期刊质量来计算。如发表文章所在期刊的年度影响因子等。
- 但是上述评价方式均存在问题。文章引用次数与

文章发表年限有关,因此很难通过他引次数将真正有影响力的文章分辨出来,而且仅统计引用数量,并不统计施引文章在引用时对该文章的评价信息。一篇文章需要经过较长时间后才会有相应的引用数量的积累;至于所谓的所在期刊的影响因子,更是至少经过 1 年之后,才能评定出该期刊上一年度的平均影响因子,影响因子统计时间不仅慢,而且无法通过影响因子了解该期刊具体单篇文章在相应学术领域的影响力。在 2016 年汤森路透出售了其知识产权和科学信息业务后,影响因子的权威性可能会在未来受到冲击。

随着 Web2.0 技术及社交网络的发展,Priem 等^[8]提出一种多元化科研人员评价体系 Altmetrics^[9],意为使用更多的社交网络参数来进行学术声望评判。Altmetrics 认为,下一步,科研评价指标将会是综合性的多元度量,即将社交网络中的多元化元素融合到科研协作平台中,通过社交网络的相关参数,优化传统的评价指标来形成新的多元化科研绩效计量体系。自从 Altmetrics 提出后,得到了大量科研人员^[10-13]的支持与肯定。国际上对科学研究人员的影响力评价体系已经逐渐从传统的以引用量、H-指数等固态指标为基础的评价系统转向以科研成果的使用(被下载)量、同行评议情况、引用量,以及 Altmetrics 量为基础的创新型综合社会化评价体系。其中 Altmetrics 量包含社交网络中的存储、连接、标签以及评述指标。

基于社交网络的用户影响力分析模型,国内外各大学术社交网络中针对其科研用户也推出各类影响力排名算法,由清华大学唐杰等开发的 Aminer^[14]研究人员社会网络,通过统计科研人员的文章数量、引用数量、H 指数、A 指数、G 指数等信息,生成专家统计信息雷达图,并可分别通过上述指数进行专家排名。截至目前,Aminer 系统已收集了 2.3 亿多论文信息,1.4 亿份研究者信息,7.5 亿论文引用关系,879 万知识实体以及 3 万多学术会议/期刊。吸引了全球 220 多个国家的 276 万多独立 IP 访问。Aminer 系统还集成了自动信息抽取、账号自动关联、重名排歧、专家发现以及跨语言联系等技术,该系统是目前较为先进的高水平科研人员搜索和发现平台。

目前,国际上较为著名的科研社交网络 ResearchGate.net,是一个可以在线分享研究成果、学术著作以及进行讨论的社交平台。其通过一个名为 RG Score 的研究者评分方式对科学家进行排名,RG Score 是一种通过研究人员的成果被同行在线认可程度来确定科研人员学术声誉的多元化度量方式。具体是通过如下几种方式来确定研究人员的 RG Score 值:

(1)学术贡献:研究人员在 ResearchGate.net 上发布自己的文章、预稿、实验结果和数据等。上传数量越

多,RG Score 值越高。

(2)同行互动:高 RG Score 值的同行对某用户的评价,会直接影响该用户的 RG Score 值。

(3)声誉传播:个人学术声誉会在整个社交网络中传播,并随着对社交网络的贡献增加而不断提高 RG Score 值。

RG Score 是一种通过在线同行认可并快速构建学术声誉的科研人员评价体系,现已成为学术领域评判科研人员声望的一个重要指标。

4 SARank 模型

将社交网络的有关技术融入到科研共享平台中,通过社会化网络将科研信息进行在线分享,这种开放型科研共享协作平台,已经成为下一步在线科学研究的发展趋势,目前国内外较为成熟的科研社交网络平台已经不少。用于用户影响力分析的计算模型也较多,但还没有一种有机结合科研领域和社交网络相关指数进行科研用户影响力分析的计算模型。这里介绍的 SARank 就是一种基于科研社交网络的多元化用户影响力分析模型。

SARank 的具体计算模型为:

(1)将科研用户影响力的影响因素分为学术影响参数 A 与社交网络影响参数 S 两部分。

(2)引入 PageRank 算法进行社交网络用户影响力分析,用于分析用户之间相互关注情况;同时引入用户间评论情况,用于不同用户间评论情况分析。

根据 PageRank 公式,SARank 模型中的 S 参数定义如下:

$$R(i) = c \sum_{j \in B(i)} \frac{R(j)}{F(j) + \lambda} \quad (2)$$

其中, R 为要计算的科研用户粉丝关注情况值; B 为该用户的关注数和粉丝数(被关注数)总和; F 为粉丝数; c 为常数。

该模型认为拥有越多高影响力粉丝的用户,该用户的学术影响力值也就越高。 λ ($\lambda = 1$) 为避免 F 过小时产生的偏差而引入的平滑因子。

$$T(i) = \sum_{j \in C(i)} \frac{G(j)}{N(j) + \lambda} \quad (3)$$

其中, T 表示某用户的用户评论情况值; G 表示好评数; N 表示差评数; C 表示所有评论数。

该公式指出,其他用户对某用户的好评越多,该用户的影响力越高;差评越多,影响力越小。为避免 N 比较小时出现对 T 的干扰和过拟合问题,在分母中引入拉普拉斯平滑因子 λ ($\lambda = 1$) 进行平滑处理。

确定社交网络影响参数为用户关注情况与评论情况之和:

$$S(i) = qR(i) + pT(i) \tag{4}$$

为鼓励科研社交网络中用户之间的相互协作与充分交流,让每一位用户都与线上的其他用户建立相互关注的关系,引入参数 $q, q = \frac{B - F}{B}$ ($0 \leq q \leq 1$),用来标识用户关注数与总关注数的比值,作为社交网络影响参数中的关注情况系数。其中 F 为粉丝数, B 为总关注数(关注数与粉丝数之和)。

为避免由于恶意评论造成的不必要影响,此处主要考虑互相关注用户的评论数,即认为双方互为好友的前提下,评论都是客观可信的。因此,引入参数 p , $p = \frac{F_c}{N}$ ($0 \leq p \leq 1$),用来标识用户评论中来自好友的评论占总评论数的比值,作为社交网络影响参数中评论情况系数。其中 F_c 为好友评论数, N 为总评论数。

(3)将学术领域用户学术声誉计算参数定义为 A 。科研领域学术评价影响因子确定为基本影响参数与合作者影响参数两部分。基本影响参数引入用户的 H-指数、总影响因子和所发文章总数三个参数。基本影响参数的具体公式定义为:

$$A_b(i) = \frac{H(i) * I(i)}{P(i)} \tag{5}$$

其中, A 为用户学术影响参数值; H 为 H-指数; I 为影响因子总和; P 为作者所发文章数量。该公式表示,在用户发的论文总数相同的情况下,作者的 H-指数和引用数和总影响因子越高,说明该用户的科研声望值越高。

SARank 将合作者影响力参数引入到研究人员影响力值中,认为文章合作者的影响力会对用户的影响力有较大的影响。最终确定公式为:

$$A(i) = A_b(i) + \sum_{k=1}^H \left(\sum_{j \in U(i)} \frac{U(j)}{L(j)} \right) \tag{6}$$

其中, U 表示合作用户的影响力值; L 表示该用户在文章中的署名位置,第一作者为 1,第二作者为 2,以此类推。由于科研用户的文章以及合作者较多,此处仅考虑用来确定该用户 H 指数的文章中相关用户的合作者影响情况。 H 为用户 i 的 H 指数。

上述公式表示用户的学术影响因子为基本学术影响参数与合作者影响参数之和。合作者影响参数确定为该用户的 H 篇文章的所有合作者影响力之和,单篇文章的合作者用户为合作者的影响力值除以在文章中的署名位置。用户影响力与合作者影响力成正比,与合作者署名位置成反比。

(4)定义科研社交网络中多元化用户影响力模型 SA,公式为:

$$SA(i) = aS(i) + bA(i) \tag{7}$$

其中, S 为数据科研社交网络中用户影响力值; S

表示用户社交网络影响参数值; A 表示学术影响参数值; a 与 b 表示两类影响因子权值, $a + b = 1$ 。

为充分体现社交网络因子在整个 SARank 模型中的重要性,暂时将 a 与 b 均设置为 0.5。

5 实验结果与分析

5.1 数据来源

为测试 SARank 的实际计算情况,又碍于目前大多学术社交网络均不公开 API,因此,实验数据是通过 python 的爬虫框架 Scrapy 从 ResearchGate. net, Aminder. org 以及 SoScholar. com 抓取大量科研人员的相关数据,然后通过 ETL 工具集 petl 来进行数据处理。为保护数据的隐私性,这里隐去科研人员姓名。

进行 SARank 验证的主要步骤分别为:

1. 获取用户数据。确定需要获得的用户数据主要包括:

(1)用户的关注与被关注数据,即该用户关注的用户数和关注该用户的用户数(粉丝数),以及每一关注和被关注用户的关注情况值 R ;

(2)用户的评论数据,即其他用户对该用户的好评数和差评数,以及来自互为好友用户的评论数据;

(3)用户的 H 指数;

(4)用户发表的文章影响因子总和;

(5)用户发表的文章总数;

(6)用户 H(H 指数)篇文章中合作者影响力值。

2. 数据归并。将通过 3 个不同社交网站获取的相应数据进行归并,将同一用户的信息进行合并,去除重复信息。归并时,这里取三个不同网络数值的平均值。

3. 通过 SARank 进行计算,得出用户 SARank 值。

5.2 结果分析

实验一:将用户的 SARank 值和 PageRank 值以及 H 指数进行比较。

PageRank 值由式(2)进行计算,即通过用户的关注数和被关注数计算用户的社交网络排名值,用来表示用户的社交排名。

H 指数为用户归并后的 H 指数平均值,用来表示用户的学术排名。

SARank 值由式(7)进行计算,这里由于用户的 SARank 值与其粉丝以及合作者的相关值有关,因此需要一个逐渐迭代计算的过程,文中暂时仅计算 2 层迭代。

分别对 2 000 用户、5 000 用户和 10 000 用户关于上述 3 个数值的平均值进行了比较,具体见表 1。

由表 1 可以看出,与 H 指数和 PageRank 值一致, SARank 值随人数变化的波动不大,具有较好的稳定性。

表 1 SARank 计算值与 H 指数及 PageRank 值对比

Users	H-index	PageRank	SARank
2 000	21	431	168
5 000	17	415	183
10 000	18	389	175

虽然将三个社交网络数据整合在一起进行计算具有一定的差异性,由于某位科研人员倾向于仅使用某一种社交网络的原因,SARank 中需要获取的某些数值可能无法获得,从而导致部分计算结果出现偏差,通过取 3 个社交网络的 SARank 的平均值,能够较好地避免由于差异性带来的数值偏差。

实验二:将 SARank 的计算值与 Researchgate. net 的 RGScore 值进行比较。

通过在 Researchgate. net 获取的用户数据使用 SARank 模型进行计算后,与 Researchgate. net 的 RGScore 值进行比较。RGScore 是 Researchgate. net 中科研人员的总体贡献分数,主要通过用户上传文章、解决其他用户提问等相关参数进行确定。

该实验采用获取数据中的 5 个用户,使用 SARank 模型进行计算后,与其相应的 RGScore 进行了对比,具体如图 1 所示。

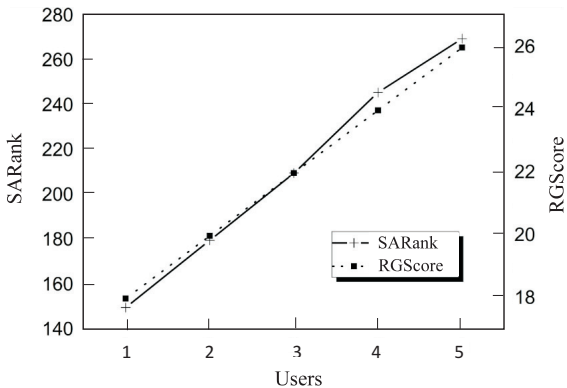


图 1 SARank 计算值与 RGScore 值对比

由图 1 可见,SARank 与 RGScore 值的走向基本一致。第四个用户中的 SARank 值偏高,是因为该用户的社交活跃性较高,S 因子影响了整个 SARank 的值。现实中,如果某位科研用户的社交活动较广,也在一定程度上扩大了其学术影响力。因此,此类现象符合实际情况。

6 结束语

结合用户在社交网络中相关信息的不同参数,对科研用户进行综合的学术影响力分析,提出了一种多元化的学术社交网络用户影响力分析模型——SARank,为现有的科研评价体系建立一种新的参考指标,并为科研管理人员提供一套可靠的评判依据,可为 万方数据

下一步研究提供有力支持。经实验测试,该模型能够得出较为稳定,并具有代表性的结果。

另外,该方法在实际应用中还有部分问题需要完善。例如,如何完善模型参数以优化计算结果;文中仅抓取了 3 个学术社交网络中的数据,仅将这 3 个网络中的数据进行融合,还不能很好地表达科研人员的相关信息;能否将同一科研人员各类其他非学术社交网络中相关信息有机整合到该模型中一并进行学术声誉度量等。这些问题还有待进一步研究。

参考文献:

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

[2] DANIE T. A Twitter analog to PageRank[EB/OL]. (2009-01-13). <http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank>.

[3] HAVELIWALA T,KAMVAR S,JEH G. An analytical comparison of approaches to personalizing PageRank[R]. Stanford;Stanford InfoLab,2003.

[4] 丁兆云,贾 焰,周 斌,等. 社交网络影响力研究综述[J]. 计算机科学,2014,41(1):48-53.

[5] WENG J,LIM E P,JIANG J,et al. TwitterRank:finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. New York,NY,USA:ACM,2010.

[6] 陈少钦,范 磊,李建华. MURank:社交网络用户实时影响力算法[J]. 信息安全与通信保密,2013(3):50-52.

[7] 陈少钦. 基于 PageRank 的社交网络用户实时影响力研究[D]. 上海:上海交通大学,2013.

[8] PRIEM J,TARABORELLI D,GROTH P. Altmetrics;a manifesto[EB/OL]. (2011-09-28). <http://altmetrics.org/manifesto/>.

[9] BHUE S,SINGH K,BISWAL S K. Altmetrics:article level metrics makes easy for user community[J]. Social Science Electronic Publishing,2016,6(2):1-7.

[10] TORRESSALINAS D,CABEZASCLAVIJO A,JIMENEZ-CONTRERAS E. Altmetrics:new indicators for scientific communication in Web 2.0[J]. Comunicar,2013,41(41):53-60.

[11] HOLBROOK J B,BARR K R,BROWN K W. Research impact:we need negative metrics too[J]. Nature,2013,497(7450):439.

[12] PRIEM J. Scholarship:beyond the paper[J]. Nature,2013,495(7442):437-440.

[13] LISTED N. The maze of impact metrics[J]. Nature,2013,502(7471):271.

[14] 唐 杰. AMiner[EB/OL]. (2006-09-06). <http://aminer.org>.