第36卷 第22期 计 算 机 工 程 2010年11月
Vol.36 No.22 Computer Engineering November 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)22—0291—03 文献标识码：A 中图分类号：TP393

# PageRank算法的分析及其改进

王德广，周志刚，梁 旭
(大连交通大学软件学院，辽宁 大连 116028)

**摘 要：**在分析PageRank算法存在偏重旧网页、主题漂移、网页权值均分、忽视用户浏览兴趣现象的基础上，对其进行改进，考虑网页修改日期、网页文本信息、网站权威度、用户兴趣度等重要因素，重新计算网页*PR*值。实验结果表明，改进算法可提高搜索引擎对网页排序的准确度，以及用户对检索结果的满意度。

**关键词：**PageRank算法；搜索引擎；文本数据挖掘；*PR*值

# Analysis of PageRank Algorithm and Its Improvement

WANG De-guang, ZHOU Zhi-gang, LIANG Xu
(Software Technology Institute, Dalian Jiaotong University, Dalian 116028, China)

**【Abstract】**This paper improves PageRank algorithm is based on analyzing the phenomenon of stressing on old pages, drifting theme, spliting page weight and neglecting user browsing interests. It considers the important factors of webpage modification data, webpage text information, website technoroti authority, user interestingness. Experimental result shows that improved algorithm can improve accuracy for webpage order and user satisfaction with search results.

**【Key words】**PageRank algorithm; search engine; text data mining; *PR* value

## 1 概述

随着互联网的快速发展，互联网上的信息越来越丰富。网络时代已经到来，上网查找资料的用户呈几何级增长，然而，面临互联网上的海量信息，大多用户都无所适从。什么信息是有用的信息、如何检索信息、如何缩短检索时间是搜索引擎面临的主要问题。传统网络搜索引擎大多是基于关键字匹配的，其查询效果不太理想。Sergey B和Lawrence P借鉴引文分析思想，提出PageRank算法，该算法通过分析网络的链接结构获得网络中的权威网页，并在搜索引擎Google中获得成功。随着对PageRank算法的研究深入，许多学者针对其不足，提出了改进。本文对PageRank算法进行了改进。

## 2 PageRank算法简介

文献[1]提出用于网络链接分析的PageRank算法，该算法建立在随机冲浪者模型上。具体来说，假设浏览者跟随链接进行若干步的浏览后转向一个随机起点网页又重新跟随其链接浏览，那么一个网页的价值程度值就由该网页被这个浏览者访问的频率决定。

PageRank算法简单描述如下：

$u$是被研究的网页，$v_i$是指向$u$的网页，$C(v_i)$是网页$v_i$的向外指出的网页的链接数，$d$是规范化因子(一般取0.85)。因此，网页$u$的*PR*值计算如下：

$$PR(u)=(1-d)+d\sum_{i=1}^{n}\frac{PR(v_i)}{C(v_i)} \tag{1}$$

## 3 PageRank算法分析

由于PageRank算法是离线计算网络的PageRank值，在用户查询时仅根据关键字匹配获得网页集合，然后排序推荐给用户，因此具有很高的响应速度，并且搜索引擎Google中的成功也证明该算法是高效、合理的。但由于仅利用了网络的链接结构，该算法还存在一些缺点：

(1)偏重旧网页：网页存在的时间越长，搜索排序的名次越靠前；

(2)主题漂移现象：无法区分网页中的超链接与网页主题相关还是不相关；

(3)平分网页权值：即一个网页被一个权威网页引用和被一个普通网页所引用，其意义与价值是不同的；

(4)忽视用户浏览兴趣：即没有把浏览者这个关键因素考虑进去。

### 3.1 PageRank算法偏重旧网页的现象

由式(1)可以得出，决定网页*PR*($u$)值高低的一个主要因素是指向该网页的链接个数。因为旧网页存在的时间长，被其他网页引用的可能性较高，而实际上新的网页通常包含更新更有价值的信息；如果一个网页刚被放到互联网，可能会由于时间短暂，许多其他网页还没有引用它，导致它的*PR*值降低。通过PageRank算法，它出现在搜索页面中的次序通常很靠后，这样可能正好与用户需求相反。因为在很多情况下，用户通常想看到新网页中的最新内容。因此，在某种程度上网页存在时间越长，通过式(1)计算出的网页*PR*值越高，但却不能很好满足用户的需求[2]。

### 3.2 PageRank算法的主题漂移现象

PageRank算法出现主题漂移现象的原因主要如下：

(1)PageRank算法无法区分网页中的链接与该网页的主

**基金项目：**辽宁省教育厅计划基金资助项目“用网页评价等级与转移概率改进PageRank算法研究”(L2010090)

**作者简介：**王德广(1968－)，男，副教授，主研方向：数据优化；周志刚，硕士；梁 旭，教授、博士

**收稿日期：**2010-05-18 **E-mail：**zzgisgod@sina.com

题是否相关，即无法判断网页内容之间的相似相关性，这样容易出现用户搜索的网页内容并不是他想要看的内容；

(2)PageRank 算法偏重以.com 结尾的网站，因为这类网站通常是综合性网站，可以比其他类型的网站获得更多链接。事实上，这类网页通常涉及的面多而不专，相比之下，某些专业网站对问题的阐述更有权威性且与搜索的主题更贴切。

### 3.3 PageRank 算法的平均网页权值现象

网页的链接分成前向链接和反向链接，而反向链接的数量和质量决定 *PR* 值。反向链接是指所考察的网页被其他网页引用，反向链接数目越多，表示该网页被引用越多，其重要性也越高[3]。但一个网页被权威网站引用和被很多垃圾网页引用，效果是完全不同的。

目前，PageRank 算法将当前网页的权值平均分配给它的全部链接。互联网中各个网页的质量价值千差万别，即使是链接在同一个网页上的各个链接，其优劣层次也差很多。所以，PageRank 算法这种平均分配权值的方法，在一定程度上影响了网页的排序质量。

### 3.4 PageRank 算法忽视用户浏览兴趣的现象

PageRank 算法在设计之初，没有考虑到用户的浏览兴趣，但一个页面能否被用户再次浏览，很大程度上取决于用户的兴趣度。

## 4 PageRank 算法的改进

### 4.1 对 PageRank 算法偏重旧网页现象的改进

考虑到大多数“旧网页”都有被引用数目多、内容陈旧、可参考性不高的特点。假设：一个网页被搜索到的时间与其最近一次被修改时间的差值越大，则网页内容的价值越低，权威性就越低。在这个假设下，引入一个与时间有关的权值函数 $W(t)$，与网页的权值呈反比：

$$W(t)=d/t$$

$$t=\begin{cases}1.0 & t\leqslant 1\text{个月}\\ 1.8 & 1\text{个月}<t\leqslant 1\text{年}\\ 0.6 & t\geqslant 1\text{年}\end{cases} \tag{2}$$

其中，$W$ 是网页的权值；$t$ 为求一个网页被搜索到的时间与其最近一次被修改的时间的差值的函数；$d$ 是一个比例常数。得到一级 *IPR* 如下：

$$IPR_1(u)=(1-d)+d\sum_{i=1}^{n}\frac{IPR_1(v_i)\times W(t_i)}{C(v_i)} \tag{3}$$

### 4.2 对 PageRank 算法主题漂移现象的改进

通过上文分析可知，传统 PageRank 算法出现主题漂移现象是因为其无法知道网页中的链接与该网页主题的相关性，所以可以采用文本数据挖掘的方法对网页的内容进行数据挖掘，文本数据挖掘是指从文本数据中抽取有价值的信息和知识的计算机处理技术。文本数据挖掘流程见图 1。

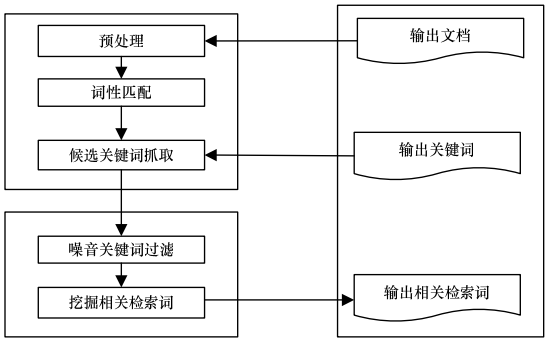


**图 1 文本数据挖掘流程**

在这里，引入一个分段函数 $F(v_i)$，用文本数据挖掘出的信息进行处理分级，据此改变其链接权重。函数 $F(v_i)$ 如下：

$$F(v_i)=\begin{cases}2.0 & \text{在网页的}head/title\text{中挖掘的关键词}\\ 1.8 & \text{通过文本挖掘出的相关关键词}\\ 1.5 & \text{在网页的}body\text{中挖掘的关键词}\\ 1.0 & \text{其他}\end{cases} \tag{4}$$

对一级 *IPR* 算法作进一步的改进：

$$IPR_2(u)=(1-d)+d\sum_{i=1}^{n}\frac{IPR_2(v_i)\times W(t_i)\times F(v_i)}{C(v_i)} \tag{5}$$

### 4.3 对 PageRank 算法平均网页权值现象的改进

PageRank 算法出现平均网页权值的现象，是因为它没有对权威网站和普通网站进行区分[4-5]，权威网站被引用的频率很高，但普通网站被引用的概率却很低。因此，引入网站权威度函数 $P(v_i)$，它是网页被指向链接与指向链接的比：

$$P(v_i)=Q\left(\frac{BackLink}{InLink}\right) \tag{6}$$

其中，*BackLink* 为引用网页 $i$ 的链接数目；*InLink* 为该网页引用其他网页的链接数目；$Q$ 为一个与 $d$ 相关的常数。在式(5)的基础上得到：

$$IPR_3(u)=(1-d)+d\sum_{i=1}^{n}\frac{IPR_3(v_i)\times W(t_i)\times F(v_i)\times P(v_i)}{C(v_i)} \tag{7}$$

### 4.4 对 PageRank 算法忽视用户浏览兴趣现象的改进

网站服务器的 Web 日志文件中记录了每个用户的访问信息，包含 time-taken 字段，该字段描述了页面访问时所用的时间。

**定义 1** 用户获取页面全部内容需要的时间 $t_s$，称为页面下载时间。

经统计，在网络畅通时，页面下载时间 $t_s\leqslant 3$ s，所以，本文设定阈值 $t_s=5$ s，若页面下载时间 $t_s>5$ s，则用户的兴趣度降低。

**定义 2** 一般人正常阅读完全部页面内容并进行评论及思考所需的时间为 $t_c$，称为页面关注时间。$t_c$ 的计算如下：

$$t_c=\frac{k(R_s+R_c\times 50+R_g\times 100)}{280} \tag{8}$$

其中，$R_s$ 是页面正文文字个数；$R_c$ 是页面图片个数；$R_g$ 是页面视频个数，为了便于计算，将图片和视频转化为文字，设定一张图片相当于 50 个文字，一个视频相当于 100 个文字，除以 280 是因为成年人的平均阅读速度仅为 280 字/min；$k$ 是评论系数，取 1.2~1.5。

兴趣度 $V$ 的计算如下：

$$V=\frac{\sum_{i=1}^{n}\left[\frac{t_r}{t_c}+(t_s-5)\times 0.1\right]}{n} \tag{9}$$

其中，$t_r$ 是用户访问此网页的实际浏览时间；$t_r/t_c$ 为基本兴趣度；$(t_s-5)\times 0.1$ 为基本兴趣度的偏移度。

在式(7)的基础上得到 $IPR_4$ 计算如下：

$$IPR_4(u)=(1-d)+d\sum_{i=1}^{n}\frac{IPR_4(v_i)\times W(t_i)\times F(v_i)\times P(v_i)\times V_i}{C(v_i)} \tag{10}$$

## 5 实验

为了验证改进算法的效果，用网络爬虫工具 Heritrix 在 Google 上采用 BroadScope 模式连续抓取网页 10 h，获得约 $2\times10^5$ 个网页。经过数据预处理，将其导入到数据库中，对 100 名研究生选取 20 个最受关注的话题(见表 1)进行对比实验。

表 1　20 个关键词列表

| 序号 | 话题 | 序号 | 话题 |
|---|---|---|---|
| 1 | 反腐倡廉 | 11 | 依法行政 |
| 2 | 医疗改革 | 12 | 三农问题 |
| 3 | 食品安全 | 13 | 安全生产 |
| 4 | 收入分配 | 14 | 城乡统筹 |
| 5 | 就业问题 | 15 | 金融危机 |
| 6 | 环保问题 | 16 | 社会稳定 |
| 7 | 住房问题 | 17 | 股市稳定 |
| 8 | 教育公平 | 18 | 灾后重建 |
| 9 | 社会保险 | 19 | 机构改革 |
| 10 | 司法公正 | 20 | 文化创新 |

本文首先根据未改进的 PageRank 算法计算出所有网页的 *PR* 值，然后对 20 个被选话题进行测试，在每个话题返回被搜索到的前 50 个结果中，统计符合被测试研究生兴趣的网页数目；然后根据改进算法计算出所有网页的 *IPR* 值，用第 1 次生成的 20 个话题同样进行测试。符合被测试研究生兴趣的网页数目如图 2 所示。

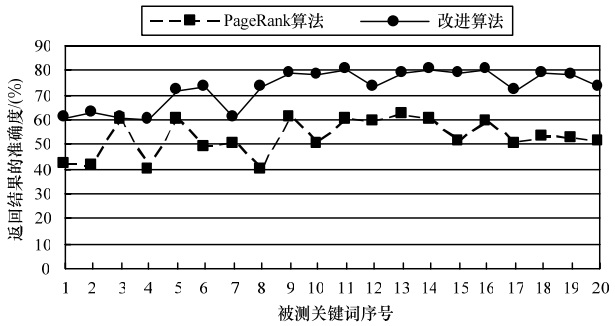


图 2　算法改进前后结果比较

由图 2 可见，用 PageRank 算法查询结果的满意度的平均值为 51.4%左右，而用改进算法把查询结果的满意度提高到 71%左右，即改进算法可以更加准确地判断网页的权威性，返回更符合用户查询条件的网页。这个实验结果与实验所选的关键词及实验调查人群的情况有关，今后将在不同背景的人群中收集数据，以进行更全面的验证。

## 6　结束语

本文通过对 PageRank 算法的研究，找出该算法的缺陷，并针对其提出具有四级改进算法，通过引入时间权值函数 *W*、分段函数 *F*、网页权值比例函数 *P* 及兴趣度 *V*，有效地解决了传统 PageRank 算法容易出现的偏重旧网页、主题漂移、网页权值均分、忽视用户浏览兴趣的缺陷，提高了网页的排序质量，从而提升搜索结果的准确性。实验证明，改进算法较 PageRank 算法在排序质量上有显著提高。

## 参考文献


[1] 张　蓉. Web 挖掘技术研究[J]. 计算机工程, 2006, 32(15): 4-6.

[2] 焦金涛. 基于 PageRank 的 Web 挖掘改进算法[J]. 计算机工程, 2009, 35(15): 284-285.

[3] 田　甜. 基于 PageRank 算法的权威值不均衡分配问题[J]. 计算机工程, 2007, 33(18): 53-55.

[4] 杨　彬. 基于概念的权重 PageRank 改进算法[J]. 情报杂志, 2006, (11): 70-72.

[5] 葛　玲. 基于共现词查询的主题爬虫研究[J]. 计算机工程, 2010, 36(8): 286-288.


编辑　陆燕菲

（上接第 290 页）

验，得到一系列先验值，计算出 RSSI 等强度梯度线，找到可读范围与车道的交界确定的那条等强线作为设定边界，并计算出该强度对应的距离，当车辆进入地感线圈，启动读写器，如果读写器读到的标签返回信号强度对应的距离大于设定边界值，则该值被丢弃，只有当读写器读到的距离小于设定边界的标签，才会被读取并存储。这样就可以有效避免误读现象，如图 3 所示。

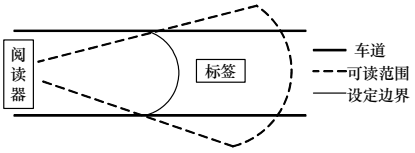


图 3　路桥项目应用演示

## 4　RSSI 的应用趋势以及改进

随着 RFID 技术的发展和普及，定位技术越来越受重视，现在已应用在矿工井下定位、机器人控制定位、门禁安全监视以及本文提到的路桥中防止误读标签等方面。目前市面上可以买到的部分芯片也具有采集 RSSI 值的功能，如 chipcon 的 cc1100、cc1000、cc2420 等芯片。

基于 RSSI 的测距在实际环境中由于受到非视距传播与多径传播的影响，定位精度不高。非视距传播是指信号由于受到障碍物阻挡不能在可视距离中直线传播。而多径传播是指信号经过 2 个或更多的途径到同一接收端的现象。这两者都是不可消弭的。而在计算过程中，由于 RSSI 是通过在数字域进行功率积分而后反推到天线口得到的，因此反向通道信号传输特性的不一致也会影响 RSSI 的精度。以上 2 点原因造成测距结果存在一定误差，一般 15 m 以内的误差在 2 m 左右。

提高精度是改进 RSSI 技术的主要方向。已有的改进主要有 2 个方面：(1)使用数学方法对模型中易受环境影响的衰减因子进行估值，根据估计值进行自动校正；(2)根据实验得到的经验值绘制拟合曲线，建立 RSSI-d 模型。

## 5　结束语

本文的系统采用六端口代替传统接收电路，简化了电路结构，降低了成本，并简化了 RSSI 求值过程。针对射频定位的基本算法，采用接收信号强度指示结合信号传输损耗模型来计算标签和阅读器之间的距离。在实际应用中，单一的 RSS 测距方法由于多径传输、非视距传输等原因导致测量精度不高，不能满足精确定位需求，但其实现简单，系统要求低，作为 RFID 系统中的辅助部分还是完全可以胜任的。

## 参考文献


[1] 陈永光. 基于信号强度的室内定位技术[J]. 电子学报, 2004, 32(9): 1456-1458.

[2] 杨大成. 移动传播环境理论基础、分析方法和建模技术[M]. 北京: 机械工业出版社, 2003.

[3] 王中云. 基于 RFID 的机器人控制与定位系统研究[D]. 武汉: 武汉理工大学, 2007.

[4] 陈德华. 零中频接收技术在 RFID 读卡机中的应用[J]. 电子产品世界, 2005, (1): 109-110.

[5] 王新稳. 微波技术与天线[M]. 2 版. 北京: 电子工业出版社, 2006.


编辑　张　帆