

# 搜索引擎 PageRank 算法的比较与改进

张毅 张冬梅  
(重庆大学软件学院 400044)

摘要:搜索引擎查询的结果按照一定的规则排序供用户查看,这种规则就是搜索引擎排序算法。目前大多数搜索引擎仍然是通过对搜索引擎的链接关系进行分析,找到相对比较重要的网页。这些算法大多是以 PageRanks 等经典算法为基础,进行改良,加入各自偏重的参数形成综合的排序模型。

关键词:搜索引擎 PageRank 排序算法 原创文章  
中图分类号:TP311.52 文献标识码:A

文章编号:1674-098X(2008)07(c)-0018-01

## 1 引言

随着信息技术的不断发展,特别是互联网应用的迅速普及,网络规模的爆炸性增长,网上的信息正以几何级的速度在增加。搜索引擎已成为互联网应用的重要组成部分,对互联网的普及正产生着极大的影响。而其中搜索引擎的核心技术——排序算法也变得极为重要,一个合理的搜索引擎排序算法可为互联网营造一个公平的竞争环境。

## 2 PageRank算法

Google 的两位创始人 Sergey Brin 和 Larry Page 于 1996 年提出了 PageRank 算法。它利用这一公式计算链接到某一网页的网站数量,然后按照从 1 到 10 分别给与表示重要度的分数。链接到这个网页的站点越多,PageRank 的分数就会越高。

PageRank 算法依靠的是网民对站点的支持率,利用大量的链接结构表明某个单独页面的价值。它就像是一个由互联网上所有其他页面发起的投票,并以此决定一个页面的。一个指向某页面的链接代表一种支持票,如果没有链接指向它,那就相当于没有支持票。

PageRank 的值定义如下:

假定页面 A 有  $T_1, \dots, T_n$  这些页面指向它(即  $T_1, \dots, T_n$  引用页面 A)。则页面 A 的 PageRank 值由下面的公式得出:

$$PR(A) = (1-d) + d \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)} \quad (2.1)$$

PageRank 值在整个网页群体中成概率分布,所以全部网页的 PageRank 值之和为 1。

## 3 PageRank算法评价

PageRank 依赖的一个基本事实是:超链接代表了一种权威认可关系。对于早期的主要依靠手工编写的网页中,这个基本事实是普遍成立的,尤其在学院派的个人主页中更是如此(它更类似与社会网络的情况)。然而随着 Web 技术的发展,很多企业的网页不再是由手工编写完成,而是通过各种模版生成或者通过关系数据库自动生成。还有一些网页,甚至采用了欺骗搜索引擎的网页编写技巧。无论如何,一个趋势就是,Web 的网页链接结构不再继续

符合基本事实的假定。

## 4 经典排序算法对网络原创文章的不公平之处

现有的经典排序算法都是根据分析网页链接来进行排序的,但它们都存在对新兴原创网站极为不利的共同问题,也即没有考虑文章内容相同的原创文章和转载文章之间的排序问题。

由于网络知识产权保护的不力及其实施的难度,原创网站的文章很有可能被其他网站随意地转载。而搜索引擎在排序的时候,却不会考虑网页的转载与被转载问题,假如转载文章的站点“经营”得好的话,转载页面得搜索引擎排名完全有可能要高于被转载页面。这样一来,网站经营者就会把原本要投入到网站内容的人力物力投入到网站经营上去,如在其它网站做广告、找专业地 SEO 公司进行优化等。这种不公平的竞争无疑会影响到现代网络的发展。

故此,为了提高现在网络竞争的公平性,更好的“保护”网络原创文章,有必要对搜索引擎的排序结果进行修正,使之更加有利于原创文章,以促进现代网络事业的健康发展。

## 5 PageRank算法的缺点与改进

PageRank 算法偏重旧网页。由公式 2.1 可以看出,决定一个网页 PR 值的主要因素是指向该网页的链接个数,如果一个网页被放到 internet 上不久,由于时间短暂,许多其它网页还没有指向它,通过公式 2.1 计算出的 PR 值也就会很低。在搜索引擎返回的结果中往往会把它排在较后的位置,这样,返回结果中新的网页反而被放在后面,可能正好与用户的需求恰恰相反,因为许多情况下,用户想首先看到最新的网页。

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,这样就很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该查询结果中,并占据很靠前的位置。而事实上,这个网页与用户的查询主题有

时并不太相关。

对于 PageRank 算法对于旧网页偏重的问题,上海交通大学的张岭博士提出了一个加速评估算法<sup>[21]</sup>,该算法使得网络上有价值的内容以更快的速度传播;相反,一些陈旧的数据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 在未来一段时间内的期望值,并把它作为搜索引擎提供检索服务的有效参数。

针对 PageRank 算法的主题漂移问题,斯坦福大选计算机科学系 Taher Haveliwala 提出了一种主题敏感(Topic-sensitive)的 PageRank 算法。该算法考虑到有些页面在某些领域被认为是重要的,但这并不表示它在其它领域也是重要的。所以,算法先根据 Open Directory 列出的 16 个基本主题向量,对每个网页离线计算出对这些主题向量的 PageRank 值。用户在查询的时候,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。

## 6 结语

本文介绍了现在搜索引擎排序算法的基础——经典搜索引擎排序算法。对 PageRank 算法的原理进行了分析和有缺点的比较,归纳了国内外对经典算法的改进方法。同时又指出了经典链接分析算法对网络原创文章不利的这一事实,提出了要通过提高原创文章的搜索引擎排名来“保护”网络原创文章这一改进思路。

## 参考文献

- [1] 田梅梅. 搜索引擎 Google 与百度的比较分析[J]. 云南档案, 2007, 01.
- [2] 张兴华, 王仕雪. 几种英文搜索引擎的性能及检索功能[J]. 现代情报, 2005, 05.
- [3] 陈洁惠. 搜索引擎排序算法的研究[D]. 淮海大学硕士学位论文, 2007, 3.
- [4] 曹军. Google 的 PageRank 技术剖析[J]. 情报杂志, 2002, 10.
- [5] 杨思洛. 搜索引擎排序技术的研究[J]. 现代图书情报技术, 2005, 1.
- [6] 宋聚平. 对网页 PageRank 算法的改进[J]. 上海交通大学学报, 2003, 3.

## 参考文献

- [1] 白中英. 计算机组成原理[M]. 科学出版社.
- [2] 李学干. 计算机系统结构[M]. 西安电子科技大学出版社.

加的新部件与硬盘的通信不会给 CPU 增加太多负担。这一结构的改变如能实现,可提高计算机系统的工作速度,并使与 CPU 近端的“提速”工作产生完美的结合,

从而提高计算机系统的整体效率。限于本人的知识有限,该体系的改变理论上和设计上是否合理有待专家、学者评判。