

# SS-ZG548: ADVANCED DATA MINING

# 02

## Incremental Data Mining



**Dr. Kamlesh Tiwari**

Assistant Professor, Department of CSIS,  
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

July 31, 2021

ONLINE

(WILP @ BITS-Pilani July-Dec 2021)

<http://ktiwari.in/adm>

# Recap: Data Mining

**Data mining** does KDD: knowledge discovery in databases

KDD involves **five** components 1) collection of data, 2) preprocessing, 3) transformation, 4) **data mining**, and 5) interpretation.

- Data mining has **three** parts **Model**, **Preference**, and **Search**
- **Two** broad categories 1) **Predictive** if we focus on new data involving classification, regression, time series analysis, and prediction 2) **Descriptive** when we want to understand/describe the data itself involving clustering, summarization, association rules, or sequence discovery

## Issues

Human interaction, Overfitting, Outliers, Large dataset, High dimension, Multimedia data, missing data, irrelevant data, noisy data, changing data.

# Association Rules

Association rules tries to do linked analysis.

**Example:** Whether same products are selling together?

- $I = \{i_1, i_2, i_3, \dots, i_m\}$ ,  $T = \{t_1, t_2, t_3, \dots, t_n\}$  and  $t_j \subseteq I$

---

<sup>1</sup>Mining association rules between sets of items in large databases, *R Agrawal, T Imielinski, and A Swami*, SIGMOD, 22(2), pp 207–216, ACM-1993

# Association Rules

Association rules tries to do linked analysis.

**Example:** Whether same products are selling together?

- $I = \{i_1, i_2, i_3, \dots, i_m\}$ ,  $T = \{t_1, t_2, t_3, \dots, t_n\}$  and  $t_j \subseteq I$
- Minimum support count should be maintained

---

<sup>1</sup>Mining association rules between sets of items in large databases, *R Agrawal, T Imielinski, and A Swami*, SIGMOD, 22(2), pp 207–216, ACM-1993

# Association Rules

Association rules tries to do linked analysis.

**Example:** Whether same products are selling together?

- $I = \{i_1, i_2, i_3, \dots, i_m\}$ ,  $T = \{t_1, t_2, t_3, \dots, t_n\}$  and  $t_j \subseteq I$
- Minimum support count should be maintained
- Can you see: Subset of frequent items is also frequent

---

<sup>1</sup>Mining association rules between sets of items in large databases, *R Agrawal, T Imielinski, and A Swami*, SIGMOD, 22(2), pp 207–216, ACM-1993

# Association Rules

Association rules tries to do linked analysis.

**Example:** Whether same products are selling together?

- $I = \{i_1, i_2, i_3, \dots, i_m\}$ ,  $T = \{t_1, t_2, t_3, \dots, t_n\}$  and  $t_j \subseteq I$
- Minimum support count should be maintained
- Can you see: Subset of frequent items is also frequent
- Apriori analysis <sup>1</sup>

---

<sup>1</sup>Mining association rules between sets of items in large databases, *R Agrawal, T Imielinski, and A Swami*, SIGMOD, 22(2), pp 207–216, ACM-1993

# Association Rules

Association rules tries to do linked analysis.

**Example:** Whether same products are selling together?

- $I = \{i_1, i_2, i_3, \dots, i_m\}$ ,  $T = \{t_1, t_2, t_3, \dots, t_n\}$  and  $t_j \subseteq I$
- Minimum support count should be maintained
- Can you see: Subset of frequent items is also frequent
- Apriori analysis <sup>1</sup>

Let's do it:

$t_1 = (1, 3, 4)$ ,  $t_2 = (2, 3, 5)$ ,  $t_3 = (1, 2, 3, 5)$ ,  $t_4 = (2, 5)$ ,  $t_5 = (1, 3, 5)$  and minimum support count be 2

---

<sup>1</sup>Mining association rules between sets of items in large databases, *R Agrawal, T Imielinski, and A Swami*, SIGMOD, 22(2), pp 207–216, ACM-1993

# Association Rules

$t_1 = (1, 3, 4)$ ,  $t_2 = (2, 3, 5)$ ,  $t_3 = (1, 2, 3, 5)$ ,  $t_4 = (2, 5)$ ,  $t_5 = (1, 3, 5)$



# Association Rules

$t_1 = (1, 3, 4)$ ,  $t_2 = (2, 3, 5)$ ,  $t_3 = (1, 2, 3, 5)$ ,  $t_4 = (2, 5)$ ,  $t_5 = (1, 3, 5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

# Association Rules

$t_1 = (1, 3, 4)$ ,  $t_2 = (2, 3, 5)$ ,  $t_3 = (1, 2, 3, 5)$ ,  $t_4 = (2, 5)$ ,  $t_5 = (1, 3, 5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

Symb	Sup
{1,2}	1
{1,3}	3 ✓
{1,5}	2 ✓
{2,3}	2 ✓
{2,5}	3 ✓
{3,5}	3 ✓

# Association Rules

$t_1 = (1, 3, 4)$ ,  $t_2 = (2, 3, 5)$ ,  $t_3 = (1, 2, 3, 5)$ ,  $t_4 = (2, 5)$ ,  $t_5 = (1, 3, 5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

Symb	Sup
{1,2}	1
{1,3}	3 ✓
{1,5}	2 ✓
{2,3}	2 ✓
{2,5}	3 ✓
{3,5}	3 ✓

Symb	Sup
{1,2,3}	1
{1,2,5}	1
{1,3,5}	2 ✓
{2,3,5}	2 ✓

# Association Rules

$t_1 = (1, 3, 4)$ ,  $t_2 = (2, 3, 5)$ ,  $t_3 = (1, 2, 3, 5)$ ,  $t_4 = (2, 5)$ ,  $t_5 = (1, 3, 5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

Symb	Sup
{1,2}	1
{1,3}	3 ✓
{1,5}	2 ✓
{2,3}	2 ✓
{2,5}	3 ✓
{3,5}	3 ✓

Symb	Sup
{1,2,3}	1
{1,2,5}	1
{1,3,5}	2 ✓
{2,3,5}	2 ✓

Symb	Sup
{1,2,3,5}	1

# Association Rule Mining

## Mathematical model of Association Rule Mining

- Let  $I = \{i_1, i_2, \dots, i_m\}$  be set of items

# Association Rule Mining

## Mathematical model of Association Rule Mining

- Let  $I = \{i_1, i_2, \dots, i_m\}$  be set of items
- Let  $T = \{t_1, t_2, \dots, t_n\}$  be set of transactions where  $t_j \subseteq I$

# Association Rule Mining

## Mathematical model of Association Rule Mining

- Let  $I = \{i_1, i_2, \dots, i_m\}$  be set of items
- Let  $T = \{t_1, t_2, \dots, t_n\}$  be set of transactions where  $t_j \subseteq I$
- $t_j$  is said to contain  $X \subseteq I$  if  $X \subseteq t_j$

# Association Rule Mining

## Mathematical model of Association Rule Mining

- Let  $I = \{i_1, i_2, \dots, i_m\}$  be set of items
- Let  $T = \{t_1, t_2, \dots, t_n\}$  be set of transactions where  $t_j \subseteq I$
- $t_j$  is said to contain  $X \subseteq I$  if  $X \subseteq t_j$

An association rule is an implication of the form  $X \Rightarrow Y$ , where  $X \subseteq I$ ,  $Y \subseteq I$  and  $X \cap Y = \phi$



# Association Rule Mining

## Mathematical model of Association Rule Mining

- Let  $I = \{i_1, i_2, \dots, i_m\}$  be set of items
- Let  $T = \{t_1, t_2, \dots, t_n\}$  be set of transactions where  $t_i \subseteq I$
- $t_i$  is said to contain  $X \subseteq I$  if  $X \subseteq t_i$

An association rule is an implication of the form  $X \Rightarrow Y$ , where  $X \subseteq I$ ,  $Y \subseteq I$  and  $X \cap Y = \phi$

- An association rule  $X \Rightarrow Y$  has a **support**  $s$  in set  $T$  if  $s\%$  of the transactions in  $T$  contains  $X \cup Y$

$$\text{support}(X \Rightarrow Y) = P(X \cup Y)$$

# Association Rule Mining

## Mathematical model of Association Rule Mining

- Let  $I = \{i_1, i_2, \dots, i_m\}$  be set of items
- Let  $T = \{t_1, t_2, \dots, t_n\}$  be set of transactions where  $t_i \subseteq I$
- $t_i$  is said to contain  $X \subseteq I$  if  $X \subseteq t_i$

An association rule is an implication of the form  $X \Rightarrow Y$ , where  $X \subseteq I$ ,  $Y \subseteq I$  and  $X \cap Y = \phi$

- An association rule  $X \Rightarrow Y$  has a **support**  $s$  in set  $T$  if  $s\%$  of the transactions in  $T$  contains  $X \cup Y$

$$\text{support}(X \Rightarrow Y) = P(X \cup Y)$$

- The association rule  $X \Rightarrow Y$  holds in the transaction set  $T$  with **confidence**  $c$  if  $c\%$  of the transactions in  $T$  that contain  $X$  also contain  $Y$

$$\text{confidence}(X \Rightarrow Y) = P(Y|X)$$

## An Example

Find **support** and **confidence** for  $X \Rightarrow Y$  in following database

(Z)
(Z)
(Z)
(X,Y)
(X,Y)
(X,Y)
(X,Z)
(X,Z)
(Z)
(Z)

- **support** =  $P(X \cup Y)$

## An Example

Find **support** and **confidence** for  $X \Rightarrow Y$  in following database

(Z)
(Z)
(Z)
(X,Y)
(X,Y)
(X,Y)
(X,Z)
(X,Z)
(Z)
(Z)

- **support** =  $P(X \cup Y)$

3/10

- **confidence** =  $P(Y|X)$

## An Example

Find **support** and **confidence** for  $X \Rightarrow Y$  in following database

(Z)
(Z)
(Z)
(X,Y)
(X,Y)
(X,Y)
(X,Z)
(X,Z)
(Z)
(Z)

- **support** =  $P(X \cup Y)$

$$3/10$$

- **confidence** =  $P(Y|X)$

$$3/5$$

## Association Rule Mining (contd..)

For a given *support* and *confidence* the problem of mining association rules is to find out all the association rules that have confidence and support greater than the corresponding thresholds.

## Association Rule Mining (contd..)

For a given *support* and *confidence* the problem of mining association rules is to find out all the association rules that have confidence and support greater than the corresponding thresholds.

It is a two-step process

- 1 Find all frequent item sets:  $\{X : support(X) \geq S_{min}\}$

## Association Rule Mining (contd..)

For a given *support* and *confidence* the problem of mining association rules is to find out all the association rules that have confidence and support greater than the corresponding thresholds.

### It is a two-step process

- 1 Find all frequent item sets:  $\{X : \text{support}(X) \geq S_{min}\}$
- 2 Generate association rules from the frequent item set: For any pair of frequent item set  $W$  and  $X$  satisfying  $X \subset W$ , of  $\text{support}(X)/\text{support}(W) \geq C_{min}$ , then  $X \Rightarrow Y$  is a valid rule where  $Y = W - X$ .



## Association Rule Mining (contd..)

For a given *support* and *confidence* the problem of mining association rules is to find out all the association rules that have confidence and support greater than the corresponding thresholds.

### It is a two-step process

- 1 Find all frequent item sets:  $\{X : \text{support}(X) \geq S_{min}\}$
- 2 Generate association rules from the frequent item set: For any pair of frequent item set  $W$  and  $X$  satisfying  $X \subset W$ , of  $\text{support}(X)/\text{support}(W) \geq C_{min}$ , then  $X \Rightarrow Y$  is a valid rule where  $Y = W - X$ .

- Second part is straight forward

## Association Rule Mining (contd..)

For a given *support* and *confidence* the problem of mining association rules is to find out all the association rules that have confidence and support greater than the corresponding thresholds.

### It is a two-step process

- 1 Find all frequent item sets:  $\{X : support(X) \geq S_{min}\}$
  - 2 Generate association rules from the frequent item set: For any pair of frequent item set  $W$  and  $X$  satisfying  $X \subset W$ , of  $support(X)/support(W) \geq C_{min}$ , then  $X \Rightarrow Y$  is a valid rule where  $Y = W - X$ .
- Second part is straight forward
  - Most of the research interest lies in solving the first part

## Association Rule Mining (contd..)

For a given *support* and *confidence* the problem of mining association rules is to find out all the association rules that have confidence and support greater than the corresponding thresholds.

### It is a two-step process

- 1 Find all frequent item sets:  $\{X : \text{support}(X) \geq S_{min}\}$
- 2 Generate association rules from the frequent item set: For any pair of frequent item set  $W$  and  $X$  satisfying  $X \subset W$ , of  $\text{support}(X)/\text{support}(W) \geq C_{min}$ , then  $X \Rightarrow Y$  is a valid rule where  $Y = W - X$ .

- Second part is straight forward
- Most of the research interest lies in solving the first part

Prior work includes *Apriori*, *DHP*, *partition based*, *TreeProjection*, *FP-Tree*, and *constraint-based* ones.

# Overview of Apriori

- Uses prior knowledge of  $k$ -item set to explore  $(k+1)$ -item set in a levelwise process.

# Overview of Apriori

- Uses prior knowledge of  $k$ -item set to explore  $(k+1)$ -item set in a levelwise process.
- The set of frequent 1-item sets  $L_1$  is initially found

# Overview of Apriori

- Uses prior knowledge of  $k$ -item set to explore  $(k+1)$ -item set in a levelwise process.
- The set of frequent 1-item sets  $L_1$  is initially found
- $L_1$  is then used by performing join and prune actions to form the set of candidate 2-items sets  $C_2$

# Overview of Apriori

- Uses prior knowledge of  $k$ -item set to explore  $(k+1)$ -item set in a levelwise process.
- The set of frequent 1-item sets  $L_1$  is initially found
- $L_1$  is then used by performing join and prune actions to form the set of candidate 2-items sets  $C_2$
- In next data scan, the set of frequent 2-item sets  $L_2$  are identified

# Overview of Apriori

- Uses prior knowledge of  $k$ -item set to explore  $(k+1)$ -item set in a levelwise process.
- The set of frequent 1-item sets  $L_1$  is initially found
- $L_1$  is then used by performing join and prune actions to form the set of candidate 2-items sets  $C_2$
- In next data scan, the set of frequent 2-item sets  $L_2$  are identified
- The whole process continues iteratively until there is no more candidate item sets



# Overview of Apriori

- Uses prior knowledge of  $k$ -item set to explore  $(k+1)$ -item set in a levelwise process.
- The set of frequent 1-item sets  $L_1$  is initially found
- $L_1$  is then used by performing join and prune actions to form the set of candidate 2-items sets  $C_2$
- In next data scan, the set of frequent 2-item sets  $L_2$  are identified
- The whole process continues iteratively until there is no more candidate item sets

## Example:

Consider  $I = \{A, B, C, D, E, F\}$  and transaction  $T = \{ t_1 = (A, B, C), t_2 = (A, F), t_3 = (A, B, C, E), t_4 = (A, B, D, F), t_5 = (C, F), t_6 = (A, B, C), t_7 = (A, B, C, E), t_8 = (C, D, E), t_9 = (B, D, E), \}$  and the minimum support be greater than 3.

# Apriori at work

Consider transactions T

$T_1 = (A, B, C)$

$T_2 = (A, F)$

$T_3 = (A, B, C, E)$

$T_4 = (A, B, D, F)$

$T_5 = (C, F)$

$T_6 = (A, B, C)$

$T_7 = (A, B, C, E)$

$T_8 = (C, D, E)$

$T_9 = (B, D, E)$

# Apriori at work

Consider transactions T

$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$

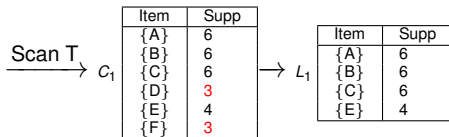
Scan T  $\rightarrow C_1$

Item	Supp
{A}	6
{B}	6
{C}	6
{D}	3
{E}	4
{F}	3

# Apriori at work

Consider transactions T

$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$



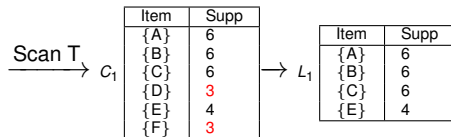
# Apriori at work

Consider transactions T

$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$

$C_2$

Item
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}
{C,E}



# Apriori at work

Consider transactions T

$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$

$C_2$	Item
	{A,B}
	{A,C}
	{A,E}
	{B,C}
	{B,E}
	{C,E}

Scan T

$C_2$	Item	Supp
	{A,B}	5
	{A,C}	4
	{A,E}	2
	{B,C}	4
	{B,E}	3
	{C,E}	3

Scan T

$C_1$	Item	Supp
	{A}	6
	{B}	6
	{C}	6
	{D}	3
	{E}	4
	{F}	3

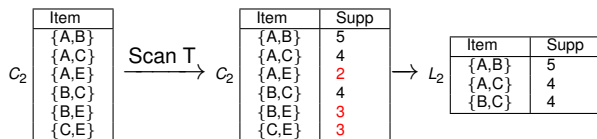
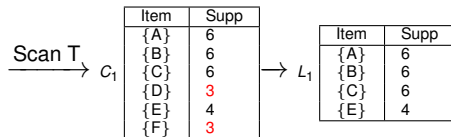
$L_1$

Item	Supp
{A}	6
{B}	6
{C}	6
{E}	4

# Apriori at work

Consider transactions T

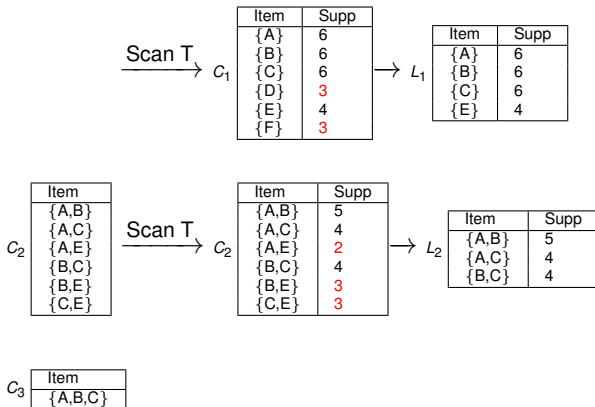
$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$



# Apriori at work

Consider transactions T

$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$

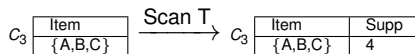
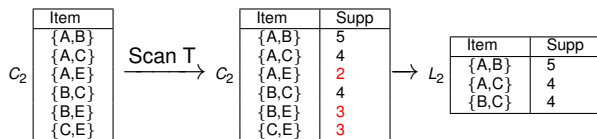
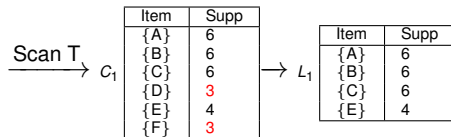




# Apriori at work

Consider transactions T

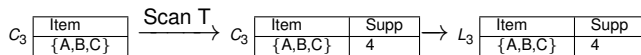
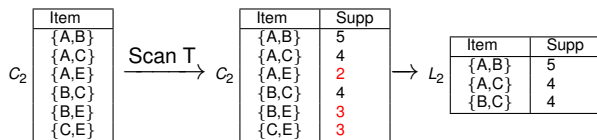
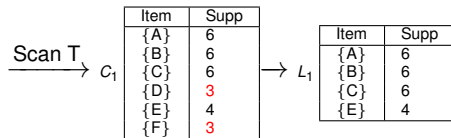
$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$



# Apriori at work

Consider transactions T

$T_1=(A,B,C)$
$T_2=(A,F)$
$T_3=(A,B,C,E)$
$T_4=(A,B,D,F)$
$T_5=(C,F)$
$T_6=(A,B,C)$
$T_7=(A,B,C,E)$
$T_8=(C,D,E)$
$T_9=(B,D,E)$



# Generate Association Rules

If  $X \subset W$  &  $support(X)/support(W) \geq C_{min}$ , then  $X \Rightarrow W - X$

# Generate Association Rules

If  $X \subset W$  &  $\text{support}(X)/\text{support}(W) \geq C_{min}$ , then  $X \Rightarrow W - X$

Transactions

$T_1=(A,B,C)$

$T_2=(A,F)$

$T_3=(A,B,C,E)$

$T_4=(A,B,D,F)$

$T_5=(C,F)$

$T_6=(A,B,C)$

$T_7=(A,B,C,E)$

$T_8=(C,D,E)$

$T_9=(B,D,E)$

# Generate Association Rules

If  $X \subset W$  &  $support(X)/support(W) \geq C_{min}$ , then  $X \Rightarrow W - X$

Transactions

$T_1=(A,B,C)$

$T_2=(A,F)$

$T_3=(A,B,C,E)$

$T_4=(A,B,D,F)$

$T_5=(C,F)$

$T_6=(A,B,C)$

$T_7=(A,B,C,E)$

$T_8=(C,D,E)$

$T_9=(B,D,E)$

- Our frequent item contains

▶  $\{A\}_6 \{B\}_6 \{C\}_6 \{E\}_4 \{A,B\}_5 \{A,C\}_4 \{B,C\}_4$   
 $\{A,B,C\}_4$

# Generate Association Rules

If  $X \subset W$  &  $support(X)/support(W) \geq C_{min}$ , then  $X \Rightarrow W - X$

Transactions

$T_1=(A,B,C)$

$T_2=(A,F)$

$T_3=(A,B,C,E)$

$T_4=(A,B,D,F)$

$T_5=(C,F)$

$T_6=(A,B,C)$

$T_7=(A,B,C,E)$

$T_8=(C,D,E)$

$T_9=(B,D,E)$

- Our frequent item contains

▶  $\{A\}_6 \{B\}_6 \{C\}_6 \{E\}_4 \{A,B\}_5 \{A,C\}_4 \{B,C\}_4$   
 $\{A,B,C\}_4$

- Possibilities are

$A \Rightarrow B, B \Rightarrow A, A \Rightarrow C, C \Rightarrow A, B \Rightarrow C, C \Rightarrow B,$

$A \Rightarrow \{B, C\}, B \Rightarrow \{A, C\}, C \Rightarrow \{B, A\},$

$\{A, B\} \Rightarrow C, \{A, C\} \Rightarrow B, \{B, C\} \Rightarrow A$

- Let's take  $C_{min} = 1.22$

# Generate Association Rules

If  $X \subset W$  &  $support(X)/support(W) \geq C_{min}$ , then  $X \Rightarrow W - X$

Transactions

$T_1=(A,B,C)$

$T_2=(A,F)$

$T_3=(A,B,C,E)$

$T_4=(A,B,D,F)$

$T_5=(C,F)$

$T_6=(A,B,C)$

$T_7=(A,B,C,E)$

$T_8=(C,D,E)$

$T_9=(B,D,E)$

- Our frequent item contains

▶  $\{A\}_6 \{B\}_6 \{C\}_6 \{E\}_4 \{A,B\}_5 \{A,C\}_4 \{B,C\}_4$   
 $\{A,B,C\}_4$

- Possibilities are

$A \Rightarrow B, B \Rightarrow A, A \Rightarrow C, C \Rightarrow A, B \Rightarrow C, C \Rightarrow B,$   
 $A \Rightarrow \{B, C\}, B \Rightarrow \{A, C\}, C \Rightarrow \{B, A\},$   
 $\{A, B\} \Rightarrow C, \{A, C\} \Rightarrow B, \{B, C\} \Rightarrow A$

- Let's take  $C_{min} = 1.22$
- Association rules that qualifies as valid rule are shown green

$A \Rightarrow B, B \Rightarrow A, A \Rightarrow C, C \Rightarrow A, B \Rightarrow C, C \Rightarrow B, A \Rightarrow \{B, C\},$   
 $B \Rightarrow \{A, C\}, C \Rightarrow \{B, A\}, \{A, B\} \Rightarrow C, \{A, C\} \Rightarrow B, \{B, C\} \Rightarrow A$

Thank You!

**Thank you very much for your attention!**

**Queries ?**