تمرینات تحویلی داده کاوی - تکس جزایری ۹۸۱۳۰۰۶ ، امیرحسین رحیمی
۹۸۱۳۰۱۳

۲

a) $s(\{e\}) = \frac{8}{10}$, $s(\{b,d\}) = \frac{2}{10}$

$s(\{b,d,e\}) = \frac{2}{10}$

b) $c(b,d \rightarrow e) = \frac{s(b,d,e)}{s(b,d)} = \frac{2/10}{2/10} = 1$

$c(e \rightarrow b,d) = \frac{s(b,d,e)}{s(e)} = \frac{2/10}{8/10} = 1/4$

خیر متقارن نیست!

c)

CID	a	b	c	d	e	
1	1	1	1	1	1	$\rightarrow \{a,b,c,d,e\}$
2	1	1	1	1	1	$\rightarrow \{a,b,c,d,e\}$
3	0	1	1	1	1	$\rightarrow \{b,c,d,e\}$
4	1	1	1	1	0	$\rightarrow \{a,b,c,d\}$
5	1	1	0	1	1	$\rightarrow \{a,b,d,e\}$

$s(e) = 4/5$

$s(b,d) = 1$

$s(b,d,e) = 4/5$

d) $c(b,d \rightarrow e) = \frac{s(b,d,e)}{s(b,d)} = \frac{4/5}{1} = 4/5$

$c(e \rightarrow b,d) = \frac{s(b,d,e)}{s(e)} = \frac{4/5}{4/5} = 1$

e) $r := X \rightarrow Y$ $\quad Z = X \cup Y$

اگر T_{TID} یک itemset مکرر
در نظر گرفتن TID به عنوان سبد باشد و T_{CID} یک itemset مکرر در نظر گرفتن
CID به عنوان سبد باشد. اگر $Z \subseteq T_{TID}$ آنگاه Z زیرمجموعه CID نظیر خریدار
T_{TID} خواهد بود.

$$S_1 = \frac{\sigma_{TID}(Z)}{10} \quad و \quad S_2 = \frac{\sigma_{CID}(Z)}{5}$$

حال فرض کنید Z در c_1 تراکنش از مشتری ۱، c_2 تراکنش از مشتری ۲ و ...

پس در این صورت

$$\sigma_{TID}(Z) = c_1 + c_2 + c_3 + c_4 + c_5$$

$$\sigma_{CID}(Z) = \frac{c_1}{c_1} + \frac{c_2}{c_2} + \frac{c_3}{c_3} + \frac{c_4}{c_4} + \frac{c_5}{c_5}$$

که در عبارت بالا $\frac{0}{0}$ را ۰ تعریف می‌کنیم.

(4) هر یک از item های {Milk, Beer, Diapers, Butter, Bread, Cookies}

$$R = 3^6 - 2^7 + 1$$

b) maximum size of itemsets is 4 since the sixth transaction has 4 items and the sixth transaction itself has the suppot greater than zero.

c) $\binom{6}{3} = \frac{6!}{3!\,3!} = 20$

d) we find an itemset of size 2.

we have $S(\text{Bread}, \text{Butter}) = 5/10$

now we prune 1-itemsets with suppot less than 0.5 which are Beer and Cookies.

Now we are left with Bread, Butter, Diapes, milk.

But $S(\text{Diapers}) = 0.7$. Hence we have to rule out the possibility of support $= 0.6$ or 0.7 to prove {Bread, Butter} is indeed the largest itemset with given conditions.

Let's prune the transaction with omitted items

We have:

1 Milk, Diapers
2 Bread, Butter, Milk
3 Milk, Diapers
4 Bread, Butter
5 Diapers
6 Milk, Diapers, Bread, Butter
7 Bread, Butter, Diapers
8 Diapers
9 Milk, Diapers, Butter, Bread

Since the third and fourth transactions are disjoint and the fifth and eigth transactions are single items the maximum support could be at most 0.6. Also the first and second transaction have a joint set of size at most 1

So 0.6 is ruled out and {Bread, Butter} is the best.

d) $c(a \to b) = \dfrac{s(a,b)}{s(a)} = \dfrac{s(a,b)}{s(b)} = c(b \to a)$

Hence $s(a) = s(b)$. Let a = Beer and b = Cookies and we are done since $s(\text{Beer}) = s(\text{Cookies}) = 0.4$

a) Let $X \subseteq Y$ and Y is frequent. Namely

$$\frac{\sigma(Y)}{N} > minsup.$$ Sine $X \subseteq Y$, $\sigma(X) \geq \sigma(Y)$.

Hence $$s(X) = \frac{\sigma(X)}{N} \geq \frac{\sigma(Y)}{N} > minsup.$$ So X is frequent

b) Proof by contradiction. Assume X is infrequent or in other words $\frac{\sigma(X)}{N} < minsup$ and $X \subseteq Y$.

Also assume Y is frequent. By the previous proposition (a) X is subset of a frequent set and should be frequent. Contradiction! Thus Y has to be infrequent.

Tracing Apriori: $\boxed{minsup = 3}$

$F_1 = \{a\}, \{b\}, \{c\}$

$C_2 =$ cand-gen(F_1) $\quad$ $F_{k-1} \times F_1$ method
lexicographic sort

$= \{a,b\}, \{a,c\}, \{b,c\}$

pruning does not chang anything since subsets of each candidate are frequent

$F_2 = \{a,b\}, \{a,c\}, \{b,c\}$ $\quad$ since $\sigma(\cdot) \geq 3$

$C_3 = \{a,b,c\}$ $\quad$ $F_2 \times F_1$ method

pruning no change $\longrightarrow$ all subsets are frequent

$F_3 = \{a,b,c\}$ $\quad$ $F_3 \times F_1$

$C_4 = \emptyset$ $\quad$ we are done!