**第二十一章读书报告**

09118223 吴亦珂

读书进度：二十二章读完

1. **问题列表**

（我提出）幂法中转移矩阵为什么可以直接写成21.22的形式？

讨论结果：根据Perron-Frobenius定理，一般PageRank的向量R是矩阵A的主特征向量，主特征值是1。所以可以直接写成这种形式。

（我提出）三种PageRank算法各有什么优点和缺点？

讨论结果：迭代算法实现起来较为简单，思路较为容易，相当于严格按照PageRank的一般定义求解。幂法先计算出一般转移矩阵，再逐步迭代求解。代数算法用到了线性代数中的知识，直接通过求解矩阵的逆矩阵来求。但矩阵的逆矩阵在计算机上求解较为复杂。

（别人提出）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？  
讨论结果：阻尼因子应该是指浏览者继续浏览网页的概率（点击了网页上的超链接），（1-d）是指随机打开一个网页的概率，所以d的大小应该有实际的情况来定。

（别人提出）周期性图中是否存在非周期的节点？

讨论结果：因为书上P417页上给出了非周期性图的定义，即所有结点都是非周期结点，则为非周期性图，其余均为周期性图。也就是说，只要包含了周期结点就为周期性图。因此，周期图中可以没有非周期结点也可以有，但是必须要有周期结点。

（别人提出）P427的规范化如何产生作用？是否会影响计算结果？  
讨论结果：因为每一次迭代都是将矩阵成上一个转移矩阵。因此如果转移矩阵使结果变大则在经过多伦迭代后，结果可能就变得非常大，从而难以计算。同理，如果过小，则可能导致结果过下，甚至发生下溢出，也会影响最终结果。

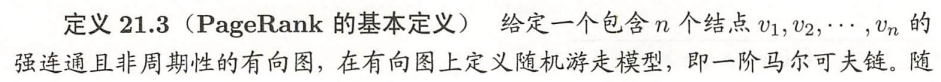
（别人提出）现实中相互链接的网页众多，这样转移矩阵阶数会很高，计算起来会不会过于复杂？  
讨论结果：时间复杂度为o(),k为迭代次数，v为网页数量。空间复杂度为o()。因此如果是海量数据，复杂度确实很高。因此可以采用一些优化方法。

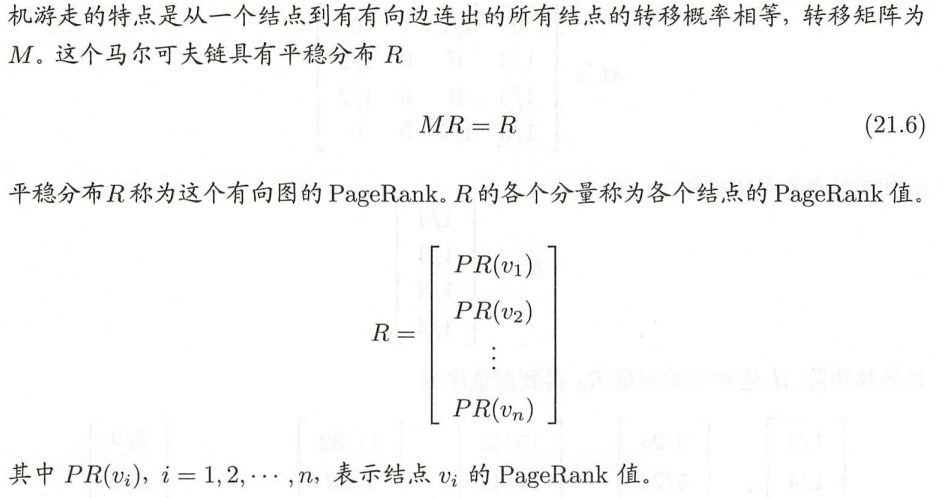
1. **读书收获**

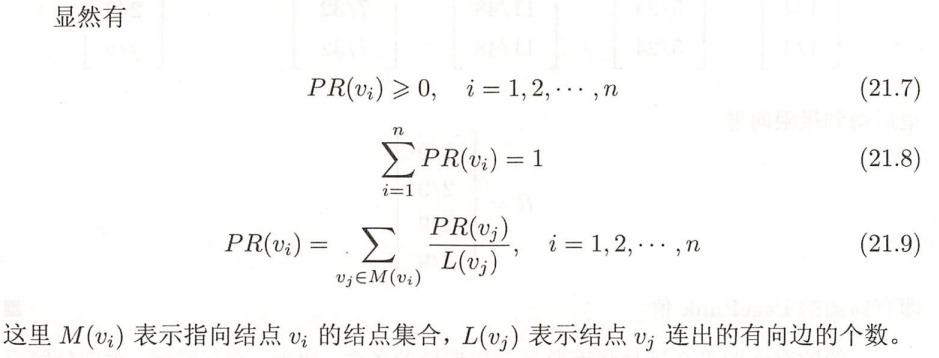
**21.1 PageRank的定义**

PageRank表示网页的重要程度，PageRank值越高，网页就越重要，在互联网排序中就越可能排在前面。简单来说，如果指向一个网页的超链接越多，那么该网页的PageRank就越高，这个网页也就越重要。

PageRank基本定义：





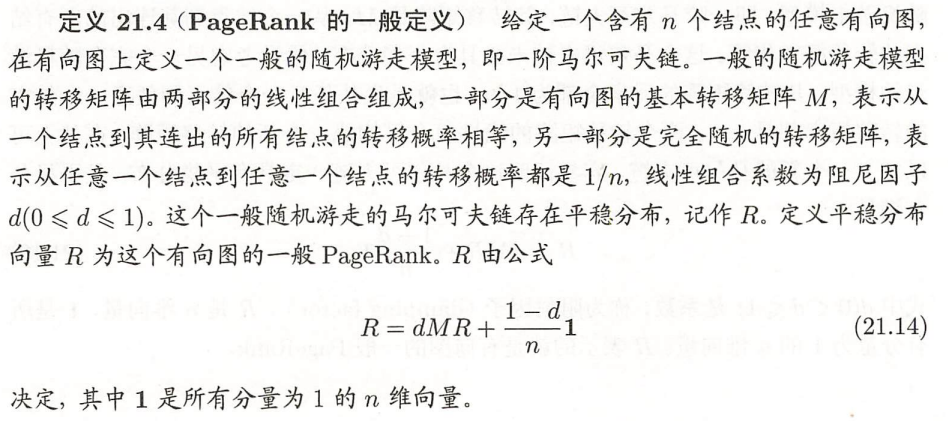


马尔可夫链平稳分布定理：不可约且非周期的有限状态马尔可夫链，有唯一平稳分布存在，并且当时间趋于无穷时状态分布收敛于唯一的平稳分布。

因此，根据该定理，强连通且非周期的有向图上定义的随机游走模型，在图上的随机游走当时间趋于无穷时状态分布收敛于唯一的平稳分布。

但是考虑到一般的有向图未必满足强连通且非周期的条件，所以PageRank基本定义一般并不适用。

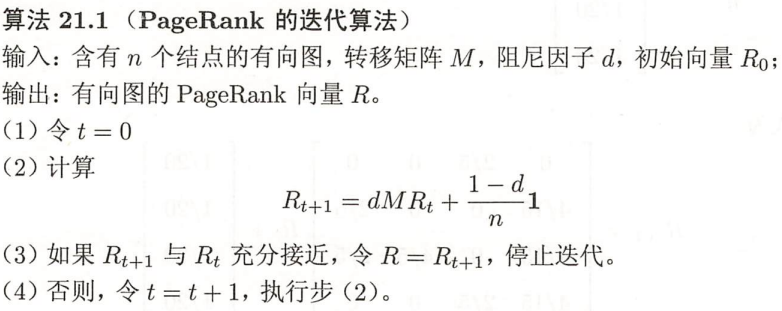
PageRank的一般定义：



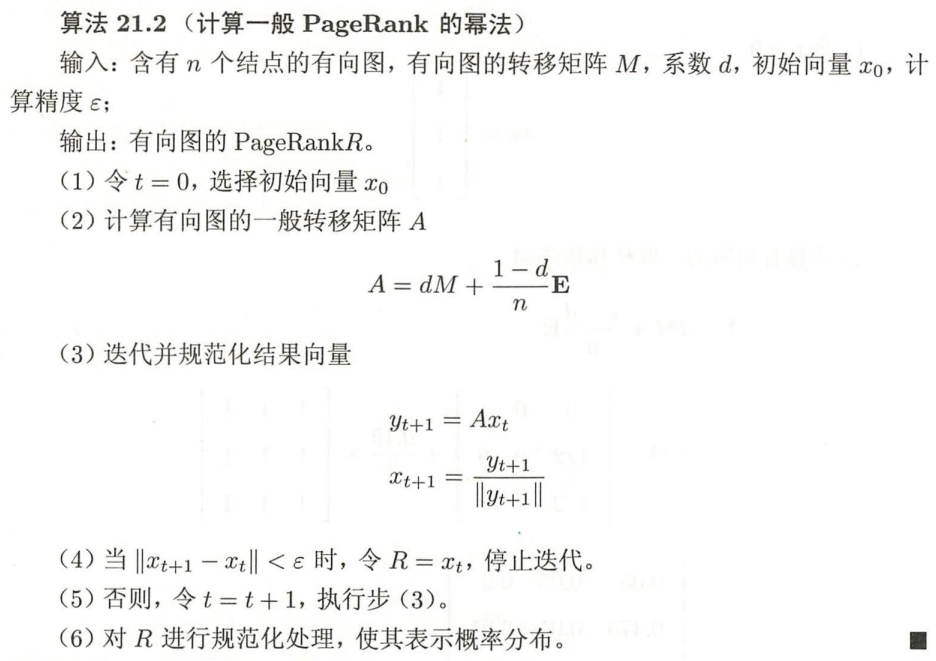
这个一般定义从直观上来看，浏览者按一下方法在网上随机游走：在任意一个网页上，浏览者或者以d的概率按照超链接等概率进行随机跳转，这就对应着公式的第一项。或者以(1-d)决定完全随机跳转，这是以等概率跳转到任意一个网页，如果有n个网页，就是1/n=的概率，这就对应着公式的第二项。

**21.2 PageRank的计算**

1. 迭代算法：直接套用一般定义的公式进行计算。



1. 幂法：



1. 代数算法：直接根据线性代数的知识，通过求解逆矩阵的方法进行求解。

