搜索引擎 PageRank 算法的改进

张延红
（浙江万里学院，宁波 315100）

摘 要：在研究搜索引擎关键技术和的基础上，剖析了 PageRank 算法，并针对 PageRank 算法的缺陷提出了改进方案.

关 键 词：搜索；PageRank；WPageRank
中图分类号：TP391.3 **文献标识码：**A **文章编号：**1671－2250（2005）04－0035－03
收 稿 日 期：2005－03－17

作 者 简 介：张延红，浙江万里学院计算机与信息学院助教.

在互联网发展初期，网站相对较少，信息查找比较容易. 然而伴随互联网爆炸性的发展，网上信息浩如烟海，普通网络用户想找到所需的资料难于大海捞针，所以迫切需要一种优异的搜索服务，将网上繁杂的内容整理成为可方便获取的信息. 搜索引擎技术为解决这一难题作出了突出贡献，然而搜索引擎提供的结果集中页面质量的好坏以及高质量的页面能否在结果集中有较好的排名，对搜索引擎用户来说具有重要意义，同时也是衡量搜索引擎技术优劣的关键指标. 所以对页面进行重要性评估并按重要性排序是搜索引擎要解决的技术核心.

1 页面的重要性指标

页面的重要性指标主要有页面级别（Pageranking）、反向链接数(In-linking-counts)、相关性(Relevance)、页面的不过时性（Up-to-Dateness）和页面的距离(Distance)等. 页面级别是指页面的重要性等级；反向链接数是单纯意义上的受欢迎度指标，反向链接是否来自推荐度高的页面是页面有根据的受欢迎指标，并且反向链接源页面的链接数是被选中的几率指标，被许多页面链接的受欢迎的页面，必定是优质的页面，所以以反向链接数作为受欢迎度的一个指标是很自然的想法；相关性是指页面与查询主题的关联程度，不只是关注查询关键词在网页上出现的次数，还对该网页的内容（以及该网页所链接的内容）进行全面检查，从而确定该网页是否满足查询要求；页面的不过时性指有越多的新建的页面指向某一个页面，则这个页面内容过时的可能性越小. 为了防止人为的级别优化，页面的距离被用来影响对链接的评价. 站内链接的权重小于站间链接的权重. 页面的距离可由页面是否在一个站内、一个服务器及物理距离等决定. 如何利用这些指标来衡量页面的重要性，排除人为因素对搜索结果的影响，为查询用户找到最重要、最有用的网页，是排名算法要解决的问题.

2 排名技术

早期的搜索工具是以在数据库中找到匹配信息的先后次序排列搜索结果，因此毫无信息关联度可言. 而现代的搜索引擎都在搜索结果排列中引入关键字串匹配程度的概念. 旨在向用户提供真正最重要、最有用的网页. 搜索引擎的搜索结果名次排列是根据网页是否确实提供了搜索对象要找的内容和提供内容的受欢迎程度来决定其排名次序. 排名算法中最有影响力当数 PageRank（页面级别）算法.

3 PageRank 算法

PageRank 算法计算出网页的 PageRank 值，从而决定网页在结果集中的出现位置，PageRank 值越高的网页，在结果中出现的位置越前.

3．1 PageRank算法定义[1]

PageRank 算法基于下面 2 个前提：

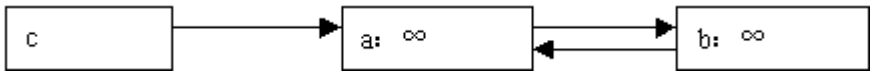
前提 1：一个网页被多次引用，则它可能是很重要的；一个网页虽然没有被多次引用，但是被重要的网页引用，则它也可能是很重要的；一个网页的重要性被平均的传递到它所引用的网页. 这种重要的网页称为权威（Authoritive）网页.

前提 2：假定用户一开始随机的访问网页集合中的一个网页，以后跟随网页的向外链接向前浏览网页，不回退浏览，浏览下一个网页的概率就是被浏览网页的 PageRank 值.

简单 PageRank 算法描述如下：u 是一个网页，$F(u)$ 是 u 指向的网页集合，$B(u)$ 是指向 u 的网页集合，$N(u)$ 是 u 指向外的链接数，显然 $N(u)=\left|F(u)\right|$ ，c 是一个用于规范化的因子（通常取 0.85），则 u 的 Rank 值计算如下：

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

如果有 2 个相互指向的网页 a，b，它们不指向其他任何网页，另外有某个网页 c，指向 a，b 中的某一个，比如 a，那么在迭代计算中，a，b 的 Rank 值不分布出去而不断的累计．如下图：



为了解决这个问题， Lawrence Page 和 Sergey Brin 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 Rank 的初始值，算法改进如下：

$$R'(u)=c\sum_{v\subset B(u)}R(v)/N(v)+cE(u) \tag{2}$$

其中，$\left\|R'\right\|=1$.

3．2 PageRank 算法特点

这个算法不以站点排序，而是对单个页面进行级别排序，页面网页级别由一个个页面各自独立决定；页面的网页级别由链向它的页面的网页级别决定，但每个链入页面的贡献的值是不同的. 如果T_i页面中链出越多，它对当前页面A的贡献就越小. A的链入页面越多，其网页级别也越高；阻尼系数的使用，减少了其他页面对当前页面A的排序贡献.

3．3 Pagerank算法的缺陷[2]

若互联网上的资源具有同一主题性，PageRank 系统可说是尽善尽美. 但互联网上的资源涵盖了上百万甚至更多的主题，而且在实际应用中，查询用户所寻找的往往是一些具有特定主题的信息. 而 PageRank 单纯根据一个网页上被链接的站点数量和质量来给该网页分配一个绝对的“ 重要性值 ”. 同时亦将链接页面的页面等级考虑在内. 指向一个网页的外部链接页的页面等级越高，则该链接页面传递给该网页的页面等级值也就越高. 但是，“ 页面等级值 ”并非针对查询词语，因而一个网页即使只是在内容中偶然提到了一个和查询主题偏离的关键词语，也会因其居高的页面等级值而获得一个比较高的排名，从而影响了搜索结果的相关性与精准性.

4 对 Pagerank 算法的改进

Pagerank 算法计算出来的页面等级值不一定能准确反映出与查询主题的相关性，Pagerank 算法需要进一步改进.

4．1 页面与主题的相关性判定[3]

为了进一步提高搜索结果页面的准确率，需要对页面进行主题相关性评价. 查询主题Q，文档D_j和主题Q的相似度按如下公式计算：

$$SIM(Q,D_j)=\sum(U_{iq},U_{ij})/\sqrt{\sum U_{iq}^2*\sum U_{ij}^2} \tag{3}$$

$U_{iq} = FREQ_{iq} * IDF_i$ ，$U_{ij} = FREQ_{ij} * IDF_i$ ，$FREQ_{iq}$ ＝项i在查询Q中的出现次数，$FREQ_{ij}$ ＝项i在文档D_j中的出现次数，IDF_i是WWW上包含项i的文档数目的估计值.

4．2　对 PageRank 算法的改进

对 PageRank 算法做如下改进：

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 WPageRank 算法. 改进公式如下：

$$WPR(A) = (1-d) + d(WPR(T_1)) \times \frac{SIM(Q,T_1)}{\sum_{1}^{k_1} SIM(Q,D_i)} + WPR(T_2) \times \frac{SIM(Q,T_2)}{\sum_{1}^{k_2} SIM(Q,D_i)} + ... + WPR(T_n) \times \frac{SIM()Q,T_n}{\sum_{1}^{k_n} SIM(Q,D_n)}) \quad (4)$$

其中，A为给定的一个网页，假设指向它的网页有T1，T2，…，Tn. k_1，k_2，…，k_n分别是网页T_1，T_2，…，T_n中所含的链接数. WPR(A)为A的WPageRank值，d为衰减因子(也设成 0.85).

WPageRank的实际意义可以这样来解释. 假设Web上有一个主题浏览者，WPageRank(即函数WPR(A))是它访问到页面A的概率. 它从初始页面集出发，按照页面链接前进，从不执行“ back ”操作. 在每一个页面，浏览者对此页面中的每个链接感兴趣的概率是和此链接与主题的相关性成比例的. 浏览者也有可能不再对本页面的链接感兴趣，从而随机选择一个新的页面开始新的浏览. 这个离开的可能性设为d. 从直观上看，如果有很多页面指向一个页面，那么这个页面的PageRank就会比较高，但WPageRank值不一定很高，除非这很多的页面中大部分都为与主题相关的页面；如果有WPageRank很高的页面指向它，这个页面的WPageRank也会很高.

5　结束语

搜索引擎技术是正在迅速发展的研究领域，虽然现阶段各种类型搜索引擎的算法很丰富，核心聚焦于对页面级别的计算，面级别精确度对搜索结果的最终排名客观公正性有重要影响. 迄今为止，没有任何一个排名算法是完美的，所以搜索引擎都停止了只使用一种有价值的算法去决定排名的做法，而是吸收多种排名算法之精华. 为了使排名结果更客观、合理，排名算法还需要在很多个方面继续做深入研究，相信在不久的将来会有更多更好的研究成果出现，能够真正帮助网络用户在 WWW 海量的信息里更快速准确地找到需要的信息，实现搜索引擎的最终目标.

参考文献：

[1]朱炜,王超,李俊,潘金贵.WEB 超链分析算法纵览[EB/OL].http://www.isedb.com/news/article/658?t=reviews&id=658, 2004.3.

[2]Serge Thibodeau. PageRank: meet Hilltop[EB/OL].http://www.isedb.com.2004.2.

[3]李盛韬.基于主题的 Web 信息采集技术研究[D].北京:中国科学院计算机技术研究所.硕士论文，2002.

Improvement of PageRank Algorithm for Search Engine

ZHANG Yan-hong
(Zhejiang Wanli University, Ningbo 315100)

Abstract： Based on the study of the search engine key technology, we analyze the PageRank algorithm and propose an improvement aimed at the deficiency of the PageRank algorithm.

Key words： search；PageRank；WPageRank