维基百科

PageRank

维基百科，自由的百科全书

Google的工具栏标示出的中文维基百科首页的PageRank

PageRank，又称网页排名、谷歌左侧排名、PR，是Google公司所使用的对其搜索引擎搜索结果中的网页进行排名的一种算法。

佩奇排名本质上是一种以网页之间的超链接个数和质量作为主要因素粗略地分析网页的重要性的算法。其基本假设是：更重要的页面往往更多地被其他页面引用（或称其他页面中会更多地加入通向该页面的超链接）[1]。其将从A页面到B页面的链接解释为“A页面给B页面投票”，并根据投票来源（甚至来源的来源，即链接到A页面的页面）和投票对象的等级来决定被投票页面的等级。简单的说，一个高等级的页面可以提升其他低等级的页面。

该算法以谷歌公司创始人之一的拉里·佩奇（Larry Page）的名字来命名。[2]谷歌搜索引擎用它来分析网页的相关性和重要性，在搜索引擎优化中经常被用来作为评估网页优化的成效因素之一。

目前，佩奇排名算法不再是谷歌公司用来给网页进行排名的唯一算法，但它是最早的，也是最著名的算法。[3][4]

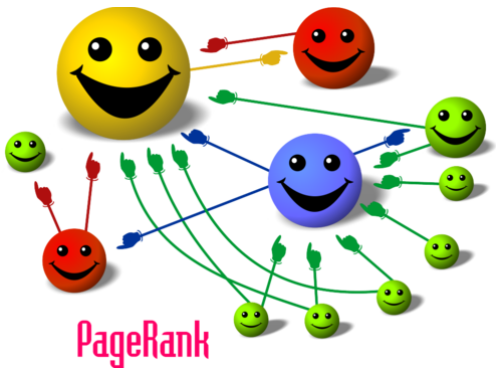
目录

- 概述
- 算法
  - 简化版本
  - 完整版本
- 缺陷
- 从谷歌工具栏中移除
- 脚注
- 参考文献
- 外部链接
- 参见

概述

PageRank是一种链接分析算法，它通过对超链接集合中的元素用数字进行权重赋值，实现“衡量集合范围内某一元素的相关重要性”的目的。该算法可以应用于任何含有元素之间相互引用的情况的集合实体。我们将其中任意元素E的权重数值称为“E的PageRank”（The PageRank of E），用符号表示为$PR(E)$。其他的因素，类似“作者排名（Author Rank）”同样可以影响到该元素的权重值。

PageRank的结果来源于一种基于图论的数学算法。它将万维网上所有的网页视作节点（node），而将超链接视作边（edge），并且考虑到了一些权威的网站，类似CNN。每个节点的权重值表示对应的页面的重要度。通向该网页的超链接称做“对该网页的投票（a vote of support）”。每个网页的权重值大小被递归地定义，依托于所有链接该页面的页面的权重值。例如，一个被很多页面的链接的页面将会拥有较高的权重值（high PageRank）。



PageRank的卡通概念图，图中笑脸的大小与指向该笑脸的其他笑脸的数目成正比.

大量关于PageRank的学术论文在Page和Brin的原版论文前就已有之。[5]在实际情况中，PageRank很容易被利用。相关的研究往往会关注那些因受到影响而出现错误的PageRank结果，以找到一种有效地避免其被错误地影响的方法（如忽略部分错误的链接）。[6] 2005年初，谷歌公司为网页链接推出一项新属性nofollow，使得网站管理员和博客作者可以创建一些不计票的链接，也就是说这些链接不算作“投票”，从而实现抵制垃圾投票的目的。

Google工具条上的PageRank指针从0到10。它似乎是一个对数标度算法，细节未知。虽然**PageRank**是谷歌的商标，其技术亦已经申请专利，但是专利权属于斯坦福大学，而非谷歌公司。

PageRank算法中的点击算法是由乔恩·克莱因伯格(Jon Kleinberg)提出的。而其他的基于链接的网页排名算法，则包括乔恩·克莱因伯格发明的HITS算法，IBM CLEVER Project，TrustRank算法以及hummingbird算法等等。

算法

PageRank算法通过输出概率分布来体现某人随机地点击某个链接的概率。PageRank值（PR）可以在任何规模的文件（document）集合中计算得出，而每个链接都指向该集合中的某个特定文件。相关研究论文指出，在初次计算前，总概率将被均分到每个文件上，使得集合中的每个文件被访问的概率都是相同的。接下来在重复多次的计算（又称为“迭代”）中，算法将根据集合的实际情况不断调整PR值，使得其越来越接近最真实的理论值。

最终的概率将通过一个在0与1之间的数值体现。概率为0.5通常意味着该事件有50%的可能性发生。因此，“PR=0.5”代表“有50%的可能性，某人点击了一个随机的链接并访问了该链接指向的文件”。

简化版本

假设一个由4个网页组成的集合：A，B，C和D。同一页面中多个指向相同的链接视为同一个链接，并且每个页面初始的PageRank值相同，最初的算法将每个网页的初始值设定为1。但是在后来的版本以及下面的示例中，为了满足概率值位于0到1之间的需要，我们假设这个值是0.25。

在每次迭代中，给定页面的PR值（PageRank值）将均分到该页面所链接的[注 1]页面上。

如果所有页面都只链接至A，那么A的PR值将是B，C及D的PR值之和，即：

$$PR(A) = PR(B) + PR(C) + PR(D)$$

重新假设B链接到A和C，C链接到A，并且D链接到A,B,C。最初一个页面总共只有一票。所以B给A,C每个页面半票。以此类推，D投出的票只有三分之一加到了A的PR值上：

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

换句话说，算法将根据每个页面连出总数 $L(x)$ 平分该页面的PR值，并将其加到其所指向的页面：

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

最后，所有这些PR值被换算成百分比形式再乘上一个修正系数 d[注 2]。由于“没有向外链接的网页”传递出去的PR值会是0，而这会递归地导致指向它的页面的PR值的计算结果同样为零，所以赋给每个页面一个最小值$(1-d)/N$：

$$PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots \right) d + \frac{1-d}{N}$$

- 需要注意的是，在Sergey Brin和Lawrence Page的1998年原版论文中给每一个页面设定的最小值是$1-d$，而不是这里的$(1-d)/N$，这将导致集合中所有网页的PR值之和为N（N为集合中网页的数目）而非所期待的1。

因此，一个页面的PR值直接取决于指向它的的页面。如果在最初给每个网页一个随机且非零的PR值，经过重复计算，这些页面的PR值会趋向于某个定值，也就是处于收敛的状态，即最终结果。这就是搜索引擎使用该算法的原因。

完整版本

这个方程式引入了**随机浏览者**（random surfer）的概念，即假设某人在浏览器中随机打开某些页面并点击了某些链接。为了便于理解，这里假设上网者不断点击网页上的链接直到进入一个没有外部链接的网页，此时他会随机浏览其他的网页（可以与之前的网页无关）。

为了处理那些“没有外部链接的页面”（这些页面就像“黑洞”一样吞噬掉用户继续向下浏览的概率）所带来的问题，我们假设：这类页面链接到集合中所有的网页（不管它们是否相关），使得这类网页的PR值将被所有网页均分。对于这种残差概率（residual probability），我们引入阻尼系数 d（damping factor），并声明 $d = 0.85$，其意义是：任意时刻，用户访问到某页面后继续访问下一个页面的概率，相对应的 $1 - d = 0.15$ 则是用户停止点击，随机浏览新网页的概率。d 的大小由一般上网者使用浏览器书签功能的频率的平均值估算得到。

所以，对于某个页面i，其对应PR值大小的计算公式如下：

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{\mathrm{PR}(p_j)}{L(p_j)}$$

- 这里，$p_1, p_2, \ldots, p_N$ 是目标页面p_i，$M(p_i)$是链入p_i 页面的集合，$L(p_j)$是页面p_j 链出页面的数量，而N是集合中所有页面的数量。

集合中所有页面的PR值可以由一个特殊的邻接矩阵的特征向量表示。这个特征向量R为：

$$\mathbf{R} = \begin{bmatrix} \mathrm{PR}(p_1) \\ \mathrm{PR}(p_2) \\ \vdots \\ \mathrm{PR}(p_N) \end{bmatrix}$$

同时，R也是下面的方程组的解：

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

- 这里的邻接函数（adjacency function） $\ell(p_i,p_j)$ 代表“从页面j指向页面i的链接数”与“页面j中含有的外部链接总数”的比值

如果p_j不链向p_i，则前面提到的“从页面j指向页面i的链接数”为零。将情况一般化：对于特定的j，应有：

$$\sum_{i=1}^{N} \ell(p_i,p_j) = 1,$$

由于上述修改后的邻接矩阵的巨大的eigengap值，几次迭代后即可在极高的精确度下估计PageRank特征向量R的值。

缺陷

PageRank算法的主要缺点在于旧的页面的排名往往会比新页面高。因为即使是质量很高的新页面也往往不会有很多外链，除非它是某个已经存在站点的子站点。这也是PageRank需要多项算法结合以保证其结果的准确性的原因。例如，PageRank似乎偏好于维基百科页面，在条目名称的搜索结果中，维基百科页面经常在大多数页面甚至所有页面之前，此现象的原因则是维基百科内部网页中存在大量的内链，同时亦有很多站点链入维基百科。[7]

Google经常处罚恶意提高网页PageRank的行为。至于其如何区分正常的链接和不正常的链接，这仍然是商业机密。但是在Google的链接规范中已清楚地说明，哪些是属于违反规范的行为。[8]

从谷歌工具栏中移除

2009年10月14日，Google员工苏珊·莫斯科（Susan Moskwa）确认该公司已将PageRank从其网站管理员工具中移除。她表示：“我们长久以来一直在告诫人们不应该过分注重PageRank；很多网站站长似乎认为PageRank是他们需要时刻关注的最重要的指标，而这几乎是错误的。”[9]然而在苏珊确认后两天，PageRank又在谷歌工具栏（Google Toolbar）上重新显示，但其指示器（indicator）在谷歌公司自家的Chrome浏览器上已不可用。

同时，公众可见的PageRank的数据更新周期也越来越长，它的最后一次更新是2013年11月份。

2014年10月7日，谷歌员工John Mueller表示：“我们可能不会继续更新PageRank，至少工具栏上的PageRank是这样。”[10]

2016年4月15日，谷歌公司停止向公众开放PageRank数据。就在几个月前，谷歌也声明将会将PageRank评分自谷歌工具栏中移除。[11] 但是，今后谷歌公司在对其搜索引擎的搜索结果进行排名时，仍然会使用PageRank中的数据。[12]

脚注

1. 此处以及下文中的链接均为单向链接，即A–>B不等价于B–>A
2. 有关该系数的明确定义请见下面的#完整版本

参考文献

1. Facts about Google and Competition. [12 July 2014]. （原始内容存档于2011-11-04）.
2. Google Press Center: Fun Facts. www.google.com. [2018-12-09]. （原始内容存档于2001-07-15）.
3. Sullivan, Danny. What Is Google PageRank? A Guide For Searchers & Webmasters. Search Engine Land. [2018-12-09]. （原始内容存档于2016-07-03）.
4. Cutts, Matt. Algorithms Rank Relevant Results Higher. www.google.com. [19 October 2015]. （原始内容存档于2013-07-02）.
5. Brin, S.; Page, L. The anatomy of a large-scale hypertextual Web search engine (PDF). Computer Networks and ISDN Systems. 1998, **30**: 107—117 [2018-12-10]. ISSN 0169-7552. doi:10.1016/S0169-7552(98)00110-X. （原始内容存档 (PDF)于2015-09-27）.
6. Gyöngyi, Zoltán; Berkhin, Pavel; Garcia-Molina, Hector; Pedersen, Jan, Link spam detection based on mass estimation, Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06, Seoul, Korea) (PDF): 439—450, 2006 [2018-12-10], （原始内容存档 (PDF)于2014-12-03）.
7. Web Structure Age and Page Quality. (PDF). [2002-05].
8. 链接方案 - Search Console帮助. support.google.com. [2016-12-16]. （原始内容存档于2016-12-21）.
9. Susan Moskwa. PageRank Distribution Removed From WMT. [October 16, 2009]. （原始内容存档于2009-10-17）.
10. Google Toolbar PageRank Finally & Officially Dead?. Search Engine Land. 2014-10-07 [2016-12-16]. （原始内容存档于2016-12-20） （美国英语）.
11. Schwartz, Barry. Google Toolbar PageRank officially goes dark. Search Engine Land. [2018-12-10]. （原始内容存档于2016-04-21）.
12. Southern, Matt. Google PageRank Officially Shuts its Doors to the Public. Search Engine Journal. [2018-12-10]. （原始内容存档于2017-04-13）.

外部链接

- Our Search: Google Technology (http://www.google.com/technology/)（页面存档备份 (https://web.archive.org/web/20080623233116/http://www.google.com/technology/)，存于互联网档案馆） by Google （英文）
- The PageRank Citation Ranking: Bringing Order to the Web (https://web.archive.org/web/20091118014915/http://ilpubs.stanford.edu:8090/422/) by Larry Page *et al.* （英文）
- The PageRank Result (https://web.archive.org/web/20150302083435/http://google-pagerank.info/) （英文）

参见

- Google
- Google轰炸
- SimRank

取自“https://zh.wikipedia.org/w/index.php?title=PageRank&oldid=66515860”

本页面最后修订于2021年7月10日 (星期六) 06:57。

本站的全部文字在知识共享 署名–相同方式共享 3.0协议之条款下提供，附加条款亦可能应用。（请参阅使用条款）
Wikipedia®和维基百科标志是维基媒体基金会的注册商标；维基™是维基媒体基金会的商标。
维基媒体基金会是按美国国内税收法501(c)(3)登记的非营利慈善机构。