# 036 | PageRank算法的核心思想是什么?

2017-12-25 洪亮劼



讲述：初明明
时长 08:00 大小 3.67M

上周我们介绍了信息搜索系统的历史进程，剖析了搜索系统的多轮打分系统，还深入探讨了倒排索引，聊了聊它的核心技术。

这周我要和你分享的是在互联网搜索引擎兴起之后的一个研发需要，那就是如何理解网页和网页之间的关系，特别是怎么从这些关系中提取网页中除文字以外的其他特性。这部分的一些核心算法曾是提高搜索引擎质量的重要推进力量。另外，我们这周要分享的算法也适用于其他能够把信息用结点与结点关系来表达的信息网络。

今天，我们先看一看用图来表达网页与网页之间的关系，并且计算网页重要性的经典算法：**PageRank**。

## PageRank 的简要历史

时至今日，谢尔盖·布林（Sergey Brin）和拉里·佩奇（Larry Page）作为 Google 这一雄厚科技帝国的创始人，已经耳熟能详。但在 1995 年，他们两人还都是在斯坦福大学计算机系苦读的博士生。那个年代，互联网方兴未艾。雅虎作为信息时代的第一代巨人诞生了，布林和佩奇都希望能够创立属于自己的搜索引擎。1998 年夏天，两个人都暂时离开斯坦福大学的博士生项目，转而全职投入到 Google 的研发工作中。他们把整个项目的一个总结发表在了 1998 年的万维网国际会议上（WWW7，the seventh international conference on World Wide Web）（见参考文献 [1]）。**这是 PageRank 算法的第一次完整表述。**

PageRank 一经提出就在学术界引起了很大反响，各类变形以及对 PageRank 的各种解释和分析层出不穷。在这之后很长的一段时间里，PageRank 几乎成了网页链接分析的代名词。给你推荐一篇参考文献 [2]，作为进一步深入了解的阅读资料。

## PageRank 的基本原理

我在这里先介绍一下 PageRank 的最基本形式，这也是布林和佩奇最早发表 PageRank 时的思路。

首先，我们来看一下每一个网页的周边结构。**每一个网页都有一个“输出链接”（Outlink）的集合。**这里，输出链接指的是从当前网页出发所指向的其他页面。比如，从页面 A 有一个链接到页面 B。那么 B 就是 A 的输出链接。根据这个定义，可以同样定义“**输入链接**”（Inlink），指的就是指向当前页面的其他页面。比如，页面 C 指向页面 A，那么 C 就是 A 的输入链接。

有了输入链接和输出链接的概念后，下面我们来定义一个页面的 PageRank。我们假定每一个页面都有一个值，叫作 PageRank，来衡量这个页面的重要程度。这个值是这么定义的，**当前页面 I 的 PageRank 值，是 I 的所有输入链接 PageRank 值的加权和。**

那么，权重是多少呢？对于 I 的某一个输入链接 J，假设其有 N 个输出链接，那么这个权重就是 N 分之一。也就是说，J 把自己的 PageRank 的 N 分之一分给 I。从这个意义上来看，I 的 PageRank，就是其所有输入链接把他们自身的 PageRank 按照他们各自输出链接的比例分配给 I。谁的输出链接多，谁分配的就少一些；反之，谁的输出链接少，谁分配的就多一些。这是一个非常形象直观的定义。

然而，有了这个定义还是远远不够的，因为在这个定义下，页面 I 和页面 J，以及其他任何页面的 PageRank 值是事先不知道的。也就是等式两边都有未知数，这看上去是个无解的问题。

布林和佩奇在他们的论文中采用了一种迭代算法。**这个算法很直观，那就是既然不知道这些 PageRank 的值，那我们就给他们一组初始值，这个初始值可以是这样的情形，所有页面有相同的 PageRank 值。**然后，根据我们上面所说的这个定义，更新所有页面的 PageRank 值。就这么一遍一遍地更新下去，直到所有页面的 PageRank 不再发生很大变化，或者说最后收敛到一个固定值为止。他们在文章中展示了实际计算的情况，往往是在比较少的迭代次数后，PageRank 值就能够收敛。

以上就是整个 PageRank 算法的基本思想和一种迭代算法。

## PageRank 算法的改进

完全按照我们上面介绍的这个最原始的 PageRank 算法，布林和佩奇很快就遇到了麻烦。

**第一个麻烦就是有一些页面并没有输出链接，比如某些 PDF 文件，或者一些图片文件。**由于没有输出链接，这些页面只能聚集从上游输入链接散发过来的 PageRank 值，而不能把自己的 PageRank 值分发出去。这样的结果就是，这些页面成为一些“悬空”（Dangling）结点。**悬空结点存在的最大问题就是会使得 PageRank 的计算变得不收敛。**这些结点成了 PageRank 值的“黑洞”，导致悬空结点的 PageRank 值越来越大，直至“吸干”其他所有输入链接的值。



要解决这个问题，就要为悬空结点“引流”，能够把这些点的值分发出去、引出去。**谢尔盖和拉里找到的一个方法是，对于每一个悬空结点，都认为这个结点能够随机到达整个网络上的其他任意一个结点。**也就相当于人工地从这个结点连接到所有页面的一个结点，让当前悬空结点的 PageRank 能够“均匀”地分散出去到其他所有的结点，这就解决了悬空结点的问题。

然而原始的 PageRank 还存在其他问题。要想保证 PageRank 的收敛性，并且能够收敛到唯一解，我们还需要第二个改进。**第二个改进就是，即便一个页面有自然的输出链接，我们也需要一个机制，能够从这个页面跳转到其他任何一个页面。**这也就是模拟假设一个用户已经浏览到了某个页面，一方面用户可以顺着这个页面提供的输出链接继续浏览下去，另一方面，这个用户可以随机跳转到其他任何一个页面。

有了这个机制以后，对于所有的结点来说，PageRank 的分配也就自然地产生了变化。在之前的定义中，每个页面仅仅把自己的 PageRank 值输送给自己原生的所有输出链接中。而现在，这是一部分的“分享”，另外一部分还包括把自己的 PageRank 值分享到所有的页面。当然，后者的总量应该比前者要少。于是，**这里可以引入一个参数，来控制有多大的**

**比例我们是顺着输出链接走，而多大的比例跳转其他页面。**通常情况下，这个参数的取值范围大约是 60%~85%。

有了这两个改进之后，整个网络上的每个页面实际上已经可以到达其他任何页面。也就是说，**整个页面网络成了一个完全联通的图，PageRank 算法就有了唯一的收敛的解。**

## PageRank 分析

PageRank 被提出后不久，就有学者开始针对 PageRank 模型和算法的性质进行分析。大家很快发现，还有一些其他的方法可以对 PageRank 进行解释。

**第一种比较流行的，也是更加正规的解释 PageRank 的方法，是把我们刚才说的这个分配等式写成矩阵的形式。**那么，整个算法就变成了一个标准的求解一个随机矩阵的“左特征向量”的过程。这个随机矩阵就是我们刚才讲的经过了两次修改后的跳转规律的矩阵形式。而刚才所说的迭代方法正好就是求解特征向量的“**乘幂法**”（Power Method）。在一定条件下的随机矩阵，经过乘幂法就一定能够得到一个唯一解。

**另外一种解释，是把刚才我们说的这个矩阵形式进行一次代数变形，也就是把等式两边的各项都移动到等式的一边，而另一边自然就是 0。**那么，整个式子就变了一个“线性系统”的求解过程。也就是说从代数的角度来解释整个 PageRank 的求解过程。

## 小结

今天我为你讲了现代搜索技术中的一个重要分支，链接分析中最重要的算法 PageRank 的核心思想 。 一起来回顾下要点：第一，我们讲了 PageRank 的一些简明历史和算法最原始的定义和思路 。第二，我们讲了 PageRank 的两种改进。第三，我们简要地介绍了针对 PageRank 的两种解释方法。

最后，给你留一个思考题，除了乘幂法，你觉得还有什么方法可以用来求解 PageRank 值?

欢迎你给我留言，和我一起讨论。

**参考文献**

1. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. Proceedings of the seventh international conference on World Wide Web 7 (WWW7), Philip H. Enslow, Jr. and Allen Ellis (Eds.). Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, 107-117, 1998.
2. Langville, Amy N.; Meyer, Carl D. Deeper Inside PageRank. Internet Math. no. 3, 335-380, 2003.

**论文链接**

The anatomy of a large-scale hypertextual Web search engine

Deeper Inside PageRank



上一篇　035 | 搜索索引及其相关技术概述

下一篇　037 | 经典图算法之HITS

**精选留言 (1)**

写留言

DasonCheng
2018-12-11

3

浏览到这里，总体感受是不够系统，多为知识点的搬运堆砌，给人一种“混”的感觉