Introduction to Itemset Mining

CSCI 347 – Data Mining

Adiesha Liyana Ralalage

FREQUENT ITEMSET MINING

- In many applications one is interested in how often two or more objects of interest co-occur, the so-called itemset.
- The prototypical application was market basket analysis, that is, to mine the sets of items that are frequently bought together at a supermarket by analyzing the customer shopping carts (the so-called "market baskets").
- Once we mine the frequent sets, they allow us to extract association rules among the itemset, where we make some statement about how likely are two sets of items to co-occur or to conditionally occur.

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, Milk, Rice, Toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, Milk, Rice, Toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

How can we find all sets of items that are frequently purchased together?

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, Milk, Rice, Toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

How can we find all sets of items that are frequently purchased together?

For example: Which sets of items are purchased at least 30% of the time?

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, Milk, Rice, Toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

For example: Which sets of items are purchased at least 30% of the time?

Brute-force approach: count the number of times each item, pair of items, triple of items, etc... appears, then report those that appear 3 or more times

Notation

- Let $\mathcal{I} = \{x_1, x_2, \dots, x_m\}$ be a set of elements called items
- Let $X \subset \mathcal{I}$ is called an itemset.
  - For example, $\mathcal{I}$ could be a set of items sold at a supermarket
  - X could be a set of items that was sold.
- Itemset of cardinality k is called a k-itemset.
- We denote the $\mathcal{I}^{(k)}$ the set of all k-itemsets, i.e., subsets of $\mathcal{I}$ with size k.
- Let $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ be another set of elements called transaction identifiers or $tids$.
- We can assume that itemsets and tids are stored in lexicographical order.
- A transaction is a tuple of the form $\langle t, X \rangle$ where $t \in \mathcal{T}$ is a unique transaction identifier.
- $\mathcal{T}$ could represent set of all customers at a supermarket.

MONTANA STATE UNIVERSITY

Notation

- Databases: There are different ways to represent this data

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

(a) Binary Database

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

(b) Transaction Database

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			

(c) Vertical Database

Notation

- Support and frequent itemsets
- The support of an itemset X in a dataset D, denoted as $Sup(X, D)$ is the number of transactions in D that contains X.

$$Sup(X, D) = |\{t \mid \langle t, i(t)\rangle \in D \; and \; X \subseteq i(t)\}| = |t(X)|$$

$$i(T) = \{x \mid \forall t \in T, t \; conains \; x\}$$

is the set of items that are common to all the transactions in the tidset $T \subseteq \mathcal{T}$

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

is the set of tids that contain all the items in the itemset X.

- An itemset X is said to be frequent in D if $Sup(X, D) \geq \mathrm{minsup}$ where $minsup$ is some user defined minimum support threshold.

Notation

- $relative\ support$ of X is the fraction of transactions that contain X.
- $rsup(X, D) = \frac{Sup(X,D)}{|D|}$

Frequent itemsets of minsup = 3

$\mathbf{D}$	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

(a) Binary Database

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

(b) Transaction Database

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

(c) Vertical Database

sup	itemsets
6	B
5	E, BE
4	$A, C, D, AB, AE, BC, BD, ABE$
3	$AD, CE, DE, ABD, ADE, BCE, BDE, ABDE$

Table 8.1: Frequent Itemsets with $minsup = 3$

$$\mathcal{F}^{(1)} = \{A, B, C, D, E\}$$
$$\mathcal{F}^{(2)} = \{AB, AD, AE, BC, BD, BE, CE, DE\}$$
$$\mathcal{F}^{(3)} = \{ABD, ABE, ADE, BCE, BDE\}$$
$$\mathcal{F}^{(4)} = \{ABDE\}$$

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, Milk, Rice, Toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

For example: Which sets of items are purchased at least 30% of the time?

Brute-force approach: count the number of times each item, pair of items, triple of items, etc… appears, then report those that appear 3 or more times

How long would this take?

How about a Brute-Force algorithm to solve this?

Algorithm 8.1: Algorithm BruteForce

BruteForce $(\mathbf{D}, \mathcal{I}, minsup)$:

$\mathcal{F} \leftarrow \emptyset$ `// set of frequent itemsets`
foreach $X \subseteq \mathcal{I}$ **do**
  $sup(X) \leftarrow$ ComputeSupport $(X, \mathbf{D})$
  if $sup(X) \geq minsup$ **then**
    $\mathcal{F} \leftarrow \mathcal{F} \cup \{(X, sup(X))\}$
return $\mathcal{F}$

ComputeSupport $(X, \mathbf{D})$:

$sup(X) \leftarrow 0$
foreach $\langle t, \mathbf{i}(t) \rangle \in \mathbf{D}$ **do**
  if $X \subseteq \mathbf{i}(t)$ **then**
    $sup(X) \leftarrow sup(X) + 1$
return $sup(X)$

How about a Brute-Force algorithm to solve this?

What about the time complexity of this algorithm?

- It's exponential in the size of the items
- If $\mathcal{I} = \{A, B, C, D, E\}$, we must look at $2^5 = 32$ combinations.
- $O\left(|\mathcal{I}||D|2^{|\mathcal{I}|}\right)$

Algorithm 8.1: Algorithm BruteForce

```
BruteForce (D, 𝓘, minsup):
𝓕 ← ∅ // set of frequent itemsets
foreach X ⊆ 𝓘 do
    sup(X) ← ComputeSupport (X, D)
    if sup(X) ≥ minsup then
        𝓕 ← 𝓕 ∪ {(X, sup(X))}
return 𝓕

ComputeSupport (X, D):
sup(X) ← 0
foreach ⟨t, i(t)⟩ ∈ D do
    if X ⊆ i(t) then
        sup(X) ← sup(X) + 1
return sup(X)
```

Apriori Algorithm idea

- When using brute force approach, we look at lot of combinations that are not useful.
- Let $X, Y \subseteq \mathcal{I}$ be any two itemsets.
- If $X \subseteq Y \;\; \rightarrow Sup(X) \geq Sup(Y)$
  - If Y is frequent, then any subset $X \subseteq Y$ must also be frequent.
  - If X is not frequent, then any superset $Y \supseteq X$ cannot be frequent.
- Apriori algorithm uses these two properties to improve the brute force algorithm.

ITEMSET MINING/MARKET BASKET ANALYSIS

- Suppose we observe the following transactions in a supermarket:

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, Milk, Rice, Toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

For example: Which sets of items are purchased at least 30% of the time?

A-Priori approach: count the number of frequent items, use those to generate frequent pairs, use those to generate frequent triplets, etc.

This will eliminate computing frequency of sets that have no chance of being frequent.

APRIORI Algorithm

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper, beans, rice, milk, baby wipes, diapers
2	Oat milk, beans, toilet paper, orange juice
3	Oat milk, milk, orange juice, toilet paper
4	Beans, toilet paper, baby wipes, diapers
5	Toilet paper, butter, baby wipes, diapers
6	Milk, toilet paper
7	Milk, rice
8	Beans, Milk, Rice, Toilet paper
9	Milk, butter, diapers
10	Beans, rice, toilet paper

→

Candidate set	Support
{Baby Wipes}	3
{Beans}	5
{Butter}	2
{Diapers}	4
{Milk}	6
{Oat Milk}	2
{Orange Juice}	2
{Rice}	4
{Toilet Paper}	8

Start frequent item sets of size $k = 1$

APRIORI Algorithm

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

→

Candidate set	Support
{1}	3
{2}	5
{3}	2
{4}	4
{5}	6
{6}	2
{7}	2
{8}	4
{9}	8

Assign an ID to each item

APRIORI Algorithm

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Candidate set	Support
{1}	3
{2}	5
{3}	2
{4}	4
{5}	6
{6}	2
{7}	2
{8}	4
{9}	8

Select frequent itemsets (those that appear 3 times or more)

APRIORI Algorithm

Frequent item sets of size 1: $\{1\},\{2\},\{4\},\{5\},\{8\},\{9\}$

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

→

Candidate set	Support
{1,2}	2
{1,4}	3
{1,5}	1
{1,8}	1
{1,9}	3
{2,4}	2
{2,5}	2
{2,8}	3
{2,9}	5
{4,5}	2
{4,8}	1
{4,9}	3
{5,8}	3
{5,9}	4
{8,9}	3

Generate new candidates of size $k + 1$

MONTANA STATE UNIVERSITY

APRIORI Algorithm

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Select frequent itemsets (those that appear 3 times or more)

Frequent item sets of size 2:
$\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Candidate set	Support
{1,2}	2
{1,4}	3
{1,5}	1
{1,8}	1
{1,9}	3
{2,4}	2
{2,5}	2
{2,8}	3
{2,9}	5
{4,5}	2
{4,8}	1
{4,9}	3
{5,8}	3
{5,9}	4
{8,9}	3

APRIORI Algorithm

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1{,}4\}, \{1{,}9\}, \{2{,}8\}, \{2{,}9\}, \{4{,}9\}, \{5{,}8\}, \{5{,}9\}, \{8{,}9\}$

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Candidate set	Support

Generate new candidates of size $k + 1$

APRIORI Algorithm

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

→

Candidate set	Support
{1,4,9}	3
{2,8,9}	3
{5,8,9}	2
{2,4,9}	2
{2,5,8}	2
{4,5,9}	2

Generate new candidates of size $k + 1$

APRIORI Algorithm

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 3: $\{1,4,9\}, \{2,8,9\},$

→

Candidate set	Support
{1,4,9}	3
{2,8,9}	3
{5,8,9}	2
{2,4,9}	2
{2,5,8}	2
{4,5,9}	2

Select frequent itemsets (those that appear 3 times or more)

APRIORI Algorithm

- Which sets of items are purchased at least 30% of the time? → $\text{"minsup"} = 3$

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

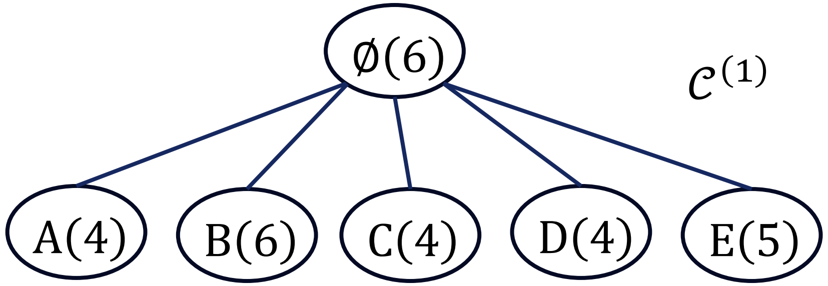
Frequent item sets of size 2:
$\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Frequent item sets of size 3: $\{1,4,9\}, \{2,8,9\}$

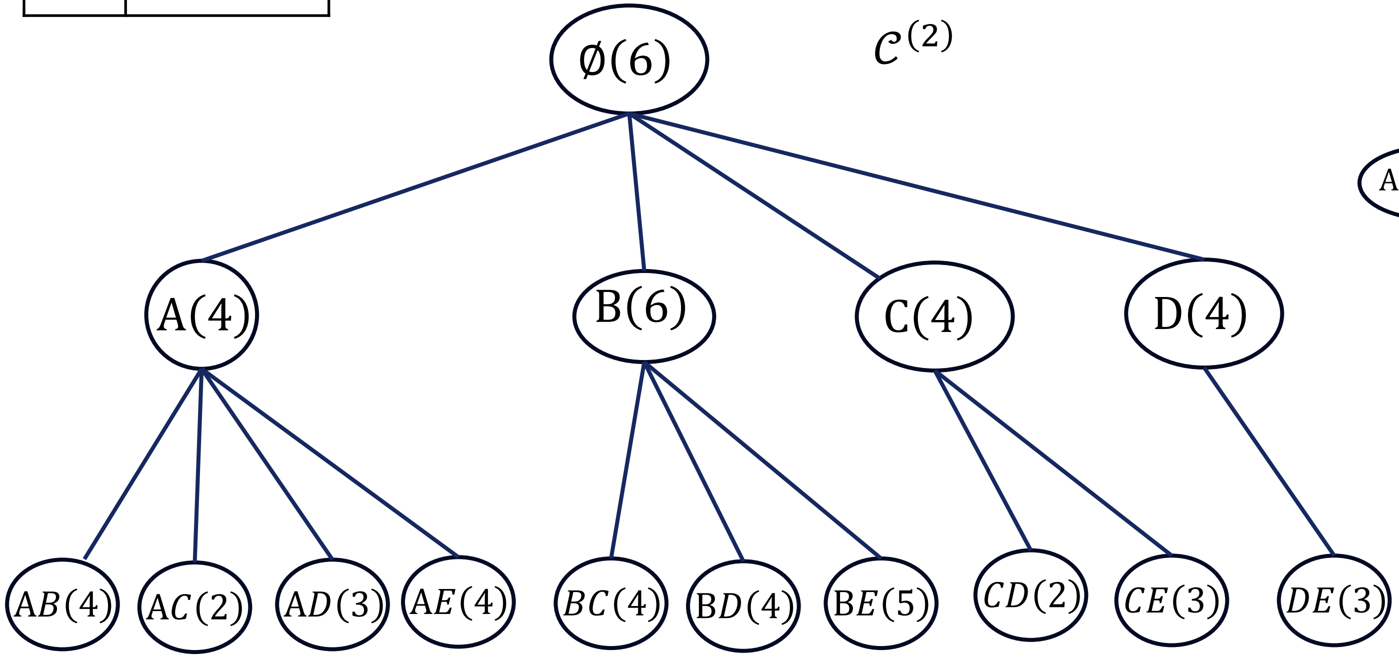
At this point, no more candidates can be generated - we've found all frequent item sets

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

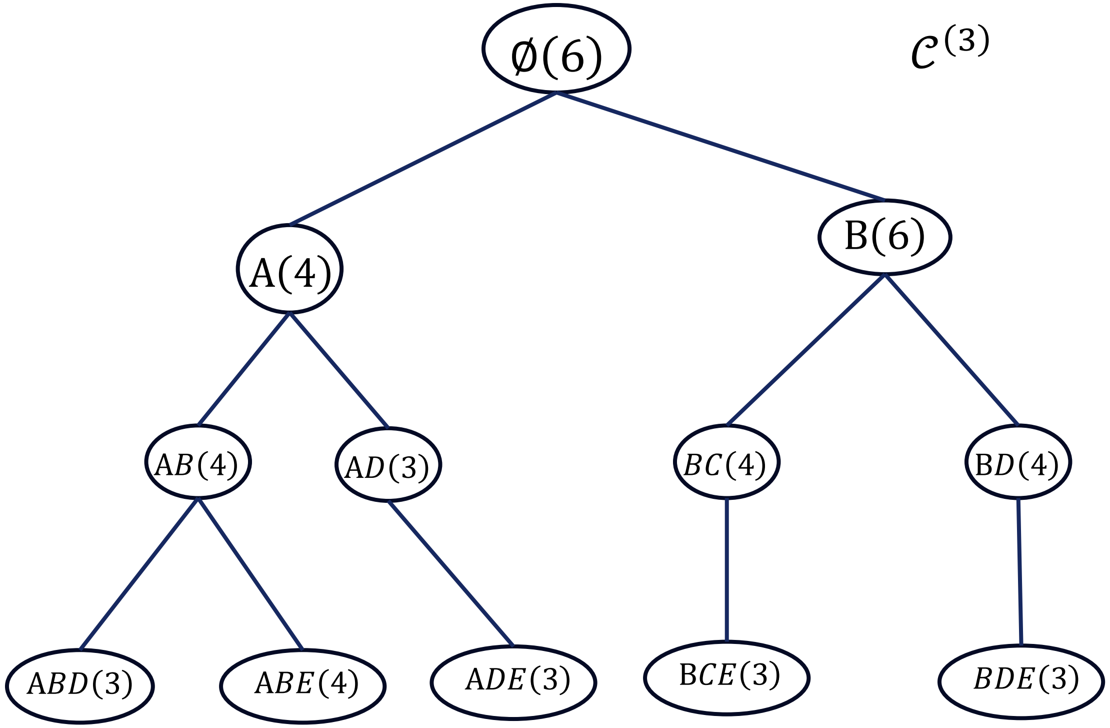
$\mathcal{C}^{(1)}$



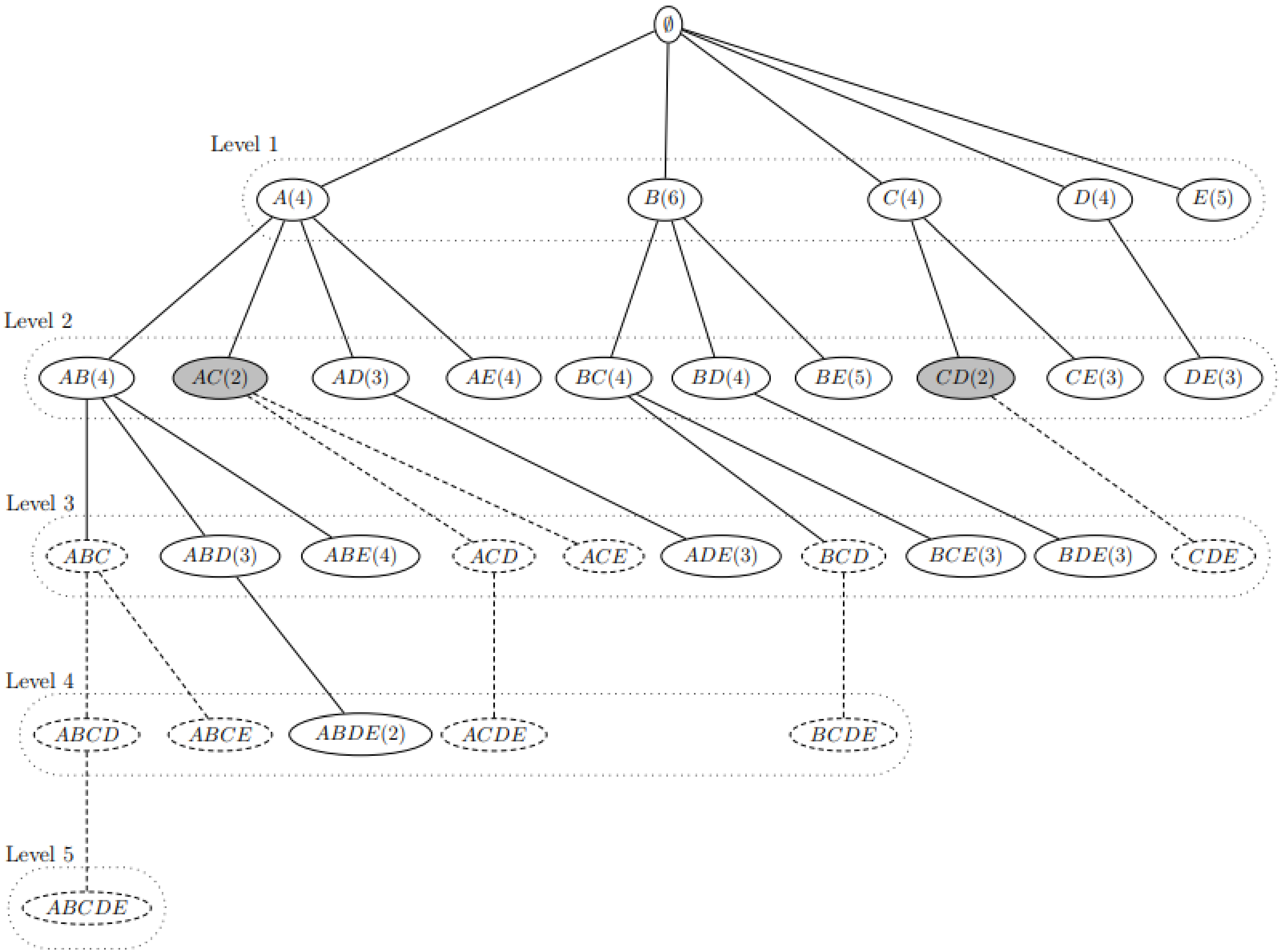
$\mathcal{C}^{(2)}$



$\mathcal{C}^{(3)}$



$\mathcal{C}^{(k)}$ denotes the prefix tree comprising all the candidate k-itemsets


∅
Level 1
A(4)
B(6)
C(4)
D(4)
E(5)
Level 2
AB(4)
AC(2)
AD(3)
AE(4)
BC(4)
BD(4)
BE(5)
CD(2)
CE(3)
DE(3)
Level 3
ABC
ABD(3)
ABE(4)
ACD
ACE
ADE(3)
BCD
BCE(3)
BDE(3)
CDE
Level 4
ABCD
ABCE
ABDE(2)
ACDE
BCDE
Level 5
ABCDE

Apriori Algorithm

$APriori(D, I, minsup)$

1. $\mathcal{F} \leftarrow \emptyset$
2. $\mathcal{C}^{(1)} = \{\emptyset\}$
3. $For\ each\ i \in I\ do$:
   - Add i as a child of $\emptyset$ in $\mathcal{C}^{(1)}$ with $\sup(i) \leftarrow 0$
4. $k \leftarrow 1$
5. While $\mathcal{C}^{(k)} = \emptyset$ do:
   - $ComputeSupport(\mathcal{C}^{(k)}, D)$
   - $For\ each\ leaf\ X \in \mathcal{C}^{(k)}\ do$:
     - $If\ \sup(X) \geq minsup\ then\ \mathcal{F} \leftarrow \mathcal{F} \cup \{(X, \sup(X))\}$
     - $Else\ remove\ X\ from\ \mathcal{C}^{(k)}$
   - $\mathcal{C}^{(k+1)} \leftarrow ExtendPrefixTree\left(\mathcal{C}^{(k)}\right)$
   - $k \leftarrow k + 1$
6. $Return\ \mathcal{F}^{(k)}$

Apriori Algorithm

$ComputeSupport\left(\mathcal{C}^{(k)}, D\right)$:

$For\ each\ leaf\ \langle t, i(t)\rangle \in D\ do$:

$For\ each\ k\_subset\ X \subseteq i(t)\ do$:

$if\ X \in \mathcal{C}^{(k)}\ then \sup(X) \leftarrow \sup(X) + 1$

- $ExtendPredixTree\left(\mathcal{C}^{(k)}\right)$:
  - $Foreach\ leaf\ X_b \in \mathcal{C}^{(k)}$ do:
    - $Foreach\ leaf\ X_b \in SIBLING(X_a), such\ that\ b > a\ do$:
      - $X_{ab} \leftarrow X_a \cup X_b$
      - //Prune candidates if there are any infrequent subsets
      - $if\ X_j \in \mathcal{C}^{(k)}, \text{forall } \mathrm{X_j} \subset X_{ab}, such\ that\ \left|X_j\right| = |X_{ab}| - 1\ then$
        - $Add\ X_{ab}\ as\ child\ of\ X_a\ with \sup(X_{ab}) \leftarrow 0$
    - $If\ no\ extensions\ from\ X_a\ then\ remove\ X_a\ from\ \mathcal{C}^{(k)}$
  - $Return\ \mathcal{C}^{(k)}$

Association Rules

- An $association\ rule$ is an expression $X \xrightarrow{s,c} Y$, where X and Y are itemsets and they are disjoint, i.e., $X, Y \subseteq \mathcal{I}, and\ X \cap Y = \emptyset$.
- Let $X \cup Y$ be denoted as XY.
- Support for the rule is the number of transactions in which both X and Y co-occur as subsets.
  - $s = Sup(X \rightarrow Y) = |t(XY)| = Sup(XY)$
- $relative\ support$ is defined as the fraction of transactions where X and Y co-occur.
  - $rsup(X \rightarrow Y) = \frac{Sup(XY)}{|D|} = P(X \wedge Y)$
- Confidence of the rule is the conditional probability that transaction contains Y given that it contains X.
  - $c = Conf(X \rightarrow Y) = P(Y|X) = \frac{P(X \wedge Y)}{P(X)} = \frac{Sup(XY)}{Sup(X)}$

Association Rules

- Support for the rule is the number of transactions in which both X and Y co-occur as subsets.
  - $s = Sup(X \rightarrow Y) = |t(XY)| = Sup(XY)$
- $relative\ support$ is defined as the fraction of transactions where X and Y co-occur.
  - $rsup(X \rightarrow Y) = \frac{Sup(XY)}{|D|} = P(X \wedge Y)$
- Confidence of the rule is the conditional probability that transaction contains Y given that it contains X.
  - $c = Conf(X \rightarrow Y) = P(Y|X) = \frac{P(X \wedge Y)}{P(X)} = \frac{Sup(XY)}{Sup(X)}$
- $lift$ is defined as the ratio of the observed joint probability of X and Y to the expected joint probability if they were statistically independent
  - $lift(X \rightarrow Y) = \frac{P(XY)}{P(X)P(Y)} = \frac{conf(X \rightarrow Y)}{rsup(Y)}$
  - Lift 1 means No association, >1 means positive association and <1 means negative association.

Association rule mining

- What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Frequent item sets of size 3: $\{1,4,9\}, \{2,8,9\}$

Association rule mining

- What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Frequent item sets of size 3: $\{1,4,9\}, \{2,8,9\}$

Example: $\{2, 8\} \rightarrow \{9\}$

Association rule mining

- What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Frequent item sets of size 3: $\{1,4,9\}$, $\{2,8,9\}$

Example: $\{2, 8\} \rightarrow \{9\}$

$\{2, 8\}$ has **support** 3 and $\{2,8,9\}$ has **support** 3

Association rule mining

- What rules can we generate of the form $X \to Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Frequent item sets of size 3: $\{1,4,9\}$, $\{2,8,9\}$

Example: $\{2, 8\} \to \{9\}$

$\{2, 8\}$ has support 3 and $\{2,8,9\}$ has support 3

So, we say that the rule $\{2, 8\} \to \{9\}$ has support 3 and confidence $\frac{3}{3} = 1$

$$Sup(\{2, 8\} \to \{9\}) = Sup(\{2,8,9\}) = 3$$

$$conf(\{2, 8\} \to \{9\}) = \frac{Sup(\{2,8,9\})}{Sup(\{2,8\})} = \frac{3}{3}$$

$$lift(\{2, 8\} \to \{9\}) = \frac{conf(\{2, 8\} \to \{9\})}{rsup(\{9\})} = \frac{1}{\left(\frac{8}{10}\right)} = \frac{10}{8}$$

Association rule mining

- What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Frequent item sets of size 3: $\{1,4,9\}$, $\{2,8,9\}$

Example: $\{9\} \rightarrow \{2,8\}$

$\{9\}$ has support 8 and $\{2,8,9\}$ has support 3

So, we say that the rule $\{9\} \rightarrow \{2,8,9\}$ has support 3 and confidence $\frac{3}{8} = 0.375$

$$Sup(\{9\} \rightarrow \{2,8\}) = Sup(\{2,8,9\}) = 3$$

$$conf(\{9\} \rightarrow \{2,8\}) = \frac{Sup(\{2,8,9\})}{Sup(\{9\})} = \frac{3}{8}$$

$$lift(\{9\} \rightarrow \{2,8\}) = \frac{conf(\{9\} \rightarrow \{2,8\})}{rsup(\{2,8\})} = \frac{\frac{3}{8}}{\left(\frac{3}{10}\right)} = \frac{10}{8}$$

Association rule mining

- What rules can we generate of the form $X \rightarrow Y$, where X and Y are itemsets, with enough support and enough confidence?

Transaction ID	Items
1	Toilet paper(9), beans(2), rice(8), milk(5), baby wipes(1), diapers(4)
2	Oat milk(6), beans(2), toilet paper(9), orange juice(7)
3	Oat milk(6), milk(5), orange juice(7), toilet paper(9)
4	Beans(2), Toilet paper(9), baby wipes(1), diapers(4)
5	Toilet paper(9), butter(3), baby wipes(1), diapers(4)
6	Milk(5), Toilet paper(9)
7	Milk(5), rice(8)
8	Beans(2), Milk(5), Rice(8), Toilet Paper(9)
9	Milk(5), butter(3), diapers(4)
10	Beans(2), rice(8), Toilet paper(9)

Frequent item sets of size 1: $\{1\}, \{2\}, \{4\}, \{5\}, \{8\}, \{9\}$

Frequent item sets of size 2: $\{1,4\}, \{1,9\}, \{2,8\}, \{2,9\}, \{4,9\}, \{5,8\}, \{5,9\}, \{8,9\}$

Frequent item sets of size 3: $\{1,4,9\}, \{2,8,9\}$

Example: $\{4\} \rightarrow \{1\}$

$\{4\}$ has support 4 and $\{1,4\}$ has support 3

So, we say that the rule $\{4\} \rightarrow \{1\}$ has support 3 and confidence $\frac{3}{4} = 0.75$

$$Sup(\{4\} \rightarrow \{1\}) = Sup(\{1,4\}) = 3$$

$$conf(\{4\} \rightarrow \{1\}) = \frac{Sup(\{1,4\})}{Sup(\{4\})} = \frac{3}{4}$$

$$lift(\{4\} \rightarrow \{1\}) = \frac{conf(\{4\} \rightarrow \{1\})}{rsup(\{1\})} = \frac{\frac{3}{4}}{\left(\frac{3}{10}\right)} = \frac{10}{4}$$