

**a)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。

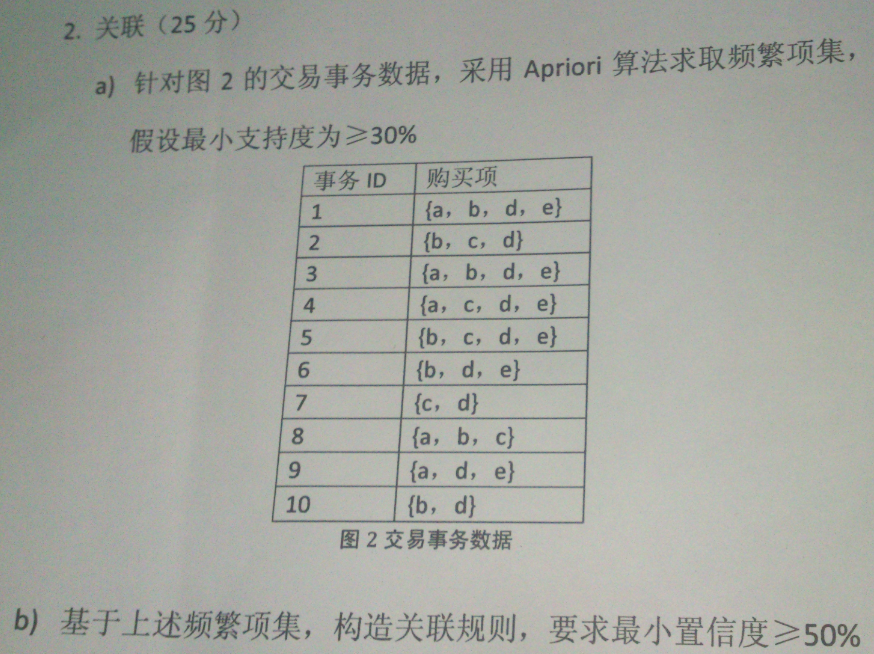
**b)** 1、[数据仓库](http://zhidao.baidu.com/search?word=%E6%95%B0%E6%8D%AE%E4%BB%93%E5%BA%93&fr=qb_search_exp&ie=utf8" \t "_blank)中海量数据对单列而言数据重复度可能会比较高，对区分度低的属性用b-tree建立索引存储开销非常大。而bitmap正好适合。

2、b-tree要求查询语句简单，返回结果少。而数据仓库中的复杂查询b-tree往往效率很低。

3、创建b-tree存储的时间复杂度和空间复杂度过高。

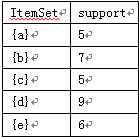
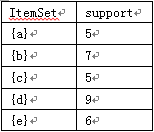
**c)**index on type

|  |  |  |  |
| --- | --- | --- | --- |
| ReID | BOOK | CD | SOFTWARE |
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 1 | 0 | 0 |
| 5 |  | 1 | 0 |
| 6 | 1 | 0 | 0 |
| 7 | 0 | 1 | 0 |
| 8 | 0 | 0 | 1 |
| 9 | 0 | 1 | 0 |
| 10 | 1 | 0 | 0 |

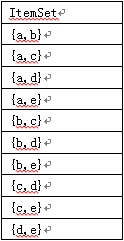
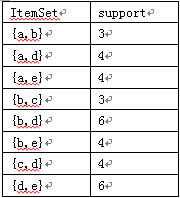


1. 由题得知，频繁项集最小个数为10\*30%=3

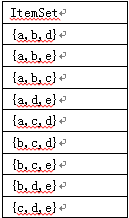
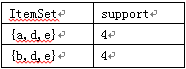
C1： L1：

C2： L2：

C3: L3：

C4: L4：为空

|  |
| --- |
| ItemSet |
| {a,b,d,e} |

b)a->b ad->e

a->e ae->d

e->a de->a

c->b bd->e

b->d be->d

d->b de->b

b->e

e->b

c->d



**a）**Age区间为15

1 4 7 8 10 13 14

2 5 11 12 16

3 6 9 15

Incoming区间为22000

10,5,3,13,4

2,8,11,7,15

1,6,9,12,14,16

**b)**第六章信息增益 I()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| age | <1000 P | 1000…2000 N | >2000 Q | 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 \* 1.449 + 5/16\*1.522 + 0.25\*1 = 1.360

Gain(age) = 1.546 – 1.360 = 0.186

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

I(8,4,4) = 1.5

E(incoming) = 0.303 + 0.226 + 0.470 = 0.999

Gain(incoming) = 1.5 – 0.999 = 0.51

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| student | <1000 P | 1000…2000 N | >2000 Q | 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 + 10/16\*1.361 = 1.226

Gain(student) = 0.32

If incoming > 56000, buy 2000+

If 34000 < incoming <=56000 and age >=52, buy 2000+

If 34000 < incoming <=56000 and age < 52, buy 1000…2000

If incoming <= 34000 and student = no, buy 1000…2000

If incoming <= 34000 and student = yes, buy <1000

c) 分类到>2000 元组

**a)**

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

**b)**3-平均 多次进行聚类，每次调整中心点

(1,4,6,7,11,12)

(2,9)

(3,5,8,10)