CiteRank算法在文献多指标排序中的应用

张 勇[1,2] 杨赛军[2] 黄 华[2]
（1.中国科学技术信息研究所，北京 100038；2.北京万方数据股份有限公司，北京 100038）

摘要：传统PageRank算法用于文献排序时主要关注引证关系，新文献被阅读的概率非常低。本文引入CiteRank算法，结合出版时间、下载次数等指标，提出一种多维度混合排序的方法，应用于万方数据搜索引擎，对3亿篇文献进行多指标混合排序。实证分析结果表明，该排序方法提高了新文献和热门文献被访问的概率，排序结果更加契合用户搜索文献的需求。

关键词：信息检索；文献排序；引证分析；PageRank；CiteRank
中图分类号：TP301 **文献标识码：**A **DOI：**10.3772/j.issn.1674−1544.2021.04.004

Application of CiteRank in Multi-index Literature Ranking

ZHANG Yong[1,2], YANG Saijun[2], HUANG Hua[2]
(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. Wanfang Data Company limited, Beijing 100038)

Abstract: PageRank algorithm only pays attention to the citation relationship when it is used in literature ranking, and the probability of new literatures being read is very low. This paper conducts research from the perspective of multi-dimensional sorting, introduces the CiteRank algorithm, uses data such as publication year, download counts, etc., to perform multi-index mixed sorting on 300 million literatures. The results show that the new sorting method improves the probability of new literatures and popular literatures being visited, and the search result is more in line with users' needs for finding literatures.

Keywords: information retrieval, literature ranking, citation analysis, PageRank, CiteRank

0 引言

信息检索系统能够帮助文献查阅者快速查找目标文献，检索系统在收到用户的检索请求时，根据用户输入的检索词，使用倒排索引技术查找到所有符合用户检索语句的文献，然后使用预先设定的指标对这些文献计算分值并排序返回。面对用户不同的文献搜索需求，检索系统需要结合引证关系、时间因子、实时热度等各类指标，计算出每一篇文献的分值，对命中文献进行实时排序。具体指标包括出版时间、被引用次数、相关度、下载次数等。搜索引擎按照一定的规则对文献进行排序，以帮助用户精准发现所需的文献。

万方数据的搜索引擎收录有近 3 亿篇的文献，包含中外文期刊、学位、会议、专利等，如果仅使用PageRank值进行排序，不是所有文献都

作者简介：张勇（1964—），男，中国科学技术信息研究所助理工程师，研究方向为知识组织、文献计量、文献排序（通信作者）；**杨赛军**（1983—），男，北京万方数据股份有限公司技术研究院总工程师，研究方向为向量搜索、数据挖掘；**黄华**（1970—），男，北京万方数据股份有限公司技术总监，研究方向为智能检索、数据分析、知识组织。

收稿时间：2021 年 3 月 26 日。

能找到引证关系，尤其是新文献，因此无法有效计算PageRank值。如何使用更有效的评分指标对海量文献进行排序，满足用户搜索文献需求，已成为万方数据搜索引擎首要解决的问题。本文研究了基于引证关系的PageRank算法和引入时间衰减因子的CiteRank算法，同时加入其他归一化的指标，包括出版年份、下载次数等，对文献进行评分，以提升新文献和热门文献被阅读的概率。

经典的文献排序算法包括PageRank和CiteRank。PageRank算法的核心是PageRank值，假设用户在浏览网页时，会随着这篇网页的链接引导一直点击进入下一层引用的页面，直到完成浏览关闭页面、或随机打开了一个新的网页。这里有两个假设：数量假设和质量假设[1]。数量假设是指在网页链接图模型中，一个页面节点接收到的其他网页被指向的入度数量越多，则这个页面就会越重要；质量假设是指页面的入度质量不同，质量高的页面会通过链接向其他页面传递更多的权重。基于以上两个假设，PageRank算法起初为网页分配相同的重要性得分，使用递归计算方法去更新各页面的PageRank值，直到该值达到稳定的状态为止。搜索引擎对于不同的查询语句，召回候选页面，优先排序PageRank值高的页面。某一个页面i的PageRank值计算公式如下：

$$PR(p_i)=(1-d)+d\times\sum_{k=1}^{n}\frac{PR(p_j)}{C(p_k)} \qquad (1)$$

在式（1）中，$PR(p_i)$ 表示页面p_i的PR值，$C(p_k)$表示引用页面p_j链出的页面总数，d为阻尼因子。

在使用PageRank算法通过引证关系计算文献分值的时候，有以下两个特点：一是网页链接可以是双向的，而引证关系是单向的，在时间上只能是后发表的文献引用先发表的文献；二是网页链接是动态的，而引文是静态的，文献的引证关系在发布后不再改变。对于发布时间久远的文献，被引次数往往大于新发表的文献。在评估新文献价值方面，孙泽锋等[2]提出了一种改进的文献排序算法，结合PageRank思想以及常用的被引量，并将出版时间作为阻尼因子，使得新文献和旧文献在价值评估时分别有不同的权重。张光前等[3]使用阅读价值衡量一篇文献的重要程度，阅读价值由文献所在期刊、文献作者、文献内容等的重要程度决定。刘松涛[4]提出了一种使用相关强度对引文进行排序的方法，用于对学术期刊文献的引用和被引用关系进行定量分析。

文献检索系统根据用户输入的查询语句，在索引中查询文献，并根据预设的评价指标对查询到的文献进行统一排序。在实际应用中，有的用户更关注新文献和热门文献，而基于PageRank的文献排序算法，更适用于发现年代久的文献，在发现新文献和热门文献方面不占优势。CiteRank算法改善了PageRank的这种局限性，在计算文献权重时引入了时间因素。其计算原理是，对于一个更合理的模型应该是优先从最近的文献开始浏览，并通过引用链接逐步浏览发布时间越来越久的文献。

本文研究了一种多指标混合排序模型，将搜索结果文献相关度得分和PageRank、CiteRank、出版时间、下载次数等指标进行归一化处理，并分配不同的权重，从而对命中文献进行混合排序，使得搜索结果更契合用户的需求。

1　研究方法

本文首先研究了引入时间衰减因子的CiteRank算法对于提升新文献被访问概率的可行性，然后提出使用TF−IDF为命中文献进行相关度评分的方法，最后提出一种基于CiteRank值、文献相关度、下载次数等指标的混合排序算法。

1.1　文献的CiteRank计算

CiteRank模型中定义了两个参数T_{dir}和α。T_{dir}是文献的特征衰减时间，即某一学科领域的热度衰减；α是每一次浏览后满意并退出的概率。那么每次不满意并点击引用链接的概率是1−α[5]。

CiteRank模型的转移矩阵如下：

$$w_{ij}=\begin{cases}\dfrac{1}{k_j^{out}} & \text{if 文献j cites 文献i}\\ 0 & \text{other wise}\end{cases} \qquad (2)$$

k_j^{out}表示文献j的被引用次数。用ρ_i表示最开始时选择文献i的概率：$\vec{\rho}=e^{\frac{-age_i}{T_{dir}}}$，通过初步选择可发现文献的概率为$\vec{\rho}$，点击该文献的引用链接发现文献的概率为：$(1-\alpha)W\cdot\vec{\rho}$。

CiteRank可用式（3）计算得出：

$$\vec{T}=I\cdot\vec{\rho}+(1-\alpha)W\cdot\vec{\rho}+(1-a)^2W^2\cdot\vec{\rho}+\cdots+(1-\alpha)^nW^n\cdot\vec{\rho} \quad (3)$$
$$=\frac{\vec{\rho}-(1-\alpha)^nW^n\cdot\vec{\rho}}{1-(1-\alpha)W}$$

t年前的一篇文献的CiteRank为$T_{tot}(t)$，它主要由以下两部分组成：

一是直接的（direct）访问$T_{dir}(t)$，表示t年前所有文献的初始选择为：

$$T_{dir}(t)=e^{\frac{-age_i}{T_{dir}}} \quad (4)$$

二是间接的（indirect）访问$T_{ind}(t)$，既通过引用列表点击跳转的访问：

$$T_{ind}(t)=(1-\alpha)\int_t^{\infty}T_{tot}(t')P_c(t',t)dt' \quad (5)$$

表明进入t年前的文献的链接可以来自所有可能的中间时间的文献。$P_c(t',t)$表示从t'年前文献中引用t年前文献的比例，这是一个经验值[2]，它可以被近似为$\frac{1}{T_c}e^{\frac{-(t-t')}{T_c}}$。

对$T_{tot}(t)=T_{dir}(t)+T_{ind}(t)$进行傅利叶变换可以得到式（6）：

$$T_{tot}(\omega)=T_{dir}(\omega)+(1-\alpha)T_{tot}(\omega)P_c(\omega) \quad (6)$$

解析$T_{tot}(\omega)$并进行傅利叶反变换得到式（7）：

$$T_{tot}(t)\sim(\tau_c-\tau_{dir})\exp\left(-\frac{t}{\tau_{dir}}\right)+(1-\alpha)\tau_{dir}exp(-\alpha t/\tau_c) \quad (7)$$

从式（7）中可以得出，对于大的α，小的τ_{dir}表示最新的文献会被更大的概率访问；对于小的α，大的τ_{dir}表示会相应提高旧文献的访问量。

1.2　文献相关度得分

文献相关度得分是评价搜索结果与查询语句匹配程度的重要指标。搜索引擎根据查询语句与命中文献的匹配程度计算相关度得分，但并不是所有的文献都包含查询语句中的词，并且每个词的重要程度不同，因此一个文献的相关度评分取决于每个查询语句在文献中的权重[6]。

万方数据的文献在索引过程中使用了倒排序索引技术[7]，搜索引擎使用布尔模型（Boolean model）查找与输入词条相匹配的文献，并用TF/IDF（term frequency/inverse document frequency）公式来计算文献相关度得分。搜索引擎使用TF表示词条t在文献d中出现的频率，即$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$。如果词或短句在某一篇文献中出现的频率比较高，并且在其他文献中很少出现，则可以认为该词或者短句具备很好的类别区分能力，适用于解决分类问题，一些通用的词语对于文献主题并没有太大的作用，反倒是一些出现频率非常少的词才能够更多地表达文献的主题，所以仅仅使用词频是不合适的。因此在计算文献相关度得分时，必须考虑词的权重。一个词预测主题的能力越强，则权重越大；反之，权重越小。

在倒排序索引中，如果文献的一些词只是在少数几篇文献中出现，那么这些词对文献主题的贡献作用会很大，这些词和短句的权重应该被设计得大一些。IDF就是词的权重，即$idf_i=\lg\frac{|D|}{|\{j:t_i\in d_j\}|}$，此时IDF为逆向文献频率（inverse document frequency），如果包含词条t的文献少，则IDF会越大，表明词条t具有非常好的类别区分能力。如果一些文献中含有词条t的文献数量为m，而其他文献中含有词条t的文献数量为k，那么可以得出所有包含t的文献数量为n = m + k。当m越大时，n也会越大，按照IDF公式得到的IDF值就会越小，说明该词条t的区分类别的能力不强。但是在实际情况中，如果一个词条或短句在一个类别的文献中经常出现，则可以说明这个词条能够很好地表示这个类别的文本特征，这些词条应该被赋予更高的权重，并用来作为该类别文本的特征词，用以区别于其他类别的文献。

在给定的文献里，词频（term frequency，TF）表示的是某一个给定的词语或短句在该文献中出现的频率。这个数值是对词数量的（term count）归一化，用于防止该值偏向比较文字较多的文献。但是在实际使用过程中，会根据不同的命中属性（标题、摘要、关键词）预先设计不同的权重w_z。如标题命中时权重会更高，摘要则会赋予相对低一些的权重。因此，可得词i命中文献j的z字段时分值为式（8）：

$$S_{i,z}=tf_{i,j}\times idf_i\times w_z=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\lg\frac{|D|}{\left|\left\{j:t_i\in d_j\right\}\right|}\times w_z \quad (8)$$

文献的相关度得分S_r即为所有词在文献所有字段中的分值之和$S_r=\sum S_{i,z}$

1.3　混合排序原理

文献的最终分值将由相关度得分S_r、CiteRank分值S_c、下载次数S_p等指标乘以相应的权重得出。

$$S=w_r\times S_r+w_c\times S_c+w_p\times S_p \quad (9)$$

其中，S_c为$T_{tot}(t)$归一化后的结果，下载次数S_p可以在文献进入索引库时预先计算为0到100之间的值，这两个指标根据数据进入索引的周期，会定期进行更新。这种预先计算分值的方式，可以避免在排序时进行大量的实时计算，使用存储空间换取计算时间，节省了大量的计算资源，提升了排序性能。在用户使用搜索引擎时，输入检索词，检索集群会为数百万命中文献进行排序。为了有效节省内存资源，将排序时的资源消耗限定在合理的范围内，搜索引擎只在内存中保存分值最大的N个值（top N）。

2　实验

2.1　数据集

本文以万方数据收录的文献和2020年被下载数据作为实验数据，对混合排序算法进行验证，文献数据集合符合标准定义[8]，包含标题、摘要、关键词、刊物收录情况、作者、作者单位等字段[9]，以及文献的引证关系[10]。本文在计算CiteRank时，使用了不同的α和τ_{dir}来预测文献被访问到的概率，并由此比对了PageRank和CiteRank分值的区别。最后基于万方数据的用户搜索行为日志，分析使用混合排序前后，用户对搜索结果前三条的点击率变化，来验证混合排序算法的有效性。

2.2　CiteRank参数估计

使用公式（7）可以计算各出版年份下文献被访问到的概率，本文使用了3组不同的τ_{dir}和α参数进行计算，对比结果如图1所示，x轴表示文献发布年份与2020年之间的差距，数字越小表示文献越新。由数据结果可以得出：对于小的τ_{dir}，大的α，表示最新的文献相比旧的文献有更高的概率被访问到；对于大的τ_{dir}，小的α，表示则会相应提高旧文献被访问的概率。为了契合部分用户访问新文献的需求，万方数据使用了CiteRank作为文献排序的一个指标。通过对比3组参数的计算结果，本文选择τ_{dir}=2.1，α=0.7作为计算CiteRank时的参数，从而可以做到以更大的概率将较新的文献排在前面。

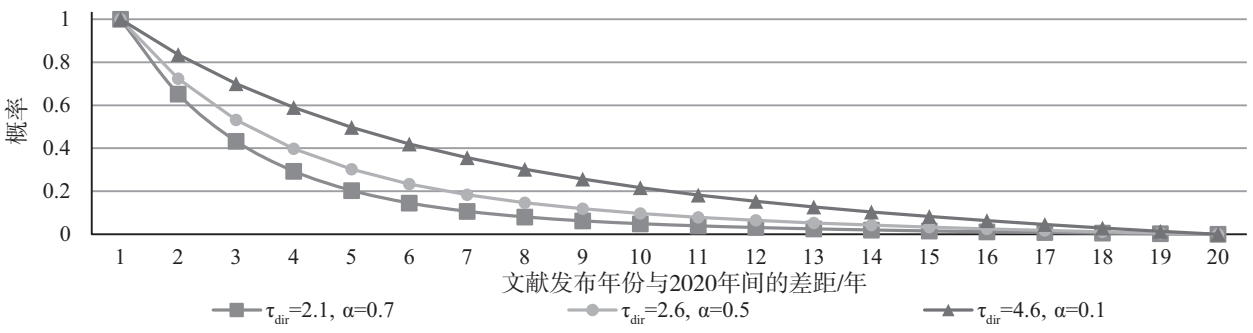

图1　α和τ_{dir}参数模拟各年份文献被访问的概率

2.3　CiteRank和PageRank对比

本文选择使用万方数据收录的2000年至2020年文献元数据和这些文献之间的引用关系作为数据集，分别计算每一篇文献的PageRank和CiteRank值。表1截取了部分近5年的数据，包含典型的权威文献和流行文献，对比了同一篇文献的CiteRank和PageRank。由于PageRank值在计算过程中仅使用文献之间的引用关系，体现了文献之间权重的贡献值，反映了文献的经典权威程度；而CiteRank为文献的贡献关系加入了出版时间作为衰减因子，出版时间越久远的文献获得的贡献值越少，反之则会获得更大的贡献值，从而反映了文献的流行度。

表1中文献按照CiteRank值从高到低进行排序，文献1和文献2的发布年份大于2019，被引数分别为360和556，文献3的年份为2015年，被引数为4 832，虽然文献3的被引次数远远大于文献1和文献2，但是根据CiteRank算法特性，年份相对久远的文献获得的贡献会根据年份进行相应的衰减，所以文献3的CiteRank值会低于文献1和文献2。对于文献的PageRank值，文献3、文献9、文献10被引数最高，PageRank值也是最高，仅仅反映了的文献的被引关系。从表1得出，PageRank可以用于经典文献优先排序，优先列出权威的文献（不考虑出版年份），CiteRank则可以用于新文献优先排序（默认排序），优先列出新的文献（考虑出版年份）。

2.4　历年出版文献被访问次数

本文统计了各出版年的文献在2020年被访问次数。数据显示，用户更倾向于选择访问新的文献，这与万方数据选择使用CiteRank作为默认的排序指标相符合，使用CiteRank可以为用户提供偏向新文献的排序结果，如图2所示。

2.5　使用混合排序前后点击率对比

本文使用单一指标排序和混合指标排序两种场景下10天内用户对搜索结果的点击行为数据进行分析。分别统计出两种排序场景下用户点击搜索结果的总点击数，以及点击前三条结果的点击数（top 3点击数），从而计算出点击率（top 3点击数/总点击数）。表2显示了这两种排序场景下的数据对比，可以得出混合排序算法能显著提升用户点击排序结果top 3文献的几率，该算

表1　PageRank值和CiteRank值对比

顺序	文献ID	CiteRank	PageRank	出版时间/年	被引数
1	zhlxbx202002003	97.988 7	1.409	2020	360
2	zgxhzz201903001	68.376 3	1.192	2019	556
3	zhsjk201504002	61.562 5	12.353 2	2015	4 832
4	zhzl201901008	58.528	1.380 8	2019	525
5	zgsynkzz201804009	53.470 2	1.660 4	2018	1 117
6	zhsjk201809005	49.466 2	1.472 8	2018	718
7	zgazyj201801001	49.061 5	1.934 7	2018	558
8	zhlxbx202002002	47.605 8	0.702	2020	158
9	zgjyxk201610002	44.972 3	6.398 2	2016	1 556
10	kcjcjf201605011	44.753 6	6.067	2016	1 101
11	zghgxyj201602013	38.720 1	2.565 7	2016	717
12	wyj201603001	37.384 1	1.990 9	2016	614
13	zhyx201814007	20.880 4	0.656	2018	258
14	sxlljy−s201803011	19.718 8	0.835 1	2018	233
15	qius201609001	19.678 2	4.290 8	2016	530
16	qius201701001	19.384 5	1.317 5	2017	191
17	qius201501001	18.917	4.693 5	2015	609

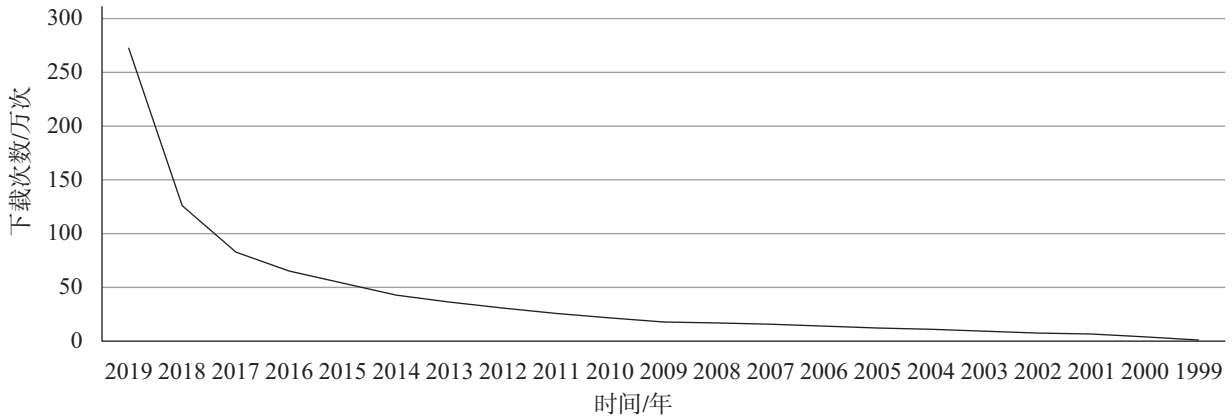

图 2　历年文献在 2020 年被下载次数统计

表 2　各排序场景下用户点击率对比

排序场景	总点击数/万次	top 3 点击数/万次	点击率/%
单一指标	519	213	41
混合指标	530	254	48

法在搜索引擎中更加契合用户默认的文献搜索需求。

3　研究结论

万方数据搜索引擎中收录有近 3 亿篇的文献，单纯地使用PageRank或CiteRank对文献进行排序，会导致权重倾向于单一的指标，不能精准地契合大数据环境下用户对新文献和热门文献的搜索需求。本文研究了基于CiteRank的混合排序算法，引入出版时间、下载次数等归一化指标进行加权平均。实验结果表明，混合排序算法能够提升新文献和热门文献在排序上的优势。最后通过分析用户的搜索行为数据，进一步验证了在搜索引擎中应用基于CiteRank的混合排序算法更能契合用户搜索文献的需求。

参考文献

[1]　李仲谋, 刘凯, 王创维.一种新的基于PageRank算法的学术文献影响力评价方法[J].数学建模及其应用, 2013(2)：46−49.

[2]　孙泽锋, 周洁, 李忠义.基于PageRank改进的文献价值排序算法[J].首都师范大学学报(自然科学版), 2020, 41(5): 1−4.

[3]　张光前, 刘欣, 冯永琴.基于阅读价值的科技文献排序方法研究[J].情报学报, 2009, 28(6): 844−850.

[4]　刘松涛.基于引文排序的科技文献检索初探[J].制造业自动化, 2010, 32(10): 129−131.

[5]　WALKER D, XIE H, YAN K K, et al. Ranking scientific publications using a simple model of network traffic[J]. Journal of statistical mechanics: theory and experiment, 2007(6): 6−10.

[6]　施聪莺, 徐朝军, 杨晓江.TFIDF算法研究综述[J].计算机应用, 2009, 29(z1): 167−170, 180.

[7]　胡宏伟, 虞萍, 周南, 等.基于Lucene的文献资料全文检索系统的设计与实现[J].重庆理工大学学报(自然科学版), 2014(11): 77−83.

[8]　冯项云, 肖珑, 廖三三, 等.国外常用元数据标准比较研究[J].大学图书馆学报, 2001(4): 15.

[9]　陈彩红.国内外元数据标准宏观比较研究[J].河北科技图苑, 2011(1): 66−67.

[10]　马袁燕. 面向发现服务的文献元数据集成整合研究[J].图书馆, 2019(1): 79.