Apriori-Algorithm

Support, confidence, Lift, Conviction

$$Support(X) = \frac{Number\ of\ transactions\ in\ which\ X\ appears}{Total\ number\ of\ transactions}$$

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} = P(Y|X)$$

$$Lift(X \rightarrow Y) = \frac{Support(X \cap Y)}{Support(X) * Support(Y)}$$

$$Coviction(X \rightarrow Y) = \frac{1 - Support(Y)}{1 - Confidence(X \rightarrow Y)}$$

If conviction is greater than 1, then this metric shows that incorrect predictions (X⇒Y) occur less often than if these two actions were independent. This metric can be viewed as the ratio that the association rule would be incorrect if the actions were independent (i.e. a conviction of 1.5 indicates that if the variables were independent, this rule would be incorrect 50% more often.)

Step 1. Set the support threshold and create the table for all items

Transactions	Onion	Potato	Burger	Milk	Beer
t_1	1	1	1	0	0
t_2	0	1	1	1	0
t_3	0	0	0	1	1
t_4	1	1	0	1	0
t_5	1	1	1	0	1
t_6	1	1	1	1	1

For this situation, we need to create

Item	Frequency of transactions
Onion	4
Potato	5
Burger	4
Milk	4
Beer	2

In this situation we set support threshold to 50%, which means we only consider items with frequency more than 3 (6*50%), so we get the following table

Item	Frequency of transactions
Onion	4
Potato	5

Burger	4
Milk	4

Step2. Get the table for pair relationships (2-itemset) and filter by threshold

In this case, order does not matter, which means AB=BA. In this situation we have

$$\frac{n!}{r!\,(n-r)!} = \frac{4!}{2!\,(4-2)!} = 6\ pairs$$

Item	Frequency of transactions
Onion & Potato	4
Onion & Burger	3
Onion & Milk	2
Potato & Burger	4
Potato & Milk	3
Burger & Milk	2

Also, we eliminate the transactions with frequency lower than 3, we get the following table

Item	Frequency of transactions
Onion & Potato	4
Onion & Burger	3
Potato & Burger	4
Potato & Milk	3

Step3. Repeat step2 to get the 3-itemset tables

Item	Frequency of transactions
Onion & Potato& Burger	3
Onion & Burger & Milk	2

Filter by threshold

Item	Frequency of transactions
Onion & Potato& Burger	3

Step4. Generate the association rule

For example

$$Support_{\text{Onion \& Potato}} = \frac{4}{6} = 66.67\%$$

$$Confidence_{\text{Onion}\rightarrow\text{ Potato}} = \frac{4}{4} = 1$$

$$Lift_{\text{Onion}\rightarrow\text{ Potato}} = \frac{\frac{4}{6}}{\frac{4}{6} * \frac{5}{6}} = \frac{6}{5} = 1.2$$

Transactions	Onion	Potato	Burger	Milk	Beer
t_1	1	1	1	0	0
t_2	0	1	1	1	0
t_3	0	0	0	1	1
t_4	1	1	0	1	0
t_5	1	1	1	0	1
t_6	1	1	1	1	1

Personal thinking

The lift value would be enlarged by irrelevant records to the rule, so it is hard to measure it's confidence. For example, if we want to analysis the rule $\{\text{Onion} \rightarrow \text{Potato}\}$, their original lift value is 1.2, but after I add irrelevant transactions like below

Transactions	Onion	Potato	Burger	Milk	Beer
t_1	1	1	1	0	0
t_2	0	1	1	1	0
t_3	0	0	0	1	1
t_4	1	1	0	1	0
t_5	1	1	1	0	1
t_6	1	1	1	1	1
t_7	0	0	1	1	0
t_8	0	0	0	0	1
t_9	0	0	0	1	1
t_{10}	0	0	1	1	0

$$Lift = \frac{\frac{4}{10}}{\frac{4}{10} * \frac{5}{10}} = 2$$

1096 * 3 = 3288

This problem shows in many traffic accident-related research. They set a very small support threshold to get ton of rules with great metrics. But for their result, the lift value is not reliable.