

搜索引擎技术分析

—— Google 的 PageRank 技术剖析

蒋得虎

(兰州商学院信息工程学院 甘肃兰州 730020)

中图分类号: O229

文献标识码: A

1. PageRank 的基本原理

PageRank 的发明者对网络超链接结构和文献引文机制的相似性进行了研究, 把引文分析思想借鉴到网络文档重要性的计算中来, 利用网络自身的超链接结构给所有的网页确定一个重要性的等级数, 当从网页 A 链接到网页 B 时, 就认为网页 A 投了网页 B 一票, 增加了网页 B 的重要性。最后根据网页的得票数评定其重要性, 以此来帮助实现排序算法的优化, 而这个重要性的量化指标就是 PageRank 值。简单地说, PageRank 就是要从链接结构中获取网页的重要性, 而网页的重要性决定着同时也依赖于其他网页的重要性。

2. PageRank 的定义

根据上面的基本原理, L. Page 等给出 PageRank 的简单定义: 令 u 为一个网页, $N(v)$ 表示从网页 v 向外的链接数目, $B(u)$ 表示链接到网页 u 的网页集合, $R(u)$ 表示网页 u 的 PageRank 值, C 为规范化因子, 作用是保证所有网页的 PageRank 总和为常量。例如为保证总的 PageRank 值为 1, 可以通过网页 PageRank 总和的倒数求得。

$$R(u) = C * \sum_{v \in B(u)} R(v) / N(v) \quad (\text{定义 1})$$

必须注意的是定义(1)有一个假设前提, 即所有的网页形成一个牢固的链接图(即每个网页能从其他网页通过超链接达到)。从定义(1)可以看出, 网页的 PageRank 是一个由网络的超链接结构所产生的一个网页重要性等级值, 所有的网页的 PageRank 值都可以根据其他网页的 PageRank 值和链接的数量来计算得到, 即所有链接到它的网页的 PageRank 值除以各自向外的链接数的商进行求和。

3. PageRank 的计算

前面给出的定义 1 本身是一个 PageRank 的计算公式, 利用这个公式, 可以计算网页集合中所有网页的 PageRank 值。假设 S 为整个网页的总和。因为所有的网页的 PageRank 值开始是未知的, 所以我们进行平均的分配, 给每个网页的 PageRank 都赋以 $1/S$ 。再根据公式 1 进行计算。然后对得到的值再次利用公式 1 计算。这样反复地计算。直到计算得到的 PageRank 值收敛于一个相对固定的数(ϵ)。也就是说, 根据超链接结构计算出的所有网页的重要性等级趋于稳定, 这时停止计算。

为了方便论述, 我们将网络结构简单化, 假设仅仅有 5 个网页。

从上面可以看出, PageRank 的总和为 1(因为采用舍入计算, 所以没有完全精确)。其 PageRank 的分布也是完全合理的, 这里网页

次数	P(1)	P(2)	P(3)	P(4)	P(5)
1	0.25	0.2	0.15	0.2	0.2
2	0.25833	0.21667	0.13333	0.19167	0.2
3	0.2625	0.22708	0.12708	0.18542	0.19792
4	0.26521	0.23417	0.12438	0.18083	0.19542
5	0.26717	0.23934	0.12316	0.17734	0.19299
6	0.26868	0.24331	0.12266	0.17459	0.19075
7	0.26989	0.24649	0.12254	0.17236	0.18873
8	0.27089	0.24909	0.12261	0.17050	0.18691
10	0.27245	0.25313	0.12306	0.16757	0.18378
12	0.27363	0.25615	0.12367	0.16535	0.18120
14	0.27456	0.25851	0.12429	0.16360	0.17903
16	0.27532	0.26043	0.12491	0.16218	0.17717
20	0.27649	0.26535	0.12603	0.15998	0.17414
24	0.27756	0.26551	0.12700	0.15836	0.17178
28	0.27803	0.26718	0.12784	0.15709	0.16986
36	0.27903	0.26962	0.12921	0.15523	0.16691
37	0.27913	0.26989	0.12936	0.15504	0.1666
38	0.27923	0.27011	0.12950	0.15486	0.16630

1 的 PageRank 值高是因为有 3 个网页链接到它, 很显然在冲浪模型中冲浪者到达网页 1 的可能性大。而网页 2 有相同的 PageRank 值是因为网页 1 链接到网页 2, 而且只有一个链接。在冲浪模型中, 冲浪者访问了网页 1 之后肯定会访问网页 2。所有到达网页 2 的可能性和到达网页 1 的可能性是一样的。

```

forall u in S: R(u)0 = 1 / |S|
while (|R(u)i - R(u)i-1| > epsilon)
{
for each u in S:
R(u)i = R(u)i-1 + sum_{v in B(u)} R(v)i-1 / N(v)
}
C = 1 / sum_{u in S} R(u)i
for each u in S:
R(u)i = C * R(u)i
}
    
```

算法 1

4. PageRank 存在的问题

PageRank 对于 google 的成功, 所起的作用是非常大的。而随着 google 的成功, 和 PageRank 类似的超链分析排序技术也纷纷被搜索引擎技术公司所采用。那么在这种排序算法已经成为主流的环境下, 商业网站为了在搜索引擎中获得好的排名, 必然会费尽心机去影响本来天然客观的网络超链接结构。

被 google 所标引的网站中每个网页都有一个 PageRank, 并非一个网站的 PageRank 就是网站的首页, 并且站内的链接也被用来计算 PageRank。网站制作者通过站点链接和网站地图等来提高站内的 PageRank 的反馈值。那么 google 会不会对这样的站点内部的链接投票值打折呢? 而这样做会不会影响这种基于链接的投票计算方法因为这本身偏偏正是 PageRank 工作的最基本的机理, 更何况庞大网络链接的计算量非常的巨大。有了这些人为的影响, 那么 google 所说的客观公正会不会只是 google 的一厢情愿另外, 网络开始显著地改变, 现在的一个链接很多是因为应用许可的需要或象交换链接这样的利益交换, 实际上并无推荐之意。所以有人认为, PageRank 在 Google 排序算法中的作用已经下降。但是有一点必须指出, 对巨大的网络超链接结构人为的影响, 其难度和 webSpamming 相比, 是不可同日而语的。