# 大数据分析第四次作业

**（最晚提交时间 2025.06.06,18:00 之前提交到微助教上）**

1. 三台计算机分别为 A，B，C，其特征如下表 1.1 所示。我们采用这些特征来定义每台计算机的向量，例如 A 的向量为[3.06,500,6]。我们可以计算任意两个向量之间的余弦距离。但是如果不对向量的分量进行放缩变换的话，那么磁盘的大小会主导距离的计算结果，而其他分量本质上几乎不起作用。假设我们分别使用 1，α，β作为处理器速度，磁盘大小和内存大小的放缩变换因子，求解以下（10 分）

（a）基于α，β，计算三台计算机的每一对向量之间的夹角余弦相似度。

（b）如果α=0.01，β=0.5，上述向量之间的夹角余弦相似度分别是多少？

表 1.1 计算机特征参数

| 特征 | A | B | C |
|---|---|---|---|
| 处理器速度 | 3.06 | 2.68 | 2.92 |
| 磁盘大小 | 500 | 320 | 640 |
| 内存大小 | 6 | 4 | 6 |

2. 如表 1.2 所示给出了一个基于 1 到 5 级评分的效用矩阵，其中有 8 个项（a 到 h），3 个用户（A,B,C），基于该效用矩阵计算如下（20 分）。

(a) 将该效用矩阵看成布尔矩阵，计算每对用户之间的 Jaccard 距离。

(b) 将该效用矩阵看成布尔矩阵，计算每对用户之间的余弦距离。

(c) 将评分 3 到 5 看成 1，评分 1 和 2 以及空白看成 0。计算每对用户之间的余弦距离。

(d) 通过减去用户非空评分的平均值对效用矩阵进行归一化。然后利用得到的归一化后的矩阵，计算每对用户之间的余弦距离。

表 1.2 效用矩阵

| | 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 |

3. 如图 1.3 所示给出了一个矩阵。可以看出矩阵的第一列加上第三列减去第二列的两倍等于 0，也即该矩阵的秩为 2。计算如下（20 分）：

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix}$$

图 1.3 矩阵 M

（a）计算矩阵$MM^T$和$M^TM$

（b）计算(a)获得矩阵的特征值

（c）计算(a)获得矩阵的特征向量

（d）基于(b)和(c)的结果计算原始矩阵 M 的 SVD 分解。注意这里只有 2 个非零特征值，所以矩阵∑只有两个奇异值，矩阵 U 和 V 只有两列。

4. 采用 80 亿位对 10 亿元素组成的集合 S 进行布隆过滤器的过滤，如果使用 3 个哈希函数，假阳率是多少？如果使用 4 个哈希函数呢？（10 分）

5. 假定某个流由整数 3、1、4、1、5、9、2、6、5 构成。给定的哈希函数形式为 h(x)=ax+b mod 32，其中 a 和 b 是给定的常数。这里的哈希结果应看成一

个 5 位的二进制整数。那么对下列的每个哈希函数，请确定每个流元素的尾长并对独立元素数目进行估计：（20 分）

(a) h(x)=2x+1 mod 32

(b) h(x)=4x mod 32

6. 计算流 3,1,41,3,4,2,1,2 的二阶矩和三阶矩(10 分)。

7. 如图 4.2 所示的窗口中，分别估计 k=5 和 k=15 时最后 k 位中的 1 的个数，并给出两个估计结果与真实值的差异（10 分）。

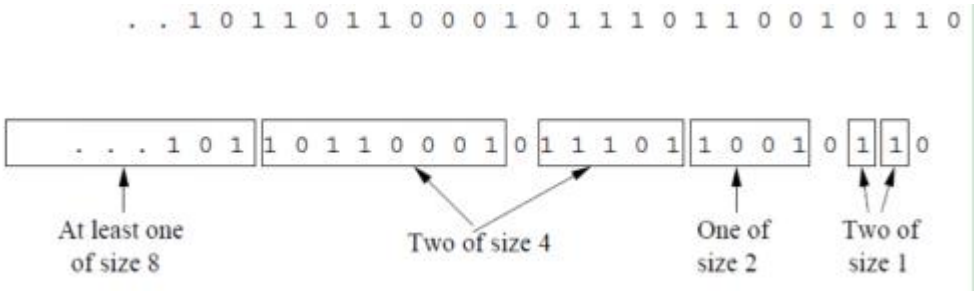


图 4.2 基于 DGIM 规则将位流划分成多个桶的例子