

## 038 | 社区检测算法之“模块最大化”

2017-12-29 洪亮劫

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 06:40 大小 3.06M



一起来回顾下本周的内容。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，就是经典的 PageRank 算法。周三我们介绍了 PageRank 的一个姊妹算法，HITS 算法，并且分析了这两种算法的内在联系，这两类算法都希望给网页赋予一个权重来表达网页的重要性。

今天，我们来看一类完全不一样的网页分析工具，那就是希望把网页所代表的图分割成小块的图，或者叫图聚类，每个小聚类代表一个“社区”。这类分析有时候被称作图上面的“社区检测”（Community Detection），意思就是从图上挖掘出潜在的社区结构。

### 社区检测的简要历史

提到社区检测就不得不提到这么一位学者，他与我们今天要介绍的算法有非常紧密的联系，而且他的研究在 2000 年~2010 年间成了社区检测研究的标杆，影响了后续的大量研究工作。这位学者就是密歇根大学（University of Michigan）的物理学教授马克·纽曼（Mark Newman）。

1991 年，纽曼从牛津大学获得理论物理博士学位。在接下来的 10 年里，他在康奈尔大学和圣塔菲学院（Santa Fe Institute）分别担任博士后研究员和研究教授。2002 年，纽曼来到密歇根大学物理系担任教授，并且一直在这里进行网络科学（Network Science）的基础研究。

2006 年，纽曼在《物理评论》杂志上发表了一个叫“**模块最大化**”（Modularity Maximization）的社区检测算法。从某种程度上来说，这个算法很快就成了社区检测的标准算法，吸引了研究领域的广泛关注，激发了大量的针对这个算法的分析和研究。对这个算法的最原始论述，请参阅参考文献 [1] 和 [2]。

今天我们就来讲一讲这个“模块最大化”算法的基本原理。

## 模块最大化的基本原理

在我们讲解模块最大化算法之前，我们先来看一看“社区”的含义。在图分析以及网络科学中，“**社区**”定义为一组结点，它们互相之间的联系比它们跟社区之外结点的联系要更加**紧密**。你可以注意到，在这个定义中，什么叫紧密，怎么来衡量“更紧密”这个关系都是没有说明的，这就为各类社区检测算法或模型带来了很大的发挥空间。

社区检测（有时候也说社区发掘）算法的核心就是要**根据给定的一组结点和它们之间的关系，在无监督的情况找到这些社区，并分配哪些结点属于哪个社区**。

我们先来谈一谈“**模块最大化**”的一个整体思路。这里，我们讨论一种简化的情况，那就是如何把一个网络分割成两个社区。首先，算法按照某种随机的初始化条件，把网络分成两个社区。然后，算法逐一检查每一个结点，看如果把这个点划归到另外一个社区的话，会不会增加“模块化”这个目标函数。最终，算法决定改变那些能够最大化模块化目标的结点的社区赋值。然后整个算法不断重复这个过程，直到社区的赋值不再发生变化。

现在我们来讨论一下模块化这个目标函数。根据上面提到的社区含义，我们希望社区里结点之间的联系紧密。**在模块化目标函数里，就表达为两个结点的连接数目减去这两个结点之间的“期望连接数”**。模块化最大化说的就是，对于同一个社区中的所有结点，我们希望这个

差值的和最大化。什么意思呢？就是说，如果我们把两个结点放到一个社区中，那它们的连接数（其实就是 1 或者 0）要足够大于它们之间的连接数的期望值，这就解决了我们刚才所说的如何来衡量“更加紧密”的难题。

那么，怎么来定义两个结点之间的“期望连接数”呢？最初纽曼在介绍模块最大化的时候，他给出了这么一个计算方法。那就是，用两个结点各自的总连接数相乘，除以整个网络的总连接数的 2 倍。直观上说，这是在衡量这两个结点之间出现任何连接的可能性。

那么，整个模块最大化的目标函数就是，针对现在网络中的所有结点，根据它们是否在同一个社区，我们计算他们两两之间的模块化数值，也就是它们之间的连接减去“期望连接数”，最后对所有的两两配对进行加和。我们希望这个目标函数最大化，这个目标函数中的未知数，就是社区的分配，也就是哪个结点属于哪个社区。一旦社区的分配已知，整个模块最大化这个目标函数的数值就可以很容易地计算出来。

那么如何得到这些社区的分配呢？和我们之前介绍的 PageRank 以及 HITS 的思路类似，纽曼使用了矩阵的表达方法对整个模块最大化进行了一个重构，经过一系列代数变形之后 [1]，纽曼得到了一个新的目标函数，那就是一个向量  $s$  的转置，乘以一个矩阵  $B$ ，然后再乘以向量  $s$ ，最后除以 4 倍的网络连接总数。这里，向量  $s$  代表了一个结点是否属于两个社区中的一个，矩阵  $B$  中的每一个元素表示了横纵坐标所代表的两个结点的模块化值。问题就是求解  $s$  的值。请注意， $s$  中的值是离散的，要么是正 1（代表属于两个社区中的一个）要么是负 1（代表属于两个社区中的另一个）。很明显，这是一个困难的离散优化问题。

**纽曼对这个复杂的离散优化问题进行了近似处理的方法。**具体来说，那就是允许  $s$  的值不仅仅是正负 1 而是实数，这样就大大简化了优化问题的难度。在设置好最优化这个新的目标函数之后，经过代数变形，我们得到了一个惊人的结论，那就是最优情况下的  $s$ ，实际就是矩阵  $B$  最大的特征值所对应的特征向量。这又和 PageRank 以及 HITS 有着极其相似的最优结构。在找最大特征向量的过程中找到  $s$  以后，我们就根据  $s$  里元素的正负号，正的属于一个社区，负的属于另外一个社区，来对整个网络中的结点进行划分。

当然，我们这里讲的仅仅是把整个网络进行二分的情况。在实际应用中，我们往往需要把整个网络划分成多个社区。纽曼在论文中 [1] 也讲解了**如何把二分法推广到多个社区的情景**。具体来说，就是先把一个网络分成两份，然后再不断地二分下去。不过，每次进行二分的时候，我们都需要检查是否对模块化目标函数起了正向的帮助，而不只是机械地进行二分。

## 小结

今天我为你讲了社区检测中一个有代表性的算法：模块最大化。一起来回顾下要点：第一，我们讲了什么是社区检测以及社区检测的一些简明历史。第二，我们讲了模块最大化的基本思想，比如模块最大化是如何定义的，又是如何把一个困难的离散优化问题转换成类似 HITS 和 PageRank 的寻找最大特征向量的问题。

最后，给你留一个思考题，如何把网页的社区信息利用到学习网页的相关度里面去呢？

欢迎你给我留言，和我一起讨论。

## 参考文献

1. M. E. J. Newman. Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA 103, 8577–8582, 2006.
2. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104, 2006.

## 论文链接

1. [Modularity and community structure in networks](#)
2. [Finding community structure in networks using the eigenvectors of matrices](#)

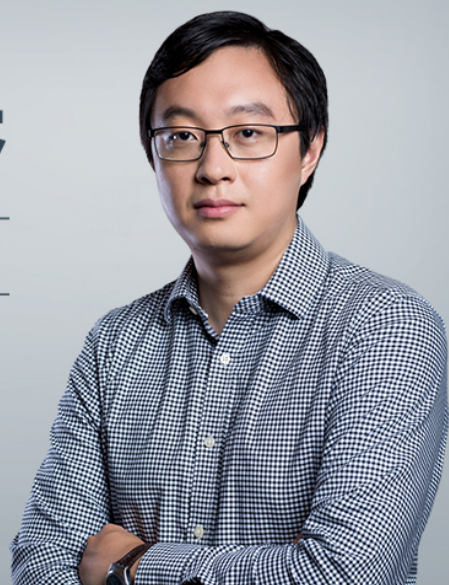



# AI 技术内参

你的360度人工智能信息助理

洪亮劫

Etsy 数据科学主管  
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

上一篇 037 | 经典图算法之HITS

下一篇 039 | 机器学习排序算法经典模型：RankSVM

## 精选留言

写留言

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。