MA206 Homework6

12110120 赵钊

1 第 1 题

假设 P、Q 是两个马尔可夫矩阵，$R = PQ$，现在证明 R 也是马尔可夫矩阵。设 $(\star)_{ij}$ 表示矩阵 $(\star)$ 第 i 行第 j 列的元素。

由 P、Q 是马尔可夫矩阵，有

$$p_{ij} \geq 0 \quad q_{ij} \geq 0$$
$$\sum_j p_{ij} = 1 \quad \sum_j q_{ij} = 1$$

那么，

$$r_{ij} = \sum_{k=1}^{n} p_{ik} q_{kj} \geq 0$$

$$\begin{aligned}
\sum_j r_{ij} &= \sum_{k=1}^{n} p_{ik} q_{k1} + \sum_{k=1}^{n} p_{ik} q_{k2} + \cdots + \sum_{k=1}^{n} p_{ik} q_{kn} \\
&= p_{i1} \sum_{k=1}^{n} q_{1k} + p_{i2} \sum_{k=1}^{n} q_{2k} + \cdots + p_{in} \sum_{k=1}^{n} q_{nk} \\
&= p_{i1} + p_{i2} + \cdots + p_{in} \\
&= 1
\end{aligned}$$

根据定义，R 也是马尔可夫矩阵

因此由数学归纳法，P 是马尔可夫矩阵，$P^2 = P \cdot P$ 为马尔可夫矩阵，$P^3 = P^2 \cdot P$ 为马尔可夫矩阵，..... 也就是，若 P 是马尔可夫矩阵，则 $\forall n$，P^n 是马尔可夫矩阵

2 第 2 题

只需要等价的证明从 i 到 j 的 m 步转移概率为 P_{ij}^m，也就是

$$P(X_{n+m} = j | X_n = i) = p^m(i, j) = P_{ij}^m$$

首先根据条件概率公式，有

$$P(X_{n+m} = j | X_n = i) = \frac{P(X_{n+m} = j, X_n = i)}{P(X_n = i)}$$

对于分子 $P(X_{n+m} = j, X_n = i)$，中间经过了 m 步，因此一个简单的想法就是把这 m 步的所有的可能的情况都列出来，也就是

$$P(X_{n+m} = j, X_n = i) = \sum_{i_1, \ldots, i_{m-1} \in S} P(X_{n+m} = j, X_{n+m-1} = i_{m-1}, \ldots, X_n = i)$$

这里的 S 表示的是所有的状态的集合.
考虑全概率公式，即

$$\begin{aligned} LHS &= P(X_{n+m}=j, X_{n+m-1}=i_{m-1}, \dots, X_n=i) \\ &= P(X_n=i)P(X_{n+1}=i_1|X_n=i)\dots P(X_{n+m}=j|X_{n+m-1}=i_{m-1}, X_{n+m-2}=i_{m-2}, \dots) \end{aligned}$$

又由于每个 X_i 都有马尔科夫性，得到

$$P(X_{n+m}=j|X_{n+m-1}=i_{m-1}, X_{n+m-2}=i_{m-2}, \dots, X_n=i) = P(X_n=i)p(i,i_1)p(i_1,i_2)\dots p(i_{m-1},j)$$

代入求和得

$$\begin{aligned} P(X_{n+m}=j, X_n=i) &= P(X_n=i) \sum_{i_1,\dots,i_{m-1}\in S} p(i,i_1)p(i_1,i_2)\dots p(i_{m-1},j) \\ &= P(X_n=i)p^m(i,j) \end{aligned}$$

因此结论得证

3 文献报告

3.1 Pagerank 概述

Pagerank, 即网页排名, 是 Google 创始人拉里 · 佩奇和谢尔盖 · 布林于 1997 年构建早期的搜索系统原型时提出的链接分析算法, 自从 Google 在商业上获得空前的成功后, 该算法也成为其他搜索引擎和学术界十分关注的计算模型。目前很多重要的链接分析算法都是在 PageRank 算法基础上衍生出来的。PageRank 是 Google 用于用来标识网页的等级/重要性的一种方法，是 Google 用来衡量一个网站的好坏的唯一标准。在揉合了诸如 Title 标识和 Keywords 标识等所有其它因素之后，Google 通过 PageRank 来调整结果，使那些更具“等级/重要性”的网页在搜索结果中另网站排名获得提升，从而提高搜索结果的相关性和质量。

3.2 基本假设

PageRank 的计算基于以下两个基本假设：

1. 数量假设：在 Web 图模型中，如果一个页面节点接收到的其他网页指向的入链数量越多，那么这个页面越重要。

2. 质量假设：指向页面 A 的入链质量不同，质量高的页面会通过链接向其他页面传递更多的权重。所以越是质量高的页面指向页面 A，则页面 A 越重要。

利用以上两个假设，PageRank 算法刚开始赋予每个网页相同的重要性得分，通过迭代递归计算来更新每个页面节点的 PageRank 得分，直到得分稳定为止。PageRank 计算得出的结果是网页的重要性评价，这和用户输入的查询是没有任何关系的，即算法是主题无关的。假设有一个搜索引擎，其相似度计算函数不考虑内容相似因素，完全采用 PageRank 来进行排序，这个搜索引擎对于任意不同的查询请求，返回的结果都是相同的，即返回 PageRank 值最高的页面。

3.3 基本思想

如果网页 T 存在一个指向网页 A 的连接，则表明 T 的所有者认为 A 比较重要，从而把 T 的一部分重要性得分赋予 A。这个重要性得分值为：$\frac{PR(T)}{L(T)}$. 其中 PR（T）为 T 的 PageRank 值，L(T) 为 T 的出链数, 则 A 的 PageRank 值为一系列类似于 T 的页面重要性得分值的累加。

即一个页面的得票数由所有链向它的页面的重要性来决定，到一个页面的超链接相当于对该页投一票。一个页面的 PageRank 是由所有链向它的页面（链入页面）的重要性经过递归算法得到的。一个有较多链入的页面会有较高的等级，相反如果一个页面没有任何链入页面，那么它没有等级。

3.4 优缺点分析

3.4.1 优点

1. Pagerank 是一个与查询无关的静态算法，所有网页的 PageRank 值通过离线计算获得；有效减少在线查询时的计算量，极大降低了查询响应时间。

3.4.2 缺点

1. 查询具有主题特征，PageRank 忽略了主题相关性，导致结果的相关性和主题性降低。
2. 旧的页面等级会比新页面高。因为即使是非常好的新页面也不会有很多上游链接，除非它是某个站点的子站点。