# 基于链接分析的科学文献个性化排序算法*

柳泉波，许 骏
（华南师范大学教育信息技术学院，广东 广州 510630）

**摘 要：** 首先分析CiteSeer引文网络的结构特征，然后深入讨论PageRank算法的本质，并在此基础上给出一种科学文献个性化排序算法；最后将该算法应用于CiteSeer引文网络，并对实验结果进行分析。个性化向量的计算是实现个性化排序的关键，本文给出等概率、伪被引次数和带时间约束的伪被引次数3种计算方法。

**关键词：** 引文网络；链接分析；排序算法；PageRank；个性化
**中图分类号：** TP391.3 **文献标识码：** A **文章编号：** 0529-6579（2008）06-0087-06

## 1 引言

基于彼此间的引用关系，大量的科学文献共同构成了一个网状的信息空间，即引文网络。为了方便讨论，作表1的符号约定。

表1 符号约定

Tab. 1 Notations

| | |
|---|---|
| $G=<V,E>$ | 有向图,引文网络的抽象模型 |
| $V$ | 端点集合 |
| $\|V\|$ | 端点个数 |
| $E$ | 有向边集合 |
| $\|E\|$ | 有向边条数 |
| $In(i)$ | 所有引用$i$的端点 |
| $\|In(i)\|$ | 端点$i$的入度 |
| $Out(i)$ | $i$引用的所有端点 |
| $\|Out(i)\|$ | 端点$i$的出度 |

引文分析[1]是指利用引用频率和模式等信息，定量地分析文献集合（如学术期刊和会议）、单篇文献或者科研人员的影响（重要程度）。例如，期刊影响因子[2]是一种重要的学术期刊评价指标。引文分析实现了科学文献按照重要程度高低的排序。

类似地，Web页面通过超链接形成一个巨型的Web图，它同样可以抽象为一个有向图。链接分析[3]是指根据超链接结构计算Web页面的相对重要程度，从而实现Web搜索结果的优化排序。

HITS[4]是一种与查询相关的链接分析算法，它首先把查询请求提交给常用的搜索引擎，选取搜索结果的前$n$个页面作为根集。接下来，添加新页面到根集，最终形成基本集。这些新增页面的特点是与根集中的某个页面存在引用或者被引用关系。最后，对于基本集中的每一个页面$i$，利用下面的公式分别计算页面的权威值$a_i$和中枢值$h_i$：

$$a_i = \sum_{\forall j \in In(i)} h_j, \qquad h_i = \sum_{\forall j \in Out(i)} a_j$$

权威值大的页面包含了与查询相关的丰富信息，而中枢值大的页面则包含了很多超链接，它们分别指向高权威值的页面。

PageRank[5]则是一种与查询无关的链接分析算法。对于每一个页面$i$，采用公式（1）计算其PageRank得分$q_i$（简称PR值）：

$$q_i = (1-\mu) + \mu \sum_{\forall j \in In(i)} \frac{q_j}{|Out(j)|} \qquad (1)$$

其中$\mu$是一个松弛系数，通常设定$\mu = 0.85$。

各种链接分析算法在Web搜索领域取得了巨大的成功，而引文网络和Web图又有很大的相似性，一种很自然的想法是把上述链接分析算法应用于引文网络，实现科学文献的排序。根据HITS算法的原理，如果一篇文献具有较高的权威值，意味着必须有高中枢值的文献引用该文献；如果一篇文献引用权威值较低的其他文献，那么它的中枢值将受到“惩罚”。以上因素决定了HITS算法并不适合直接作为一种引文分析方法，而PageRank算法不存在类似问题。文献PR值是一种相对“客观”的度量——对于同一个引文网络，不同用户将获得相同的文献PR值。在实际应用中，科研人员希望实现对文献的个性化排序，即排序结果能够反映科研人员的研究兴趣等个人偏好。

与Web图相比，引文网络具有下列两个显著

* 收稿日期：2008-9-10
基金项目：国家自然科学基金资助项目（70671044）；广东省科技计划基金资助项目（2006B15001004）
作者简介：柳泉波（1976年生），男，博士，讲师；通讯联系人：许骏；E-mail：xuj@scnu.edu.cn

特征：①引用关系具有稳定性：正式发表的文献对其他文献的引用不会再发生改变；②文献引用具有时间效应：一篇文献只能引用以前已正式发表的其他文献，引文网络不包含有向环；一篇文献的被引次数随着时间的推移而增长，最后基本稳定在某一个数值。

为了满足科研人员的个性化需求，本文以PageRank算法为基础，提出一种适合引文网络特点的个性化排序算法。

## 2 引文网络特征分析

本研究采用CiteSeer① 提供的文献元数据构建引文网络。首先，从数据文件中抽取每篇文献的关键信息，包括标识符、作者、发表时间以及引用的其他文献的标识符等；接下来，根据引用关系构造引文网络的邻接矩阵表示，在此过程中，将比较文献及其引用文献的发表时间，确保引文网络不会出现有向环。最终得到的引文网络包含716800篇文献，1331934次引用。

（1）度分布

度分布包括入度分布和出度分布两种，入度表示文献的被引次数，而出度表示文献引用其他文献的次数。CiteSeer引文网络的入度满足幂率分布：入度为 $m$ 的所有文献占总文献数量的百分比与 $1/(m^{\lambda})$ 成正比，$\lambda>1$。分别求出入度值及其对应文献所占比例的对数值，然后进行一次多项式拟合，得到的直线斜率就是幂率分布的指数。入度分布的幂率指数 $\lambda=2.08$，如图1所示。

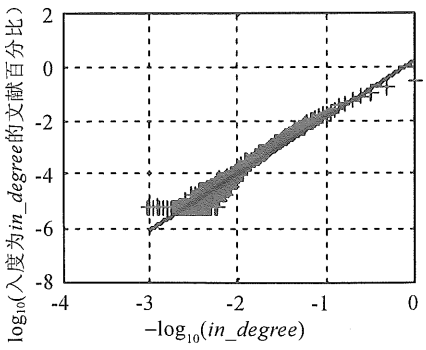


图1 CiteSeer引文网络的入度分布，幂率指数是2.08

Fig. 1 The in-degree distribution in the citation graph subscribes to the power law with exponent = 2.08

采用相同的处理方法，可以看到CiteSeer引文网络的端点出度同样满足幂率分布，幂率指数 $\lambda=2.98$，如图2所示。

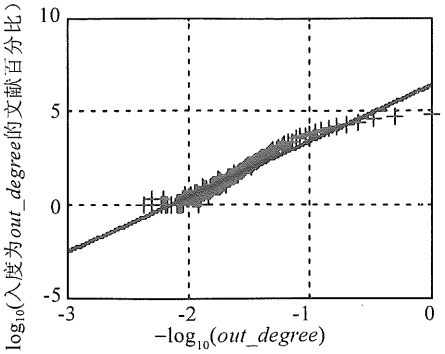


图2 CiteSeer引文网络的出度分布，幂率指数是2.98

Fig. 2 The out-degree distribution in the citation graph subscribes to the power law with exponent = 2.98

Chen和Xie的研究[6]表明：CiteSeer引文网络的端点入度满足幂率分布，幂率指数是2.3；端点出度并不满足幂率分布。上述结论与本研究的结果不同，原因在于文献［6］的研究对象是与关键词“citation analysis”对应的引文子图，该子图仅包含4401篇文献和9959次引用。

（2）连通性

引文网络是一个有向图，它的强连通部件 $SCC$ 是指满足如下条件的最大子图：任取一对端点 $<u, v>\in SCC$，必然存在一条从 $u$ 到 $v$ 的有向路径。实验结果证明：CiteSeer引文网络共有716800个强连通部件，且大小均为1。这是CiteSeer引文网络不包含有向环所导致的必然结果。

科研人员检索到感兴趣的文献后，会根据引用和被引关系进一步查找其他文献。如果把CiteSeer引文网络的所有有向边变成无向边（包含了引用和被引两种关系），此时得到的最大连通子图被称为弱连通部件，如表2所示。

表2 CiteSeer无向图的连通部件

Tab. 2 The results of weakly connected component experiments on the citation graph

| 序号 | 1 | 2 | 3 |
|---|---|---|---|
| 连通部件大小 | 326595 | 48 | 47 |
| 百分比/% | 45.56 | 0.0067 | 0.0066 |

CiteSeer引文网络含有一个很大的弱连通部件，它包含的科学文献大约占文献总数的45.56%。同时，引文网络还包含378211个小连通部件，其大小从几十个到几个不等。

（3）引用次数的时间变化规律

把时间划分成51个区间。第一个区间表示一

① CiteSeer元数据：http://citeseer.ist.psu.edu/oai.html

年以内，第二个区间表示一年到两年之间，……，最后一个区间表示 50 年以上。根据文献及其引用文献的发表时间间隔，计算落在每一个时间区间内的引用次数，结果如图 3 所示。

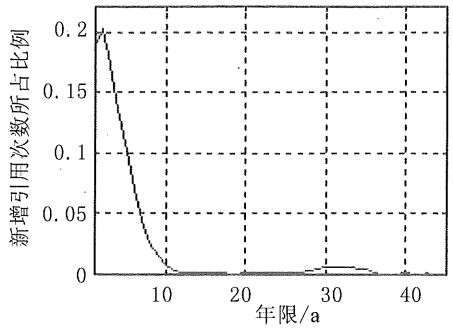


图 3　新增引用次数所占比例与年限的关系

Fig. 3　Relation between ratio of newly accrued citations and years

图 3 表明，新增引用次数在文献发表后一到两年内达到最高，所占比例约为 20.22%；然后逐年下降；10 年后，新增引用所占比例稳定在一个比较小的数值。使用引用次数作为文献重要性的度量，必须考虑到时间因素的影响。

## 3　个性化排序算法

首先介绍 PageRank 算法的基本思想，然后根据文献检索的过程模型，提出 3 种计算个性化向量的方法，最终实现文献的个性化排序。

（1）PageRank 算法

引文网络可抽象为一个有向图 $G = <V, E>$，其中端点集合 $V$ 代表所有的科学文献，有向边集合 $E$ 代表了所有的引用链接。有向图 $G$ 可进一步表示为一个 $n \times n$ 的稀疏矩阵 $\boldsymbol{H}$，其中 $n = |V|$。任取 $h_{ij} \in \boldsymbol{H}$，如果文献 $i$ 引用了文献则 $j$，则 $h_{ij} = 1$；否则，$h_{ij} = 0$。任取一个端点 $i \in V$，它的出度 $|Out(i)| = \sum_j h_{ij}$。定义有向图 $G$ 的转移矩阵 $\boldsymbol{P}$：任取 $p_{ij} \in \boldsymbol{P}$，如果 $|Out(i)| > 0$，$p_{ij} = h_{ij}/|Out(i)|$；如果 $|Out(i)| = 0$，$p_{ij} = 0$。显然，转移矩阵 $\boldsymbol{P}$ 具有下列特点：每一个行向量的各个分量之和要么是 1（称之为行随机的），要么是 0。

科研人员检索文献的过程满足“随机冲浪模型”。假设某位科研人员沿着引文网络的链接进行文献查找。在第 $k$ 步，他检索到了文献 $i$；在第 $k+1$ 步，他等概率地从 $Out(i)$ 中选取某一篇文献 $j$。从这个意义上讲，所有文献共同组成了某个马尔科夫链的 $n$ 个状态，$\boldsymbol{P}$ 是该马尔科夫链的转移矩阵。用 $q_i(k)$ 表示科研人员在第 $k$ 步检索到文献 $i$ 的概率，于是得到

$$q_j(k+1) = \sum_{i \in In(j)} q_i(k) / |Out(i)| = \sum_{i \in V} p_{ij} q_i(k)$$

即在第 $k+1$ 步，所有文献被检索到的概率分布

$$\boldsymbol{q}(k+1) = \boldsymbol{P}\boldsymbol{q}(k) \tag{2}$$

所有文献 PR 值共同组成一个向量，该向量是公式（2）的一个驻点。然而，公式（2）存在一个问题：转移矩阵 $\boldsymbol{P}$ 含有取值为 0 的行向量，即引文网络包含这样的文献，它没有引用其他任何文献。具有上述特征的文献端点被称为悬挂端点。悬挂端点是一个吸引子：在检索过程中，到达该端点后再也不会离开，这会导致公式（2）的求解不稳定。

要消除悬挂端点的不利影响，一种方法是直接从引文网络中删除悬挂端点，但这有可能造成新的悬挂端点；另外一种办法是修改转移矩阵 $\boldsymbol{P}$，添加从悬挂端点到所有端点的人工链接。用一个 $n \times 1$ 列向量 $\boldsymbol{d}$ 来标识悬挂端点，如果端点 $i$ 是悬挂端点，则 $d_i = 1$；否则，$d_i = 0$。令一个 $1 \times n$ 行向量 $w$ 表示访问所有端点的某个概率分布（$\sum w_i = 1$），例如说均匀分布，即 $w = (1/n, 1/n, \cdots, 1/n)$。于是得到修改后的转移矩阵

$$\boldsymbol{S} = \boldsymbol{P} + \boldsymbol{d}w$$

显然，新转移矩阵 $\boldsymbol{S}$ 的每一个行和都是 1，因此 $\boldsymbol{S}$ 是一个非负随机矩阵，它也真正成为马尔科夫链的转移矩阵。

以转移矩阵 $\boldsymbol{S}$ 为基础，构造 Google 矩阵

$$\boldsymbol{G} = \mu \boldsymbol{S} + (1-\mu)\boldsymbol{e}v,$$

其中 $\mu$ 是一个标量且 $0 \leqslant \mu < 1$，$e$ 是分量取值均为 1 的列向量，$v$ 是一个表征概率分布的行向量，通常称之为个性化向量或瞬间移动向量。在很多研究中，$0.85 \leqslant \mu \leqslant 0.99$，$v = (1/n, 1/n, \cdots, 1/n)$。于是公式（2）变成

$$\boldsymbol{q} = \boldsymbol{G}\boldsymbol{q} \tag{3}$$

其中，$\boldsymbol{q}$ 是所有文献对应的 PR 向量。

下面证明公式（3）存在唯一的解。矩阵 $\boldsymbol{G}$ 是两个非负随机矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{e}v$ 的凸性组合，所以 $\boldsymbol{G}$ 也是一个非负随机矩阵，即所有的行和以及列和均为 1。矩阵 $\boldsymbol{G}$ 是不可约的①，即它对应一个强连通的有向图。对于非负不可约矩阵 $\boldsymbol{G}$，根据 Frobenius 定理[7]，矩阵的谱半径 $\rho(\boldsymbol{G}) > 0$ 是 $\boldsymbol{G}$ 的特征值；存在一个正向量 $x$，使得下面的等式成立：

$$\boldsymbol{G}x = \rho(\boldsymbol{G})x \tag{4}$$

① 非负矩阵 $\boldsymbol{A} \in \boldsymbol{R}^n$ 是不可约的，当且仅当 $(\boldsymbol{I} + \boldsymbol{A})^{n-1} > 0$，其中 $\boldsymbol{I}$ 是恒等矩阵；$\boldsymbol{A} \in \boldsymbol{R}^n$ 是本原矩阵，当且仅当存在某个 $m \geqslant 1$，使得 $\boldsymbol{A}^m > 0$。

非负矩阵的谱半径介于最小行（或列）和与最大行（或列）和之间，于是有 $\rho(G)=1$，代入公式（4），得到

$$\boldsymbol{Gx}=\boldsymbol{x}$$

由此可见公式（3）有解。同时，$\boldsymbol{G}$ 还是一个本原矩阵，这决定了公式（3）的解是唯一的。与 $\boldsymbol{q}$ 的初始值无关，$\boldsymbol{q}$ 最终会收敛于一个稳定的文献 PR 值向量。

从初始转移矩阵 $\boldsymbol{P}$ 到最终的 Google 矩阵 $\boldsymbol{G}$，这种数学处理不仅保证了解的存在性和唯一性，它还很好地反映了现实生活中人们浏览网页或者检索文献的过程。假设在第 $k$ 步，科研人员检索到文献 $i$。接下来，有两种可能的操作：①按照概率分布 $\boldsymbol{S}_i$ 选择一篇被 $i$ 引用的文献；②按照概率分布 $\boldsymbol{v}$ 随机地选择一篇文献。科研人员执行操作①的概率是 $\mu$，执行操作②的概率是 $(1-\mu)$。要实现表征科研人员个人偏好的个性化排序算法，关键在于个性化向量 $v$ 的设定。

（2）确定个性化向量

科研人员检索文献的流程主要包括3个步骤。

**步骤1**：科研人员检索 Google 学术搜索、Scopus、ISI WoS 及 CiterSeer 等综合数据库，或者是 ACM DL、IEEE Xplore 等专业数据库，对搜索结果进行评级，例如分成“非常相关”、“中等相关”和“一般相关/不相关”三类，分别记作 $V_h$、$V_m$ 和 $V_l$。用 $V_u=V-V_h-V_m-V_l$ 表示科研人员没有评级的所有其他文献。相关度较高的文献共同组成了初始的核心文献集 $V_h\cup V_m$。

**步骤2**：科研人员执行下列两种操作之一：①从核心文献集移出某一篇文献，重点关注它引用以及引用它的其他文献。对结果文献进行评价，如果相关度较高，则把文献添加到核心文献集；②重新检索各种文献库，把相关度高的新文献添加到核心文献集。

**步骤3**：重复步骤2，直到获得满意的文献检索结果为止。

显然，上述文献检索流程满足“随机冲浪模型”，引文网络对应的 Google 矩阵 $\boldsymbol{G}$ 能够很好地表征检索过程。在步骤1中，文献等级是由科研人员设定的，它反映了科研人员的个人兴趣，应该体现在个性化向量 $v$ 的取值上；而步骤2涉及的文献评级，则是由 PageRank 算法经迭代计算而得。

与文献总量相比，科研人员在步骤1中能够给出明确评级的文献数量非常少，一般从几十到几百不等。令 $v_i$ 表示在个性化向量 $v$ 中与文献 $i\in V$ 对应的分量，即文献 $i$ 的被检概率。对于任意一篇未被明确评级的文献 $j\in V_u$，如何确定 $v_j$ 的取值，进而确定个性化向量 $v$ 的取值，这是实现个性化排序的关键所在。本文提出3种计算 $\boldsymbol{v}$ 的方法，包括等概率法、伪被引次数法和带时间约束的伪被引次数法。

等概率法是指为未被评级的文献与不相关文献设定相等的被检概率，然后确定“非常相关”、“中等相关”和“一般相关/不相关”三类文献被检概率的比例关系，例如 100∶10∶1。求解下面的方程式

$$\sum_{i\in V_h}v_i+\sum_{j\in V_m}v_j+\sum_{k\in V_l}v_k+\sum_{l\in V_u}v_l=$$
$$100v\cdot|V_h|+10v\cdot|V_m|+v\cdot|V_l\cup V_u|=1 \tag{5}$$

计算得到的 $v$ 是不相关以及未被评级文献的被检概率，而中等相关和非常相关文献的被检概率分别是 $v$ 的10倍和100倍。

伪被引次数法是指根据被引次数确定未被评级文献的被检概率的大小，同时把不相关文献视为被引次数为0的文献；把中等相关文献视为被引次数较高的文献，例如100次；把非常相关文献视为被引次数非常高的文献，例如500次。用 $cc(i)$ 表示文献 $i$ 的真实被引次数，则可以定义文献 $i$ 的伪被引次数

$$pcc(i)=\begin{cases}cc(i) & i\in V_u\\ \max(500,cc(i)) & i\in V_h\\ \max(100,cc(i)) & i\in V_m\\ 0 & i\in V_l\end{cases} \tag{6}$$

于是，文献 $i$ 的被检概率

$$v_i=pcc(i)\Big/\sum_{j\in V}pcc(j) \tag{7}$$

所有 $v_i$ 一起构成个性化向量 $v$。

图3表明，文献的被引次数与其已发表年限具有相关性，一篇文献的被引次数大概在发表10年后才会比较稳定。定义文献 $i$ 的时限被引次数

$$tcc(i)=cc(i)/\tau(age(i))$$

其中 $age(i)$ 表示文献 $i$ 的已发表年限，函数

$$\tau(x)=\begin{cases}0.2 & 0<x\leqslant 1\\ 0.4 & 1<x\leqslant 2\\ 0.56 & 2<x\leqslant 3\\ 0.7 & 3<x\leqslant 4\\ 0.8 & 4<x\leqslant 5\\ 1 & x>5\end{cases}$$

注意函数 $\tau$ 的定义与引文网络的结构紧密相关，这里的 $\tau$ 符合 CiteSeer 引文网络的特征。用 $tcc(i)$ 替换公式（6）中的 $cc(i)$，得到时限伪被引次数 $tpcc(i)$ 的定义。所谓带时间约束的伪被引次数

法，就是用 $tpcc(i)$ 替换公式（7）中的 $pcc(i)$ 并计算个性化向量 $v$ 的各个分量 $v_i$。

（3）实验

实验是这样设计的：首先从 CiteSeer 引文网络随机选取 100 篇文献，文献的发表时间在 2002 年 1 月 1 日到 2005 年 6 月之间，且每篇文献的引用次数和被引次数均不为零。接着按照一定的比例给文献设置等级，其中“非常相关”、“中等相关”和“一般相关”分别占 5%、10% 和 85%。然后分别用常规 PageRank 算法、等概率法、伪被引次数法以及带时间约束的伪被引次数法（以下分别称之为算法 1、2、3 和 4），计算各篇文献的 PR 值。松弛系数 $\mu=0.85$，收敛阈值是 $10^{-8}$，初始 PR 向量 $\boldsymbol{q}(0)$ 满足均匀分布，即每个分量值均为 1/716800。

由于 CiteSeer 引文网络不存在有向环，四种算法的收敛速度都很快，分别经过 18、18、17 和 17 步迭代。初始文献集 $V_h \cup V_m$ 包含的 15 篇相关文献的排名如表 3 所示。

表 3　15 篇相关文献的排名

Tab. 3　Ranking of the publications in the initial set

| 文献号 | 算法 1 | 算法 2 | 算法 3 | 算法 4 |
|---|---|---|---|---|
| 1 | 132782 | 301 | 444 | 479 |
| 2 | 94636 | 295 | 445 | 480 |
| 3 | 120928 | 299 | 443 | 478 |
| 4 | 116077 | 298 | 442 | 477 |
| 5 | 129834 | 300 | 441 | 475 |
| 6 | 163697 | 8933 | 4583 | 4881 |
| 7 | 149890 | 8862 | 4587 | 4885 |
| 8 | 145977 | 8834 | 4588 | 4886 |
| 9 | 92761 | 8337 | 4585 | 4883 |
| 10 | 140033 | 8794 | 4582 | 4880 |
| 11 | 93264 | 8342 | 4577 | 4864 |
| 12 | 160611 | 8914 | 4589 | 4887 |
| 13 | 73997 | 8106 | 4558 | 4834 |
| 14 | 102167 | 8472 | 4584 | 4882 |
| 15 | 148678 | 8846 | 4586 | 4884 |

表 3 的结果说明设置个性化向量能够很好地表征研究人员的个人偏好，大幅提高对应文献的排名。比较四种算法排序结果的前 1000 篇文献。令 $c_{ij}$ 表示算法 $i$ 和 $j$ 的排序结果包含的相同文献数，则有 $c_{12}=989$，$c_{13}=829$，$c_{14}=821$；$c_{23}=830$，$c_{24}=822$；$c_{34}=984$。算法 1 和算法 2 的结果比较接近，主要原因是起始文献集的基数与总的文献量相比非常小，导致二者的个性化向量区别也较小。算法 3 和算法 4 的结果也非常接近，说明对于 CiteSeer 引文网络而言，文献发表年限对于科学文献个性化排序的影响并不显著。

## 4　相关研究工作

SCEAS[8] 是一种科学文献排序算法，文献 $i$ 的评级计算公式是

$$r_i = (1-\mu) + \mu \sum_{\forall j \in In(i)} \frac{r_j + b}{|Out(j)|} a^{-1} \qquad (8)$$

其中 $b$ 是直接引用加强因子，而 $a$ 决定了间接引用的影响收敛到零的速度。当 $b=0$ 且 $a=1$ 时，公式（8）变成公式（1），即标准的 PageRank 算法。Prestige[4] 是一种适合于引文网络的排序算法，它是 HITS 算法的变形。Fiala 等人[9] 对 PageRank 算法进行改进，使之能够充分利用引用关系和共同作者关系提供的信息。

个性化 PageRank 算法是一个研究热点。实现个性化主要有两种途径：一是定制个性化向量 $v$ 的取值；二是为外出链接添加权重参数，以反映个性化需求。主题敏感的 PageRank 算法[10] 首先计算网页与不同主题对应的 PR 值，然后利用加权汇总得到与某个查询相关的总 PR 值。基于使用信息的 PageRank 算法[11] 分析用户的访问日志，并根据分析结果设定个性化向量以及权重参数的取值。Fogaras 等人[12] 提出一种新算法，通过预先计算一个紧凑数据库，支持对任意个性化查询的在线响应。

带时间约束的 PageRank 算法[13] 是指为公式（1）中的 $q_j$ 添加一个时间衰减因子。对于文献 $i$，时间衰减因子

$$w_i = \tau^{(now - r_i)/12}$$

其中 $now$ 代表当前时间，$r_i$ 表示文献 $i$ 的出版时间，二者均以月为单位；$\tau=0.5$ 代表衰减率。

需要指出的是，采用添加参数实现个性化，其实质是改变了转移矩阵的内容和性质，此时得到的 PageRank 计算公式是否存在唯一解，研究者必须给出相应的证明。

## 5　结论与展望

设置不同的 PageRank 个性化向量，能够表征研究人员的个人偏好，实现对科学文献的个性化排序。与正式发行的引文库相比，CiteSeer 引文网络包含的文献信息和引用关系存在信息不完整等不足，这在一定程度上会影响到研究结果的准确性。因此，接下来的首要工作是构建一个信息更加完整和准确的引文库，以便后续研究。要实现科学文献的个性化排序，除了引用信息，还应该充分利用文

献的内容以及各种元数据（如作者、发表的刊物名称和等级等）[14]。此外，社会化协作也能够为个性化排序提供支持。如何把引用、内容和协作三方面结合起来，构造以网络环境为基础的科研协作平台，是下一步研究的主要目标。

**参考文献：**


[1] GARFIELD E. Citation Analysis as a Tool in Journal Evaluation [J]. Science, 1972, 178: 471 - 479.

[2] GARFIELD E. The History and Meaning of the Journal Impact Factor [J]. Journal of the American Medical Association, 2006, 295(1): 90 - 93.

[3] BORODIN A, ROBERTS G O, ROSENTHAL J S. Link analysis ranking: algorithms, theory, and experiments [J]. ACM Transactions on Internet Technology, 2005, 5(1): 231 - 297.

[4] KLEINBERG J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604 - 632.

[5] BRIN S, PAGE L. The Anatomy of a Large - scale Hypertextual Web Search Engine [J]. Computer Networks and ISDN Systems, 1998, 30(1 - 7): 107 - 117

[6] CHEN T T, XIE L Q. Identifying Critical Focuses in Research Domains [C] // Proceedings of the Ninth International Conference on Information Visualisation. London, England, 2005: 135 - 142.

[7] 胡茂林. 矩阵计算与应用 [M]. 北京：科学出版社，2008: 352.

[8] SIDIROPOULOS A, MANOLOPOULOS Y. A citation-based system to assist prize awarding [J]. SIGMOD Record, 2005, 34(4): 54 - 60.

[9] FIALA D, ROUSSELOT F, JEZEK K. PageRank for bibliographic networks [J]. Scientometrics, 2008, 76(1): 135 - 158.

[10] HAVELIWALA T H. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784 - 796.

[11] EIRINAKI M, VAZIRGIANNIS M, KAPOGIANNIS D. Web path recommendations based on page ranking and Markov models [C] // Proceedings of the 7th annual ACM international workshop on Web information and data management, 2005: 2 - 9.

[12] FOGARAS D, CZ R B, CSALOG N Y K. Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments [J]. Internet Mathematics, 2005, 2(3): 333 - 358.

[13] YU P S, LI X, LIU B. Adding the temporal dimension to search - A case study in publication search [C] // Proceedings of The 2005 IEEE/WIC/ACM International Conference on Web Intelligence. Compiegne, France, 2005: 543 - 549.

[14] STROHMAN T, CROFT W B, JENSEN D. Recommending Citations for Academic Papers [C] // Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Amsterdan, Netherlands, 2007: 705 - 706.


## Personalized Ranking of Scientific Publications Using Link Analysis


*LIU Quan-bo, XU Jun*

(Department of Educational Technology, South China Normal University, Guangzhou 510630, China)



**Abstract**: Inspired by the PageRank algorithm, a personalized ranking algorithm for scientific publications is discussed in this paper. Three methods for computing the personalization vector are presented. These methods are based on not only user ratings but also citation counts and publication dates. The personalized algorithm is then applied to the citation graph of CiteSeer digital library and the experiment result is analyzed. In addition, the structure of the citation graph is fully explored.



**Key words**: citation graphs; link analysis; ranking algorithm; PageRank; personalized