一种改进的 PageRank 算法①

古文丽，陈　玮，陈　娇，陆晓野

(上海理工大学 光电信息与计算机工程学院，上海 200090)

摘　要：研究了现有的基于链接结构的 PageRank 算法。结合网页链接分析和网页内容相关性分析提出了一种改进的 PageRank 算法，从分析网页内容相关性的角度解决相关性需求，从网页链接分析的角度解决权威性需求，并且实验证明，改进的 PageRank 算法优于传统的 PageRank 算法的排序结果。

关键词：PageRank；网页排序；链接分析；文件相关性

Improved PageRank Algorithm

GU Wen-Li, CHEN Wei, CHEN Jiao, LU Xiao-Ye

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200090, China)

Abstract: This paper researched the PageRank algorithm based on the existing link structure and proposed a modified PageRank algorithm, combining the analysis of webpage analysis and Web content analysis. The relevance and authority algorithm demands are met by analyzing the similarity of the contents of Web pages and the link structure respectively, and experiments show that the improved PageRank algorithm is superior to the traditional approach in terms of ordering results.

Key words: PageRank; Web page ranking; link analysis; document relevance; information retrieva

随着 Web 数据的急剧增长，搜索引擎成为用户获取信息的重要工具。Internet 已成为世界上最丰富和最密集的信息来源。然而，这些大量无序的信息也给信息检索带来了很多的问题。如何让用户先获得最权威和查询最相关的网页，是目前迫在眉睫的问题，也是搜索引擎商业运行成功的关键。如何把查询最权威、最相关的搜索结果页中的网页排到最前列是解决这个问题的关键。要解决这个问题，就要充分利用网络的各种信息，包括网络链接，网页内容和用户访问离开的信息。受欢迎的传统的 PageRank[1]和 HITS[2]算法是简单的从排序分析的角度，而忽略了网站和其他的信息，故很难得到较好的排序结果。

1　PageRank算法介绍

PageRank 的基本思想：如果一个页面被其他许多页面引用，则这个页面可能是重要页面；一个页面尽管没有被多次引用，但被一个重要页面引用，那么这个页面可能也是重要页面；一个页面的重要性被平均分配并传递到它所引用的页面。PageRank 技术基于整个 Web 的链接结构来计算各网页的重要性，它认为用户能够通过网页之间的超链接访问到整个网络。PageRank 计算公式可以表示如下：

$$R(u)=c\sum_{v\in B(u)}R(v)/N(v) \quad (1)$$

构造有向 Web 图 G=（V，E），其中顶点 V 为所有网页集合，边 E 为网页间的链接集合，网页 A 中有指向网页 B 的链接表示顶点 A、B 间存在一条边，则公式(1)中 B（u）表示直接指向网页 u 的网页集合（u 的入链网页的集合；N(v)表示网页 v 出链网页的数量），R(v)/N(v)是指网页 v 把自己的 PageRank 值平均分配给自己网页中的出链网页，c=1[3]。

互联网中可能存在这样的情况：有一组网页互相

① 收稿时间:2011-05-27;收到修改稿时间:2011-06-09

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，因此传递过来的 PageRank 值就一直滞留在这组网页内部，不能传递出去，这就是 PageRank 值沉淀现象（LinkSink）。为了避免沉淀现象，对公式(1)引入一个阻尼系数 d，使其变为：

$$R(u) = (1-d) + d\sum_{v\in B(u)} R(v)/N(v) \qquad (2)$$

PageRank 计算公式可以从概率的角度解释为一个随即的网页浏览者随机选择一个网页后，不断的点击网页上的连接，但是从不返回；除非最后厌烦了才随机选择另一个页面,随机网页浏览者访问某个页面的随机概率就是该页面的 PageRank 值；阻尼系数 d 就是随机浏览者在某个页面会厌烦，然后选择一个新页面的概率，取值为 0~1，一般取 0.85,页面的 PageRank 值越高，则网页浏览者发现它的概率也越高,PageRank 算法的优点：它是与查询无关的静态算法，所有网页的 PageRank 值均可以通过离线计算获得，加速了查询相应时间[3]。

2 PageRank算法分析

由于 PageRank 算法是离线计算网络的 PR 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，因此具有很高的相应速度，并且搜索引擎 Google 的成功也验证该算法是合理、高效的。

但是，只有网络的链接结果的使用，此算法也有不小的缺点：1)PageRank 算法更注重.com 结尾的网站，.com 结尾的网站一般是综合性的网站，自然能比其他类型的网站获得更多的联系，但实际上一些对这个问题的论述更具有权威性的专业网站也许更有权威性；2)PageRank 算法不会区分网页和网络链接相关或不相关的主题，那就是无法确定网页内容的相似性，所以很容易造成主题漂移的问题。谷歌，雅虎作为互联网上最受欢迎的网站，自然具有较高的 PR 值。因此，如果用户输入一个查询关键词，这样网页通常会出现在查询结果集中，并将占据相当靠前的位置。但是事实上有时会与用户查询的主题不太相关；3)PageRank 算法是偏重于旧的网页，因为其他网页上的旧链接的可能性会更大，而事实上，在新网站上会找到更多的有信息价值的资料[4]。

3 对PageRank改进的算法算法

从 PageRank 算法以及随机漫游模型中，可以看出在迭代过程中权值是按当前网页的出度平均分配的，但实际上，从链接结构上看网页按入度和出度的不同是具有相对重要性的，入度和出度较大的网页比较重要，因此应该分得较高的权重。基于此方面考虑 Xing[5] 提出加权 PageRank 算法，其中网页的重要性和网页的入度、出度成正比。

网页间的链接反应的是一种认可关系，网页 A 中有链接指向网页 B，说明网页 B 的内容与 A 相关或者具有一定的价值，同一网页中不同链接指向的网页的内容与当前网页内容的相关程度是有差别的。基于此思想，Ingongngam 等人[6]提出了以主题为中心的 PageRank 算法，算法指出网页权值的分配应用和网页的内容的相似度成正比，被链接的网页内容与当前网页的内容越相似分配到的权值比重就越大。

4 本文改进PageRank算法

目前国内外对 PageRank 算法改进最多的就是基于超链接的算法改进，主要是对链接的链入和链出的改进，并且在链接的权重上做出的一些研究。使一个网页的链接对另一个网页的权重影响更合理。在大量研究中，都是通过链接分析来如何准确分析该网页的权值，但是具有权威性的网页不一定是我们要查找的网页，这称作“主题漂移”。所谓主题漂移就是已经查找的网页和用户所要查找的主题相关性不大。加权 PageRank 算法在分配权值时以网页重要性为比例，因此知名网站会获得更高的权重，所以在一定程度上加剧了主题漂移的发生。在以主题为中心的 PageRank 算法中根据网页的相关性来分配权值可以有效解决主题漂移现象，但确忽略了排序中对权威性的需求。

本文综合上述两方面的思想，从链接分析的角度解决权威性的问题，从内容相关性分析的角度解决相关性问题，对 PageRank 算法进行改进。本文改进的 PageRank 算法（Extended PageRank）算法公式如下：

$$ER(u) = (1-d) + d\sum_{v\in B(u)} ER(v)(W_{a(v)} \times W_{r(v)}) \qquad (3)$$

其中：Wa(v), Wr(v)分别表示网页 V 对 U 的权威性和相关性。Wa(v) ×Wr(v)表示不同权威网页上内容的相

关性。即：权威网页上的内容相关性要大于非权威网页上内容相关性。

在链接关系的基础上，加入页面与查询主题的相关性权重，以使得所产生的 PageRank 值高的页面是针对用户查询主题的，这就形成了加权 PageRank 算法。在加权 PageRank 算法中有：

$$Wa(v)=Win(v,u)\times Wout(v,u) \tag{4}$$

其中，$Win(v,u)=\dfrac{Iv}{\sum_{q\in G(u)} Iq}$，

$$Wout(v,u)=\frac{Ov}{\sum_{q\in G(u)} Oq} \tag{5}$$

Win(v,u)和 Wout(v,u)分别表示基于出度和入度的权重因子，Iv 和 Ov 分别是网页 v 的入度和出度，Wa(v)可以有衡量网页权威性的其他任何算法计算，Wr(v)是用网页内容间相关的程度来衡量相关性的比例，假设 W(v,u)表示网页 v 和网页 u 的相关性程度值，那么有：

$$W_{r(v)}=\frac{W(v,u)}{\sum_{q\in G(u)} W(q,u)} \tag{6}$$

加权 PageRank 的实际意义可以解释为：假设网页上有一个主题查询者，它从初始页面出发，按照页面链接前进，从不执行后退操作。对于没个页面来说，浏览者对此页面中的每个链接感兴趣的概率是和此链接主题的相关性成正比的。如果有很多页面指向一个页面，那么这个页面的 PageRank 值就会高，但加权的 PageRank 不一定高，和页面中大部分都为主题相关的页面有关；如果加权的 PageRank 很高的页面指向它，这个页面的加权 PageRank 也会很高。

在以主题为中心的 PageRank 算法中根据页面内容将其归类为不同的主题，然后针对不同的主题进行相似度的计算，主要步骤为：首先确定主题的类别，将常见查询内容归纳总结，创建主题列表。然后把数据库中的网页和主题列表中的不同主题匹配，匹配过程是通过采用 VSM 空间向量的模型进行计算，通过计算可以得到该网页相对于各个主题的相似度得分，从而可以明显提高主题相关度。两个文档的相似性则由表示文档的向量内积来进行计算。假设网页 u 和 v 的文档向量分别表示为 $U=(u_1,u_2,\ldots,u_m)$，$V=(v_1,v_2,\ldots v_m)$，那么他们的相关程度可以表示为

$$W(v,u)=\frac{U\bullet V}{|U|\times|V|}=\frac{\sum_{i=1}^{m} u_i v_i}{\sqrt{\sum_{i=1}^{m} u_i^2 \sum_{i=1}^{m} v_i^2}} \tag{7}$$

其中 u_i 和 v_i 是关键词 i 在网页 u 和 v 中的权值，一般按照经典的 TF-IDF 算法计算，定义关键词 i 在文档 j 中的权值 w_{ij}，则有 $w_{ij}=tf_{ij}lg(N/df_i)$，其中 tf_{ij} 表示关键词 i 在文档 j 中出现的次数，df_i 是包含关键词 i 的文档数量，N 表示文档总数。

5　实验结果分析

利用网络蜘蛛程序在 http://news.sohu.com 在网络上爬行一段时间后，获取了 25784 张新闻网页。网页内容可以分为国内、国际、财经、体育、历史等。本实验中改进的 PageRank 算法与经典的 PageRank 算法，加权 PageRank 算法，以主题为中心的 PageRank 算法进行比较分析。我们输入保障性住房，利比亚，足球，个人所得税，抗日战争作为关键词。取搜索前 200 个作为标准结果集。通过人工评价查询的网页与要查找的主题的内容相关性。实验结果数据见图 1。

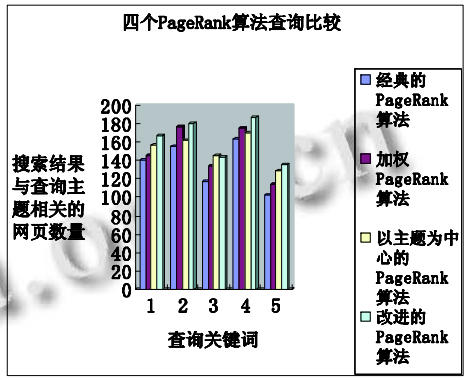

图 1　各关键词在不同算法查询下的相关网页数

其中横坐标 1 2 3 4 5 分别代表的关键词：保障性住房、利比亚、足球、个人所得税、抗日战争。纵坐标表示的是搜索结果中与要查询的主题相关的网页数量。两条条形柱分别表示在改进的 PageRank 算法和经典的 PageRank，加权 PageRank 算法，以主题为中心的 PageRank 算法下的实验结果。

同时为了比较查询结果的优劣，本文还考察了查准率，由于一般情况下用户只会关注前二十页的查询结果，所以本文仍然取前 200 项查询结果来考察查准率。

可以用如下公式计算：

$$R = R/D \quad (8)$$

其中 R 是搜索到的相关文档数，D 是检索到的文档数。实验结果见表 1.

表 1 不同的查询关键词在每种算法下的查全率

关键词 / 算法	保障房	利比亚	足球	个人所得税	抗日战争
经典的 PageRank 算法	0.7	0.78	0.59	0.82	0.51
加权的 PageRank 算法	0.725	0.89	0.67	0.875	0.57
以主题为中心的 PageRank 算法	0.785	0.81	0.725	0.855	0.645
改进的 PageRank 算法	0.84	0.9	0.72	0.94	0.68

通过实验我们发现改进的 PageRank 算法优于经典的 PageRank 算法，加权 PageRank 算法和以主题为中心的 PageRank 算法。实验结果表明改进的 PageRank 算法很好的达到了算法改进的初衷。

6 总结与展望

本文结合网页链接分析和网页内容相关性分析两个方面提出一种改进的 PageRank 算法，本论文改进算法较好的满足了用户的查询需求，既解决了权威性的要求，又解决了主题相关性的要求。算法只是基于网页的结构与内容挖掘，下一步的工作可以研究关于用户的使用挖掘，以便更好的满足个性化的主题搜索。

参考文献

1 Page L, Brin S, Motwani R, et al. The pagerank citation ranking: Bringing order to the web.Stanford Digital Libraries SIDL-WP-1999-0120.1999.

2 Kleinberg J. Authoritative sources in a hyperlinked.Environment. Proc. of the Ninth Annual ACMSIAM Symposium on Discrete Algorithms. San Francisco, California, 1998: 668–677.

3 王冬,雷景生.一种基于 PageRank 的页面排序改进算法.微电子学与计算机,2009,26(4):57–60.

4 彭聪,吴强等.一种改进型的网页排序算法.微计算机信息,2010,11(3):72–74.

5 Xing WP, Ghorbani A. Weighted PageRank Algorithm. Communication Networks and Services Research. Proc. of Secnod Annual Conference. 19-21 May 2004: 305–314.

6 Havelieala TH. Topic-sensitive PageRank. Proc. of the 11th International World Wide Web Conference. Hawaii, 2002: 517–526.

（上接第 248 页）

该算法只有一个控制参数，通过随机变量的取值实现对目标粒子的逼近。实验结果表明，该模型是有效的可行的，为粒子群算法的改进提供了一种新途径。

参考文献

1 Kennedy J, Eberhart RC. Particle swarms optimization. Proc. of IEEE international conference on Neural Networks.USA: IEEE Press,1995:1942–1948.

2 郭文忠,陈国龙.求解 VLSI 电路划分问题的混合粒子群优化算法.软件学报,2011,22(5):833–842.

3 Lin SW, Ying KC, Chen SC, et al. Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Systems with Applications, 2008,35(4):1817–1824.

4 Cai XJ, Cui ZH, Zeng JC, et al. Dispersed particle swarm optimization. Information Processing Letters, 2008, 105(6): 231–235.

5 Liu Y, Qin Z, Shi ZW, et al. Center particle swarm optimization. Neurocomputing, 2007,70(4–6):672–679.

6 张英杰,邵岁锋.一种基于云模型的云变异粒子群算法.模式识别与人工智能,2011,24(1):90–96.

7 朱海梅,吴永萍.一种高速收敛粒子群优化算法.控制与决策,2010,25(1):20–24.

8 方伟,孙俊,谢振平,须文波.量子粒子群优化算法的收敛性分析及控制参数研究.物理学报,2010,59(6):3686–3694.

9 赵淑清,郑薇.随机信号分析.哈尔滨:哈尔滨工业大学出版社,1999.54–56.