PageRank 算法研究综述

李稚楹 杨 武 谢治军
(重庆理工大学计算机科学与工程学院 重庆 400054)

摘 要 网页排序是搜索引擎的关键技术之一。介绍了著名的 PageRank 算法,针对其存在主题漂移、偏重旧网页等不足,分析了各种改进算法的基本思想和技术特点,希望为以后的研究工作提供基础性支持。

关键词 PageRank,主题漂移,偏重旧网页

Research on PageRank Algorithm

LI Zhi ying YANG Wu XIE Zhi jun
(Department of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract Pages ranking is one of the key technologies used in the design of a search engine. This paper described the famous PageRank algorithm. For the problems such as topic drift and emphasis on the old page, the approaches and features of a variety of improved algorithms were analysed in detail, hoping that can provide the basic support to the research in the future.

Keywords PageRank, Topic drift, Emphasis on the old page

1 引言

互联网的急速发展、网络信息量的暴涨,使得搜索引擎成为人们必不可少的信息检索工具。中国互联网信息中心(CNNIC)在《第 26 次中国互联网络发展状况统计报告》[1]中指出:截至2010 年 6 月,搜索引擎在网民中的使用率达 76.3%,用户规模达3.2 亿人。搜索引擎使用频率的增加以及用户规模的扩大,使如何让用户快速准确地检索出有用信息成为近几年研究者所共同关注的问题。

网页排序作为搜索引擎的关键技术之一,它的好坏将直接影响用户对信息的准确查找。目前,有许多排序算法,不过应用最成功、最具研究价值的是由斯坦福大学的 Larry Page 和 Sergey Brin 于 1996 年首次提出 PageRank 算法[2]。其思想是通过分析网络的链接结构来获得网络中网页的重要性排名。该算法虽在 Google 搜索引擎的应用中获得成功,但其也存在一些缺陷。研究者们在算法存在的主题漂移、偏重旧网页、忽视用户个性化等问题基础上,提出了各自的改进算法。

2 PageRank 算法

2.1 PageRank 算法简介

PageRank 算法是基于网页链接分析对关键词匹配搜索结果进行处理的。它借鉴传统引文分析思想:当网页 A 有一个链接指向网页 B,就认为 B 获得了 A 对它贡献的分值,该值的多少取决于 A 本身的重要程度,即网页 A 的重要性越大,网页 B 获得的贡献值就越高。由于网络中网页链接的相互指向,该分值的计算为一个迭代过程,最终网页根据所得分值进行检索排序。

一个网页的 PageRank 值(下文用 PR 表示),可由式(1)[3]计算所得:

$$PR(p)=(1-d)+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \quad (1)$$

式中,$PR(p)$ 表示网页 p 的页面级别;T_i($i=1, 2, \cdots, n$)表示指向网页 p 的其他网页;d 为用户随机到达一个网页的概率,介于 0 到 1 之间(通常为 0.85);$C(T_i)$ 为网页 T_i 向外指出的链接数目;$\frac{PR(T_i)}{C(T_i)}$表示网页 p 的链入网页 T_i 给予 p 的 PR 值。通常,我们设每个网页的初始 PR 值为 1,由公式递归计算各个网页的 PR 值,直到该值趋于稳定。算法采用用户完全随机访问网络的行为模型,使得一个网页的 PR 值被均分给它的链出网页。

由于互联网的链接结构是自发、无序形成的,因此可能存在这样的情况:在一组组内相互彼此链接,组外无链接的网页中,一旦有组外网页链接到组内的网页,由于在组内不存在对外的链接,因此传递进来的 PR 值就一直滞留在这组网页内部,不能传递出去,导致 PR 值沉淀(LinkSink)。

如图 1 所示:网页 a,b 内部相互链接,无向外的链接。若网页 c 链接到网页 b,则在迭代计算中,网页 a,b 的 PR 值不能分布出去而累计。

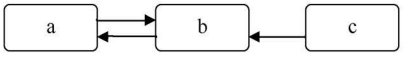

图 1 PR 值沉淀现象

为了避免沉淀现象,Sergey Brin 和 Lawrence Page 改进了算法[4],引入衰退因子 $E(A)$,$E(A)$ 对应网页集的某一向量,对应 PR 的初始值,改进算法为:

$$PR'(A)=(1-d)+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}+dE(A)$$

其中,$\|PR'\|=1$。

李稚楹(1986-),女,硕士生,主要研究方向为信息检索,E-mail:zhiyingll@cqut.edu.cn。

2.2 PageRank 算法的优点

PageRank 算法通过网页间的链接来评价网页的重要性，在一定程度上避免和减少了人为因素对排序结果的影响；采用与查询无关的离线计算方式，使其具有较高的响应速度；一个网页只能通过别的网页对其引用来增加自身的 PR 值，且算法的均分策略使得一个网页的引用越多，被引用网页所获得的 PR 值就越少。因此，算法可以有效避免那些为了提高网站的搜索排名而故意使用链接的行为。

2.3 PageRank 算法的缺点

算法在 Google 搜索引擎的成功运用，说明其是高效、可行的。但由于完全基于链接分析，且链接信息相对静态，没有考虑网页使用的动态信息，因此算法还存在一些缺陷，主要可归纳为：

(1) 主题漂移问题

PageRank 算法仅利用网络的链接结构，无法判断网页内容上的相似性；且算法根据向外链接平均分配权值使得主题不相关的网页获得与主题相关的网页同样的重视度，出现主题漂移。

(2) 偏重旧网页问题

决定网页 PR 值的主要因素是指向它的链接个数的多少。一个含有重要价值的新网页，可能因为链接数目的限制很难出现在搜索结果的前面，而不能获得与实际价值相符的排名。算法并不一定能反映网页的重要性，存在偏重旧网页现象。

(3) 忽视用户个性化问题

PageRank 算法在设计之初，没有考虑用户的个性化需要。个性化搜索引擎的兴起，对 PageRank 排序算法提出新的挑战。

3 PageRank 算法的改进

算法的改进可归纳为两类：其一，基于算法理论的改进。可转化为求解矩阵特征向量和大型稀疏矩阵线性方程组的问题。比如，采用幂法迭代思想计算矩阵特征向量的 Power 算法；通过建立、变换搜索子空间寻求最优解的 GMRES 算法；以及结合 Krylov 子空间和 Power 算法优点的 Power Arnoldi 算法，等等。其二，针对互联网实际特点的改进。比如，利用互联网存在大量超链接指向的网页和互联网按块分布的特点。本文主要针对第二种改进思路进行研究阐述。

3.1 主题漂移问题

3.1.1 Topic Sensitive PageRank 算法

为了提高查询结果的主题相关性，斯坦福大学的 Taher H. Haveliwala[5] 提出主题敏感(Topic Sensitive) PageRank 算法(TSPR)。众所周知，在某些领域认为是重要的网页，并不表示它在其它领域也是重要的。算法首先根据 Open Directory Project(ODP)提出 16 个基本分类标准，对网页进行分类，并针对每个网页的 ODP 分类，给出一个主题敏感值，然后将此值引入算法，离线计算出网页在 ODP 分类下的 PR 值。再根据用户输入的查询主题及查询上下文，计算网页的综合得分，进而进行网页排序。

算法的形式化表示为：

$$R(u) = M' \times R(u) = cM \times R(u) + (1-c)P_u$$

式中，P_u 是网页 u 的主题敏感向量。

算法基于网页分类思想，根据用户的查询请求和相关上下文判断用户查询相关的主题，返回与查询主题相关的网页，从而提高排序结果的准确性。这样可有效地避免一些明显的主题漂移现象，但遗憾的是并没有利用主题的相关性来提高链接得分的准确性；当查询缺少上下文时，很难确定查询的分类，此时将无法抑制主题漂移现象。

3.1.2 结合链接分析和文本内容的 PageRank 算法

华盛顿大学的 Matthew Richardson 和 Pedro Domingos[6] 认为用户会受到当前网页内容和查询主题的影响，而从一个网页跳转到另一个网页，为此提出一种结合链接和内容信息的 PageRank 算法(MP PageRank 算法)。算法的计算形式为：

$$PR_q(j) = (1-d)p'_q(j) + d\sum_{i\in B_j} PR_q(i)h_q(i,j)$$

式中，$h_q(i,j)$ 表示用户在查询主题 q 下从网页 i 跳转到网页 j 的可能性。$p'_q(j)$ 表示用户在网页没有链出时，跳转到 j 的可能性。$PR_q(j)$ 表示网页 j 在查询主题 q 下的 PR 值。对于 $h_q(i,j)$ 和 $p'_q(j)$，采用查询主题 q 和网页 j 的相关函数 $R_q(j)$ 计算得出，公式如下：

$$p_q'(j) = \frac{R_q(j)}{\sum_{k\in W} R_q(k)} \qquad h_q(i,j) = \frac{R_q(j)}{\sum_{k\in F_i} R_q(k)}$$

式中，W 表示网络中的网页集合，F_i 表示网页 i 的链出网页集。相关函数 $R_q(j)$ 原则上可以是任意的，但一般取值为在网页 j 的文本中查询主题 q 的关键字出现的次数。

算法除去 PageRank 算法假设用户完全随机访问网络的这一前提，增加考虑用户从一个网页直接跳转到与之内容相关的一个非直接相邻网页的情况。在传递网页 PR 值时，不仅充分利用网页的链接结构，还考虑到网页间的主题相关性，使得 PR 值的传递更加精确。

3.1.3 二阶相似度改进算法

黄德才，戚华春[7] 在 MP PageRank 算法的理论基础上，提出一个针对两个网页进行相似度描述的二阶相似度概念。将其定义为某个网页在网络链接结构中出现的次数 t，并利用这个二阶相似度形成网络的相似度矩阵 S，用 $S_{i,j} = t$ 表示网页 i 有指向网页 j 的链接；若无链接，则 $S_{i,j} = 0$。相似度矩阵 S 中网页 p 对网页 T_i 的相似度值用 S_{p,T_i} 表示，$S(T_i) = \sum_{u\in B_{T_i}} S_{T_i,u}$，$B_{T_i}$ 是网页 T_i 的链出集。

算法的具体实现为：

$$PR(p) = (1-d) + d\times \sum_{i=1}^{n} \frac{PR(T_i)\times S_{p,T_i}}{S(T_i)}$$

该算法吸收了 MP PageRank 算法的基本思想，即认为用户从一个网页跳转到另一个网页是受到当前网页内容和查询主题的影响。在此基础上，进一步计算两个网页之间的相似度，使得网页的 PR 值在具有相似主题的网页间传播，减少在主题无关网页上的扩散，表现出较好的查全率，主题漂移的现象得到有效遏制。

3.1.4 基于分块主题的 PageRank 算法[8]

算法考虑同一网页内属于不同分块的出链有着不同的重要性，赋予不同分块的出链以相应权重。利用基于视觉特征的 VIPS 算法对网页集中的所有网页进行划分，再用重要性模型对划分后的网页进行分块处理，得到各块的重要性等级并计算重要性权值，以此权重计算得到更加准确合理的 PR 值。

算法的计算形式为：

$$PR(A)=(1-d)+d\sum_{i=1}^{n}PR(T_i)\cdot f(T_i,A)$$

式中，d 取 0.85，$f(T,A)$ 是网页 T 分给网页 A 的 PR 值，其形式为：

$$f(T,A)=W(T,A)/\sum_{i=0}^{n}W(T,A_i)$$

式中，$W(T,A)$ 表示网页 T 引用网页 A 的重要性权值，$A\sim A_n$ 表示被网页 T 所引用的所有网页。

算法对 PR 值分配策略进行了有效修改，经过实验证实在多关键词搜索中，该改进算法与传统改进算法相比更能有效地解决主题漂移现象。算法充分利用网页间的相互引用关系，考虑网页的文本结构和内容，更合理、公正、有效地计算网页的 PR 值，使搜索结果具有较高的精确度。

本文所提到的前两种改进算法都是需要利用查询主题以外的信息(利用上下文或网页的文本信息)来提高对网页的辨识能力，以减少主题漂移现象的发生；而后两种算法则是针对 PageRank 算法的平均分配策略进行改进，根据网页间相关性的大小来分配权值。除上述所列算法外，还有 Diligenti[9] 等人提出的将访问某链接的概率分成两部分(网页主题相关和链接主题相关)来计算的 Double Focused PageRank 算法；黄德才，戚华春[10] 提出的基于虚拟文档的主题相似度模型和基于主题相似度模型的 TS PageRank 算法；王冬[11] 提出的不平均分配 PR 值的改进算法 NPR，等等。

3.2 偏重旧网页问题

3.2.1 加速评估算法

为了让高质量的新网页能快速地在网络上传播，上海交通大学的张玲博士[12] 提出了基于时间序列分析的加速评估算法。算法的核心思想是通过分析基于时间序列的 PR 值变化情况，预测 URL 在未来一段时期内的期望值。在用户检索时，将此预测值作为决定一个 URL 在检索结果中排名的有效参数。

定义一个 URL 的加速因子 AR 为：

$$AR=PR\times sizeof(D)$$

式中，PR 是 URL 的 PR 值，$sizeof(D)$ 为整个网页集的网页总量。

加速后的 PageRank 表达式为：

$$PR_{accelerate}=\frac{AR_{last}+B\times D}{M_{last}}$$

AR_{last} 是 URL 最近一次的 AR 值，B 是该 URL 一段时间内 PR 值的二次拟合曲线的斜率，D 为离最近一次网页下载的时间间隔天数，M_{last} 是最近一次下载的文档集内的文档数目。

算法使得网络上有价值的内容以更快的速度传播，那些已陈旧的网页的 PR 值将加速下滑，从而帮助用户发掘有价值的网页，保证信息检索中的优胜劣汰，确保向用户提供高质量的 URL 链接。

3.2.2 具有时间反馈的 PageRank 算法

PageRank 算法没有考虑网页使用的动态信息，越旧的网页排名越靠前，偏重旧网页。因此，研究者们提出了具有时间反馈的改进算法。

考虑到“旧网页”链接数目多、内容陈旧、参考性不强的特点。我们认为[13]，如果一个网页发布的时间与被检索到的时间间隔越长，其网页内容的价值、权威性就越低。在这个前提下，引入一个与网页的权值呈反比的时间权值函数 $W(t)$，其表达式为：

$$W(t)=\frac{d}{t},\ t=\begin{cases}1.0, & t\leqslant 1\text{ 个月}\\ 1.8, & 1\text{ 个月}<t\leqslant 1\text{ 年}\\ 0.6, & t\geqslant 1\text{ 年}\end{cases}$$

式中，W 是网页的权值；t 为一个网页分布的时间与被检索到的时间间隔的函数；d 是一个比例常数。

改进的 PageRank 计算公式为：

$$PR(p)=(1-d)+d\sum_{i=1}^{n}\frac{PR(T_i)\times W(t_i)}{C(T_i)}$$

该算法有利于时效性强的查询，增加用户对返回结果的满意度。但是，由于现在很多网页是由程序自动生成的，且网络中大量的 HTML 网页格式不规范，很难提取网页的发布时间。因此，戚华春[14] 等人把网页的存在时间通过搜索引擎搜索的周期数来表示，用一个网页被搜索引擎访问的周期次数来代替时间反馈权值，提出 PageRank Time 算法。

其核心思想是基于这样一个事实：一般而言，搜索引擎的搜索周期为半个月到一个月，如果一个网页存在的时间较长，那它将在每个搜索周期里都被搜索到(在同一个搜索周期里，不管搜索到该网页几次，都算作 1 次)，即页面的存在时间正比于搜索引擎搜索到该页面的次数。网页的时间反馈因子 $W_t=e/T$。其中，T 为一个网页被搜索引擎访问的周期次数；e 为常数，它的取值受到式(1)中 d 的影响，且也与搜索引擎的搜索周期相关。

改进后的 PageRank 算法的公式为：

$$PR(p)=(1-d)+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}+W_t$$

算法引入时间反馈因子，克服了提取网页发布时间的困难，能更准确地判断网页日期。网页 PR 值的大小受网页的发布时间长短的影响，有利于旧网页的下沉、新网页的迅速上浮。

3.3 用户个性化问题

PageRank 算法在用户查询前做好计算，与用户的兴趣爱好及反馈没有联系，随着个性化搜索引擎的发展，使得算法不能更好地根据用户的查询特点对其做针对性的推荐。对于此缺陷，研究者提出基于个人兴趣和反馈技术的 PageRank 算法研究[15]，在传统的 PageRank 算法基础上，添加用户兴趣爱好值和相关反馈值，为用户建立搜索行为记录，便于对特定用户再次查询时提供准确的搜索。李凯[16] 等人提出使用用户点击信息来评价网页重要性，将网页被点击次数作为集体个性化向量应用于 PageRank 算法中，以提高搜索引擎查询结果的精度。

结束语　改进后的 PageRank 算法，可有效避免某些网页的恶意欺骗行为，使搜索结果具有较高的精确度；在一定程度上减少主题无关网页对 PR 值的扩散；更能体现搜索引擎排名的人性化和个性化特点。虽然该算法在商用搜索引擎的应用中获得巨大的成功，但如何更有效地利用网页的文本信息，或更加深入研究网络链接结构的内在特性以及结合用户反馈信息来改进算法还有很长的路要走。在以后的研究中，还需要更加完善和优化该算法，提高搜索匹配程度，尽可能准确、快速地将用户最需要的信息排在返回结果的最前端。

参 考 文 献

[1] 中国互联网络信息中心(CNNIC)[R]. 第 26 次中国互联网络发展状况报告，2010，7. http://www.cnnic.net.cn/html/Dir/

2010 /07 /15 /5921. html
[2] Kamvar S. Extrapolation Methods for Accelerating PageRank computations[D]. USA: Stanford University, 2003
[3] Haveliwala T H. Efficient Computation of PageRank[R]. Stanford. 1999
[4] Soon I Y, Koh S N. Speech enhancement using 2-D fourier transform[J]. IEEE Transactions on Speech and Audio Processing, 2003, 11(6): 717-724
[5] Haveliwala T H. Topic sensitive PageRank[C] //Proceedings of the Eleventh International World Wide Web Conference. Hoho Lulu Hawaii, 2002
[6] Richardson M, Domingos P. The intelligent surfer: Probabilistic combination of link and content information[J]. PageRank Advances in Neural Information Processing Systems, 2002, 14: 673-680
[7] 黄德才, 戚华春. PageRank 算法研究[J]. 计算机工程, 2006(2)
[8] 白似雪, 刘华斌. 基于页面分块模型的 PageRank 算法研究[J]. 南昌大学学报, 2008, 6
[9] Diligenti M, Gori M, Maggini M. Web page scoring systems for horizontal and vertical search[C] //The 11th Int'l World Wide Web Conference. Honolulu, Hawaii, USA, 2002
[10] 黄德才, 戚华春, 钱能. 基于主题相似度模型的 TS PageRank 算法[J]. 小型微型计算机系统, 2007(03)
[11] 王冬, 雷景生. 一种基于 PageRank 的页面排序改进算法[J]. 微电子学与计算机, 2009(4)
[12] 张岭, 马范援. 加速评估算法: 一种提高 Web 结构挖掘质量的新方法[J]. 计算机研究与发展, 2004, 41(1): 98-103
[13] 王德广, 周志刚, 梁旭. PageRank 算法的分析及其改进[J]. 计算机工程, 2010(11)
[14] 戚华春, 黄德才, 等. 具有时间反馈的 PageRank 改进算法[J]. 浙江工业大学学报, 2005, 33(3): 272-275
[15] 王小玲, 胡平. 基于个人兴趣和反馈技术的 PageRank 算法研究[M]. 合肥: 合肥工业大学出版社, 2006(3)
[16] 李凯, 赫枫龄, 左万利. PageRank Pro 一种改进的网页排序算法[J]. 吉林大学学报, 2003(4)

(上接第 180 页)

表 2 新算法与文献[5-9]运行结果的比较表

example	Reference	Time(s)	Optimal value	ε_f
例 1	[8]	2	-83.249728406	0
	[6]	<1	-83.249728406	10^{-9}
	This paper	1.049	-147.666666666667	0
例 2	[6]	2	623249.87529475	10^{-9}
	[8]	3	623249.876118100	0
	This paper	12.487	469739.033786623	0
例 3	[6]	1	1.177124327	0
	[5]	28.6160	1.177643	0
	This paper	4.6200	1	0
例 4	[6]	<1	0.2276.9428	0
	[5]	14.9040	0.2107658815	0
	This paper	10.7820	0.2100624439623	10^{-9}
例 5	[6]	0	60	0
	[5]	3.3047	60	0
	This paper	2.4500	60	0
例 6	[4]	-	24.319	0
	[5]	173.2500	24.8368215	0
	This paper	138.8478	25.486030875	0
例 7	[9]		-0.8660254037	0
	[5]	481.9150	-0.841353568	0
	This paper	479.375110	-0.57484653390	10^{-9}
例 8	[9]	-	8853.967480648	-
	[5]	481.9150	8931.5135060	-
	This paper	527.8373	9178.21534156	10^{-9}

从表 2 可以看出, 新算法在 50 次的搜索过程中, 基本上找到了最优解, 同时对最优解的寻找也提高了一定的搜索精度, 特别对于例 1, 例 2, 例 3, 例 4, 新算法找到的最优解和最优值都比文献[5-9]要好, 显示了新的算法具有更高的稳定性和更好的鲁棒性; 例 6、例 7、, 例 8 是高维复杂的约束优化问题, 通常被认为是用进化算法很难优化的复杂函数, 从表 1 可以看出新算法基本上找到了最优解, 但是我们可以看出新算法在求解高维的约束优化问题时容易陷入局部极值, 并且运行的时间也比较长, 如何克服这一缺点提高算法的寻优性能也是我们今后进一步研究的课题。

结束语 对解决混合约束优化问题, 本文提出了基于乘子法的混合粒子群算法来求解约束优化问题, 该方法主要是利用乘子法来处理在算法迭代中出现的不可行粒子, 设计出新的混合的粒子群优化算法。通过乘子法的引入来处理约束条件, 可以减少适应值函数的复杂性, 提高计算结果的可靠性。数值实验显示了新算法的有效性、通用性和稳健性, 是一种有潜力的全局优化算法, 但对高维函数的优化, 效果仍然不是很理想, 如何能克服这一缺点, 提高算法的寻优性能是我们今后进一步研究的课题。

参 考 文 献

[1] Kennedy J, Eberhart R C. Particle swarm optimization[C] // Proc. IEEE International Conference on Neural Networks. Perth, Australia, IEEE ServiceCenter, Piscataway, NJ, IV: 1942-1948
[2] Shi Y, Eberhart R C. A Modified Particle Swarm Optimizer[C] // Proc. of the IEEE CEC. 1998: 69-73
[3] He Qie, Wang Ling. A hybrid particle swarm optimization with a feasible based rule for constrained optimization[J]. Applied Mathematics and Computation, 2007, 186: 1407-1422
[4] Lu Hai yan, Chen Wei qi. Self adaptive velocity particle swarm optimization for solving constrained optimization problems[J]. Journal of global optimization, 2008, 41(3): 427-445
[5] 高岳林, 李会荣. 非线性约束优化问题的混合粒子群算法[J]. 计算数学, 2010, 32(2): 135-146
[6] Shen Pei ping, Jiao Hong wei. A new rectangle branch-and-pruning approach for generalized geometric programming[J]. Applied Mathematics and Computation, 2006, 183: 1207-1038
[7] Shen Pei ping, Jiao Hong wei. Linearization method for a class of multiplicative programming with exponent[J]. Applied Mathematics and Computation, 2006, 183: 328-336
[8] Shen Pei ping, Zhang Ke cun. Global optimization of signomial geometric programming using linear relaxation[J]. Applied Mathematics and Computation, 2004, 150: 99-114
[9] Zhang Min, Luo Wei jian, Wang Xu fa. Differential evolution with dynamic stochastic selection for constrained optimization [J]. Information Sciences, 2008, 178: 3043-3074
[10] 王凌, 刘波. 微粒群优化与调度算法[M]. 北京: 清华大学出版社, 2008: 39-40