Frequent Itemset

8/31

Let $\mathcal{I} = \{x_1, \ldots, x_n\}$ items (products in the grocery)

$X \subseteq \mathcal{I}$ is called an item set $\{milk, eggs\}$ ← 2-itemset

itemset of size k k-itemset

$\mathcal{I}^{(k)}$ set of all k-itemsets

$\mathcal{T} = \{t_1, \ldots, t_n\}$ a set of (transaction) ids tids (txn)

$T \subseteq \mathcal{T}$ is called tidset
(assumed sorted lex order)

Let t be a unique txn identifier
X be an itemset

A transaction (t, X)

A set of txs
{
$(abc123, \{milk, eggs\})$
$(abd1, \{carrots, peaches, bananas\})$
⋮
}

Database Rep

binary DB $D \subseteq T \times I$

tid $t \in T$

item $x \in I$

$(t,x) \in D \iff$ for some txn (t,X) $x \in X$

t contains item $x \iff (t,x) \in D$

E.g $I = \{A,B,C,D,E\}$
$T = \{1,2,3,4,5,6\}$

$(1, ABDE)$
$(2, BCE)$

Binary DB

	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1

For a set X let 2^X be the power set of X (set of all subsets of X)

Let $i: 2^T \to 2^I$

$T \subseteq T$

$i(T) = \{x \mid \forall t \in T,\ t \text{ contains } x\}$

x: all items in all txns in T

E.g $i(\{2,3,4\}) = \{B,E\}$

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

Sometimes write txnDB as tuples $(t, i(t))$

Running Example

binary data base

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

t	$\mathbf{i}(t)$
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

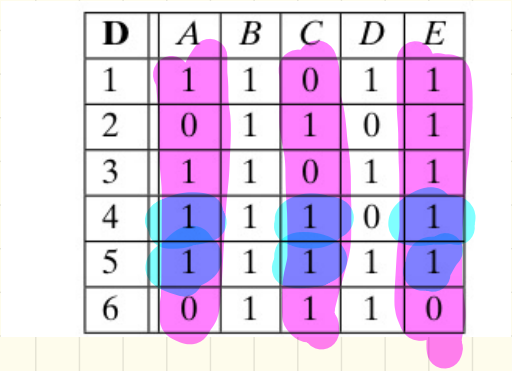
let $t: 2^{\mathcal{I}} \to 2^{\mathcal{T}}$

$X \subseteq \mathcal{I}$

$t(X) = \{t \mid t \in \mathcal{T} \text{ and } t \text{ contains } X\}$

↑ all of the tids containing all items of itemset

$t(\{A, C, E\}) = \{4, 5\}$



for item $x \in \mathcal{I}$

for all trns w/ item x

$t(x)$

Vertical DB

for item $x \in I$
tuples $(x, t(x))$

Binary DB

D	*A*	*B*	*C*	*D*	*E*
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

Vertical DB

x	*A*	*B*	*C*	*D*	*E*
$\mathbf{t}(x)$	1	1	2	1	1
	3	2	4	3	2
	4	3	5	5	3
	5	4	6	6	4
		5			5
		6			

Support of itemset X in D $sup(X, D)$ → $sup(X)$ when database is clear from context

$$sup(X, D) = \left| \{ t \mid (t, i(t)) \in D \text{ and } X \subseteq i(t) \} \right| = |t(X)|$$

↑ # of txns containing itemset X

relative support of X

$$rsup(X, D) = \frac{sup(X, D)}{|D|}$$

↳ $rsup(X)$ when D is clear from context

(est of joint prob of items containing X)

given user defined min support threshold (minsup)

itemset X **frequent** in D

if $sup(X, D) \geq minsup$ if $minsup \in \mathbb{Z}^+$

$rsup(X, D) \geq minsup$ if $minsup \in [0, 1)$

F is the set of all frequent itemsets

$F^{(k)}$ is the set of all frequent k-itemsets

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

t	$\mathbf{i}(t)$
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

minsup = 3

$F^{(1)}$ = ?

$F^{(2)}$ = ?

$F^{(5)}$ = ?

↖ think about sup(ABCDE)