课后作业与答案

### 第一次作业

1. 5.1.1 假定不采用“抽税”法，计算图5-7中每个网页的Pagerank值。

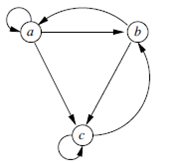


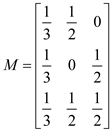
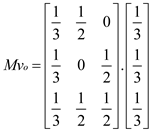
图 5-7 习题示例图

**参考答案：**PageRank是一种为每个网页分配数值的方法，其用于计算网站链接的数量和质量，以便大致了解其重要性，其假设的潜在前提是，更受欢迎的网站似乎更适合从大量不同的领域获得连接。

在给定的有向图中，节点表示为网站页面，每条弧线表示页面之间的一个链接。图5-7表示三个网站页面和其之间的链接关系。页面a链接到所有三个页面;页面b只链接到页面a和c;页面c只链接到页面b和c。

因此，页面a链接到页面a、b和c，那么如果当前处于页面a，之后登陆这些页面的概率是1/3。页面b链接到页面a和页面c，那么之后登陆到页面a和页面c的概率是1/2，而从页面b登陆页面a的概率是0。页面c链接到页面b和页面c，那么访问页面b和页面c的概率分别是1/2，而访问页面a的概率是0。

因此，有向图的转移矩阵M为:

，，

假设一个人初始上相同的概率访问n (n=3)个网站中的任意一个网站，那么原始向量v0的每个元素的值都是1/n。如果M是网络的转移矩阵，那么其下一个访问的网站的概率的分布将是的Mv0，两步后的M(Mv0) = M2v0，以此类推。例如，可以通过将其起始向量v0乘以M总i次来得到其在i次跳转后的分布。

最终稳定后得到vi=[3/13,4/13,6/13]T。即a的Pagerank值为3/13，b的Pagerank值为3/13，c的Pagerank值为6/13。

1. 5.1.2 假定 ，计算图5-7中每个网页的Pagerank值。

**参考答案：**在5.1.1的基础上，使用vi+1=βMvi+(1-β)e/n进行迭代计算，其中e=[1,1,1]。稳定后的vi约为[0.26,0.31,0.43]T。

### 第二次作业

1. 5.3.1 在假设随机跳转集合分别为如下集合的情况下，计算图5-15所对应的面向主题的PageRank值：
   1. 仅包含A；
   2. 包含A和C。

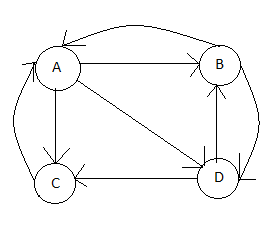
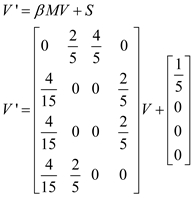


图5-15习题示例图

**参考答案：**a) 要确定面向主题的PageRank值，首先考虑图5.15，得到转移矩阵M，并解方程（β=0.8）：

A black background with a black square

Description automatically generated with medium confidence**,**

得到：



b) 解方程：

A black background with a black square

Description automatically generated with medium confidence

得到：



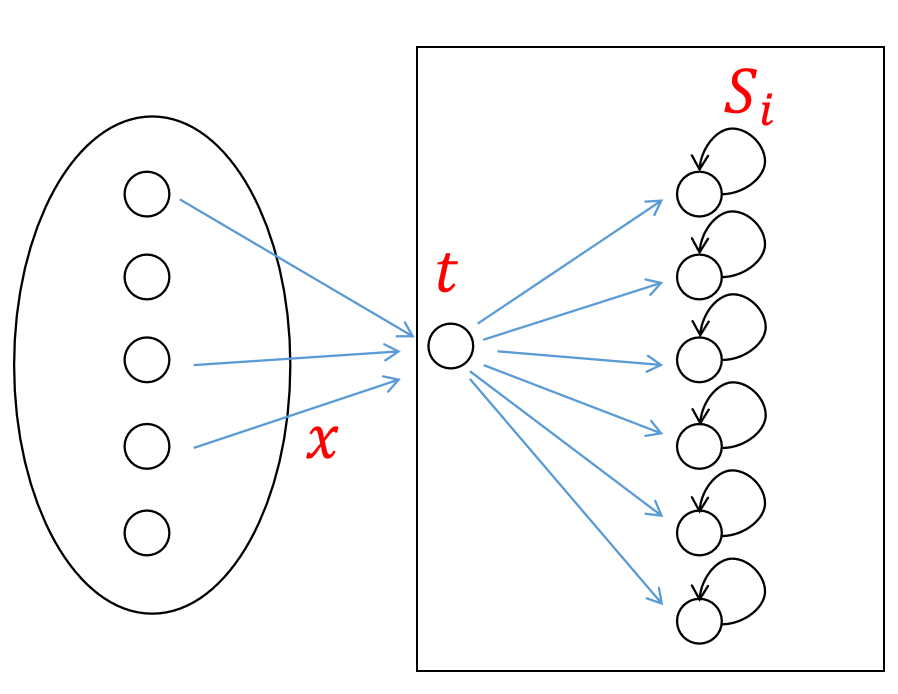
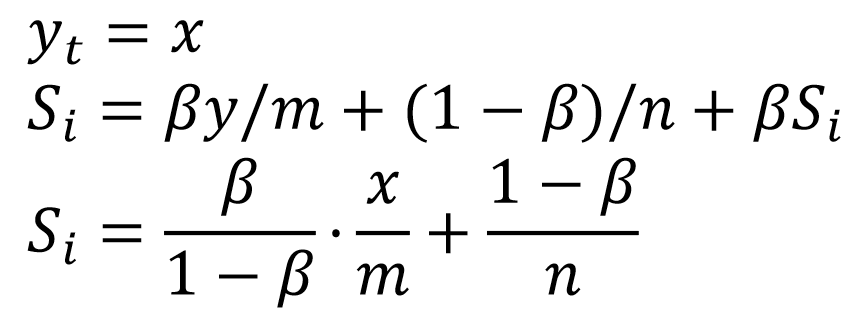
1. 5.4.1 在 5.4.2 节中，我们分析了图 5-16所示的垃圾农场，其中每个支持页都链向目标页。对于下列垃圾农场,重复5.4.2节的分析过程：

a)每个支持页只链向自己,而不是链向目标页；

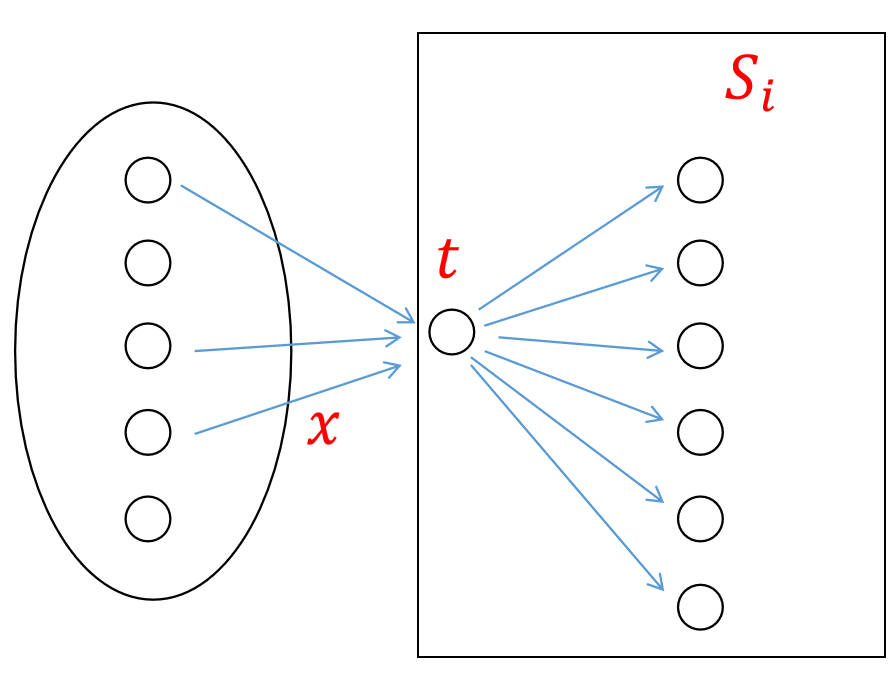
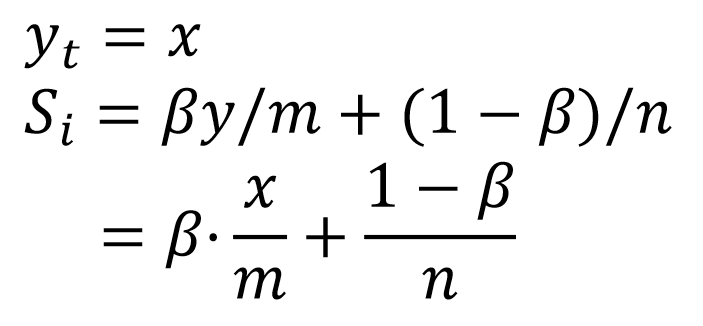
b)每个支持页不链向任何网页；

c)每个支持页同时链向自己和目标页。

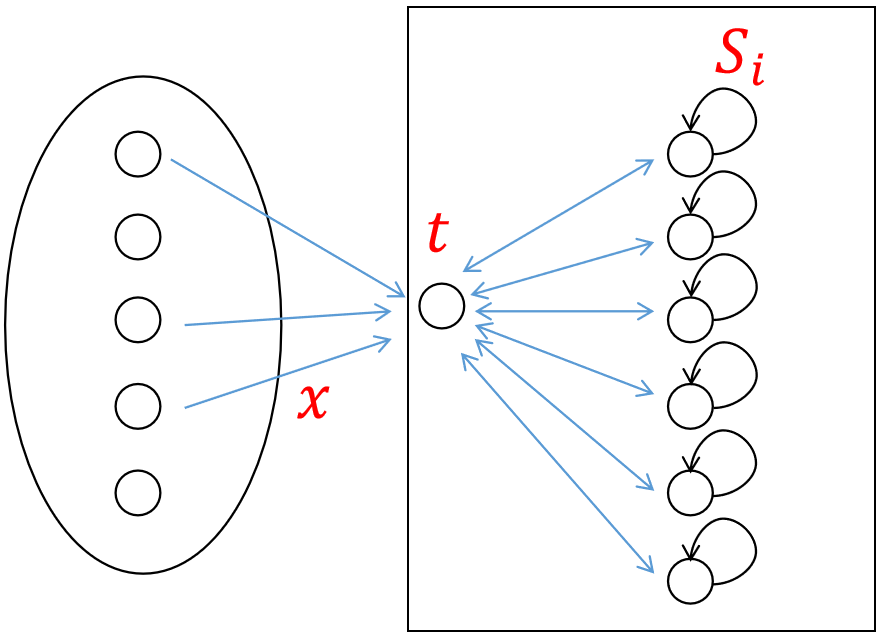
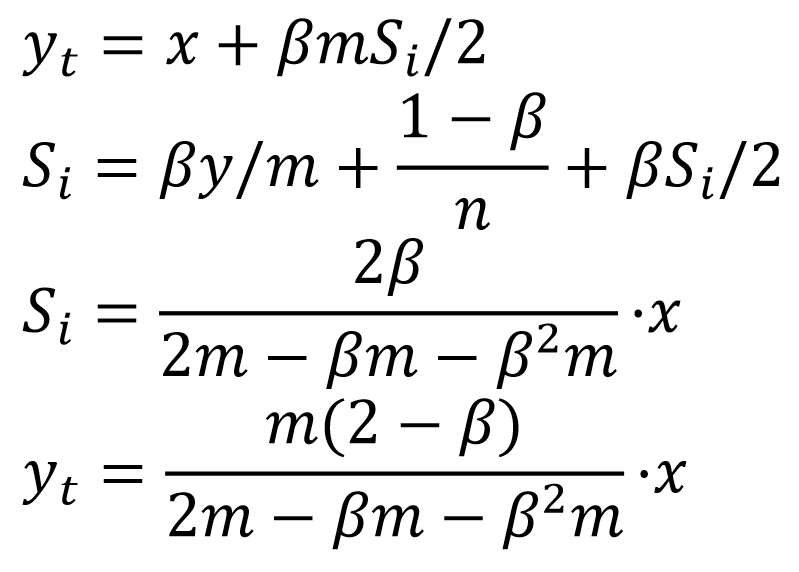
**参考答案：**a)

，

b)

,

c)

,

1. 5.4.2 对于图 5-1 给出的原始Web图，假定只有B是可靠网页：

a)计算每个网页的TrustRank；

b)计算每个网页的垃圾质量。

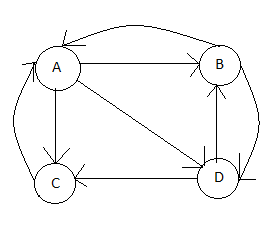


图5-15习题示例图

**参考答案：**TrustRank是一种用于进行链接分析，将有用的网页从几个垃圾网页中区分出来的算法。参照图5.6的web图，假设只有B页面是可信页面，使用例5.10中所示的方法确定TrustRank, 然后从例5.12中确定每个页面的Spam mass:

|  |  |  |  |
| --- | --- | --- | --- |
| **Node** | **PageRank** | **TrustRank** | **Spam mass** |
| **A** | **1/3** | **0.269** | **0.192** |
| **B** | **2/9** | **0.358** | **-0.610** |
| **C** | **2/9** | **0.158** | **0.290** |
| **D** | **2/9** | **0.215** | **0.033** |

### 第三次作业

1. 2.3.1 设计MapReduce算法来实现下列功能,其中每种功能的输入都是整数构成的大文件,而输出则是下面的情形之一：

a)最大整数；

b)所有整数的平均值；

c)整数集合,但是每个整数只出现一次；

d)输入中不同整数的出现次数。

上面每一种情况下,都假设每个元组对的键被忽略或抛弃。

**参考答案：**a）Map:对于主文件中出现的第m个整数j输出键值对(𝑚，𝑗)。Reduce:读取每个键值对，将关联值存储为列表[𝑗1，𝑗2,…，𝑗𝑛]。Reduce输出max([𝑗1，𝑗2,…，𝑗𝑛])。

b) Map:对于主文件中出现的第m个整数j输出键值对(𝑚，𝑗)。Reduce:读取每个键值对，将键设为Hash(m) mod R，并将关联值存储为列表[𝑗1，𝑗2,…，𝑗𝑛]。 Reduce输出average([𝑗1，𝑗2,…，𝑗𝑛])以及[𝑗1，𝑗2,…，𝑗𝑛]的大小。

c) Map:对于主文件中出现的第m个整数j输出键值对(𝑚，𝑗)。Reduce:读取每个键值对，将键设为Hash(m) mod R，并将关联值存储为集合{𝑗1，𝑗2,…，𝑗𝑛}。 Reduce输出{𝑗1，𝑗2,…，𝑗𝑛}。

d) Map:对于主文件中出现的第m个整数j输出键值对(𝑚，𝑗)。Reduce:读取每个键值对，并存储{j,count}。每次读取得到j，则将count增加1。Reduce输出[(𝑗1,count1),…，(𝑗𝑛,countn)]。

1. 3.1.1 计算三个集合{1,2,3,4}、{2,3,5,7}和{2,4,6}两两之间的Jaccard相似度。

**参考答案：**三个集合分别设为A={1,2,3,4}, B={2,3,5,7}, C={2,4,6}。因此，A和B集合的Jaccrad相似度SIM(A,B)=(|A∩B|)/(|A∪B|)=2/6。类似的，SIM(A,C)=2/5。SIM(B,C)=1/6

1. 3.3.2 使用图3-4的数据，计算增加如下两个哈希函数之后得到的签名矩阵：
2. h3(x)=2x+4 mod 5；
3. h4(x)=3x-1 mod 5。

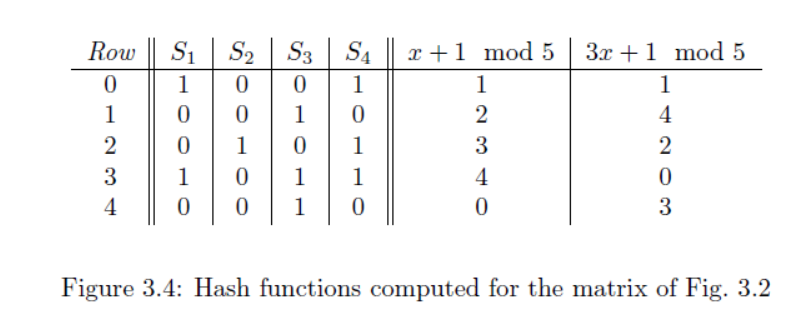


图3-4 习题示例图

**参考答案：**初始时，签名矩阵为:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| h1 |  |  |  |  |
| h2 |  |  |  |  |
| h3 |  |  |  |  |
| h4 |  |  |  |  |

首先，考虑原始矩阵中第0行。h1(0)=1;h2(0)=1,h3(0)=4;h4(0)=4。此外，只有S1和S4列的值为1，因此签名矩阵中这两列的值需要修改，修改后的签名矩阵为：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| h1 |  |  |  |  |
| h2 |  |  |  |  |
| h3 |  |  |  |  |
| h4 |  |  |  |  |

接下来，考虑原始矩阵中第1行。h1(1)=2；h2(1)=4;h3(1)=1;h4(1)=2. 此外，有S3列的值为1，因此签名矩阵中这一列的值需要修改，修改后的签名矩阵为:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| h1 |  |  |  |  |
| h2 |  |  |  |  |
| h3 |  |  |  |  |
| h4 |  |  |  |  |

接下来，考虑原始矩阵中第2行。h1(2)=3；h2(2)=2;h3(2)=3;h4(2)=0. 此外，只有S2和S4列的值为1，因此签名矩阵中这两列的值需要对比修改为更小的值，修改后的签名矩阵为:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| h1 |  |  |  |  |
| h2 |  |  |  |  |
| h3 |  |  |  |  |
| h4 |  |  |  |  |

接下来，考虑原始矩阵中第3行。h1(3)=4；h2(3)=0;h3(3)=0;h4(3)=3. 此外，有S1，S3和S4列的值为1，因此签名矩阵中这三列的值需要对比修改为更小的值，修改后的签名矩阵为:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| h1 |  |  |  |  |
| h2 |  |  |  |  |
| h3 |  |  |  |  |
| h4 |  |  |  |  |

接下来，考虑原始矩阵中第4行。h1(4)=0；h2(4)=3;h3(4)=2;h4(4)=1. 此外，有S3列的值为1，因此签名矩阵中这列的值需要对比修改为更小的值，修改后的最终签名矩阵为:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| h1 |  |  |  |  |
| h2 |  |  |  |  |
| h3 |  |  |  |  |
| h4 |  |  |  |  |

### 第四次作业

1. 6.1.1 假定有100个项，编号是1到100，同时有100个购物篮，编号也是1到100。

当且仅当 b能被i整除时，项i放入购物篮 b中。因此，项1放在所有的购物篮中，项2只放在50个偶数编号的购物篮中，其余以此类推。所有能整除12的数所对应的项组成第12号篮中项集{1，2，3, 4，6，12}。请回答下列问题：

a)如果支持度阈值是5，那么哪些项是频繁的？

b)如果支持度阈值是5，那么哪些项对是频繁的？

c)所有购物篮中项的数目之和是多少？

**参考答案：**a)1在篮子里:1、2、3、…、100(总共100个篮子),项目2在篮子里:2、4、6、…、100(总共50个篮子),项目3在篮子里:3、6、9、…100(总共33个篮子), … 因此，只有20个项目在5个以上的篮子里。这些项的列表是:{1,2,3,4,5，…，20}。b） {1,2}将出现50次，{1,3}将出现33次，以此类推，直到20，从阈值5开始，{1,20}将出现5次。然后,{2,3}出现16次;{2,4}出现25次;{2,5}出现10次;{2,6}出现16次;{2,7}出现7次;{2,8}出现12次;{2,9}出现5次;{2,10}出现10次;{2,12}出现8次;{2,14}出现7次;{2,16}出现6次;{2,18}出现5次;{2,20}出现5次;同样地，第一个项目3，将形成一对，{3,4}出现了8次;{3,5}出现6次;{3,6}出现16次;{3,9}出现了11次;{3,12}出现8次;{3,15}出现6次;{3,18}出现5次;同样地，取第一对4，{4,5}出现5次;{4,6}出现了8次;{4,8}出现12次;{4,10}出现5次;{4,12}出现了8次;{4,16}出现6次;{4,20}出现5次;同样地，取第一对5，{5,10}出现10次;{5,15}出现6次;{5,20}出现5次;同样地，取第一对6，{6,9}出现5次;{6,12}出现8次;{6,18}出现5次;同理，对于7,8,9,10，只有一对存在，{7,14},{8,16},{9,18},{10,20}。

c) 计算所有篮子的长度之和:100/1+100/2+100/3+…+100/100=100+50+33+…+1=482。因此，篮子指数从1到100的因子长度之和为482。

1. 6.2.6 将A-priori算法应用到下列数据集，其中支持度阈值为5：

a)习题6.1.1中的数据集；

b)习题6.1.3中的数据集。

**参考答案：**a)执行A-priori算法，首先确定大小为1、支持阈值为5的频繁项集，如下所示: 1,2,3,4,…,18,19,20。然后再次执行A-Priori算法，确定大小为2，支持阈值为5的频繁项集，如下所示: {1,2},{1,3},{1,4},{1,5},{1,6},…,{1,18},{1,19},{1,20},{2,3},{2,4},{2,5},{2,6},{2,7},{2,8},{2,9},{2,10},{2,12},{2,14},{2,16},{2,18},{2,20},{3,4},{3,5},{3,6},{3,9},{3,12},{3,15},{3,18},{4,5},{4,6},{4,8},{4,10},{4,12},{4,16},{4,20},{5,10},{5,15},{5,20},{6,9},{6,12},{6,18},{7,14},{8,16},{9,18},{10,20}。然后再次执行A-Priori算法，确定大小为3的频繁项集，支持阈值为5，如下所示: {1,2,3},{1,2,4},{1,2,5},{1,2,6},{1,2,8},{1,2,9},{1,2,10},{1,2,12},{1,2,14},{1,2,16},{1,2,18},{1,2,20},{1,3,4},{1,3,5},{1,3,6},{1,3,9},{1,3,12},{1,3,15},{1,3,18},{1,4,5},{1,4,6},{1,4,8},{1,4,10},{1,4,12},{1,4,16},{1,4,20},{1,5,10},{1,5,15},{1,5,20},{1,6,9},{1,6,12},{1,6,18},{1,7,14},{1,8,16},{1,9,18},{1,10,20},{2,3,4},{2,3,6},{2,3,9},{2,3,12},{2,3,18},{2,4,5},{2,4,6},{2,4,8},{2,4,10},{2,4,12},{2,4,16},{2,4,20},{2,5,10},{2,5,20},{2,6,9},{2,6,12},{2,6,18},{2,7,14},{2,8,16},{2,9,18},{2,10,20},{3,4,12},{3,5,12},{4,5,10},{4,5,20},{4,6,12},{5,10,20},{6,9,18}。然后再次执行A-Priori算法，确定大小为4，支持阈值为5的频繁项集，如下所示: {1,2,3,4},{1,2,3,6},{1,2,3,9},{1,2,3,18},{1,2,4,5},{1,2,4,6},{1,2,4,8},{1,2,4,10},{1,2,4,12},{1,2,4,12},{1,2,4,20},{1,2,5,10},{1,2,5,20},{1,2,6,9},{1,2,6,12},{1,2,6,18},{1,2,7,14},{1,2,8,16},{1,2,9,18},{1,2,10,20},{1,3,4,12},{1,3,5,15},{1,4,5,10},{1,4,5,20},{1,4,6,12},{1,5,10,20},{1,6,9,18},{2,3,4,12},{2,4,5,10},{2,4,5,20},{2,4,6,12},{2,5,10,20},{2,6,9,18},{4,5,10,20}。然后再次执行A-Priori算法，确定支持阈值为5的大小为5的频繁项集，如下所示: {1,2,3,4,12},{1,2,4,5,10},{1,2,4,5,20},{1,2,4,6,12},{1,2,5,10,20},{1,2,6,9,18},{1,4,5,10,20},{2,4,5,10,20}。然后再次执行A-Priori算法，然后确定支持阈值为5的大小为6的频繁项目集，如下所示: {1,2,4,5,10,20}。没有大小为7的频繁项目集。因此，A-Priori算法到此为止。

b)过程同a)。大小为1：12,16,18,20,24,28,30,32,36,40,42,44,45,48,50,52,54,56,60,63,64,66,68,70,72,75,76,78,80,81,84,88,90,92,96,98,99,100。

大小为2：{36,72},{40,80},{42,84},{45,90},{48,96},{50,100}。

大小为3：{28,56,84},{30,60,90},{32,64,96}。

大小为4：{24,48,72,96}。

大小为5：{18,36,54,72,90},{20,40,60,80,100}。

大小为6：{16,32,48,64,80,96}。

没有大小为7的频繁项目集。因此，A-Priori算法到此为止。

1. 6.3.1 下面给出了12个购物篮组成的集合。每个购物篮都由项1到项6的3个项组成:{1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6}{1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4,5}{3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6}。假定支持度阈值为4。在PCY算法的第一次扫描中，我们使用一个具有11个桶的哈希表，集合{i，j)会哈希到桶i\*j mod 11。

a)不管采用什么方式,计算每个项及每个项对的支持度。

b)哪个项对会哈希到哪个桶中？

c)哪些桶是频繁的？

d)在PCY算法的第二次扫描中，哪些项对会被计数？

**参考答案：**a) 每个item的支持度如下：

|  |  |
| --- | --- |
| **item** | **count** |
| 1 | 4 |
| 2 | 6 |
| 3 | 8 |
| 4 | 8 |
| 5 | 6 |
| 6 | 4 |

每一对items的支持度如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **item** | **pair** | **Hash**  **Bucket** | **count** |
| 1 | {1,2} | 2 | 2 |
| 2 | {1,3} | 3 | 3 |
| 3 | {1,4} | 4 | 2 |
| 4 | {1,5} | 5 | 1 |
| 5 | {1,6} | 6 | 3 |
| 6 | {2,3} | 6 | 3 |
| 7 | {2,4} | 8 | 4 |
| 8 | {2,5} | 10 | 2 |
| 9 | {2,6} | 1 | 1 |
| 10 | {3,4} | 1 | 4 |
| 11 | {3,5} | 4 | 4 |
| 12 | {3,6} | 7 | 2 |
| 13 | {4,5} | 9 | 3 |
| 14 | {4,6} | 2 | 3 |
| 15 | {5,6} | 8 | 2 |

b) 桶0:

桶1:{3,4},{2,6}

桶2:{1,2},{4,6}

桶3:{1,3}

桶4:{3,5},{1,4}

桶5:{1,5}

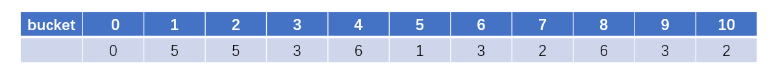
桶6:{1,6},{2,3}

桶7:{3,6}

桶8:{2,4},{5,6}

桶9:{4,5}

桶10:{2,5}

c) 每个桶及其计数为:

因此，给定的支持阈值为4，频繁桶是支持阈值大于4的桶。因此，频繁的存储桶是1、2、4和8。

d) 桶1、2、4和8的支持阈值大于4，它们的对分别是:{2,6}、{3,4}、{1,2}、{4,6}、{1,4}、{3,5}、{2,4}、{5,6} 。

### 第五次作业

1. 9.3.1 图9-8是一个效用矩阵,代表了3个用户A、B和C对8个项目a到h的评分,评分范围是1到5星。从该矩阵的数据计算以下内容:  
   a) 将效用矩阵视为布尔矩阵,计算每对用户之间的Jaccard距离；  
   b) 重复(a),但使用余弦距离；  
   c) 将3、4、5星视为1,将1、2星和空白视为0。计算每对用户之间的Jaccard距离  
   d) 重复(c),但使用余弦距离；  
   e) 对矩阵进行归一化处理,即对每个非空条目中减去该用户的平均值；  
   f) 使用(e)中的归一化矩阵,计算每对用户之间的余弦距离。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 4 | 5 |  | 5 | 1 |  | 3 | 2 |
| B |  | 3 | 4 | 3 | 1 | 2 | 1 |  |
| C | 2 |  | 1 | 3 |  | 4 | 5 | 3 |

图9-8 效用矩阵

**参考答案：**a）A与B交集{b,d,e,g}， A与B并集{a,b,c,d,e,f,g,h}，J(AB)=1-4/8= 4/8。

A与C交集{a,d,g,h}，A与C并集{a,b,c,d,e,f,g,h}，J(AC)= 1-4/8= 4/8。

B与C交集{c,d,f,g}，B与C并集{a,b,c,d,e,f,g,h}，J(AC)= 1-4/8= 4/8。

b）  
bool向量A:[1,1,0,1,1,0,1,1]  
bool向量B:[0,1,1,1,1,1,1,0]  
bool向量C:[1,0,1,1,0,1,1,1]  
cos( A, B )=1-4/6=2/6; cos( A, C)= 1-4/6=2/6; cos(B, C )= 1-4/6=2/6;

c)阈值处理后

A与B交集{b,d}，A与B并集{a,b,c,d,g}，J(AB)=1-2/5= 3/5。

A与C交集{d,g}，A与C并集{a,b,d,f,g,h}，J(AC)= 1-2/6= 4/6。

B与C交集{d}，B与C并集{b,c,d,f,g,h}，J(AC)= 1-1/6= 5/6。

d) cos( A, B )=; cos( A, C)=1-1/2=1/2; cos(B, C )=;

e) 标准化后

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 0.67 | 1.67 |  | 1.67 | -2.33 |  | -0.33 | -1.33 |
| B |  | 0.67 | 1.67 | 0.67 | -1.33 | -0.33 | -1.33 |  |
| C | -1 |  | -2 | 0 |  | 1 | 2 | 0 |

f）  
A向量：[0.67,1.67,0,1.67,-2.33,0,-0.33,-1.33]  
B向量：[0,0.67,1.67,0.67,-1.33,-0.33,-1.33,0]  
C向量：[-1,0,-2,0,0,1,2,0]

cos(A,B)= 0.416

cos(A,C)= 1.115

cos(B,C)= 1.739

### 第六次作业

1. 4.3.1 对于8billion bits的位数组,集合S有1billion个成员,如果使用三个哈希函数,计算误报率是多少?如果使用四个哈希函数又会是多少?

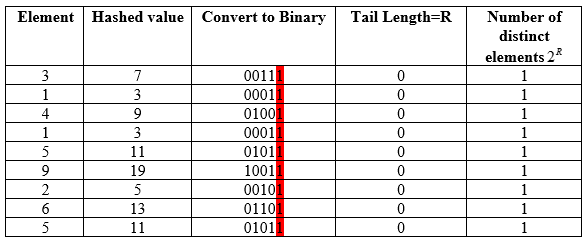
**参考答案：**假阳率公式,其中k 表示哈希函数的个数,n 表示位数组的大小(布隆过滤器的长度),m 表示集合 S 中成员的总数

那么k=3时，

k=4时，

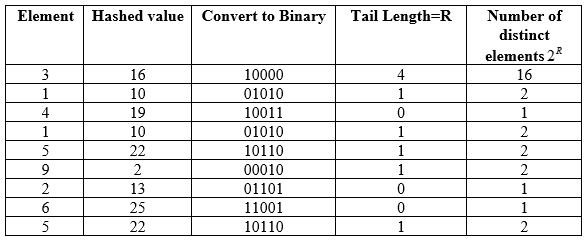
1. 4.4.1 假设我们的数据流包含整数 3，1，4，1，5，9，2，6，5。我们的哈希函数都采用形式 h(x) = ax + b mod 32,对于某些 a 和 b。结果视为一个 5 位的二进制整数。确定每个流元素的尾长度,并根据以下哈希函数估算不同元素的数量:   
   a) h(x) = 2x + 1 mod 32   
   b) h(x) = 3x + 7 mod 32  
   c) h(x) = 4x mod 32

**参考答案：**a) **第1步:** 将元素x的值代入给定公式h(x)=2x+1 mod 32计算哈希值。  
**第2步:**将计算出的哈希值转换为二进制数字。然后若有需要，在前面补0使其成为5位二进制整数。  
**第3步:** 计算尾长度R。当用户将哈希函数h应用于流元素a时,哈希值h(a)将以某些0结尾,可能没有0。  
**第4步:**最后,将尾长度R的值代入公式，其中,计算不同元素的数量估计值。 通过上述步骤并结合Flajolet-Martin算法进行计算:



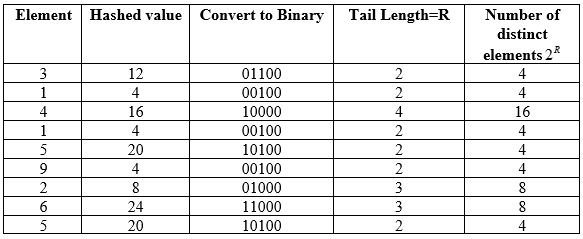
通过上述得R=0, 不同元素数量估计为

b)



通过上述得R=4, 不同元素数量估计为。

c)



通过上述得R=4, 不同元素数量估计为1。

1. 4.5.1 计算数据流3,1,4,1,3,4,2,1,2的奇异数(第二阶矩)。这个数据流的第三阶矩是多少?

**参考答案：**这是一个长度为9的流，其中元素3出现2次，元素4出现2次，元素1出现3次，元素2出现2次。

因此奇异数=3\*2^2+3^2=21

而三阶矩=3\*2^3+3^3=51

1. 4.6.1假设窗口如图4-2所示。估计最后k个位置中1的数量,对于k=  
   a)5   
   b)15  
   在每种情况下,你的估计值与正确值相差多少?

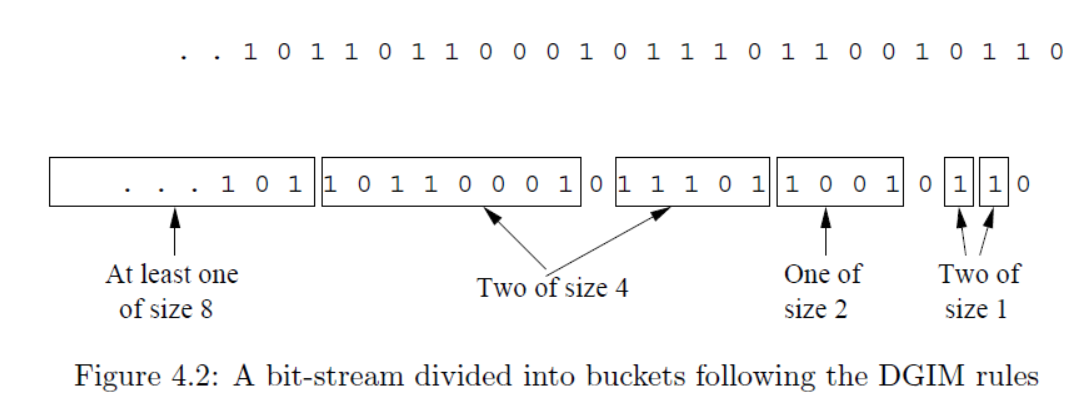


图4-2 习题示例图

**参考答案：**a) k=5时，预估1的数量=2\*桶大小为1+桶大小为2的一半=3。   
真实1的数量=3，预测与真实值一致。

b)k=15时，预估1的数量=2\*桶大小为1+桶大小为2+桶大小为4+桶大小为4的一半=10。 真实1的数量=9，预测值与真实值差1。

### 第七次作业

1. 7.2.2 如果我们对两个cluster之间的距离使用以下定义,那么图7.2中的聚类结果会发生怎样的变化?  
   a) 两个cluster中任意一对点之间的最小距离；  
   b) 两个cluster中所有点对之间距离的平均值。

**参考答案：**a）根据图7.2的信息，先计算任意两个数据之间的欧氏距离得到一个相似度矩阵(proximity matrix)如下【备注是对称矩阵，因此只写了其中一部分信息】

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | (2,2) | (3,4) | (4,8) | (4,10) | (5,2) | (6,8) | (7,10) | (9,3) | (10,5) | (11,4) | (12,3) | (12,6) |
| (2,2) | 0 |  |  |  |  |  |  |  |  |  |  |  |
| (3,4) |  | 0 |  |  |  |  |  |  |  |  |  |  |
| (4,8) |  |  | 0 |  |  |  |  |  |  |  |  |  |
| (4,10) |  |  |  | 0 |  |  |  |  |  |  |  |  |
| (5,2) |  |  |  |  | 0 |  |  |  |  |  |  |  |
| (6,8) |  |  |  |  |  | 0 |  |  |  |  |  |  |
| (7,10) |  |  |  |  |  |  | 0 |  |  |  |  |  |
| (9,3) |  |  |  |  |  |  |  | 0 |  |  |  |  |
| (10,5) |  |  |  |  |  |  |  |  | 0 |  |  |  |
| (11,4) |  |  |  |  |  |  |  |  |  | 0 |  |  |
| (12,3) |  |  |  |  |  |  |  |  |  |  | 0 |  |
| (12,6) |  |  |  |  |  |  |  |  |  |  |  | 0 |

一开始每个点构成一个簇，其簇质心就是点本身。现在簇间距离定义为两个簇上点之间的最短距离，两个点分别来自不同的簇。在所有的点对之间的距离最近（10，5）和（11,4），或（11,4）和（12,3），都是。因此随机选择一组合并。例如（11,4）和（12,3），将他们当做一个新的簇。那么接下来更新簇到簇之间的距离，这儿定义为两个簇上点对之间的最短距离，且这两个点来自不同的簇。因此更新相似度矩阵如下：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | (2,2) | (3,4) | (4,8) | (4,10) | (5,2) | (6,8) | (7,10) | (9,3) | (10,5) | (11,4),(12,3) | (12,6) |
| (2,2) | 0 |  |  |  |  |  |  |  |  |  |  |
| (3,4) |  | 0 |  |  |  |  |  |  |  |  |  |
| (4,8) |  |  | 0 |  |  |  |  |  |  |  |  |
| (4,10) |  |  |  | 0 |  |  |  |  |  |  |  |
| (5,2) |  |  |  |  | 0 |  |  |  |  |  |  |
| (6,8) |  |  |  |  |  | 0 |  |  |  |  |  |
| (7,10) |  |  |  |  |  |  | 0 |  |  |  |  |
| (9,3) |  |  |  |  |  |  |  | 0 |  |  |  |
| (10,5) |  |  |  |  |  |  |  |  | 0 |  |  |
| (11,4),(12,3) |  |  |  |  |  |  |  |  |  | 0 |  |
| (12,6) |  |  |  |  |  |  |  |  |  |  | 0 |

那么接下来再去检查簇之间的最短距离，发现（10，5）聚类刚才的簇（11,4），（12,3)距离最近，因此将该点聚类到一起。因此更新相似度矩阵如下：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | (2,2) | (3,4) | (4,8) | (4,10) | (5,2) | (6,8) | (7,10) | (9,3) | (11,4),(12,3),(10,5) | (12,6) |
| (2,2) | 0 |  |  |  |  |  |  |  |  |  |
| (3,4) |  | 0 |  |  |  |  |  |  |  |  |
| (4,8) |  |  | 0 |  |  |  |  |  |  |  |
| (4,10) |  |  |  | 0 |  |  |  |  |  |  |
| (5,2) |  |  |  |  | 0 |  |  |  |  |  |
| (6,8) |  |  |  |  |  | 0 |  |  |  |  |
| (7,10) |  |  |  |  |  |  | 0 |  |  |  |
| (9,3) |  |  |  |  |  |  |  | 0 |  |  |
| (11,4),(12,3),(10,5) |  |  |  |  |  |  |  |  | 0 |  |
| (12,6) |  |  |  |  |  |  |  |  |  | 0 |

那么接下来再去检查簇之间的最短距离，发现（4，10）聚类点(4,8)距离最近，因此将这两个点聚类到一起。因此更新相似度矩阵如下：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | (2,2) | (3,4) | (4,8),(4,10) | (5,2) | (6,8) | (7,10) | (9,3) | (11,4),(12,3),(10,5) | (12,6) |
| (2,2) | 0 |  |  |  |  |  |  |  |  |
| (3,4) |  | 0 |  |  |  |  |  |  |  |
| (4,10),(4,8) |  |  | 0 |  |  |  |  |  |  |
| (5,2) |  |  |  | 0 |  |  |  |  |  |
| (6,8) |  |  |  |  | 0 |  |  |  |  |
| (7,10) |  |  |  |  |  | 0 |  |  |  |
| (9,3) |  |  |  |  |  |  | 0 |  |  |
| (11,4),(12,3),(10,5) |  |  |  |  |  |  |  | 0 |  |
| (12,6) |  |  |  |  |  |  |  |  | 0 |

如此循环。

b）

初始时最先也是先选两个最近的点合并。例如（11,4）和（12,3），将他们当做一个新的簇。接下来，采用簇上点对的平均距离时，需要更详细地计算不同簇之间的所有点对的距离再求平均值作为簇间距离。其他步骤类似a)。

1. 7.3.1 对于图7-8中的点,如果我们使用首先随机选择一个初始点，之后每一个初始点与已选择初始点的最小距离是所有未被选择点中最大的方式选择三个起始点,并且第一个选择的点是(3,4),那么其他被选择的点是哪些？

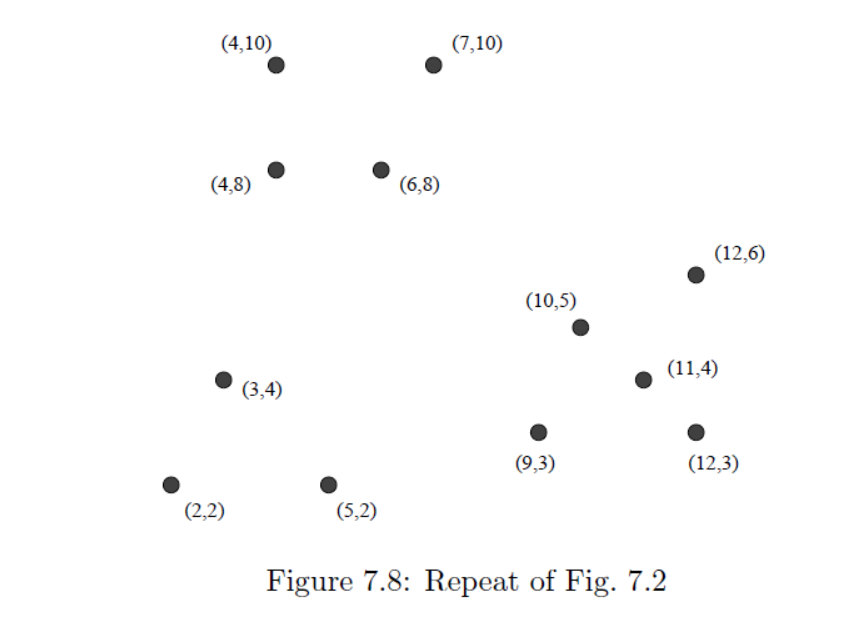


图7-8 习题示例图

**参考答案：**在图中除了（3,4）点以外的其他11个点到（3,4）的距离分别是：

|  |  |
| --- | --- |
|  | （3,4） |
| （2,2） |  |
| （4,8） |  |
| （4,10） |  |
| （5,2） |  |
| （6,8） |  |
| （7,10） |  |
| （9,3） |  |
| （10,5） |  |
| （11,4） |  |
| （12,3） |  |
| （12,6） |  |

根据表格可知，距离（3,4）最远的点是（12,6），因此第二个初始点为（12,6）。

在剩下的10个点中，离（3,4）或（12,6）的距离分别是：

|  |  |  |  |
| --- | --- | --- | --- |
|  | (3,4) | (12,6) | 该点的最终得分  （两个距离中取最小的那个值） |
| （2,2） |  |  |  |
| （4,8） |  |  |  |
| （4,10） |  |  |  |
| （5,2） |  |  |  |
| （6,8） |  |  |  |
| （7,10） |  |  |  |
| （9,3） |  |  |  |
| （10,5） |  |  |  |
| （11,4） |  |  |  |
| （12,3） |  |  |  |

从该表格可知，到点（3,4）或（12,6）的最短距离最大的点是（7,10）。综上，我们选择（3,4），（12,6），（7,10）作为三个初始点。

1. 7.3.4 对于图7-8中的三个簇:   
   a) 计算每个簇的BFR算法表示,即计算N、SUM和SUMSQ。   
   b) 计算每个簇在两个维度上(x和y)的方差和标准差。

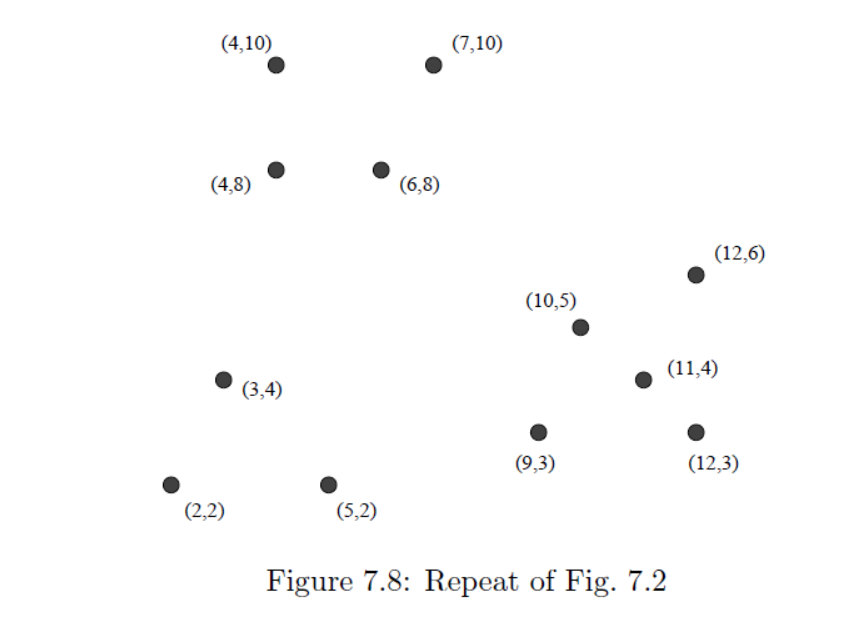


图7-8 习题示例图

**参考答案：**a)根据图7-8，图中存在三个簇：

Cluster 1: (4,10), (7,10), (4,8), (6,8)  
Cluster 2: (12,6), (10,5), (11,4), (12,3), (9,3)  
Cluster 3: (2,2), (5,2), (3,4)

Cluster 1:  
N = 4  
SUM\_x = 21, SUM\_y = 36  
SUMSQ\_x = 117, SUMSQ\_y = 328

Cluster 2:  
N = 5  
SUM\_x = 54, SUM\_y = 21  
SUMSQ\_x = 590, SUMSQ\_y = 95

Cluster 3:  
N = 3  
SUM\_x = 10, SUM\_y = 8  
SUMSQ\_x = 38, SUMSQ\_y = 24

b)

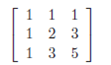
Cluster 1:  
Variance\_x = 27/16, Variance\_y = 1  
Standard Deviation\_x = 3 / 4\*√3, Standard Deviation\_y = 1

Cluster 2:  
Variance\_x = 34/25, Variance\_y = 34/25  
Standard Deviation\_x = √34 / 5, Standard Deviation\_y = √34 / 5

Cluster 3:  
Variance\_x = 14/9, Variance\_y = 8/9  
Standard Deviation\_x = √14 / 3, Standard Deviation\_y = 2/3 √2

### 第八次作业

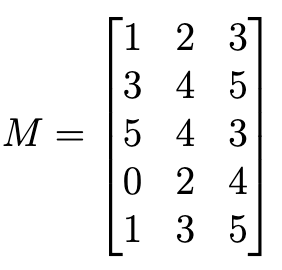
1. 11.1.4 请找出下列矩阵的特征值和特征向量:

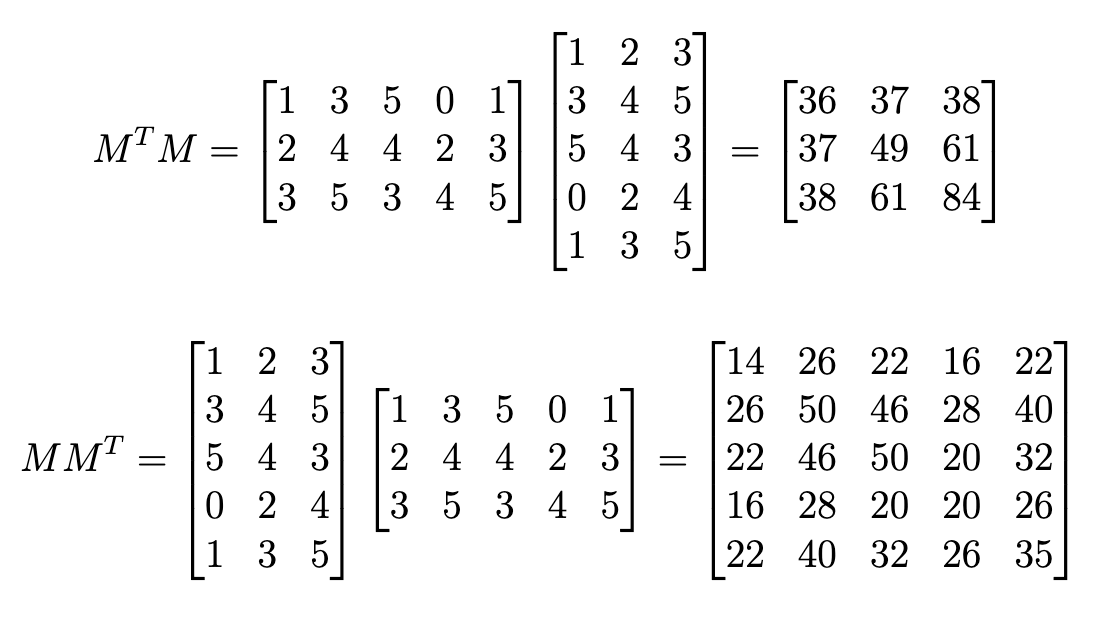


**参考答案：**根据得到，得到。

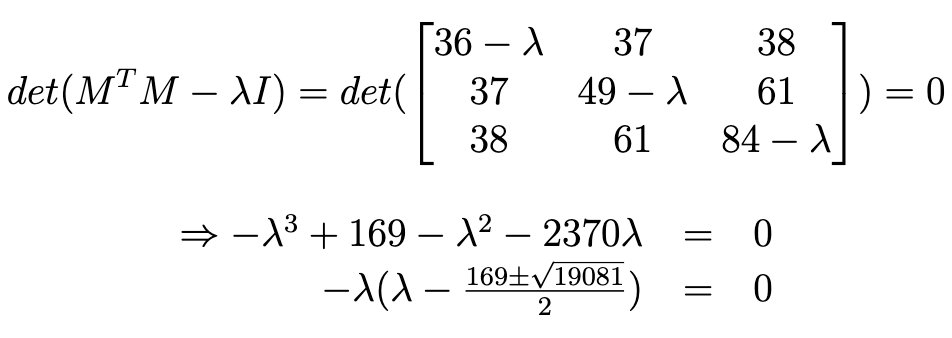
对应特征向量

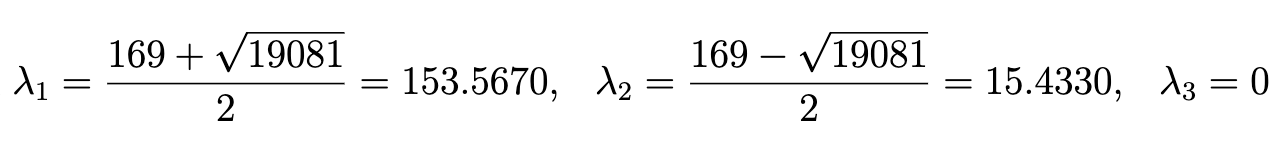
1. 11.3.1 给出了一个矩阵M。通过观察可以看出,第一列加上第三列再减去两倍的第二列等于0,因此这个矩阵的秩为2。  
   a) 计算矩阵M^T M和MM^T  
   b) 求出(a)部分矩阵的特征值   
   c) 求出(a)部分矩阵的特征向量  
   d) 根据(b)和(c)部分的结果,求出原始矩阵M的奇异值分解(SVD)。注意只有两个非零特征值,所以你的∑矩阵应只有两个奇异值,而U和V只有两列。  
   e) 将较小的奇异值设为0,计算图1中矩阵M的一维近似。   
   f) 一维近似保留了原始奇异值能量的多少部分?



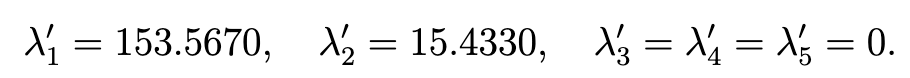
**参考答案：**a）  


b）的特征值可通过以下方式计算

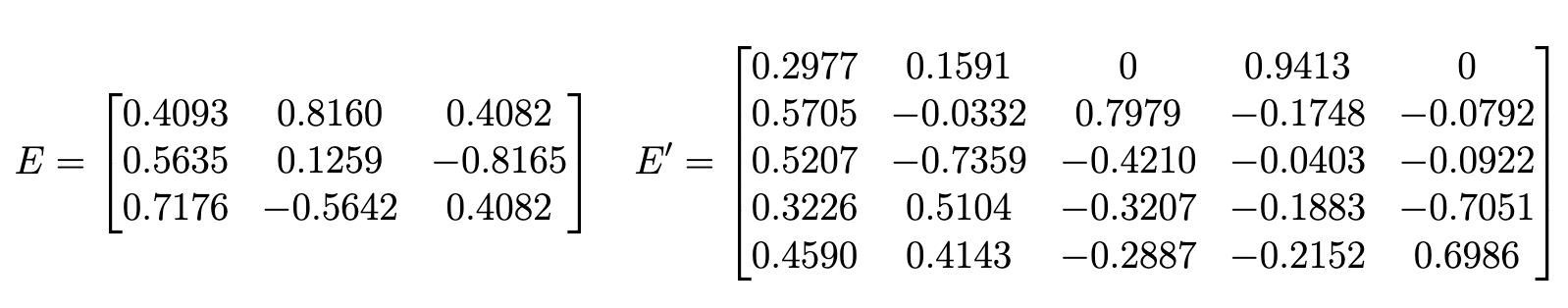




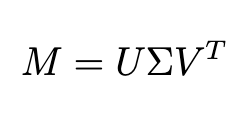
由于dim()<dim()，所以的特征值是的特征值加上额外的0，即



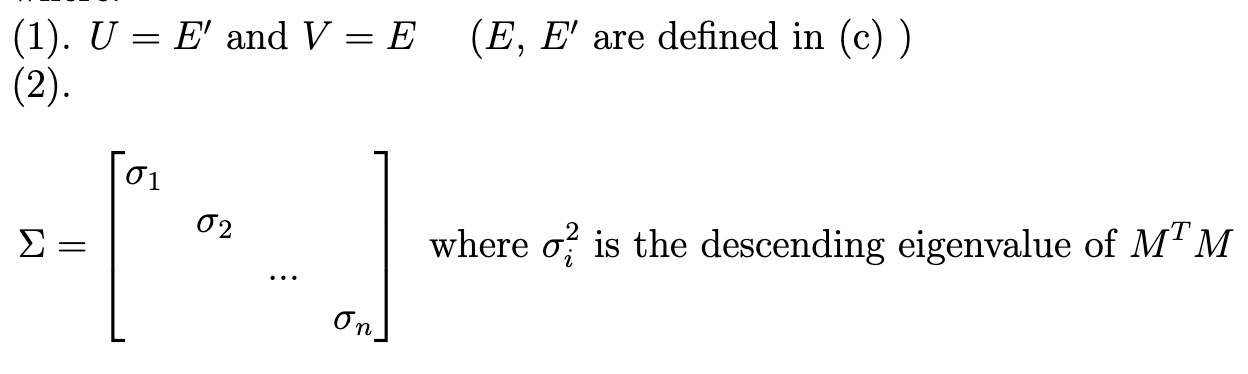
c）使用power iteration算法分别得到和的特征向量。让为矩阵，其中第i列为的第i个特征向量，同样有对于



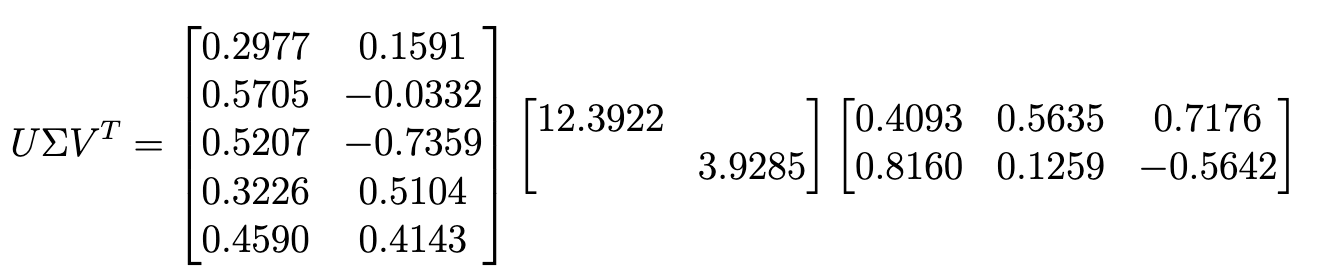
d）根据SVD定义



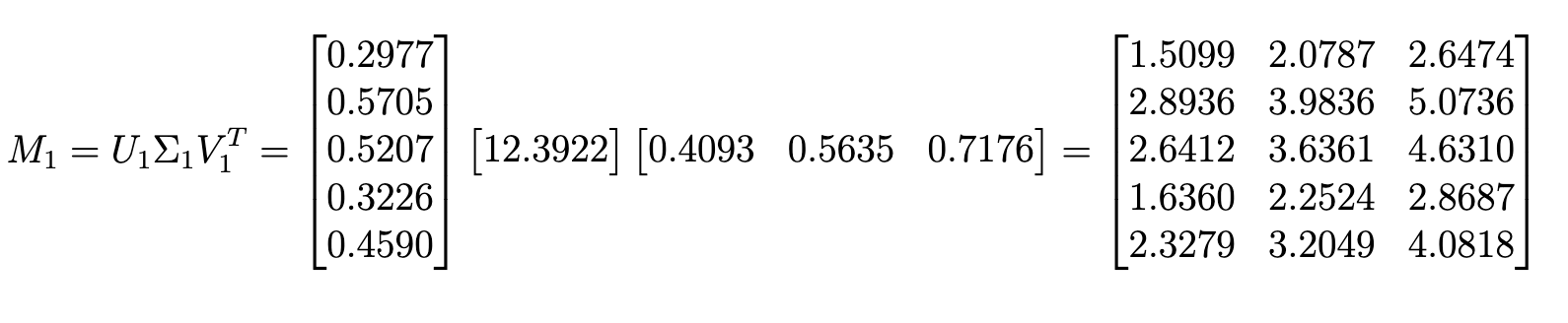
其中：



最终有



e）



f）

