·情报、信息与共享 ·

BlogRank 算法及其在图书馆博客中的应用

邱均平[1] 徐 蓓[1] 李江[2]
(1. 武汉大学科学评价研究中心,武汉,430072; 2. 南京大学信息管理系,南京,210093)

[摘要] BlogRank 算法是基于博客计量学和 PageRank 算法提出的,它对 PageRank 算法中的链接作了实质性链接与非实质性链接的区分。本文简要介绍了 BlogRank 算法产生的背景和 BlogRank 算法在图书馆博客中的计算实例,从链接分析的角度论述了 PageRank 算法在图书馆博客的评价以及在图书馆博客搜索中的应用。

[关键词] PageRank 算法 BlogRank 图书馆博客 博客计量
[中图分类号] G350 [文献标识码] A [文章编号] 1003-2797(2008)01-0068-04

[Abstract] The BlogRank algorithm is proposed based on the PageRank algorithm and Bolgmetrics, and it makes a distinct between the substantive link and non-substantive link. This paper briefly describes the background of the BlogRank algorithm, an example of the calculation process and application in the library blog evaluation and blog search engine.

[Key words] PageRank algorithm BlogRank Library blogs Blogmetrics

1 BlogRank 算法的产生背景

1.1 博客计量学[1]

“博客”来源于英文的“Blog”或“Web Blog”,是一种特别的网络出版和发表文章的方式。2002 年 7 月正式引入中国后,得到了迅速的发展。它以网络日志和超文本链接作为基本的构成要素,通常按照年份和日期进行排列。用户可以通过浏览相关的主题博客来获得所需要的相关资源,也可以通过浏览不同博客对某些信息的评论来拓宽自己的视野。

随着博客传播的越来越广泛,对博客的研究也越来越多,如博客现象研究(即“博客文化”)、博客计量研究(即“博客计量学”)等。在文献[2-5]中,都涉及到博客计量学及博客影响因子的相关介绍。名为“信息大管家”的博客在其日志中提出了博客影响因子的概念,并分别将博客计量学的基本概念、研究基础、研究方法及研究目标作了系统阐述。尽管这篇博客日志并不足以使博客计量成为一门学问——博客计量学,但可以从中看出博客在网络上的发展态势,即寻求文化本体及科学理论基础。事实上,博客计量学的基本概念(如博客影响因子)是依照网络计量学的基本概念提出的,其研究方法、研究目的、研究工具等都与网络计量学基本相同,所以可认为,所谓的“博客计量学”只不过是网络计量学以博客为研究对象的一个应用而已。网络计量学自 1997 年诞生至今仅十余年的历程,“博客计量学”的提出是其研究活跃的表现。

1.2 PageRank 算法的缺陷

PageRank 最早是由 Sergey Brin 和 Lawrence Page 在《大规模超文本网络搜索引擎的剖析》一文中首先提出的 Google 的一种算法,它对网页进行评价,为每个网页赋予一个衡量其重要性的值,并应用于检索结果排序。PageRank 的基本思想主要来自传统的文献计量学中的文献引文分析。

传统的引文分析认为,一篇学术论文的重要性及质量可以通过其他学术论文对其进行引用的数量来衡量,即被其他学术论文引用得越多,则这篇文章就显得越重要。PageRank 应用传统的文献引文分析思想,提出了一个假设,即网页的重要性和质量可以通过其他网页对其超文本链接的数量来衡量。具体来说,假如网页 A 有一个指向网页 B 的链接,则意味着网页 A 认为网页 B 是重要的。假如有 10 个网页指向网页 A,而指向网页 B 的链接却只有 2 个,

[基金项目] 本文系国家自然科学基金项目“网上学术信息的分布与变化规律研究及其应用”(70673071)的研究成果之一
[作者简介] 邱均平,男,1947 年生,教授、博士生导师;徐蓓,女,1983 年生,硕士生;李江,男,1982 年生,博士生。

则说明网页 A 比网页 B 更加重要。但多位学者研究证明,这一假设前提成立的可能性仅为 27%左右,远远低于引文分析中同类假设前提成立的可能性。这主要是因为网络中链接动机更复杂,且结构性链接过多,多数都不能代表"推荐"或"认可",不能代表对被链接页面质量的肯定。因此笔者考虑针对 PageRank 算法的这一缺陷提出 BlogRank 算法。

2 BlogRank 算法概述

1998 年,在 Ingwerson 提出网络影响因子算法的同时,Sergey Brin 和 Lawrence Page 提出了 PageRank 算法[6],公式如下:

PageRank(A) = (1 - D) + D(PageRank(T1)/C(T1) + … + PageRank(Tn)/C(Tn))

其中 PageRank(A)表示给定 Page(A)的 PageRank 得分; D 为阻尼因子,一般设为 0.85; PageRank(T1)表示一个指向 Page(A)的 Page 的 PageRank 得分; C(T1)表示该 Page 所拥有的实质性链接数量; PageRank(Tn)/C(Tn)表示为每一个指向 Page(A)的 Page 重复相同的操作步骤。这是一个基于 Markov 过程的迭代算法,其基本理论是:若 B 网页设置有指向 A 网页的链接(B 为 A 的导入链接)时,说明 B 认为 A 有链接价值,是一个"重要"网页。当 B 网页级别(重要性)比较高时,则 A 网页可从 B 网页这个导入链接分得一定的级别(重要性),并平均分配给 A 网页上的导出链接。一般而言,PageRank 值是由导入链接的数量及其级别(重要性)所决定的。

尽管如此,PageRank 算法并非是完美的。当前,学者们纷纷提出链接分析研究中需对链接类型作区分,以提高链接分析的精度,而 PageRank 算法并未对链接类型作区分。笔者在保留 PageRank 算法中 Markov 迭代过程的同时,对链接作了实质性链接与非实质性链接的区分。实质性链接包括引用链接(主要为网页内容中引用了其他网页内容,并将其设为超链接形式,这种链接类型如同期刊文献中的引用)、兴趣链接(又可分为友情链接、资源链接等)等多种类型;非实质性链接包括非 http 链接、结构性链接以及广告链接等。将实质性链接用于链接分析算法的计算,以 Blog 为例,在 PageRank 算法基础上提出 BlogRank 算法,公式如下:

BlogRank(A) = (1 - D) + D(BlogRank(T1)/C(T1) + … + BlogRank(Tn)/C(Tn))

其中 BlogRank(A)表示给定 Blog(A)的 BlogRank 得分; D 为阻尼因子,一般设为 0.85; BlogRank(T1)表示一个指向 Blog(A)的 Blog 的 BlogRank 得分; C(T1)表示该 Blog 所拥有的实质性链接数量; BlogRank(Tn)/C(Tn)表示为每一个指向 Blog(A)的 Blog 重复相同的操作步骤。仅从公式上看,BlogRank 与 PageRank 相比唯一的区别在于 C(T1)的变化,即由"该 Blog 所拥有的导出链接数量"变为"该 Blog 所拥有的实质性导出链接数量",这一细微变化将导致整个 Blog 群计算过程发生变化。计算 BR(BlogRank)值之前,首先需要对链接类型作出识别,即区分每个链接属实质性链接还是非实质性链接。然后,在 BR 值的计算过程中,计算导出链接数量时,需将非实质性导出链接剔除。

3 BlogRank 算法在图书馆博客中的计算实例

3.1 数据来源

CSSCI 数据库共收录 16 种期刊(图书情报知识、中国图书馆学报、情报学报、大学图书馆学报、图书情报工作、现代图书情报技术、情报资料工作、情报理论与实践、图书馆、图书馆杂志、图书与情报、图书馆理论与实践、情报科学、图书馆论坛、图书馆工作与研究、情报杂志),它们为数据来源刊。

首先,我们提取这 16 种期刊的所有收录论文的所有关键词,并认为这些关键词可以展现整个的研究状况(二八律[7])。从期刊所收录的 27004 篇论文中,我们提取出共 95877 个关键词(28283 种,未做同义或近义归并处理,因为数据量足够大,不会因同义或近义而淹没研究热点的关键词)。

第二步,对关键词按词频高低排序,剔除含义太泛的关键词,如"信息"、"网络"等。

第三步,取 TOP100 关键词后在"百度博客搜索"中按关键词搜索博客,通过一一浏览后,得到 453 个涉及图书情报学学术性内容的博客。

最后,一一打开各个博客,通过 Maxthon 下载博客上的所有链接。

3.2 计算过程

BR 值是本文提出的一个新概念,计算过程较为复杂,具体计算过程可分为以下三步:

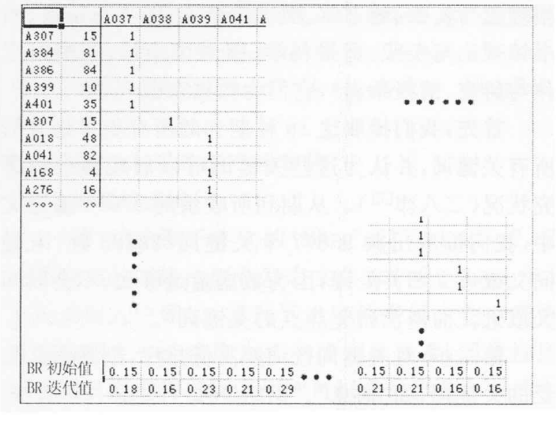
第一步,从所有链接中挑选出实质性链接,并剔除自链,得到链接关系如下页表 1。具体操作是:首先去掉所有非 http://链接,非 http://链接主要为:①javascript;②file://C:/Documents%20and%20Settings/Administrator/My%20Documents/My%20Pictures/LearningLife1.JPG 等;再从剩余的 http://链接中去掉所有结构性链接,如"首页"、"登录"、"总目录"、"索引"、"回复"、"评论"、"收藏"、"加入"、"RSS"、"查看全文"等链接,其识别方式主要是从 URL 上识别,

如 http://youmeng.bokee.com/到 http://youmeng.bokee.com/5983128.html 的链接;从 URL 上看,"youmeng"是该博客主人(Blogger)的名称,所以可以断定 http://youmeng.bokee.com/5983128.html 为该博客的内部链接;最后从剩余的链接中去掉"商城"、"汽车"、"广告服务"等广告链接。

表 1　链接关系表

博客编号	出链	博客编号	出链	博客编号	出链	博客编号	出链
A001	A307	A002	A307	A004	A283	A005	A403
A001	A384	A003	A018	A005	A379	A005	A042
A002	A386	A003	A041	A005	A384	…	…
A002	A399	A003	A168	A005	A386	…	…
A002	A401	A003	A276	A005	A399	…	…

第二步,将表 1 中的链接关系转化为矩阵关系,列在 EXCEL 中,如下图所示:



BR 值计算过程示意图

图中,第一列为博客编号,第二列为该博客上的实质性链接数,第一行是出链中属于 BSI 收录范围的博客编号,单元格中 1 代表从列到行上存在链接关系(列上的单元格指向行上的单元格),最后两行分别是 BR 初始值(根据 BR 值计算公式,初始值赋为 0.15)和 BR 迭代值,BR 迭代值是根据 BR 值计算公式计算而得到的。

第三步,上述的 BR 迭代值只是迭代过程的第一步,最终的 BR 值需要将迭代过程反复执行,以至迭代前后两次的差值极小(在限定的阈值范围内)。事实上,因为 BSI 收录的博客形成的链接网络中,存在较多博客没有入链,这便导致了 BR 值迭代过程中各博客的 BR 值不发生变化,因此,各博客最终的 BR 值就是上图中的各 BR 迭代值。

3.3　结果分析

笔者将涉及的 453 个博客逐一打开,通过浏览博客内容,对博客内容进行关键词标引,对涉及图书馆及图书馆学内容的博客访问量进行统计,并按 BlogRank 值和访问量的大小排序,列出表 2。

表 2　博客的访问量与 BlogRank 值排序

编码	博客名称	访问量	BlogRank 值
A051	山高水长	2507927	0.15283
A387	花生壳	980126	0.16572
A041	图谋博客	493086	0.21105
A037	老槐也博客	459887	0.17798
A068	程焕文的博客	327416	0.24677
A383	超平的博客	244777	0.17933
A042	学林望"道"	174927	0.16040
A237	游园惊梦	171303	0.29177
A377	图林老姜的 BLOG	100610	0.19100
A047	思考的乐趣——雨禅的博客	82060	0.16029
A150	书间道	73431	0.23118
A289	信息空间	72639	0.15000
A379	数图研究笔记	65878	0.15000
A148	蓝天白云	64076	0.19943
A039	年心博客	54313	0.16275
A427	图林丫枝	51656	0.22949
A088	编目精灵 II——On the Fly	51468	0.21365
A046	伊丽莎白的书屋	49810	0.15000
A368	图有其表的 BLOG	46516	0.16144
A382	建中读书	41890	0.17540

(注:访问量的数据截止于 2007 年 10 月 8 日)

很明显,BlogRank 值的排序与访问量的排序存在较大的差异,博客的 BlogRank 值高而其访问量却不一定高。在表 2 中,"山高水长"访问量最高,其博文数仅 76 篇,但是引用却达到 84449 次,最新的博文"中国以图书馆界名人命名的奖学金初步统计"是在 2006 年 12 月发表的,其博客的内容涉及海峡两岸图书资讯学学术研讨会、图书馆事业的发展以及图书馆学的教学与教育的发展等,文章专业性较强,所涉及的内容在图书馆及图书馆学界比较受关注。访问量仅次"山高水长"的是"花生壳",其博文数达到了 417 篇,更新的频率较高,博文分为 12 个类目,内容涉及国内外数字图书馆的的研究进展、现代信息技术以及个人的工作体会和生活感悟等。

访问量在一定程度上说明了博客被关注的热度,但访问量的统计在一定程度上也受到如博客建立时间的长短、博文的数量、博文内容所涉及的范围以及博客的更新频率等因素的影响。表 2 中,博客建立时间最早的是在 2004 年,如"花生壳"和"老槐也博客"的第一篇博文发表的时间都是 2004 年 11 月;大多数的博客建立的时间都是分布在 2005 至 2006 年。

BlogRank 值是从链接分析的角度说明了博客被

关注的程度,是通过其他网页对其超文本链接的数量以及链接网页的重要性来衡量的。BlogRank 值较高的“游园惊梦”,博文数达到 597 篇,分为图林图学、情报科学和杂谈杂评三个类目,内容涉及图书馆事业的管理和发展、图书馆学的思考、国内外图书馆的发展以及个人的工作体会和生活感悟等,博文内容的范围比较广;其博客的链接(除去结构性链接)涉及图书馆学类博客链接、情报科学类博客链接、网上社区、学术批评、英文网志、网游天下和友情链接等 7 个类目。而“程焕文的博客”其 BlogRank 值仅次于“游园惊梦”,其博文涉及图书馆精神、图书馆智慧、图书馆故事、图书馆札记、图书馆疑惑和图书馆百态,总博文数达 240 篇;其博客的链接主要为专业博客推荐、专业网站链接、文化博客和好友的博客链接。

BlogRank 值相近其访问量也会存在明显的差距。如“学林望道”和“思考的乐趣”,其 BlogRank 值相近,两者之间的访问量却存在明显的差距。其中“学林望道”博文数达 400 多篇,内容涉及图书馆事业以及图书馆学的发展等。而“思考的乐趣”博文数为 250 多篇,内容涉及数字图书馆以及图书馆学的研究进展等。

4 BlogRank 算法在图书馆博客评价中的应用

博客因其操作简单、即时发布以及可以根据自己的喜好来组织等特点,为图书馆人发表这一学科领域的最新知识、最新的学术动态以及见解提供了一个很好的平台。图书馆界最早的博客是由美国图书馆员珍妮·利维在1995 年创建的[8]。近几年来,国内图书馆员的博客发展很迅速,表 2 所列都是具有代表性的博客站点。

目前,对于博客的评价主要依据博客访问量(即人气)这一指标。如新浪博客、搜狐博客等,依据博客的访问量以及文章的点击数发布了博客的总排行和分类排行。访问量作为网络中网站、网页人气的衡量指标,在博客评价中固然必不可少,但仅以这一个指标来评价博客,容易导致舞弊。链接分析法作为网络信息资源的定量评价方法,也可用于博客评价。在 Webleon's bolg 中关于《如何对博客进行量化的评价》一文中,提出了利用 Google 反向链接、Technorati 搜索、Bloglines 链接引用等工具来对博客进行定量的评价。可见,在对博客的评价中,对其进行链接分析正受到越来越多的关注。

虽然 BlogRank 的排名在初期与点击量的排名并不吻合,但是经过一段时间的相互影响,两者会越来越靠近。趋于稳定后的 BlogRank 排名将成为真正的博客排名。通过 BlogRank 排名我们可以寻找出图书馆博客中的核心博客,评定博客等级,并绘制博客关系地图。

5 BlogRank 算法在图书馆博客搜索中的应用

当前,图书馆学个人博客几乎都是由作者自己选择建立在不同的博客服务平台上。如:新浪博客、搜狐博客、博客网、Donews 等,其位置比较分散。虽然也出现了一些图书馆学博客聚合,如厦门大学图书馆网志聚合、上海大学图书馆新闻聚合系统等;“程焕文的博客”的作者也建立了图林博客圈,共有 295 个图书情报领域的个人博客加入,但都只是为浏览者提供一个链接,而对于博客文章所涉及的内容却并不是很了解。

搜索引擎是对网络上庞杂的信息资源进行有效的搜集、整理、归类和排序,使其变得有序化,从而便于用户能够快速准确地获取到自己所需要的信息资源。目前绝大多数的搜索引擎都是采用 PageRank 算法与其他算法相结合来确定网页与搜索关键词的匹配程度,即网页与搜索关键词之间的相关度,从而确定网页在检索结果相关度排序时的位置。现在,国外的博客搜索引擎中如 Google Blog Search、Sphere、Feedster 等以及中文的搜索引擎中如 Souyo、BOOSO、奇虎博客搜索、百度博客搜索、中文 RSS 搜索引擎等都是支持关键词检索,其检索结果按相关度排序。本文所提出的 BlogRank 算法对链接作了实质性链接与非实质性链接的区分,去掉了如“首页”、“登录”等结构性链接以及商业广告链接等,只将实质性链接用于链接分析算法的计算,使得网页与搜索关键词之间的相关度更高,检索结果的检准率也会更高。

6 结语

图书馆博客所发表的博文内容大多都是作者本身比较关注和感兴趣的。对于博客身份的不同,专家教授博客记录的多是对学科理论的思考、学术会议的报道、行业热点的评述等内容;图书馆馆员和专业期刊编辑则多记录与工作相关的个人活动、或对于某些问题的看法评论等内容;在校学生以及关注图书馆和图书馆学动态的博客多是对专业理论的思考、生活学习的记录以及个人爱好兴趣的关注等,都具有很强的时效性。通过博客使得各种信息能更为及时有效地传递和交流;而博客所具有的评论功能,使得网友们能够就各种话题展开及时有效深入的交流,从而促进理论、学科和事业的完善和发展。

随着图书馆博客的迅猛发展, (下转第 77 页)

采用 h 指数作为学者评价的参考依据。有理由相信,在不久的将来,h 指数和 h 型指数可能成为重要评价工具和核心评价参数,因此,对 h 指数和 h 型指数进行研究具有重要意义。

致谢:感谢潘有能博士对本文写作的协助。

参考文献

1 Seglen P. O. The skewness of science. Journal of the American Society for Information Science, 1992(9)

2,10 Hirsch J. E. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the USA, 2005(46)

3 Rousseau, R. 著,刘俊婉译. Hirsch 指数研究的新进展. 科学观察,2006(4)

4,17,24 Egghe L. Theory and practice of the g-index. Scientometrics, 2006(1)

5,7,25 Jin, B. et al. The R- and AR-indices: complementing the h-index. Chinese Science Bulletin, 2007(6)

6 Jin B. The AR-index: complementing the h-index. ISSI Newsletter, 2007(1)

8 Kosmulski M. A new Hirsch-type index saves time and works equally well as the original h-index. ISSI Newsletter, 2006(3)

9 Batista P. D. et al. Is it possible to compare researchers with different scientific interests? Scientometrics, 2006(1)

11 Saad G. Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively. Scientometrics, 2006(1)

12 Oppenheim C. Using the h-index to rank influential British researchers in Information Science and Librarianship. Journal of the American Society for Information Science and Technology, 2007(2)

13 Anthony F. J. Van R. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research group. Scientometrics, 67(3):491-502

14 Cronin B, Meho L. Using the h-index to rank influential information scientists. Journal of the American Society for Information Science and Technology, 2006(9)

15 Rousseau 著,刘俊婉译. 案例研究:美国信息学会会刊 h 指数的时间序列变化. 科学观察, 2006(1)

16 Imperial J, Rodriguez-Navarro A. Usefulness of Hirsch's h-index to evaluate scientific research in Spain. Scientometrics, 2007(2)

18 Liang L. M. h-index sequence and h-index matrix: constructions and applications. Scientometrics, 2006(1)

19 万锦堃等. h 指数及其用于学术期刊评价. 评价与管理, 2006(3)

20 姜春林等. H 指数和 G 指数 ——期刊学术影响力评价的新指标. 图书情报工作, 2006(12)

21 叶鹰. h 指数和 h 型指数的机理分析与实证研究导引. 大学图书馆学报, 2007(5)

22 Egghe L, Rousseau R. An informetric model for the Hirsch-index. Scientometrics, 2006(1)

23 Egghe L. Dynamic h-index: the Hirsch index in function of time. Journal of the American Society for Information Science and Technology, 2007(3)

26 Glänzel W. On the h-index ∕ A mathematical approach to a new measure of publication activity and citation impact. Scientometrics, 2006(2)

27 Glänzel 著,刘俊婉译. 也谈 h 指数的机会和局限性. 科学观察, 2006(1)

28 Braun T. et al. A Hirsch-type index for journals. Scientometrics, 2006(1)

29 Moed, H. F. 著,刘俊婉译. 指数构建有创意 用于评价要慎重. 科学观察, 2006(1)

(收稿日期:2007-07-04)

(上接第 71 页)势必会影响到图书馆的服务理念、服务手段、服务内容、管理机制以及图书馆学科的发展。

参考文献

1 引用信息计量学知识发展博客计量学和博客影响因子. http://blog.donews.com/limer/archive/2005/12/24/669596.aspx

2 邱均平,李江. 搜索引擎用于网络影响因子测定时的不一致性比较及原因分析. 情报学报, 2006(6)

3 http://www.gdnetlib.edu.cn/blog/article.asp?id=77 (2007-01-01)

4 孟继红. 新颖独特的引文索引. 四川图书馆学报,1996(6)

5 兰州大学图书馆编. 中文自然科学引文索引:1980-1983. 兰州:兰州大学图书馆,1985.

6 Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web, 1998. http://dbpubs.stanford.edu:8090/pub/showDoc.F(2006-10-26)

7 http://cssci.nju.edu.cn/cssci_qkff2.htm(2007-01-02)

8 http://www.dlf.net.cn/newsshow2.asp?articleid=339

9 李江. 链接分析工具研究\[硕士论文\]. 武汉大学,2007.

10 崔新蕊. 论博客在图书馆创新实践中应用. 情报探索, 2007(5)

(收稿日期:2007-07-18)