第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·软件技术与数据库· 文章编号：1000—3428(2007)18—0053—03 文献标识码：A 中图分类号：TP393

基于 PageRank 算法的权威值不均衡分配问题

田 甜，倪 林
(中国科技大学电子工程与信息科学系，合肥 230027)

摘 要：PageRank 对所链接的网页的“权威值”存在平均分配的思想，由于互联网的网页是千差万别的，因此这种方法存在一定的局限性。该文利用了 Web 链接结构，提出了一种权威值不均衡分配的方法(IPR)，通过与 PageRank 算法相比，IPR 的排序结果比 PageRank 提高了近 90%的相关度。

关键词：网页结构挖掘；网页排序；改进的 PageRank

Problem of Unequal Authorities Assignment Based on PageRank Algorithm

TIAN Tian, NI Lin
(Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027)

【Abstract】 There is equitable distribution thinking in authorities of other pages linked to PageRank. As pages are different, the method has some limitation. This paper takes advantage of Web link structure, proposes an unequal way to treat the different pages when distributing authorities and call it improved PageRank. Experimental results show that IPR improves relation degree by about 90%, comparing to PageRank.

【Key words】 Web structure mining; page ranking; improved PageRank(IPR)

随着互联网信息的增长，用户不得不利用搜索引擎去查找所需的信息，由于网上信息量的增长，与用户相关的信息也随着增加，因此搜索引擎可以找到大量与用户查询相对应的网页。如何把贴近用户的信息放在搜索结果的前面，逐渐成为用户和学者共同关注的问题，据 iProspect 的调查报告，2006 年 62%的用户只点击搜索结果页第 1 页的结果，而高达 90%的用户只点击搜索结果页的前 3 页里的结果。而在 2002 年，这两个数字分别为 48%、81%。另外，根据 iResearch 的《个人门户发展趋势研究报告》，高达 57.9%的网民表达了对搜索引擎结果中冗余信息多的不满。这些变化说明，用户对搜索引擎的要求越来越高，他们希望花在寻找结果上的时间越来越少。因此，排序查找到的结果比搜索本身更为重要。

研究搜索引擎的排序算法的改进，逐渐成为热点问题，随着深入的研究，越来越多的人意识到排序质量不令人满意的原因为：不是网页提供给的信息太少，而是可用的信息太少，或所用信息不恰当。

笔者对 PageRank 算法进行了改进。PageRank 算法的基本思想是：把当前网页的权威值平均分配给它的全部链接。但互联网的网页是千差万别的，即使链接在同一个网页上的链接，也是有差别的。PageRank 算法这种平均分配权威值的方法，在一定程度上影响了网页的排序质量。笔者提出了一种按照不同网页分配权威值的方法，能弥补这一缺陷，通过实验表明，改进算法具有更加优良的排序质量。同时，改进算法的框架是在 PageRank 算法的基础上提出的，并可以方便地被移植。

1 PageRank 算法

PageRank[1]主要思想是把网页的链接分成前向链接(in-links)和反向链接(back-links)，也称为引用链接和被引用链接，反向链接的数量和质量决定PageRank值。PageRank值由反向链接所决定，反向链接表示所考察的网页可被其他网页引用，反向链接数目越多，则说明该网页被引用较多，其可能是很重要的网页，因此，可以凭借反向链接的数目来确定该网页的重要程度。如果一个网页被很多垃圾网站引用，和一个网站被很多重要网站引用，效果是不同的，在PageRank算法中，如果一个网页被很重要的网页引用，则该网页的PageRank值会提高，如果网页被许多垃圾网站或者权威值不高的网站引用，对于提高其权威值，同样是没有帮 助的。

按上述思想，可以得到 PageRank，即

$$PR(u) = c \sum_{v \in B(u)} PR(V) / N_v \quad (1)$$

其中，$B(u)$ 表示直接指向网页 u 的网页集(u 的反向链接的集合)；Nv 表示网页 v 前向链接的数量；$PR(V)/N_v$，是指网页 v 把自己的权威值平均分配给从自己网页中指出的链接，$c = 1$。

PageRank算法为了解决权威值沉积问题，采用下面的模式，以适合通常情况下的排序算法[2]，即

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \quad (2)$$

PageRank 算法对一个网页的 PageRank 值分配是没有针对性的，它把自己的权威值平均分发给各个网页，为了进一步地体现公平公正和有效的网络秩序，以下将介绍 IPR 算法。

基金项目：国家自然科学基金资助项目(60372059)；安徽省自然科学基金资助项目(03042206)
作者简介：田 甜(1982－)，男，硕士研究生，主研方向：Web 数据挖掘；倪 林，副教授
收稿日期：2006-09-24 **E-mail：**tian123@mail.ustc.edu.cn

2 改进的 PageRank 算法

改进的 PageRank 跟 PageRank 是基于同一个事实的，即：(1)被引用得越多的网页，其权威值越高；(2)被权威网页引用过的网页，其权威值越高。这 2 点是 PageRank 取得成功的重要原则，跟现实生活中比较吻合。

改进的 PageRank 算法还注意以下因素：(1)权威值非常高的网页被引用的次数会大大多于“引用其他网页”的次数。(2)如果引用一个权威网页，就应该多分配一些权威值；如果引用普通网页，应该少分配给他一些权威值。

根据上述思想，IPR 对权威网页进行了重新描述：最权威的网页是被多个网页引用，或者能被权威网页引用的网页，同时被引用的次数大大超过本身引用其他网页的次数。如果缺少一个条件，就被评为次等权威网页。

2.1 算法描述

IPR 可以用 PageRank 算法类似表达式加以描述，即

$$IPR(u) = c \sum_{v \in B(u)} a_v IPR(v) \tag{3}$$

其中，a_v表示网页v分配给网页u权威值的比重，同式(1)，$c = 1$，a_v可以通过如下 2 步得到。

(1)计算某个网页 j 的反向链接和前向链接之比，即

$$IO_j = \frac{BackLink_j}{InLink_j} \tag{4}$$

其中，$BackLink_j$ 表示网页 j 反向链接数目，相当输入了权威值；$InLink_j$ 表示前向链接数目，相当于从网页输出了权威值，称之为输入输出比值 IO；IO 表示该网页从别的网页获取权威值的能力，IO 相对其他网页越大，则更容易获取到大的权威值，但是，这里将产生一个问题，当前向链接的数目为零时，IO 将出现无穷大的情况。

为了解决这一问题，首先把 IO 为无穷大(即 $InLink_j$ 为零)的网页设置为一个比较小的常数。

$$IO(\infty) = m \ (m \text{ 为常数}) \tag{5}$$

(2)设指向u的网页即在 $v \in B(u)$ 中，有网页 $v_1, v_2, v_3, ..., v_n$，先计算网页v_i的全部n个前向链接 $S_1, S_2, S_3, ..., S_n$ 的反向链接总数和前向链接总数的比值IO，设分别为 $IO_1, IO_2, ..., IO_n$，再根据这个比值重新分配权威值，按照反向和前向链接的比率来分配PageRank值，即式(3)中a_v的值(设这个网页为 $v \in B(u)$ 中的第j个，现在用a_j表示a_v)，则a_v应该表示为

$$a_v = a_j = \frac{IO_j}{\sum_{i=1 \cdots n} IO_i} \tag{6}$$

网页U和S从网页V_1得到不同权威值，如图 1 所示。

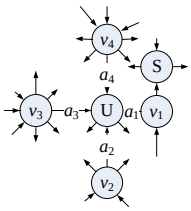


图 1 网页U和S从网页V_1得到不同权威值

网页U从 4 个网页获取权威值，获取的比例分别是 4 个网页所分配的 a_1, a_2, a_3, a_4，某个页V_i分配给网页U比例，是按照网页U的输入输出比占全部V_i所指向网页的输入输出比之和的大小所决定的。如对于网页V_1而言，它指向两个网页，全部IO比之和为网页S的IO(大小为 2/2)和网页U的IO (4/2)的和，因此网页V_1分配给网页U的权威值比重a_1为

$$a_1 = \frac{2}{1+2} = \frac{2}{3}$$

网页V_1把自己权威值的 2/3 赋给了网页U，如果是PageRank算法，网页V_1只把自己权威值的 1/2 赋给网页U (因为网页V_1有 2 个前向链接，所以可以进行平均分配)。

2.2 通常情况下的 IPR 算法

对于 IPR，同样存在权威值沉积问题，经过修正后的改进 PageRank 算法表达式为

$$IPR(u) = (1-d) + d \sum_{v \in B(u)} a_v IPR(v) \tag{7}$$

各参数的意义同PageRank算法相同，为了保持和PageRank算法的一致性，采用的同样为随机漫游模型，取d=0.85。IPR的计算复杂度与PageRank算法的复杂度可以相提并论(复杂度稍偏大)，PageRank算法的另外一种改进方法，WPR(Weighted PageRank)描述如下[3]：

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)}$$

WPR 算法中引入了乘法，而且 $W^{in}_{(v,u)}$ 和 $W^{out}_{(v,u)}$ 的计算也比较繁杂，尤其在迭代中，算法复杂度将会提高。

3 试验

3.1 试验条件

为了验证改进算法的排序结果，本文将试验建立在国内著名的新闻网站网易：http://news.163.com 中。试验方法如下：

(1)网页采集工作，自行设计的一个多线程网页抓取 Spider 程序，共抓取了 15 398 张网页。

(2)利用开元项目 Lucene 提供的源代码(http://lucene.apache.org)，由于针对的是中文网页，因此首先进行中文分词的处理，然后对网页的 title 域、Meta 域、H1 域建立索引，并存储好网页的位置，以提供基于文本的搜索和结果返回。

(3)实现 PageRank 和 IPR 算法。将得到的排序值存入 MySql 数据库中，两种算法采用的是迭代计算(40 次)。迭代算法的具体实现可参考文献[2]。IPR 和 PageRank 具有完全相同的算法框架。

(4)针对不同的查询要求，找到 Lucene 对应的索引文件，读出数据库中相应网页的排序值，按照排序值降序输出结果。

(5)评估 PageRank 和 IPR 的结果。整个开发过程是用 Java 完成，PageRank 以及改进的 PageRank 算法不针对查询主题， PageRank 排序独立于查询主题。但在本试验中，设计的排序是针对于查询主题的，原因是：网页数量相对比较大，不容易对网页进行评估。对于某个查询关键词，找到这个关键词的所有网页，然后按照排序值降序排列，看看所得到的结果网页是否真正与查询项保持一致(即相关程度)，如果相关程度比较好，则认为排序的质量优越，反之，则认为质量不佳。这个道理跟搜索引擎的排序是相同的，只是没有结合网页相关性和排序之间的关系，而是只要网页存在关键词，就将这个网页列入查询结果中。如果排序算法比较优越的话，相关度高的网页会自动地往结果的前面靠齐。

3.2 实验结果

对于某个查询关键词，出现的结果可能会比较多，但网页不一定跟查询相关性完全相关，为此，笔者设计了一个相关度等级函数：

$$P_i = \begin{cases} 1.0 & \text{网页与查询项完全相关} \\ 0.3 & \text{网页与查询项部分相关} \\ 0.1 & \text{网页与查询项微弱相关} \\ 0 & \text{网页与查询项毫不相干} \end{cases} \tag{8}$$

为了保证衡量准则的有效性，为试验建立了一个评估小组(共 5 个人)，当大多数人认为该出现某种结果时，则把网页归结为这种结果，如果出现无法断定的情况，重新评估该

网页。

对于某个查询条件，最终结果的评定是按照排序靠前的15个网页相关性之和来确定的，由于排序的优劣不单单与该网页是否相关，而且取决于相关网页出现的位置，因此相关函数设计成与位置和相关程度的大小有关的函数，即

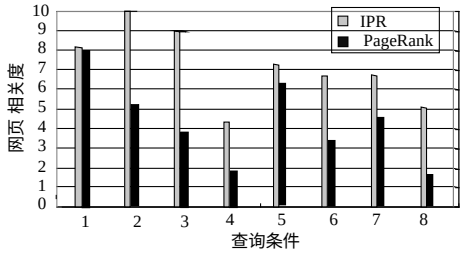
$$R = \sum_{i=1\_to\_15} P_i \bullet (15 - i + 1) \tag{9}$$

对8个主题的查询见表1。为了直观地了解PageRank和IPR 算法之间的排序效果，按关键词从上至下的顺序，得到图2。

表1　相同查询主题条件下 PageRank 和 IPR 的排序结果比较

查询关键词	网页数量	<完全相关，部分相关，弱相关>		网页相关度	
		IPR	PageRank	IPR	PageRank
朝鲜核问题	38	<9,3,1>	<9,5,0>	8.18	8.08
中日关系	179	<11,1,2>	<5,3,2>	10.4	5.19
联合国改革	68	<8,3,2>	<4,2,6>	9.01	3.89
国企改革	19	<4,4,2>	<5,4,1>	4.35	1.95
环境保护	124	<8,1,4>	<6,2,6>	7.27	6.22
住房问题	22	<6,2,1>	<7,1,1>	6.72	4.54
…	…	…	…	…	…

图2　相同主题下 PageRank 和 IPR 的排序结果



从表1和图2可以看出来，相对于PageRank，IPR具有明显的优势，通过计算可知，在8个查询条件下，从上至下各个查询项相关度分别提高了0.012倍、1.0004倍、1.316倍、1.231倍、0.169倍、0.959倍、0.480倍、2.024倍。通过计算，IPR相对PageRank，使排序的相关度平均增长了89.937 5%。IPR能提供比PageRank性能更优越的排序算法。

IPR具有如下2个优点：(1)具有更优越的排序结果；(2)与PageRank共用算法框架，能方便地在现有的PageRank算法应用中，并移植算法。

4　总结

本文通过对PageRank算法研究，提出了一种不均衡分配权威值的改进PageRank算法，通过计算前向链接和反向链接数量的比值关系来确定权威值的划分，从而细致地划分了网页的权威级别，提高了网页的排序质量。这些改变和提升，并没有以大幅度提高算法的复杂度为代价，这些优点对网页的排序带来了非常好的前景。IPR和PageRank具有相似的算法框架，可以很好地融入其他PageRank的改进算法中，如PowerRank[4]算法等，从而提升搜索结果的准确性和高效性。

致谢　向为本次试验做数据测试的钱功伟、黄晶、曹荣、赵小青等表示感谢。

参考文献

1 Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking Bringing Order to the Web[EB/OL]. (1998-04). http://www-db.stanford.edu/~backrub/pageranksub.ps.

2 Haveliwala T H. Efficient Computation of PageRank[EB/OL]. (1999-10). http://www.stanford.edu/~taherh/papers/efficient-pr.pdf.

3 Xing W, Ghorbani A. Weighted PageRank Algorithm[C]//Proceedings of the 2nd Annual Conference on IEEE Communication Networks and Services Research. 2004.

4 Lu Yizhou. The Powerrank Web Link Analysis Algorithm[C]//Proc. of the 13th International World Wide Web Conference on Alternate Track Papers & Posters. 2004-05: 254-255.

(上接第52页)

基于位置的空间数据融合方法对点状对象之间的融合是比较有效的。但是对于形状更为复杂的对象(例如线状的对象和多边形对象)来说就不实用了，而实际应用中对于复杂形状对象之间的融合，甚至是不同类型对象之间的融合是更为常用的，因此，基于位置的空间数据融合方法还不能完全满足实际需要。

4　总结

尽管基于特征的技术简单易行，但是在实际情况中，有很多空间对象的特征属性是缺失或者错误的，这就大大影响了它的效果。而基于本体的技术应该是未来空间数据融合技术乃至更广范围内数据融合技术的重要发展方向，它能够从根本上解决数据融合时，无法确定表示同一概念的实体的问题，但是从目前来看，基于本体的技术依赖于本体库的建立，当前在本体库稀少的情况下还很难得以广泛地应用。

基于位置的技术是目前最可行的空间数据融合技术，因为它的融合依据是空间数据所必须具备的特性——位置属性，但是基于位置的技术不能保证可以完全找出所有正确的融合集，如何提高该技术的正确性将是研究的方向。另外，目前基于位置的技术只考虑点状对象之间的融合，对于形状更加复杂的空间对象(例如线状对象和多边形对象)之间的融合只是简单地用这些对象的中心点等来进行模拟。在实际情况中，有很多对象是具有相同或者相近的模拟点，却代表完全不同的实体(例如 2 条垂直相交的道路)。研究点状对象之间的融合技术是远远不够的，还有更多关于复杂空间对象之间的融合技术需要研究，例如多边形对象之间的融合技术，甚至是多边形对象和点对象的融合技术。如何将基于空间属性的技术与现有地理信息系统集成技术结合起来，形成完整的“地理信息集成技术”是一个值得研究的方向。

参考文献

1 Fonseca F T, Egenhofer M J, Agouris P. Using Ontologies for Integrated Geographic Information Systems[J]. Transaction on GIS, 2002, 6(3): 231-257.

2 Samal A, Seth S, Cueto K. A Feature Based Approach to Conflation of Geospatial Sources[J]. International Journal of Geographical Information Science, 2004, 18(5): 459-489.

3 Guarino N. Formal Ontology and Information Systems[M]. Amsterdam, Netherlands: IOS Press, 1998

4 Beeri C, Kanza Y, Safra E, et al. Object Fusion in Geographic Information Systems[C]//Proceedings of the 30th VLDB Conference. 2004: 816-827.

5 Sonka M, Hlavac V, Boyle R. Image Processing, Analysis, and Machine Vision[EB/OL]. (1999-02). http://www.icaen.uiowa.edu/~sonka/ps-files/CFAI.pdf.