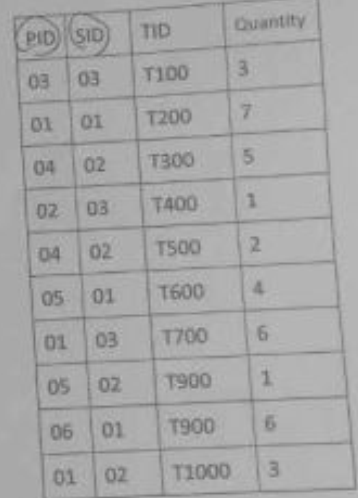
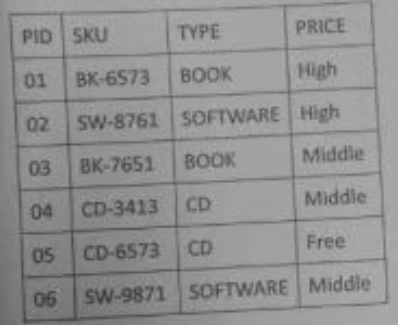
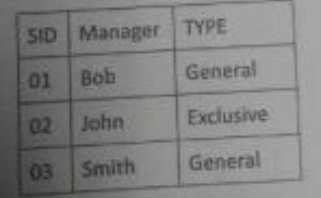
1. 数据仓库及其实现技术
2. 试采用bitmap索引方式对图一的维度表进行索引
3. 试采用join index对图一的事实表和维表进行索引

产品维度表



销售事实表

商店维度表

1. 给定图2中的目标集（DOG）和对比集（CAT），使用信息增益计算各个属性与当前概念描述任务之间的相关性。并采用T=0.1作为阈值，对属性进行筛选。



|  |  |
| --- | --- |
| 事务ID | 购买项 |
| 1 | {a,b,d,e} |
| 2 | {b,c,d} |
| 3 | {a,b,d,e} |
| 4 | {a,c,d,e} |
| 5 | {b,c,d,e} |
| 6 | {b,d,e} |
| 7 | {c,d} |
| 8 | {a,b,c} |
| 9 | {a,d,e} |
| 10 | {b,d} |

针对图3的交易事务数据，采用Fp增长算法求频繁项集，最小支持度为>=30%

{abde}

基于上述频繁项集，构造关联规则，要求最小置信度>=50%

1. 数据预处理与分类

（1）针对图4的训练数据集进行离散化处理。要求采用等宽分桶的方式将age和income属性离散到3个区间。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | age | income | student | Class:buys\_MP |
| 1 | 23 | 68000 | no | >2000 |
| 2 | 49 | 36000 | no | 1000…2000 |
| 3 | 55 | 22000 | no | 1000…2000 |
| 4 | 34 | 30000 | yes | <1000 |
| 5 | 38 | 15000 | yes | <1000 |
| 6 | 57 | 75000 | no | >2000 |
| 7 | 21 | 52000 | no | 1000…2000 |
| 8 | 31 | 45000 | yes | 1000…2000 |
| 9 | 66 | 58000 | no | 1000…2000 |
| 10 | 34 | 12000 | yes | <1000 |
| 11 | 40 | 40000 | yes | 1000…2000 |
| 12 | 50 | 78000 | no | >2000 |
| 13 | 29 | 20000 | yes | 1000…2000 |
| 14 | 25 | 70000 | no | <1000 |
| 15 | 61 | 55000 | no | >2000 |
| 16 | 45 | 65000 | no | >2000 |

1. 依据训练集，采用朴素贝叶斯方法分类未知元组（24,75000,yes），对分类属性Class:buys\_MP进行预测
2. 聚类
3. 针对图5的数据，采用曼哈顿距离作为距离函数，给出对应的相异矩阵

|  |  |  |
| --- | --- | --- |
| ID | x | y |
| 1 | 3 | 8 |
| 2 | 2 | 7 |
| 3 | 4 | 8 |
| 4 | 3 | 4 |
| 5 | 4 | 5 |



1. 采用凝聚式层次式方法对该数据集进行聚类，聚类间的距离使用聚类中的数据之间的最大距离进行度量