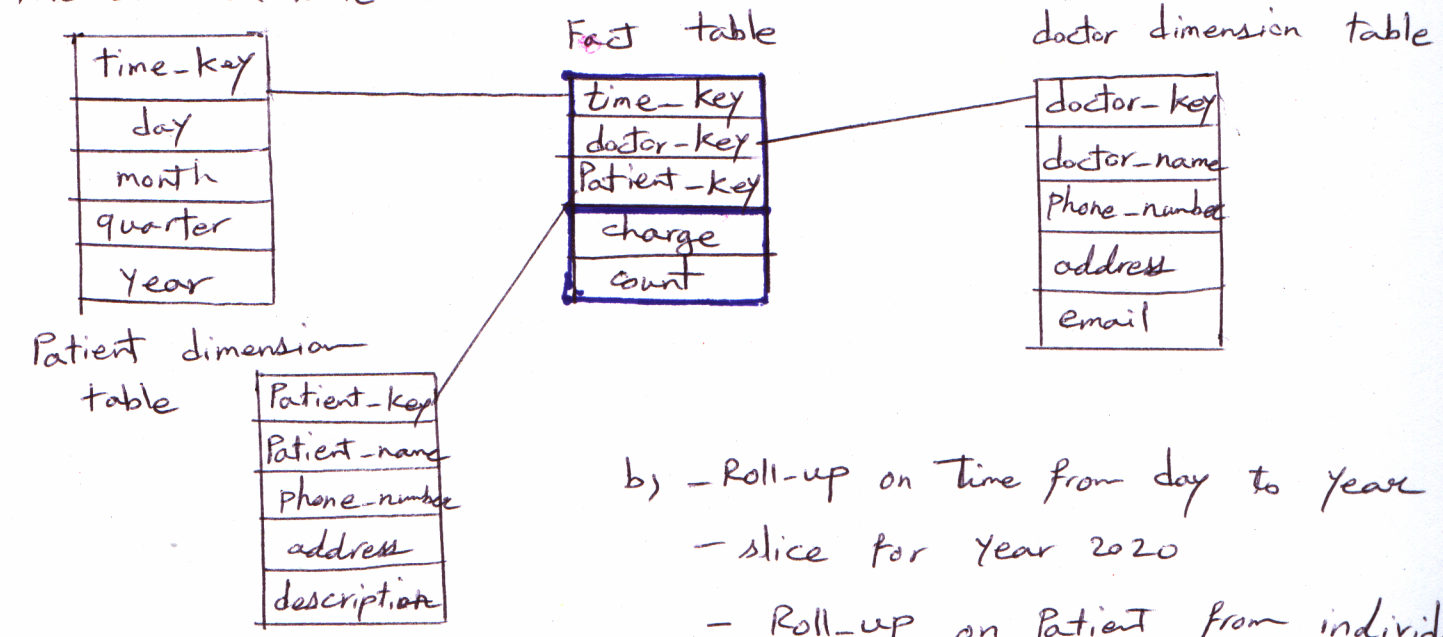
Answers - Assignment 3 Comp 541

1)

a) star schema diagram:



b) - Roll-up on time from day to year
- slice for year 2020
- Roll_up on Patient from individual Patient to all

---

2) min_sup = 0.6 x 5 = 3

a) Apriori Alg.

C_1

itemset	sup_count
{M}	3
{O}	3
{N}	2
{K}	5
{E}	4
{Y}	3
{D}	1
{A}	1
{U}	1

Scan D for count of each candidate →

Compare candidate support_count with minimum support_count →

L_1

itemset	
{K}	5
{E}	4
{O}	3
{M}	3
{Y}	3

C_2

itemset	sup_count
{K, E}	4
{K, O}	3
{K, M}	3
{K, Y}	3
{E, O}	3
{E, M}	2
{E, Y}	2
{O, M}	1
{O, Y}	2
{M, Y}	2

Generate C_2 candidates From L_1 → & scan D For count of each candidate

Compare candidate support count with min_support count →

L_2

itemset	sup_count
{K, E}	4
{K, O}	3
{K, M}	3
{K, Y}	3
{E, O}	3

C_3

itemset	sup_count
{K, E, O}	3

Generate C_3 candidates From L_2 → scan D for count of each candidate

Compare candidate support count with min_support count →

L_3

itemset	sup_count
{K, E, O}	3

$C_4 = \{\} \Rightarrow L_4 = \{\}$

Final result:

$\{\{K\}, \{E\}, \{O\}, \{M\}, \{Y\}, \{K, E\}, \{K, O\}, \{K, M\}, \{K, Y\}, \{E, O\}, \{K, E, O\}\}$

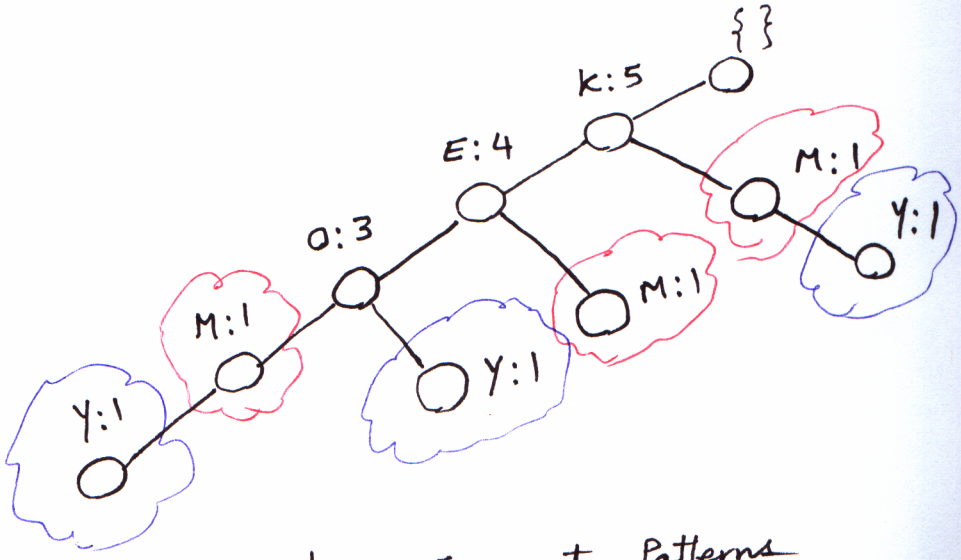
2) FP_tree First scan database & find frequent 1 itemset.

$L = \{ \{K:5\}, \{E:4\}, \{O:3\}, \{M:3\}, \{Y:3\} \}$

Then sort frequent items in frequency descending order

TID	items_bought
T100	{K, E, O, M, Y}
T200	{K, E, O, Y}
T300	{K, E, M}
T400	{K, M, Y}
T500	{K, E, O, O}

scan database again & construct FP_tree



item	Conditional Pattern Base (CPB)	Conditional FP_tree min_sup = 3	Frequent Patterns Generated (FPS)
Y	{K, E, O, M : 1} {K, E, O : 1} {K, M : 1}	{K : 3}	{K, Y : 3}
M	{K, E, O : 1} {K, E : 1} {K : 1}	{K : 3}	{K, M : 3}
O	{K, E : 3}	{K, E : 3}	{K, E, O : 3} {E, O : 3} {K, O : 3}
E	{K : 4}	{K : 4}	{K, E : 4}

FP_tree result:

{ {K:5}, {E:4}, {O:4}, {M:3}, {Y:3}, {K, Y:3},
{K, M:3}, {K, O:3}, {K, E:4}, {E, O:3}, {K, E, O:3} }

FP_growth is more efficient because it reduces the size of data set to be searched.

---

3) a)

$$\text{support} = \frac{2000}{5000} = 0.4 = 40\% > 25\%$$

$$\text{confidence} = \frac{2000}{3000} = 0.66 = 66.7\% > 50\%$$

Therefore association rule is strong

b)

$$\text{Lift} = \frac{P\{\text{hot dogs, hamburgers}\}}{P\{\text{hot dogs}\}\ P\{\text{hamburgers}\}} = \frac{0.4}{0.6 \times 0.5} = 1.3 > 1$$

they are positively correlated.

$$P(\{\text{hot dogs}\}) = \frac{3000}{5000} = 0.6$$

$$P(\{\text{hamburgers}\}) = \frac{2500}{5000} = 0.5$$

$$P(\{\text{hot dogs, hamburgers}\}) = \frac{2000}{5000} = 0.4$$

	hot dogs	$\overline{\text{hot dogs}}$	Σ row
hamburgers	2000 (1500)	500 (1000)	2500
$\overline{\text{hamburgers}}$	1000 (1500)	1500 (1000)	2500
Σ col	3000	2000	5000

2 x 2 table

Freedom = (2-1)(2-1)

Freedom = 1

$$c_{11} = \frac{\text{count}(\text{hot dogs}) \times \text{count}(\text{hamburgers})}{n} = \frac{3000 \times 2500}{5000} = 1500$$

$$c_{12} = \frac{\text{count}(\overline{\text{hot dogs}}) \times \text{count}(\text{hamburgers})}{n} = \frac{2000 \times 2500}{5000} = 1000$$

$$c_{21} = \frac{\text{count}(\text{hot dogs}) \times \text{count}(\overline{\text{hamburgers}})}{n} = \frac{3000 \times 2500}{5000} = 1500$$

$$c_{22} = \frac{\text{count}(\overline{\text{hot dogs}}) \times \text{count}(\overline{\text{hamburgers}})}{n} = \frac{2000 \times 2500}{5000} = 1000$$

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} =$$

$$\frac{(2000-1500)^2}{1500} + \frac{(500-1000)^2}{1000} + \frac{(1000-1500)^2}{1500} + \frac{(1500-1000)^2}{1000} = 833.33$$

$\chi^2 > 10.828 \Rightarrow$ They are strongly correlated.