SS-ZG548: ADVANCED DATA MINING

Lecture-02: Incremental Mining

Dr. Kamlesh Tiwari
Assistant Professor
Department of Computer Science and Information Systems Engineering,
BITS Pilani, Rajasthan-333031 INDIA

Jan 07, 2018 (WILP @ BITS-Pilani Jan-Apr 2018)

Recap: Data Mining

- **Data mining** is a process supporting knowledge discovery in databases (KDD). KDD involves collection of data, preprocessing, transformation, data mining, and interpretation.

Recap: Data Mining

- **Data mining** is a process supporting knowledge discovery in databases (KDD). KDD involves collection of data, preprocessing, transformation, data mining, and interpretation.
- Three parts **Model**, **Preference**, and **Search**

Recap: Data Mining

- **Data mining** is a process supporting knowledge discovery in databases (KDD). KDD involves collection of data, preprocessing, transformation, data mining, and interpretation.
- Three parts **Model**, **Preference**, and **Search**
- Two broad categories

Recap: Data Mining

- **Data mining** is a process supporting knowledge discovery in databases (KDD). KDD involves collection of data, preprocessing, transformation, data mining, and interpretation.
- 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

Recap: Data Mining

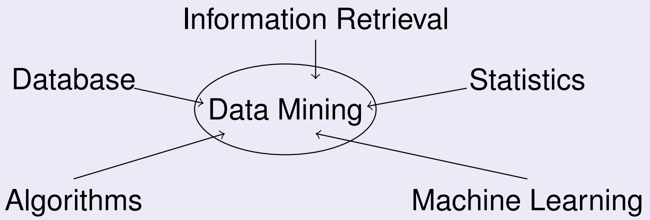
- **Data mining** is a process supporting knowledge discovery in databases (KDD). KDD involves collection of data, preprocessing, transformation, data mining, and interpretation.
- Three parts **Model**, **Preference**, and **Search**
- Two broad categories 1) **Predictive** if we focus on new data involving 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

Recap: Data Mining

- **Data mining** is a process supporting knowledge discovery in databases (KDD). KDD involves collection of data, preprocessing, transformation, data mining, and interpretation.
- 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
- Differs from traditional query processing
  - *Query* may not be well formed
  - *Data* may be preprocessed and modified
  - *Output* could be an analysis that may not be a subset of data

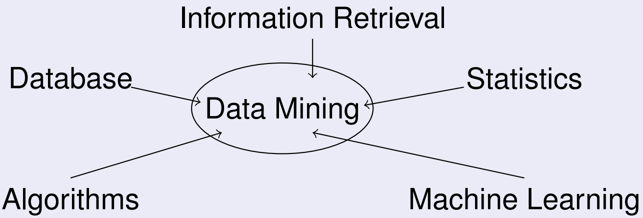
Recap: Data Mining

Discipline that includes many fields



Recap: Data Mining

Discipline that includes many fields

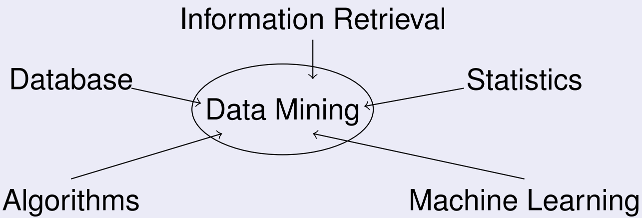


Issues:

Human interaction,

Recap: Data Mining

Discipline that includes many fields

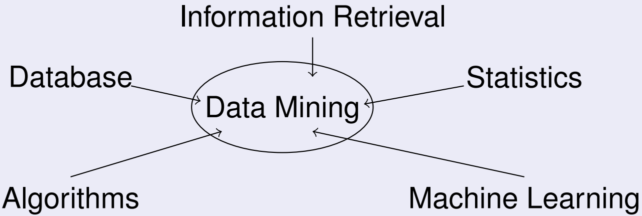


Issues:

Human interaction, Overfitting,

Recap: Data Mining

Discipline that includes many fields

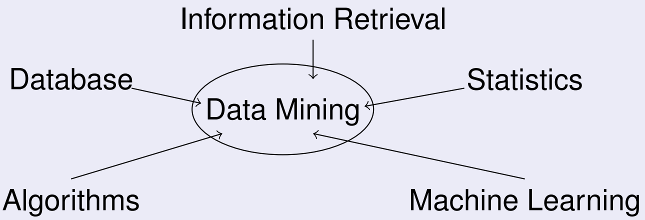


Issues:

Human interaction, Overfitting, Outliers,

Recap: Data Mining

Discipline that includes many fields

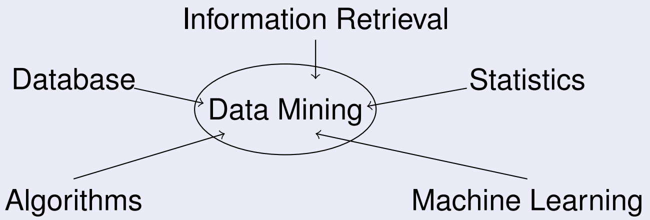


Issues:

Human interaction, Overfitting, Outliers, Large dataset,

Recap: Data Mining

Discipline that includes many fields

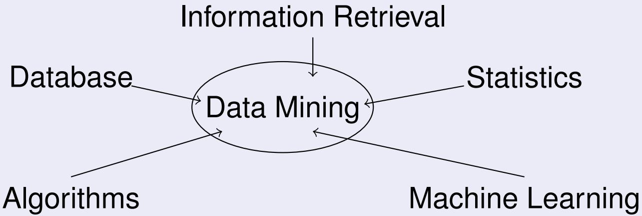


Issues:

Human interaction, Overfitting, Outliers, Large dataset, High dimension,

Recap: Data Mining

Discipline that includes many fields

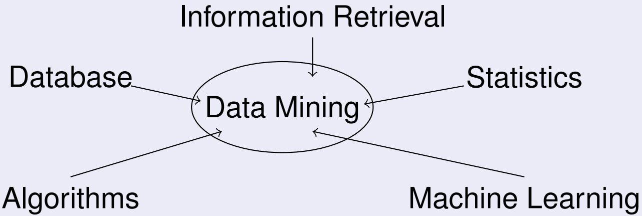


Issues:

Human interaction, Overfitting, Outliers, Large dataset, High dimension, Multimedia data,

Recap: Data Mining

Discipline that includes many fields

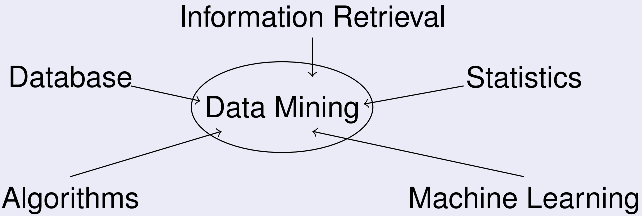


Issues:

Human interaction, Overfitting, Outliers, Large dataset, High dimension, Multimedia data, missing data,

Recap: Data Mining

Discipline that includes many fields

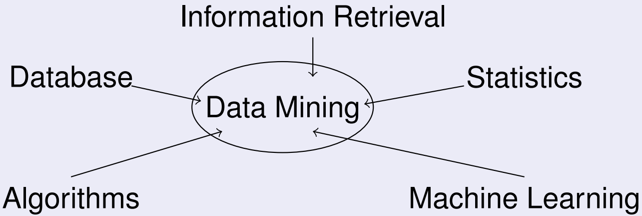


Issues:

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

Recap: Data Mining

Discipline that includes many fields

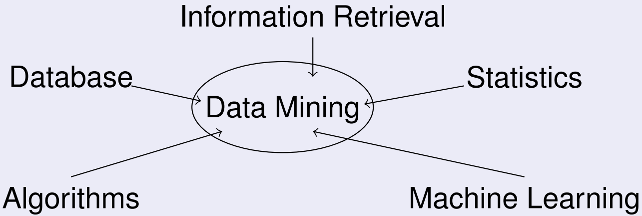


Issues:

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

Recap: Data Mining

Discipline that includes many fields



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 sames products are selling together?

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

Association Rules

Association rules tries to do linked analysis.

Example: Whether sames products are selling together?

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

Association Rules

Association rules tries to do linked analysis.

Example: Whether sames products are selling together?

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

Association Rules

Association rules tries to do linked analysis.

Example: Whether sames products are selling together?

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

Association Rules

Association rules tries to do linked analysis.

Example: Whether sames products are selling together?

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

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

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 $I = \{i_1, i_2, ..., i_m\}$ be set of items

Association Rule Mining

Mathematical model of Association Rule Mining

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

Association Rule Mining

Mathematical model of Association Rule Mining

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

Association Rule Mining

Mathematical model of Association Rule Mining

- Let $I = \{i_1, i_2, ..., i_m\}$ be set of items
- Let $T = \{t_1, t_2, ..., t_n\}$ 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 $X \cap Y = \phi$

Association Rule Mining

Mathematical model of Association Rule Mining

- Let $I = \{i_1, i_2, ..., i_m\}$ be set of items
- Let $T = \{t_1, t_2, ..., t_n\}$ 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 $X \cup Y$

$$support(X \Rightarrow Y) = P(X \cup Y)$$

Association Rule Mining

Mathematical model of Association Rule Mining

- Let $I = \{i_1, i_2, ..., i_m\}$ be set of items
- Let $T = \{t_1, t_2, ..., t_n\}$ 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 $X \cup Y$

$$support(X \Rightarrow Y) = P(X \cup Y)$$

- 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.

$$confidence(X \Rightarrow Y) = P(Y|X)$$

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:* $\{X : support(X) \geq S_{min}\}$
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 $Y = W - X$.

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

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:* $\{X : support(X) \geq S_{min}\}$
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

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 then 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)

$\xrightarrow{\text{Scan T}}$ C_1

Item	Supp
{A}	6
{B}	6
{C}	6
{D}	3
{E}	6
{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)

Scan T $\longrightarrow C_1$

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

$\longrightarrow L_1$

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

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 $\longrightarrow$ C_1

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

$\longrightarrow$ L_1

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

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)

$\xrightarrow{\text{Scan T}}$ C_1

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

$\longrightarrow$ L_1

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

C_2

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

$\xrightarrow{\text{Scan T}}$ C_2

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

Scan T $\longrightarrow$ C_1

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

$\longrightarrow$ L_1

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

C_2

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

Scan T $\longrightarrow$ C_2

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

$\longrightarrow$ L_2

Item	Supp
{A,B}	5
{A,C}	4
{B,C}	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)

Scan T $\longrightarrow$ C_1

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

$\longrightarrow$ L_1

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

C_2

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

Scan T $\longrightarrow$ C_2

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

$\longrightarrow$ L_2

Item	Supp
{A,B}	5
{A,C}	4
{B,C}	4

C_3

Item
{A,B,C}

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 $\longrightarrow C_1$

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

$\longrightarrow L_1$

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

C_2

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

Scan T $\longrightarrow C_2$

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

$\longrightarrow L_2$

Item	Supp
{A,B}	5
{A,C}	4
{B,C}	4

C_3

Item
{A,B,C}

Scan T $\longrightarrow C_3$

Item	Supp
{A,B,C}	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)

$\xrightarrow{\text{Scan T}} C_1$

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

$\longrightarrow L_1$

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

C_2

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

$\xrightarrow{\text{Scan T}} C_2$

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

$\longrightarrow L_2$

Item	Supp
{A,B}	5
{A,C}	4
{B,C}	4

C_3

Item
{A,B,C}

$\xrightarrow{\text{Scan T}} C_3$

Item	Supp
{A,B,C}	4

$\longrightarrow L_3$

Item	Supp
{A,B,C}	4

Thank You!

Thank you very much for your attention!

Queries ?