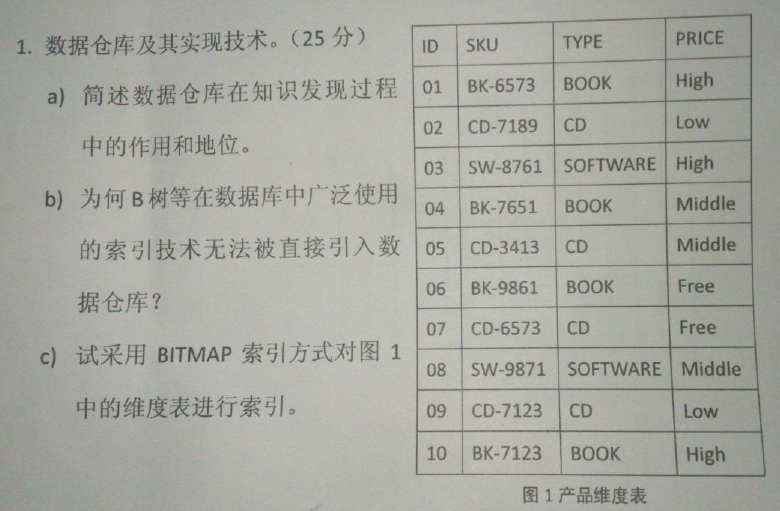
**2012年**



**a)简述数据仓库在知识发现过程中的作用和地位。**

【**作用**】数据仓库是一个面向主题的、集成的、时变的、非易失的数据集合。数据仓库通过数据清理、变换、继承、装入和定期刷新等方法，从一个或多个数据源收集信息，存放在一个一致的模式下。数据仓库能够提供大量的、按照实际要求集成的不同主题的数据，通过OLAP引擎对其进行数据挖掘，发现知识。

【**地位**】数据仓库是知识发现过程中不可或缺的一环，它是进行数据挖掘的必要基础。数据仓库能够提供非冗余的有效数据，这些数据都是面向主题的，因此能够大大提高知识发现的能力和效率，没有数据仓库，知识发现就没有数据源。

**b)为何B树等在数据库中广泛使用的索引技术无法被直接引入数据仓库？**

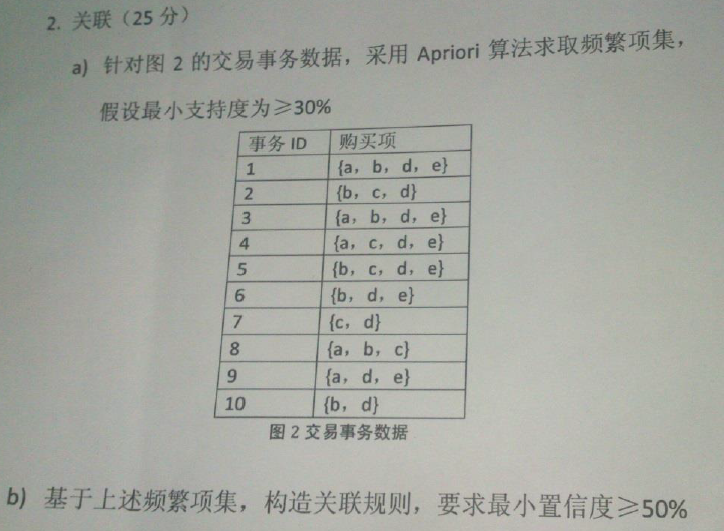
1.数据仓库中海量数据对单列而言数据重复度可能会比较高，对区分度低的属性用B树建立索引存储开销非常大，而bitmap正好合适。

2.B树的查询语句简单，返回结果少。而数据仓库中的复杂查询用B树往往效率很低。

3.创建B树存储的时间复杂度和空间复杂度过高。

**c)试采用BITMAP索引方式对图1中的维度表进行索引。**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ReID** | **BOOK** | **CD** | **SOFTWARE** |  | **ReID** | **Free** | **Low** | **Middle** | **High** |
| 01 | 1 | 0 | 0 | 01 | 0 | 0 | 0 | 1 |
| 02 | 0 | 1 | 0 | 02 | 0 | 1 | 0 | 0 |
| 03 | 0 | 0 | 1 | 03 | 0 | 0 | 0 | 1 |
| 04 | 1 | 0 | 0 | 04 | 0 | 0 | 1 | 0 |
| 05 | 0 | 1 | 0 | 05 | 0 | 0 | 1 | 0 |
| 06 | 1 | 0 | 0 | 06 | 1 | 0 | 0 | 0 |
| 07 | 0 | 1 | 0 | 07 | 1 | 0 | 0 | 0 |
| 08 | 0 | 0 | 1 | 08 | 0 | 0 | 1 | 0 |
| 09 | 0 | 1 | 0 | 09 | 0 | 1 | 0 | 0 |
| 10 | 1 | 0 | 0 | 10 | 0 | 0 | 0 | 1 |

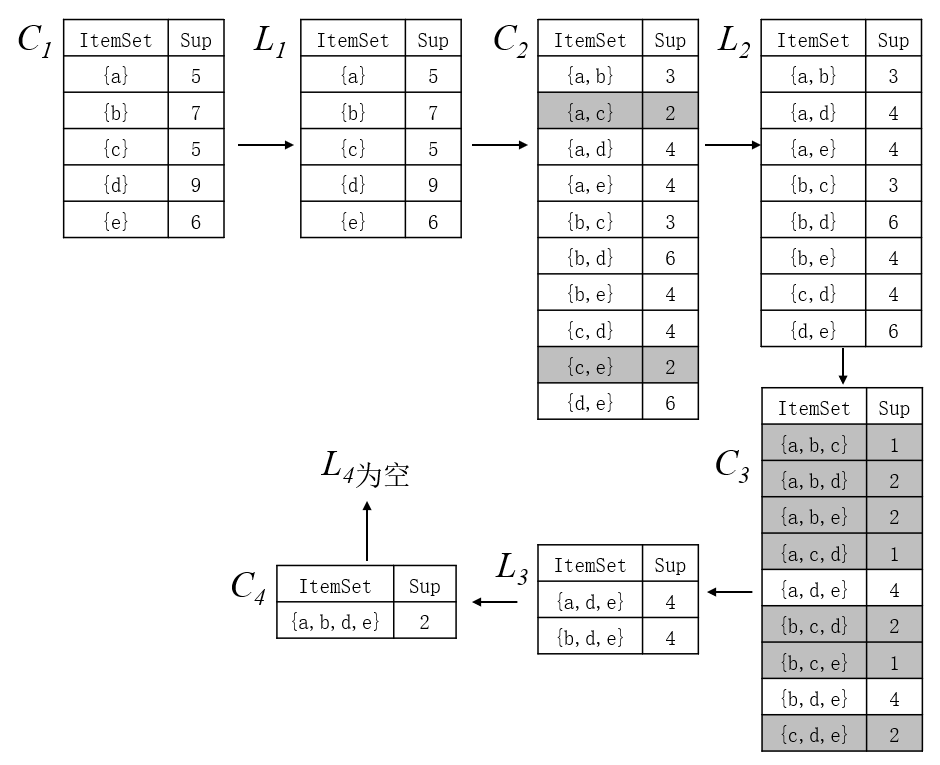


**a)针对图2的交易事务数据，采用Apriori算法求取频繁项集，假设最小支持度为≥30%。**

**由题可知，频繁项集最小个数为10\*30%=3**

**C代表候选集**

**L代表频繁项集**

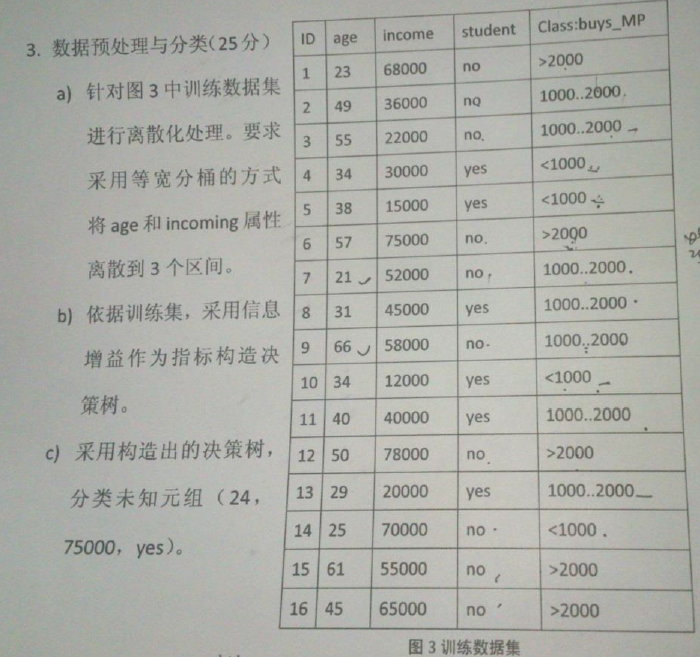


**b)基于上述频繁项集，构造关联规则，要求最小置信度为≥50%。**

confidence(a→b) = support(a∪b)/support(a)

a→b a→d a→e b→d b→e c→b c→d d→b d→e e→a e→b e→d

a→de e→ad b→de e→bd ad→e ae→d de→a bd→e be→d de→b



**a)针对图3中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incoming属性离散到3个区间。**

age区间为15。（max-min）/N = (66-21)/3 = 15

1 4 7 8 10 13 14

2 5 11 12 16

6 9 15

income区间为22000。 (78000-12000)/3 = 22000

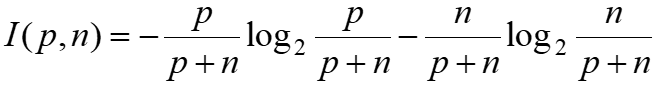
3 4 5 10 13

2 7 8 11 15

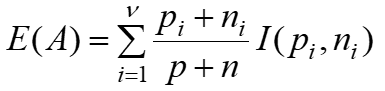
1 6 9 12 14 16

**b)依据训练集，采用信息增益作为指标构造决策树。**

用来判断任一元素属于P还是N的信息量



用来区分的信息量





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **age** | **P (<1000)** | **N (1000..2000)** | **Q (>2000)** | **I(P,N,Q)** |
| <=36 | 3 | 3 | 1 | 1.449 |
| 37…51 | 1 | 2 | 2 | 1.522 |
| 52…66 | 0 | 2 | 2 | 1 |

I(4,7,5) = 1.546

E(age) = 7/16\*I(3,3,1)+5/16\*I(1,2,2)+4/16\*I(0,2,2) = 1.360

Gain(age) = I(4,7,5)-E(age) = 0.186

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **incoming** | **P (<1000)** | **N (1000..2000)** | **Q (>2000)** | **I(P,N,Q)** |
| <=34000 | 3 | 2 | 0 | 0.971 |
| 34001…56000 | 0 | 4 | 1 | 0.722 |
| 56001…78000 | 1 | 1 | 4 | 1.252 |

I(4,7,5) = 1.546

E(incoming) = 5/16\*I(3,2,0)+5/16\*I(0,4,1)+6/16\*I(1,1,4) = 0.999

Gain(incoming) = I(4,7,5)-E(incoming) = 0.547

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **student** | **P (<1000)** | **N (1000..2000)** | **Q (>2000)** | **I(P,N,Q)** |
| yes | 3 | 3 | 0 | 1 |
| no | 1 | 4 | 5 | 1.361 |

I(4,7,5) = 1.546

E(student) = 6/16\*I(3,3,0)+10/16\*I(1,4,5) = 1.226

Gain(student) = I(4,7,5)-E(student) = 0.32

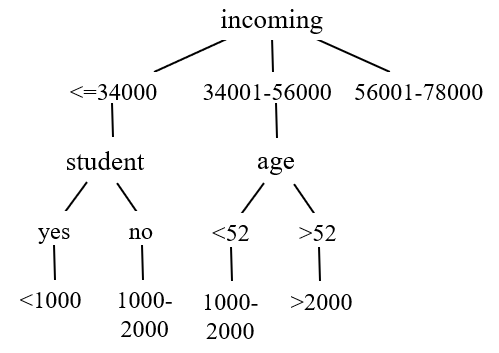
IF incoming=”56001-78000”, THEN buy=”>2000”

IF incoming=”34001-56000” and age=”<52”, THEN buy=”1000-2000”

IF incoming=”34001-56000” and age=”>52”, THEN buy=”>2000”

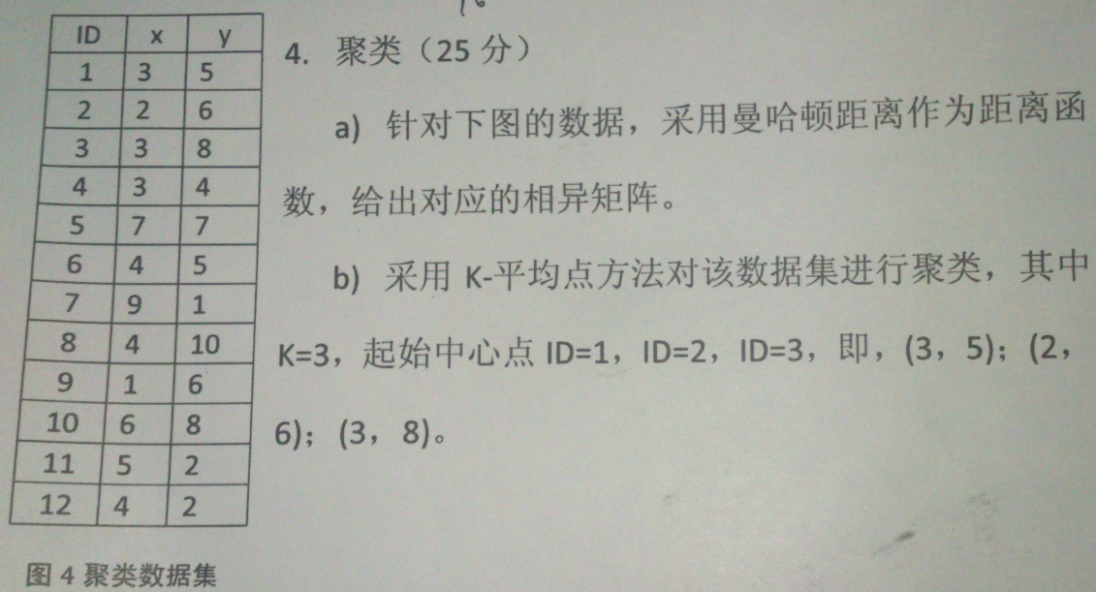
IF incoming=”<=34000” and student=”yes”, THEN buy=”<1000”

IF incoming=”<=34000” and student=”no”, THEN buy=”1000-2000”



**c)采用构造出的决策树，分类未知元组（24，75000，yes）。**

分类至buy=”>2000”元组。



**a)针对下图的数据，采用曼哈顿距离作为距离函数，给出对应的相异矩阵。**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| **1** | 0 |  |  |  |  |  |  |  |  |  |  |  |
| **2** | 2 | 0 |  |  |  |  |  |  |  |  |  |  |
| **3** | 3 | 3 | 0 |  |  |  |  |  |  |  |  |  |
| **4** | 1 | 3 | 4 | 0 |  |  |  |  |  |  |  |  |
| **5** | 6 | 6 | 5 | 7 | 0 |  |  |  |  |  |  |  |
| **6** | 1 | 3 | 4 | 2 | 5 | 0 |  |  |  |  |  |  |
| **7** | 10 | 12 | 13 | 9 | 8 | 9 | 0 |  |  |  |  |  |
| **8** | 6 | 6 | 3 | 7 | 6 | 5 | 14 | 0 |  |  |  |  |
| **9** | 3 | 1 | 4 | 4 | 7 | 4 | 13 | 7 | 0 |  |  |  |
| **10** | 6 | 6 | 3 | 7 | 2 | 5 | 10 | 4 | 7 | 0 |  |  |
| **11** | 5 | 7 | 8 | 4 | 7 | 4 | 5 | 9 | 8 | 7 | 0 |  |
| **12** | 4 | 6 | 7 | 3 | 8 | 3 | 6 | 8 | 7 | 8 | 1 | 0 |

**b)采用K-平均点方法对该数据集进行聚类，其中K=3，起始中心点ID=1，ID=2，ID=3，即，（3，5）；（2，6）；（3，8）。**

求第4-第12个点到三个起始中心点的距离，给每个点分配距离最近的起始中心点。

聚类结果：

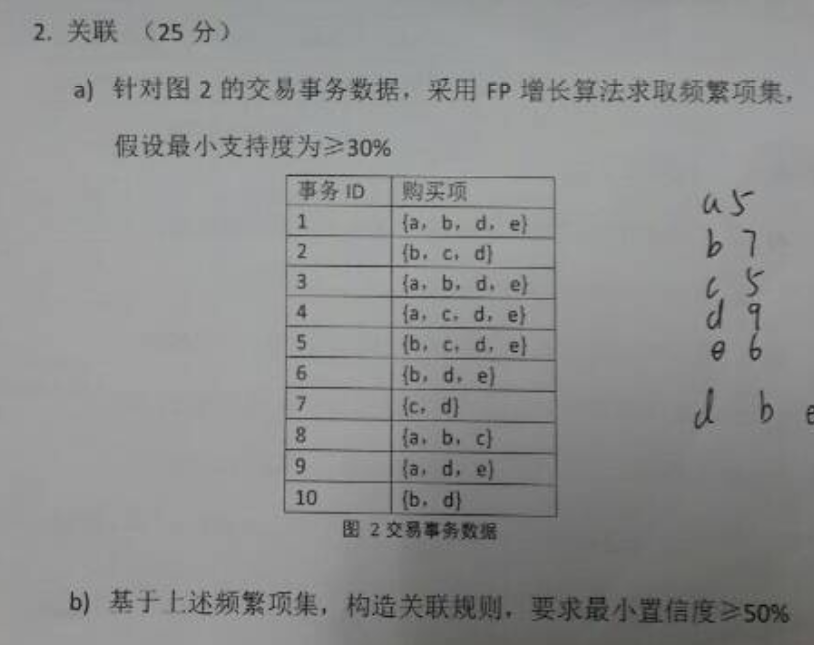
（1，4，6，7，11，12）

（2，9）

（3，5，8，10）

2013

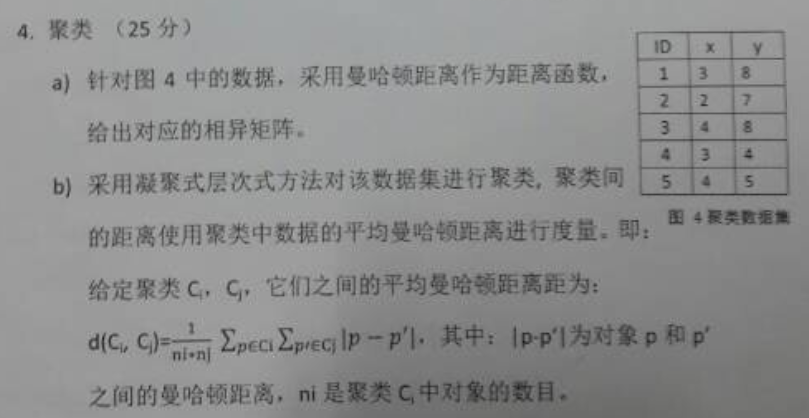
第一题同2012年第一题。



a)针对图2的交易事务数据，采用FP增长算法求取频繁项集，假设最小支持度为≥30%

b)同2012年第二题的b)





2015



