统计学习读书报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. 周期性图中是否存在非周期的节点？

讨论后的理解：根据书中的定义，只要一个图中存在一个周期节点，那么就是周期性图，这个定义从侧面说明了周期性图可能存在非周期性节点。我自己的想法是可以构造一个周期性节点有多条路径的图，那么只有这个节点是周期性的，而其他路径上的点则不一定是周期性的。或者局部有一个子图呈现周期性。

1. 阻尼因子的作用是什么？其过大或者过小造成的影响是什么？

讨论后的理解：阻尼因子的值代表了打开超链接的概率，而1-d代表随机跳转的概率，那么过大说明接近于原始的pagerank定义，而过小的话说明随机性占主导，这个d的值一般根据实际情况来确定，可能需要统计才能设定较好的阻尼因子。

1. 427的规范化如何产生作用？是否会影响计算结果？

讨论后的理解：如果不进行规范化的话，可能导致得到的R向量在计算过程中变得过大或者过小，这样就会影响进一步的运算和归一化，而每一步的规范化就是相当于每一次都检查一下算法的运算情况，保证这个向量比较好计算，因为在这种情况下向量是一个倍数关系，所以规范化不会影响最终结果。

1. 别人提出的问题的理解
2. 幂法中转移矩阵为什么可以直接写成21.22的形式？

自己的理解：这个可以理解成p427说到的perron-frobenius定理证明的内容，此处这个矩阵本身应该具有的性质就是主特征值为1，而其对应的特征向量应该是pagerank的一个倍数，它能写成这个形式我的理解是是由pf定理推到出来的，而书上没有给出详细的证明。而这也是为什么这个等式得出的r向量也需要进行归一化处理，因为其是一个pagerank向量的倍数而不是实际的pagerank向量。

1. 三种PageRank算法各有什么优点和缺点？

自己的理解：迭代法的优势应该是思路比较简单，而且实现起来应该很容易。幂法的话也是很常用的矩阵求解方法，和迭代法差不多，都是逐步靠近。而解方程的方法直接但是难度很大，可能不太适合由计算机来做，而且太复杂的方程可能也难以直接求解

1. 从一个结点到其余每一个结点的概率为什么都相等?

自己的理解：这个相等不是一个事实，而是pagerank算法提出时给出的假设，而假设的话可以假设相等，也可以不相等，这是比较自由的。

1. 阻尼因子d在应用中应该如何选择，是否可以理解为阻尼因子越大，PageRank越不稳定？

自己的理解：这个d值应该根据实际的搜索情况来确定，也就是根据经验。当d值过大时，说明越倾向原始的pagerank，随机性应该越小，而阻尼因子越小，越随机。

1. 读书计划

1、本周完成的内容章节：复习21章。

2、下周计划：阅读speech and processing第7章内容。

四、读书摘要及理解

1. Pagerank的定义：

Pagerank算法是计算互联网网页重要度的非常重要的算法，其是定义在网页集合上的一个函数。其基础是有向图和随机游走模型，当随机游走导致每个节点被访问的概率收敛到稳定值时，可以将其作为网页的重要程度。其本质上是一阶马尔科夫链模型。

2. Pagerank的原始定义和一般定义：

原始定义下，pagerank的要求是一个强连通的非周期有向图，并且定义一个转移矩阵M，最终通过计算找出MR=R，也就是马尔可夫链的平稳分布。其中R代表的就是每个节点的重要性。初始定义随机游走的模型时，假设一个节点到其连接的每一个节点的转移概率都相等。原始定义可以通过迭代的方式求解。但是考虑到并不是所有的网页都有超链接，也就是强连通的条件不一定满足，所以pagerank的基本定义并不适合所有情况。

一般定义下，引入了阻尼因子和随机向量，从而保证了不会出现节点重要度为0的情况，这和拉普拉斯平滑很类似。其中阻尼因子的值需要根据经验来确定。

3. Pagerank的求解方法：

迭代法：由于平稳分布的条件，可以通过给定阻尼因子，初始分布和转移矩阵来迭代求解R，其好处在于简单方便，易于实现。

幂法：幂法和迭代法是很类似的，其使用到了特征值和特征向量的性质，根据perron-Frobenius定理，可以证明通过幂法可以得到pagerank向量。其优势在于在某种程度上，幂法更加简单，同时收敛的速度快一些，得出的向量也比较好处理。

解方程法：这个是因为pagerank最终满足的就是一个等式，那么只要求解这个方程就可以得出结果，其问题在于对于高维度的矩阵运算来说，这是非常困难的事情，对于计算机来说，直接求解不仅实现复杂，而且效率很差，所以上面两种近似方法效果更好一些。

思考：

Pagerank算法最早是由google提出的，该算法可以说改变了互联网的格局，以往的搜索引擎效果很差，依托于关键词检索，而pagerank则是真正将网络看成了一个图的结构，从而准确的估计了网站的重要程度，也帮助用户可以找到需要的内容，这也使得google的搜索引擎最终脱颖而出。除此之外，pagerank还可以用在估计图节点重要性，社交网络节点重要性分析等热门研究中，是一种十分实用的算法。