第 32 卷 第 4 期 计 算 机 工 程 2006 年 2 月
Vol.32 № 4 Computer Engineering February 2006
·网络与通信· 文章编号：1000—3428(2006)04—0145—02 文献标识码：A 中图分类号：TP 391

PageRank 算法研究

黄德才，戚华春
(浙江工业大学信息学院，杭州 310014)

摘 要：深入剖析了著名搜索引擎 Google 的关键技术 PageRank 算法，介绍分析了该算法的当前发展现状。并针对 PageRank 算法容易出现主题漂移现象，利用提出的二阶相似度改进算法。实验表明，改进的算法有利于减少主题漂移现象，提高用户对检索结果的满意度。
关键词：PageRank；搜索引擎；排序算法；Google

PageRank Algorithm Research

HUANG Decai, QI Huachun
(Information College, Zhejiang University of Technology, Hangzhou 310014)

【**Abstract**】This paper researches the PageRank algorithm of search engine Google, and introduces its development and the improved algorithm. It aims at that the PageRank algorithm always comes forth the topic drift, so this paper uses the new idea to improve this algorithm. The experiment indicates that the improved algorithm works more efficient than traditional one, reduces the topic drift and improves the satisfactory of the surfers obviously.
【**Key words**】PageRank; Search engine; Ranking algorithm; Google

由于互联网的快速发展，互联网上包含的信息越来越丰富。1994 年，最早的搜索引擎 World Wide Web Worm 就已经标引了 11 万网页；到 1998 年，互联网包含了大约 3.5 亿个文档[1]，并且还以每天 1 百万的文档数量在增长[2]；而著名的搜索引擎 Google 目前拥有 10 亿个网址，30 亿个网页，3.9 亿张图像，而且还在不断的增长中[3]。

面对如此海量丰富的信息，人们只能依靠网络搜索引擎来获得自己需要的信息。但由于传统的网络搜索引擎大多是基于关键字匹配的，其查询的质量不尽如人意。直到 1998 年，斯坦福大学的博士研究生 Sergey Brin 和 Lawrence Page，借鉴引文分析思想，提出 PageRank 算法[4]，算法通过分析网络的链接结构来获得网络中的权威网页，并在商用搜索引擎 Google 中大获成功，掀起网络链接分析的高潮。

1 PageRank 算法

1998 年，斯坦福大学的博士研究生 Sergey Brin 和 Lawrence Page 提出了网络链接分析的一个新算法 PageRank，该算法是建立在随机冲浪者模型上的。具体来说，假设冲浪者跟随链接进行了若干步的浏览后转向一个随机的起点网页又重新跟随链接浏览，那么一个网页的价值程度值就由该网页被这个随机冲浪者所访问的频率所决定。

PageRank 算法简单描述如下：u 是一个网页，$F(u)$是页面 u 指向的网页集合，$B(u)$是指向 u 的网页集合，$N(u)=|F(u)|$是 u 指向外的链接数，c 是规范化因子(一般取 0.85)。

那么网页 u 的 PageRank 值可以利用下面的公式计算：

$$R(u)=c\sum_{v\in B(u)}R(v)/N(v)$$

该算法的矩阵描述形式为：

设 A 为一个方阵，行和列对应网页集的网页。如果网页 u 有指向网页 v 的一个链接，则$A_{u,v}=1/N_u$，否则$A_{u,v}=0$。设 R 是对应网页集的 PageRank 值向量，则有$R=cAR$，可见 R 为 A 的特征根为 c 的特征向量。实际上，只需要求出最大特征根的特征向量，就是网页集对应的最终 PageRank 值。

但在上述算法中有一个比较大的漏洞。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。见图 1，有一组网页互相之间是彼此链接的，但都没有对组外网页的链接。这样，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 PageRank 值就一直滞留在这组网页内部，不能传递出去，这就是所谓的 PageRank 值沉淀现象。

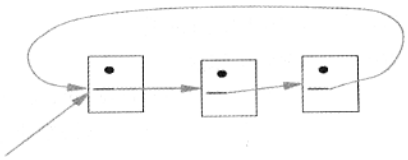

图 1 PageRank 值沉淀现象

为了解决这个问题，Sergey Brin 和 Lawrence Page 改进了算法，引入了衰退因子 $E(u)$，$E(u)$是对应网页集的某一向量，对应 PageRank 的初始值，算法改进如下：

$$R'(u)=c\sum_{v\in B(u)}\frac{R(v)}{N(v)}+cE(u)$$

其中，$\|R'\|_1=1$，对应的矩阵形式为 $R'=c(AR'+E)$。

2 PageRank 算法分析

由于 PageRank 算法是离线计算这个网络的 PageRank 值，在用户查询时仅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。

但由于仅仅利用了网络的链接结构，该算法还存在不少

作者简介：黄德才(1964—)，男，博士、教授，主研方向：人工智能与数据挖掘，数据库技术及其应用，算法设计与分析，系统工程理论方法；戚华春，硕士
收稿日期：2005-02-17 **E-mail**：qihuachun@zj165.com

的缺点：(1)比如 PageRank 算法偏重旧网页，因为旧网页被其它网页链接到的可能性更高，而事实上新的网页可能会具有更好的信息价值；(2)PageRank 算法偏重以.com 结尾的网站，因为这类网站往往是综合性网站，自然可以比其它类型的网站获得更多链接，而事实上某些专业网站对问题的阐述更具有权威性；(3)PageRank 算法无法区分网页中的超链接是和网页主题相关还是不相关，即无法判断网页内容上的相似性，这样就容易导致出现主题漂移问题。比如，Google，Yahoo 是互联网上最受欢迎的网页，拥有很高的 PageRank 值。这样，如果用户输入一个查询关键字时，这些网页往往也会出现在该查询的结果集中，并会占据很靠前的位置，而事实上这个网页与用户的查询主题有时并不太相关。

3 PageRank 的改进算法

继 Lawrence Page 提出 PageRank 算法以后，很多研究者对 PageRank 进行了改进，主要有以下几个方面。

3.1 加速评估的 PageRank 改进算法

针对缺点(1)，上海交通大学的张玲博士[8]提出了一个加速评估算法，该算法使得网络上有价值的内容以更快的速度传播；相反，一些已经陈旧的数据的页面评估值也将加速下滑。算法的核心思想是通过分析基于时间序列的 PageRank 值变化情况，预测某个 URL 在未来一段时期内的期望值并把它作为搜索引擎提供检索服务的有效参数。

算法定义了一个 url 的加速因子 AR 为

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

其中，$sizeof(D)$为整个页面集合的文档数。

加速后的 PageRank 表达式为

$$PR_{acclerate} = \frac{AR_{last} + BD}{M_{last}}$$

其中，AR_{last}是 url 最近一次的 AR 值，B 是该 url 一段时间内 PageRank 值的二次拟合曲线的斜率，D 为离最近一次页面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。

用户检索时，搜索引擎将按照预测的 PageRank 值的高低决定一个 URL 在检索结果中的位置。

3.2 主题敏感的 PageRank 算法

斯坦福大学计算机科学系 Taher Haveliwala[6]提出了一种主题敏感(Topic-sensitive)的 PageRank 算法解决了上文提到的缺点(3)。该算法考虑到有些页面在某些领域被认为是重要的，但并不表示它在其它领域也是重要的。所以，算法先根据 Open Directory[7]列出 16 个基本主题向量，并对每个网页离线计算出对这些基本主题向量的 PageRank 值。在用户查询时，算法根据用户输入的查询主题或查询的上下文，计算出该主题与已知的基本主题的相似度，在基本主题中选择一个最接近的主题代替用户的查询主题。算法的形式化表示如下：

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

其中 Pu 是网页 u 的主题敏感向量。

该算法可以有效地避免一些明显的主题漂移现象，比如在查询“美洲虎”时，有上下文的指引，算法能明确地区分用户是在查询的是：(1)美洲虎汽车；(2)美洲虎橄榄球队；(3)美洲虎产品；(4)哺乳动物美洲虎，提供高质量的推荐结果集。

3.3 结合链接分析和文本内容的 PageRank 算法

华盛顿大学计算机科学与工程系的 Matthew Richardson 和 Pedro Dominggos[8]认为用户从一个网页跳到另一个网页是受到当前网页内容和正在查询的主题的影响的，所以提出了一个结合链接和内容信息的 PageRank 算法。该算法的计算形式为

$$P_q(j) = (1-\beta)P_q'(j) + \beta \sum_{i \in B_j} P_q(i) P_q(i \to j)$$

这里 $P_q(i \to j)$指用户在查询主题 q 下从网页 i 跳转到 j 的可能性。$P_q'(j)$表示用户在网页中没有链出链接时，跳转到 j 的可能性。而 $P_q(j)$就是在查询 q 下的网页 j 的 PageRank 值。虽然 $P_q(i \to j)$和 $P_q'(j)$是任意分布的，为了统一这两个值，作者采用一个查询 q 和网页 j 的相关函数 $P_q(j)$分别计算这两个值：

$$P_q'(j) = \frac{R_q(j)}{\sum_{k \in W} R_q(k)}, \qquad P_q(i \to j) = \frac{R_q(j)}{\sum_{k \in Fi} R_q(k)}$$

其中，W 是整个网络的网页集合，Fi 是网页 i 的链出网页集合。有关函数 $P_q(j)$是任意的，一般取在网页 j 的文本中查询 q 的关键字出现的次数。

3.4 其它的改进算法

由于网页采用了半结构化的 HTML 语言，其包含有丰富的结构信息，上海交通大学的宋聚博士[9]认为在抽取网页的主题内容时应加以利用。对位于〈head〉、〈t it le〉、〈m eta〉以及〈a h ref=〉等标记之内的关键词无疑应该重视，计算时应赋予较大的权重系数。并对权重系数作出如下定义：

$$W_{ij} = tf_{ij} \lg\left(\frac{N}{df_j} + 0.5\right) \cdot func(t_j)$$

$$func(t_j) = \begin{cases} 3.0 & \text{关键字在链接文字中} \\ 2.0 & \text{关键字在} head / title / H_1 / H_2 \text{标记中} \\ 1.8 & \text{关键字在meta标记中} \\ 1.0 & \text{其它} \end{cases}$$

北京大学计算机系[10]也提出了利用 HTTP 协议，记录每个页面最近一次的修改时间，在运行分析算法的时候把页面修改时间作为控制参数，给予新修改的页面以较高的权值，而给予老页面以较低权值。

3.5 本文的改进

本文仔细分析了 PageRank 算法的随机冲浪模型，认为主题漂移现象发生主要在于传统的 PageRank 对 PageRank 值是平均分配的，这就导致主题无关网页获得它本不该获得 PageRank 值，使 PageRank 值在无效网页扩散。所以本文吸收了 Matthew Richardson 和 Pedro Dominggos[8] 的思想，即认为用户从一个网页跳到另一个网页是受到当前网页内容和正在查询的主题的影响，提出一个对两个网页进行相似度描述的二阶相似度概念。由于网页的相似性仅仅表现在两个网页之间，因此二阶相似度概念就定义为某个网页对在网络连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，$S_{i,j} = t$ 如果 i 有到 j 的链接，否则 $S_{i,j} = 0$。新算法描述为

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

其中 $S_{p.T_i}$是在相似度矩阵 S 中网页 p 对 T_i 的相似度值，$S(T_i) = \sum_{u \in B_{T_i}} S_{T_i,u}$，$B_{T_i}$是网页 Ti 的链出连接集合。

通过新算法就可以使网页的 PageRank 值在具有相似主题的网页上传播，减少主题无关网页对 PageRank 值的扩散。

为验证算法的有效性，我们对 http//ent.sina.con.cn 网站进行爬行，获得 10 万张有效网页，分别利用传统的 PageRank 算法和本文改进的算法进行 PageRank 值计算。并模拟查询 10 个不同的主题，每次取获得结果集的前 100 项。同时，为取得标准的结果集，我们利用 Google 搜索引擎的高级搜索功

(下转第 162 页)

一取 37db)，计算它们的 Waston 视觉感知距离，结果如表 1 示，从中可以看出在嵌入水印量相同的情况下，小波域 FS 算法的 Waston 视觉感知距离[6]更小，水印透明性更强。

表 1 小波域 IC 算法和小波域 FS 算法的水印透明性比较

图像	水印透明性（Waston 视觉感知距离）	
	小波域 IC 算法	小波域 FS 算法
Lena	4.54	3.71
Baboon	3.26	3.06
Pepper	4.27	3.26
Bridge	3.78	3.39

3.2 算法鲁棒性分析

我们采用 USC-SIPI 作为实验图像数据库对小波域 FS 算法进行分析，结果表明小波域 FS 算法具有良好的鲁棒性。算法鲁棒性的衡量标准采用提取水印与原始水印的相关性。

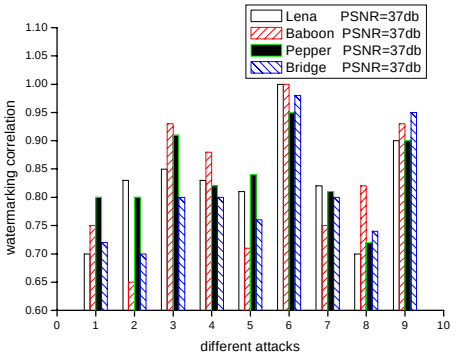

图 5 小波域 FS 算法鲁棒性测试

选用代表了不同的纹理特性的典型载体图像 Lena、Baboon、Pepper 和 Bridge(512×512 大小)为实验载体图像，水印为 64×64 大小二值 UP 图像，调整算法的嵌入强度，使嵌入水印后的载体图像 PSNR=37db。对图像进行各种攻击，算法的抗攻击性检测结果如图 5 示，图中的横坐标表示不同的攻击处理，从 1~9 分别为 3×3 低通滤波、3×3 中值滤波、3×3 高通滤波、叠加高斯噪声(0,0.001)、叠加椒盐噪声(0,0.01)、缩放、旋转、25%JPEG 压缩和 50%JPEG 压缩处理。从中可以看出遭受各种攻击后小波域 FS 算法提取的水印相关性依然很高(Corr>0.65)，说明小波域 FS 算法具有良好的鲁棒性。

4 结束语

本文提出了一种新的基于模糊理论的小波域鲁棒性水印算法，该算法首先建立一个模糊系统，以图像块的熵和标准差特征值为系统的输入参数，通过选择合适的输入输出隶属度函数、制定合理的模糊规则，最终输出具有一定连续性的图像块水印嵌入强度。实验结果表明，与基于图像块分类的小波域算法相比，该算法在保持良好鲁棒性的同时，进一步增强了水印的透明性。

参考文献

1 陆哲明，姜守达，董寒丽. 基于人类视觉系统的自适应水印嵌入算法[J]. 哈尔滨工业大学学报, 2003, 35(2): 138-141.

2 汪春生，程义民，王以孝. 一种基于块分类的自适应数字水印算法[J]. 计算机工程与应用, 2002, 38(21): 106-110.

3 易开祥，石教英. 自适应二维数字水印系统[J]. 中国图象图形学报, 2001, 6(5): 444-449.

4 易开祥，王　铁，石教英. 基于 DCT 域的自适应二维数字水印系统[J]. 计算机应用, 2000, 20(增刊): 12-15.

5 王慧琴，李人厚，王志雄. 一种模糊自适应图像水印算法的研究[J].西安交通大学学报, 2002, 36(2): 182.

6 Cox I J. 王　颖译. 数字水印[M]. 北京: 电子工业出版社, 2003.

(上接第 146 页)

能，单独在 http://ent.sina.con.cn 网站查询刚才的 10 个主题，取每次查询结果的前 100 项为标准结果集。见图 2，其中系列 1 是改进算法获得结果集的查全率，系列 2 是传统算法获得结果的查全率。横坐标是 10 个不同的查询主题。

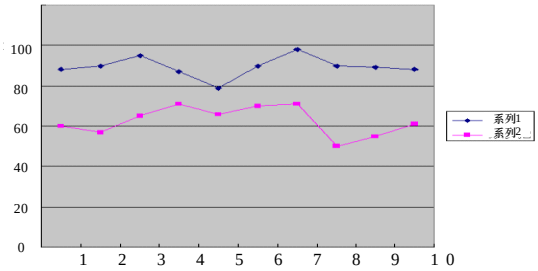

图 2 算法获得结果集的查全率

可见本文改进的算法表现出较好的查全率，主题漂移的现象得到了有效遏制。

4 结论

本文详尽地介绍了 PageRank 算法及其发展现状，并对它做了改进。但由于该算法仅仅只对网络的链接结构进行分析，不可避免地存在主题漂移的问题，尤其当网页的主题不止一个时，漂移的现象更加严重。因此如何有效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来改进算法解决主题漂移问题还有很长的路要走。

参考文献

1 Yates R B, Neto B R. Moderm Information Retrieval[M]. New York, USA: Addison Wesley, 1999.

2 Chakrabarti S, Dom B, Gibson D. Hypersearching the Web[Z]. http://www.sciam.com/, 1999-06.

3 Brin S, Page L. The Anatomy of a Large-scale Hypertextual Web Search Engine[C]. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998: 107-117.

4 Page L, Brin S. The PageRank Citation Ranking: Bringing Order to the Web[EB/OL]. http://www.db.stanford.edu/~backub/PageRanksub.ps, 1998~2001.

5 Kleinberg J. Authoritative Sources in a Hyperlinked Environment[J]. Extended Version in Journal of the ACM, 1999, 46(5): 604-632.

6 Haveliwala T H. Topic-sensitive PageRank[C]. Proceedings of the Eleventh International World Wide Web Conference, Hoho Lulu Hawaii, 2002.

7 The Open Directory Project: Web Directory for over 2.5 Million Urls[EB/OL]. http://www.dmoz.org/.

8 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.

9 宋聚平，王永成. 对网页 PageRank 算法的改进[J]. 上海交通大学学报, 2003, 37(3): 397-400.

10 Chakrabarti S, Dom B, Gibson D, et al. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text[C]. Proceedings of the 7th ACM-WWW International Conference. Brisbane: ACM Press, 1998: 65-74.