037 | 经典图算法之HITS

2017-12-27 洪亮劼

AI技术内参

进入课程 >

讲述：初明明

时长 07:28　大小 3.43M

这周我们分享的内容是如何理解网页和网页之间的关系。周一我们介绍了用图（Graph）来表达网页与网页之间的关系并计算网页的重要性，就是经典算法 PageRank。今天我来介绍一下 PageRank 的姊妹算法：**HITS 算法**。

HITS 的简要历史

HITS 是 Hypertext-Induced Topic Search 算法的简称。这个算法是由康奈尔大学计算机科学教授乔·克莱恩堡（Jon Kleinberg）于 1998 年发明的，正好和我们周一讲的布林和佩奇发表 PageRank 算法是同一年。

这里有必要简单介绍一下乔这个人。乔于 1971 年出生在马萨诸塞州波士顿。1993 年他毕业于康奈尔大学获得计算机科学学士学位，并于 1996 年从麻省理工大学获得计算机博士学

位。1998 的时候，乔正在位于美国西海岸硅谷地区的 IBM 阿尔玛登（Almaden）研究院做博士后研究。HITS 的工作最早发表于 1998 年在旧金山举办的第九届 ACM-SIAM 离散算法年会上（详细论述可参阅参考文献）。

乔目前是美国国家工程院（National Academy of Engineering）和美国自然与人文科学院（American Academy of Arts and Sciences）院士。顺便提一下，乔的弟弟罗伯特·克莱恩堡也在康奈尔大学计算机系任教职。

HITS 的基本原理

在介绍 HITS 算法的基本原理之前，我们首先来复习一下网页的网络结构。每一个网页都有一个“输出链接”（Outlink）的集合。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，我们来看“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。

要理解 HITS 算法，我们还需要引入一组概念：**“权威”（Authority）结点**和**“枢纽”（Hub）结点**。这两类结点到底是什么意思呢?

HITS 给出了一种“循环”的定义：**好的“权威”结点是很多“枢纽”结点的输出链接，好的“枢纽”结点则指向很多好的“权威”结点**。这种循环定义我们在 PageRank 的定义中已经见识过了。

很明显，要用数学的方法来表述权威结点和枢纽结点之间的关系就必须要为每一个页面准备两个值。因为从直觉上来说，不可能有一个页面完全是权威，也不可能有一个页面完全是枢纽。绝大多数页面都在这两种角色中转换，或者说同时扮演这两类角色。

数学上，对于每一个页面 I，**我们用 X 来表达这个页面的“权威值”，用 Y 来表达这个页面的“枢纽值”**。那么，一个最直观的定义，对于 I 的权威值 X 来说，它是所有 I 页面的输入链接的枢纽值的总和。同理，I 的枢纽值是所有 I 页面输出链接的权威值的总和。这就是 HITS 算法的原始定义。

我们可以看到，如果 I 页面的输入链接的枢纽值大，说明 I 页面经常被一些好的“枢纽”结点链接到，那么 I 自身的权威性自然也就增加了。反之，如果 I 能够经常指向好的“权威”结点，那 I 自身的“枢纽”性质也就显得重要了。

当然，和 PageRank 值一样，X 和 Y 在 HITS 算法里也都是事先不可知的。因此，**HITS 算法的重点就是要求解 X 和 Y**。如果把所有页面的 X 和 Y 都表达成向量的形式，那么 HITS 算法可以写成 X 是矩阵 L 的转置和 Y 的乘积，而 Y 是矩阵 L 和 X 的乘积，这里的矩阵 L 就是一个邻接矩阵，每一行列表达某两个页面是否相连。进行一下代数变形，我们就可以得到 X 其实是一个矩阵 A 乘以 X，这里的 A 是 L 的转置乘以 L。Y 其实是一个矩阵 B 乘以 Y，这里的 B 是 L 乘以 L 的转置。

于是，惊人的一点出现了，那就是 HITS 算法其实是需要求解矩阵 A 或者矩阵 B 的主特征向量，也就是特征值最大所对应的特征向量，用于求解 X 或者 Y。这一点和 PageRank 用矩阵表达的形式不谋而和。也就是说，尽管 PageRank 和 HITS 在思路和概念上完全不同，并且在最初的定义式上南辕北辙，但是经过一番变形之后，我们能够把两者都划归为**某种形式的矩阵求解特征向量的问题**。

实际上，**把图表达为矩阵，并且通过特征向量对图的一些特性进行分析是图算法中的一个重要分支**（当然，我们这里说的主要是最大的值对应的特征向量，还有其他的特征向量也有含义）。既然我们已经知道了需要计算最大的特征向量，那么之前计算 PageRank 所使用的“乘幂法”（Power Method）在这里也是可以使用的，我们在这里就不展开了。

如何把 HITS 算法用于搜索中呢？最开始提出 HITS 的时候是这么使用的。

首先，我们根据某个查询关键字构建一个“相邻图”（Neighborhood Graph）。这个图包括所有和这个查询关键字相关的页面。这里，我们可以简化为所有包含查询关键字的页面。这一步在现代搜索引擎中通过“倒排索引”（Inverted Index）就可以很容易地得到。

有了这个相邻图以后，我们根据这个图建立邻接矩阵，然后就可以通过邻接矩阵计算这些结点的权威值和枢纽值。当计算出这两组值之后，我们就可以根据这两组值给用户展现两种网页排序的结果，分别是根据不同的假设。

值得注意的是，PageRank 是“查询关键字无关”（Query-Independent）的算法，也就是说每个页面的 PageRank 值并不随着查询关键字的不同而产生不同。而 HITS 算法是“查询关键字相关”（Query-Dependent）的算法。从这一点来说，HITS 就和 PageRank 有本质的不同。

HITS 算法的一些特点

HITS 算法依靠这种迭代的方法来计算权威值和枢纽值，你一定很好奇，这样的计算究竟收敛吗？是不是也需要像 PageRank 一样来进行特别的处理呢？

答案是 HITS 一定是收敛的。这点比原始的 PageRank 情况要好。然而，HITS 在原始的情况下，不一定收敛到唯一一组权威值和枢纽值，也就是说，解是不唯一的。因此，我们其实需要对 HITS 进行一部分类似于 PageRank 的处理，那就是让 HITS 的邻接矩阵里面所有的结点都能够达到其他任何结点，只是以比较小的概率。经过这样修改，HITS 就能够收敛到唯一的权威值和枢纽值了。

HITS 算法的好处是为用户提供了一种全新的视角，对于同一个查询关键字，HITS 提供的权威排序和枢纽排序能够帮助用户理解自己的需求。

当然，**HITS 的弱点也来自于这个依赖于查询关键字的问题。**如果把所有的计算都留在用户输入查询关键字以后，并且需要在响应时间内计算出所有的权威值和枢纽值然后进行排序，这里面的计算量是很大的。所以，后来有研究者开始使用全局的网页图，提前来计算所有页面的权威值和枢纽值，然而这样做就失去了对某一个关键字的相关信息。

小结

今天我为你讲了 HITS 算法的核心思想 。 一起来回顾下要点：第一，我们讲了 HITS 的一些简明历史。第二，我们讲了 HITS 最原始的定义和算法，并且联系 PageRank，讲了两者的异同之处。第三，我们分析了 HITS 的一些特点。

最后，给你留一个思考题，有没有办法把权威值和枢纽值所对应的两个排序合并成为一个排序呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (September 1999), 604-632，1999.

论文链接

Authoritative sources in a hyperlinked environment

上一篇　036 | PageRank算法的核心思想是什么?

下一篇　038 | 社区检测算法之“模块最大化”

精选留言 (3)

写留言

黄德平
2018-12-16

理一下思路，L表示连接矩阵，Lij是矩阵i行j列的元素，这个值取1当且仅当节点有链接指向节点j，否则为0。L的转置用M表示，根据权威值X和枢纽值Y的定义，我们可以得到
X=MY
Y=LX
进一步可以得到...
展开

xxw
2018-05-21

感觉可以适量列些公式。用文字表达公司有点闷逼

展开

白杨

2018-05-16

某种意义上，可以把权威理解为精度，枢纽理解为广度，然后用F值的思想去合并