Assignment #3

1) S'pose that a data warehouse consists of 3D: time, doctor, patient w/ 2 measures: count, charge

where:

charge = fee doctor charges a patient for a visit

a) Draw a star schema

Visits

Fact table

- time key
- doctors key
- patients key

- charge
- count

Time

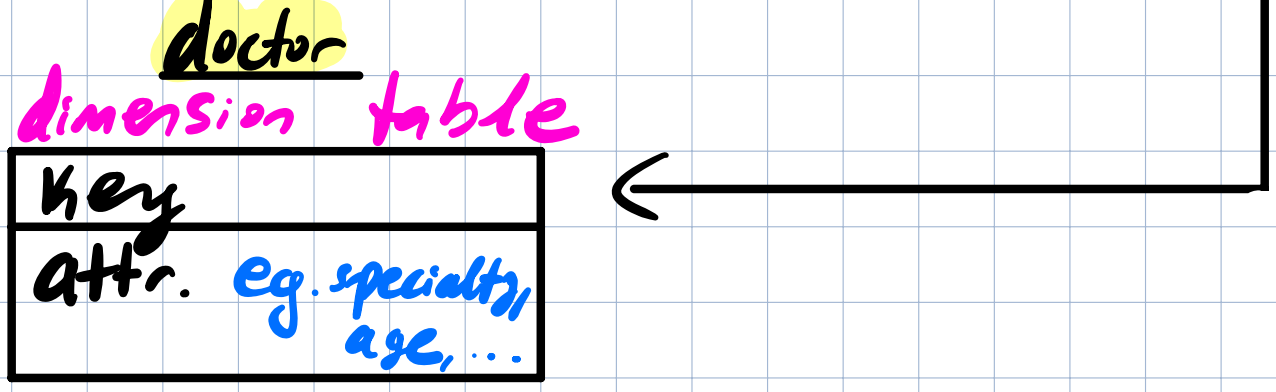
dimension table

key
attr. eg. day, month, ...

Patient

dimension table

key
attr. eg. ailment, age, ...



- The measures belong to the main fact table.
- The main fact table holds the keys to the dims provided.
- b/c the measure given describes a visit & relies on the data from the dims, we name the fact table "Visits".

b) Starting w/ a base cuboid [day, doctor, patient], what OLAP ops should be performed in order to list the total fee collected ∀ doctors in 2020?

time

1) Take day & perform a Roll-Up on day → year.

2) Our data cube now holds info: [year, doctor, patient].

We Pivot so the data cube

is shaped [doctor, year, patient]

3) W/ this new view, we are able to Slice & Dice the time/year dimension to just 2020.

4) The resulting data cube has the form [doctor, 2020, patient]. We now use the measure charge on this data cube to see the fees collected ∀ doctors in 2020.

5) Perform a sum & OLAP op operation to aggregate the charges.

2) A db has 5 transactions

min_sup = 60%.

min_conf = 80%.

TID	items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

Find all frequent itemsets using Apriori & Fp-Growth.

Compare efficiency.

$|D| = 5.$

min_sup = 60% · 5 transactions

= 3 transactions

a) Apriori Alg

C_1

Item	Supp. Ct.
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

compare w/ min_sup →

L_1

Item	Supp. Ct.
M	3
O	3
K	5
E	4
Y	3

i.e. ∀ items

supp_ct ≥ min_sup

C_2

Item Set	Supp. Ct
M O	1
M K	3
M E	2
M Y	2
O K	3
O E	3
O Y	2
K E	4
K Y	3
E Y	2

compare w/ min sup →

L_2

Item Set	Supp Ct.
MK	3
OK	3
OE	3
KE	4
KY	3

C_3

Item Set	Supp. Ct
MOK	1
MOE	1
MKE	2
MKY	2
OKE	3
OKY	2
KEY	2

compare w/ min sup →

L_3

Item Set	Supp. Ct
OKE	3

b) FP-growth (Tree)

Recall min_conf = 80%

TID	items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

MONKEY

M => ONKEY = 1/3 ✗

O => MNKEY = 1/3 ✗

N => MOKEY = 1/2 ✗

K => MONEY = 1/5 ✗

E => MONKY = 1/4 ✗

Y => MONKE = 1/3 ✗

ONKEY => M = 1/2 ✗

MNKEY => O = 1/1 ✓

MOKEY => N = 1/1 ✓

MONEY => K = 1/1 ✓

MONKY => E = 1/1 ✓

MONKE => Y = 1/1 ✓

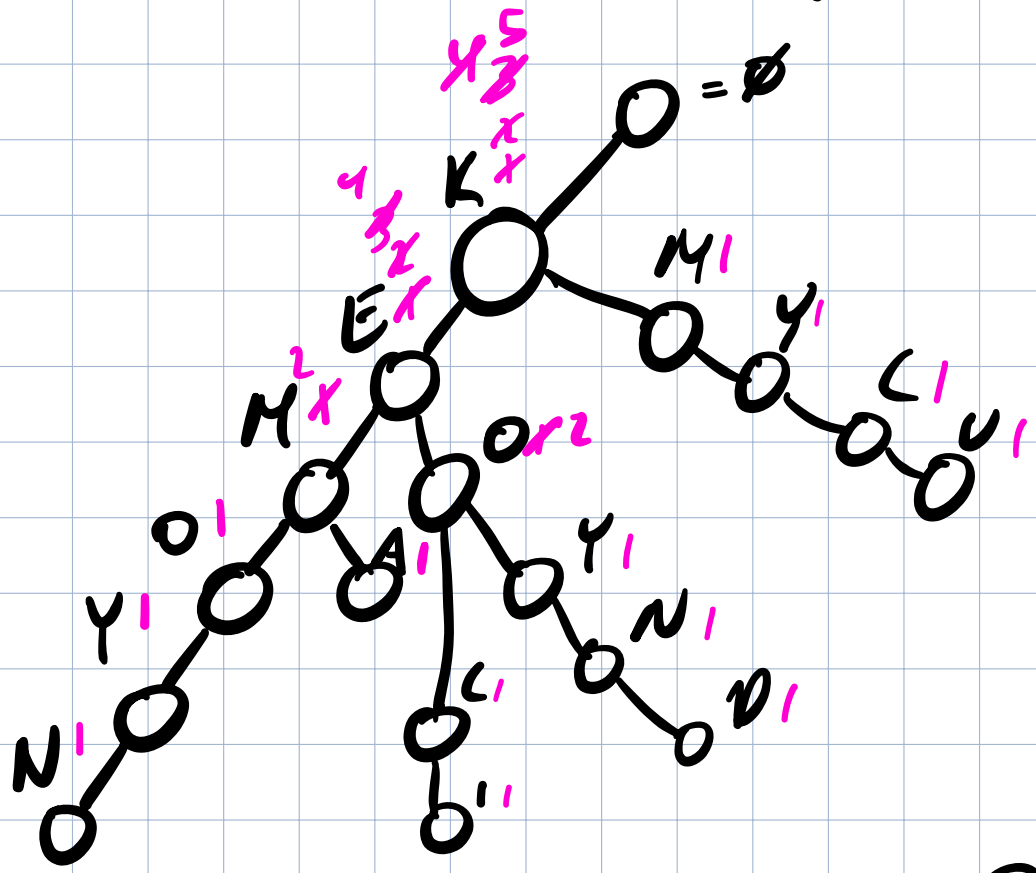
⋮

FP Tree

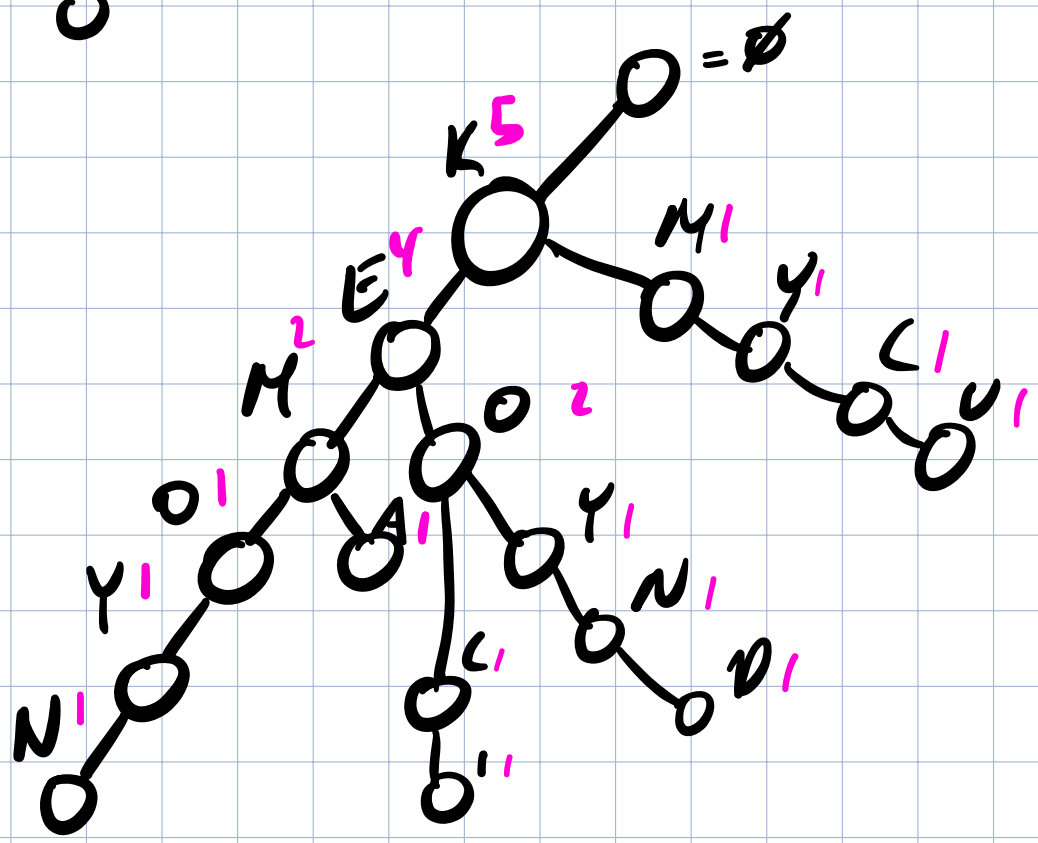
Sort C_1 by supp_ct $\forall$ items

=> KEMOYNCDAUI

~~DONKEY~~
~~MAKE~~
~~MUCH~~
COOKIE



=>



FPTree is easier to make.

3) Contingency table given

	hot dogs	$\overline{\textit{hot dogs}}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{\textit{hamburgers}}$	1000	1500	2500
Σ_{col}	3000	2000	5000

item $\in$ {hotdogs, hamb}

item refers to transactions containing item

a) Assoc Rule: hotdogs $\Rightarrow$ hamburgers.

min_sup = 25%, min_conf = 50%

Strong?

✓ $\frac{2000}{5000} = \frac{2}{5} \geq \frac{1}{4} = 25\% = \text{min\_sup}$

✓ $\frac{2000}{3000} = \frac{2}{3} \geq \frac{1}{2} = 50\% = \text{min\_conf}$

$\therefore$ The given Assoc Rule is strong w/ the provided min_sup & min_conf.

b) Calculate lift & Chi-Square.

Recall lift $= \dfrac{P(A \cup B)}{P(A) \cdot P(B)}$

$$P(\text{hot dogs}) = \frac{3000}{5000} = \frac{3}{5}$$

$$P(\text{hamburgers}) = \frac{2500}{5000} = \frac{1}{2}$$

$$P(\text{hot dogs, hamburgers}) = \frac{2000}{5000} = \frac{2}{5}$$

$$\text{lift} = \frac{\left(\frac{2}{5}\right)}{\left(\frac{3}{5}\right) \cdot \left(\frac{1}{2}\right)} = \frac{\left(\frac{2}{5}\right)}{\left(\frac{3}{10}\right)} = \left(\frac{2}{\cancel{5}}\right) \cdot \left(\frac{\cancel{10}^{\,2}}{3}\right) = \frac{4}{3} > 1$$

$\therefore$ hot dogs & hamburgers are positively corr.

Recall χ^2 formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

	hot dogs	$\overline{\textit{hot dogs}}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{\textit{hamburgers}}$	1000	1500	2500
Σ_{col}	3000	2000	5000

$$C_{11} = \frac{3000 \cdot 2500}{5000} = 1500$$

$$C_{12} = \frac{2000 \cdot 2500}{5000} = 1000$$

$$C_{21} = \frac{3000 \cdot 2500}{5000} = 1500$$

$$C_{22} = \frac{2000 \cdot 2500}{5000} = 1000$$

$$\chi^2 = \frac{(2000-1500)^2}{1500} + \frac{(500-1000)^2}{1000} + \frac{(1000-1500)^2}{1000} + \frac{(1500-1000)^2}{1000}$$

$$\Rightarrow \chi^2 = \frac{2500}{3} = 833.\overline{3}$$

Recall Degree of Freedom $= (\dim 1 - 1) \cdot (\dim 2 - 1)$

$\Rightarrow DF = 1.$

max prob. is 10.828

$\Rightarrow$ we reject the hypothesis that hot dogs & hamburgers are independent.

$\Rightarrow$ they're dependent.

c) What kind of corr. relationship exists b/n purchase of hot dogs & purchase of hamburgers.

They're positively corr. as found by lift & verified by χ^2 in part b).