OALP: (a1, a2, a3, a4, a5, a6), (b1, b2, a3, a4, a5, a6), and (c1, c2, a3, a4, a5, a6)

cuboids in the full data cube: 2^6 = 64

aggregated = *

Apex cuboid: 0-D cube

Base cuboid: n-D cube

(*,*,*,*) 4-D apex cube

(a1, a2, a3, a4) 4-D base cube

Datacube measures

Distributive: can divide and conquer 分布求 Mean

Algebraic: can be computed by function

Holistic: no bound on the storage data size.

Drill Down 细分 Roll Up Generalize

Slice and dice 选几个 sample pivot 换 XY 轴

distinct aggregated cells

(d1, *)(*, d2)(d1, d2) * 2^4 = 48 *3 = 144

(*, *) 2^4 = 16 - (a1,a2…)(b1, b2….)(c1, c2….)

distinct aggregated cells iceberg cube count >= 3 (*, *, a3, a4, a5, a6) = 2^4

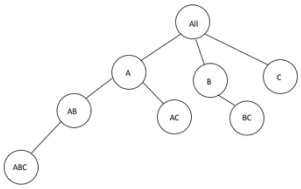
Closed cell: no subset has the same count

(a0,b0,c0) : 1 (a0,b1,c0) : 1 (a0,b2,c0) : 1 (a0,b3,c0) : 1

(a0,b0,c1) : 1 (a0,b1,c1) : 1 (a0,b2,c1) : 1 (a0,b3,c1) : 1

(a0,b0,c2) : 1 (a0,b1,c2) : 1 (a0,b2,c2) : 1 (a0,b3,c2) : 1

trace trees of expansion



Min support = 4 B->A->C

All (* , * , *) : 12 - expansion ————————

B (* ,b0,*):3

B (* ,b1,*):3

B (* ,b2,*):3

B (* ,b3,*):3 ————————

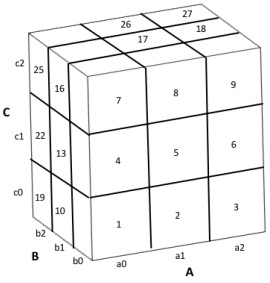
A (a0, * , *) : 12 - expansion ————————

AC (a0,* ,c0) : 4

AC (a0,* ,c1) : 4

AC (a0,* ,c2) : 4 ————————

C (* ,* ,c0):4 C (* ,* ,c1):4 C (* ,* ,c2):4



A, B, and C are 300, 100, and 200 respectively.

minimum memory for holding all the related 2-D planes?

AB = 300 * 100 * 3 BC = 100 * 200 * 1 AC = 300 * 200 * 9 Total = AB + AC + BC = 650000

Trans.	Items
1	b,d,f,g,l
2	f,g,h,l,m,n
3	b,f,h,k,m
4	a,f,h,j,m
5	d,f,g,j,m

Apriori with relative min sup = 0.6 = 5*0.6 = 3

get rid of non-frequent 1-itemsets

C1 = {a,b,d,f,g,h,j,k,l,m,n}; Do db-scanning on C1 to get L1; L1 = {f : 5, g :3, h : 3, m : 4}.

generate all candidate 2-itemsets self-joining on L1 to get C2; C2 = {fg, fh, fm, gh, gm, hm}.

L2={fg:3,fh:3,fm:4,hm:3}.

X is closed if X is frequent and no super- pattern X' ⊃ X that has the same support as X.

All closed pattern: f,fm,fg,fhm.

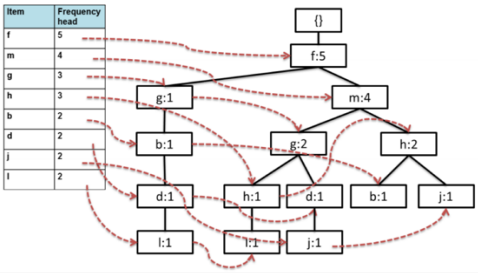
X is a max pattern if X is frequent and no super-pattern X' ⊃ X that is also frequent.

All max pattern: fg,fhm.

Tid	Items	Ordered frequent items
1	b, d, f, g, l	f, g, b, d, l
2	f, g, h, l, m, n	f, m, g, h, l
3	b, f, h, k, m	f, m, h, b
4	a, f, h, j, m	f, m, h, j
5	d, f, g, j, m	f, m, g, d, j

relative min sup = 0.4 = 2

Ordering the items by their frequency will lead to more shared nodes and therefore smaller FP- trees.



Generate Conditional Pattern Bases and Conditional FP-trees for items m,h,b,j based on the FP-tree, and list the frequent patterns computed based on each of the Conditional FP-trees.

Item	Cond. Pattern Base	Frequent Patterns
m	f:4	m,fm
h	fm:2,fmg:1	h,fh,mh,fmh
b	fg:1,fmh:1	b,fb
j	fmh:1,fmgd:1	j,fj,mj,fmj

Association rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

must be bigger than min support and min confidence

Beer > Diaper (60%, 100%)

Diaper > Beer (60%, 75%)

Support = T(X ∪Y)/T(Total)

Confidence = T(X ∪Y)/T(X)

PCA maximize variance, minimize covariance

Zero-mean normalize the data, to have covariance matrix

每个减去 mean

X	0.552	-1.448	0.252	-0.088	1.152	0.352	0.052	-0.948	-0.448	0.572
Y	0.616	-1.384	0.316	0.176	0.916	0.416	-0.024	-0.984	-0.484	0.436

Covariance Matrix

$$\frac{1}{10} * A * A^T = \begin{bmatrix} 0.535 & 0.501 \\ 0.501 & 0.483 \end{bmatrix}$$

calculate eigenVector

$$\det(A-\lambda I) = \begin{vmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1$$
$$= \lambda^2 - 6\lambda + 8 = 0 \implies \lambda_1 = 2,\ \lambda_2 = 4$$

- Find eigenvectors with eigenvalue $\lambda_1 = 2$:
  $A - \lambda_1 I = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
  Solutions to $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \boldsymbol{x} = \boldsymbol{0}$ have basis $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.
  So: $\boldsymbol{x}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ is an eigenvector with eigenvalue $\lambda_1 = 2$.

$$B = P*A = \begin{bmatrix} -0.067 & 0.0070 & -0.056 & -0.188 & 0.129 & -0.059 & 0.053 & 0.061 & 0.043 & 0.078 \\ -0.824 & 2.003 & -0.40 & -0.057 & -1.466 & -0.542 & -0.021 & 1.365 & 0.658 & -0.715 \end{bmatrix}$$

放到新的basis上

new Covariance Matrix

$$\frac{1}{10} * A * A^T = \begin{bmatrix} 0.008 & 0 \\ 0 & 1.011 \end{bmatrix}$$

Since 1.011 > 0.008, it is the most important value.

Second row of B

Attribute types

Nominal: categorize Numerical: not categorized

Symmetry: equally important, asymmetric(opposite)

Ordinal: have meaningful order

Discrete: has only finite or comfortably finite set

Continuous: has real number as attribute value

经验公式: mean – mode = 3 * (mean - median)