Association Rules and Sequence Mining

Instruction 8

IDS

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Recap

Association rule mining can be seen as two steps:

1. Find all frequent item sets using:

- **Apriori algorithm**
- **FP-Growth algorithm**

2.Generate strong association rules with the given criteria from the frequent item sets:

- **By definition, this rules must satisfy minimum support and minimum confidence.**

Why FP-Growth algorithm?

Disadvantages of Apriori algorithm:

- **Find candidate sets in an expensive way. If frequent items are large in amount, so the combination would be huge and it would be an expensive operation.**

So Apriori algorithm is a slow algorithm.

FP-Growth Algorithm

TID	Item sets
T_1	$\{I_1, I_2, I_3, I_4, I_5, I_6\}$
T_2	$\{I_{7,} I_2, I_3, I_4, I_5, I_6\}$
T_3	$\{I_1, I_8, I_4, I_5\}$
T_4	$\{I_1 ,I_9 ,I_{10}, I_4, I_6\}$
T_5	$\{I_{10} , I_2, I_4 , I_{11}, I_5\}$

Item	Support count
I_1	3
I_2	3
I_3	2
I_4	5
I_5	4
I_6	3
I_7	1
I_8	1
I_9	1
I_{10}	2
I_{11}	1

FP-Growth Algorithm

- **Min-support =3**

Item	Support count
I_1	3
I_2	3
I_4	5
I_5	4
I_6	3

Write in descending order →

L

item	Support count
I_4	5
I_5	4
I_1	3
I_2	3
I_6	3

FP-Growth Algorithm

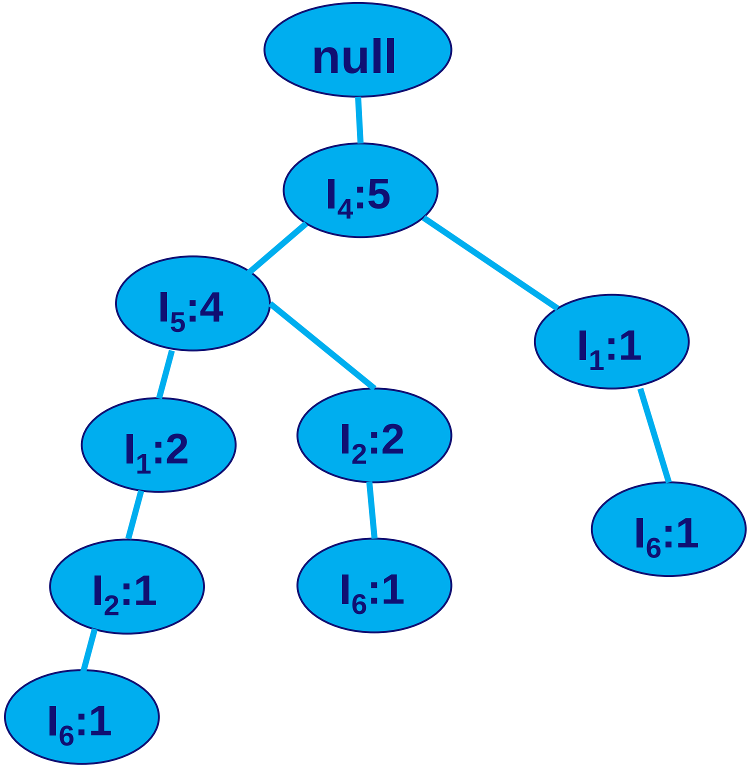
item	Support count
I_4	5
I_5	4
I_1	3
I_2	3
I_6	3

TID	Item sets	Ordered item set
T_1	$\{I_1, I_2, I_3, I_4, I_5, I_6\}$	$\{I_4, I_5, I_1, I_2, I_6\}$
T_2	$\{I_7, I_2, I_3, I_4, I_5, I_6\}$	$\{I_4, I_5, I_2, I_6\}$
T_3	$\{I_1, I_8, I_4, I_5\}$	$\{I_4, I_5, I_1\}$
T_4	$\{I_1, I_9, I_{10}, I_4, I_6\}$	$\{I_4, I_1, I_6\}$
T_5	$\{I_{10}, I_2, I_4, I_{11}, I_5\}$	$\{I_4, I_5, I_2\}$

Chair of Process and Data Science

FP-Growth Algorithm

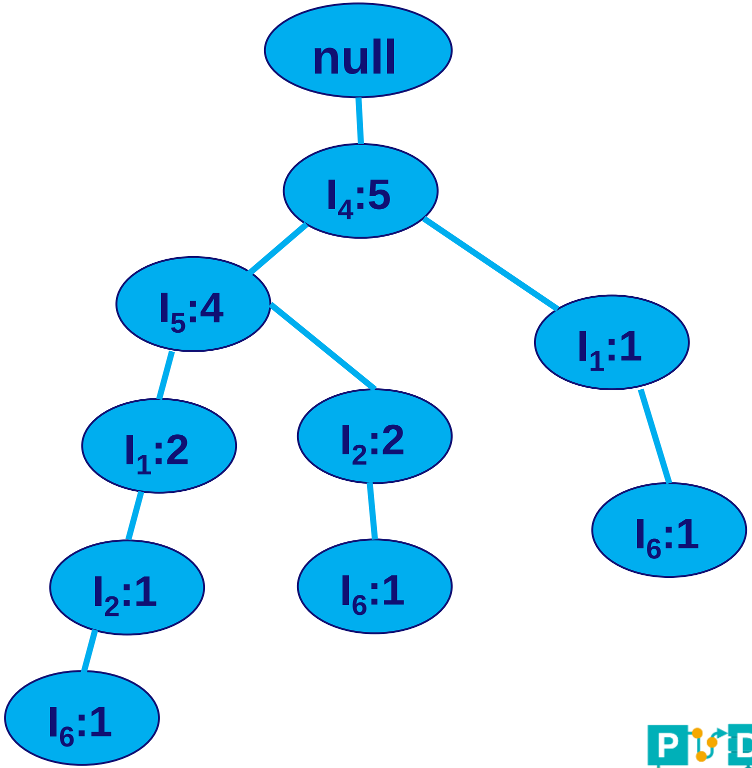
TID	Ordered item set
T_1	$\{I_4, I_5, I_1, I_2, I_6\}$
T_2	$\{I_4, I_5, I_2, I_6\}$
T_3	$\{I_4, I_5, I_1\}$
T_4	$\{I_4, I_1, I_6\}$
T_5	$\{I_4, I_5, I_2\}$



Chair of Process and Data Science

FP-Growth Algorithm

Item	Conditional pattern base
I_6	$(\{I_4, I_5, I_1, I_2\}:1, \{I_4, I_5, I_2\}:1, \{I_4, I_1\}:1)$
I_2	$(\{I_4, I_5, I_1\}:1, \{I_4, I_5\}:2)$
I_1	$(\{I_4, I_5\}:2, \{I_4\}:1)$
I_5	$(\{I_4\}: 4)$
I_4	-



Chair of Process and Data Science

FP-Growth Algorithm

Item	Conditional pattern base	Conditional FP tree	Frequent pattern generated
I_6	({I_4, I_5, I_1, I_2}:1, {I_4, I_5, I_2}:1,{I_4, I_1}:1)	[I_4:3]	<I_4, I_6: 3>
I_2	({I_4, I_5, I_1}:1, {I_4, I_5}:2)	[I_4, I_5:3]	<I_4, I_2:3> <I_5, I_2:3><I_2, I_5, I_4:3>
I_1	({I_4, I_5}:2, {I_4}:1)	[I_4:3]	<I_4, I_1:3>
I_5	({I_4}: 4)	[I_4:4]	<I_4, I_5:4>
I_4	-	-	-

Next step: association rule mining… what is confidence and support for $I_4 \Rightarrow I_6$ and $I_6 \Rightarrow I_4$?

Chair of Process and Data Science

Association Rules

- **T is the set of transactions.**
- **I is the set of all possible item sets composed by items in T.**
- **$A \subseteq I$ and $B \subseteq I$ are two item sets/sub-item sets from T.**
- **A => B is an association rule.**

Usually, we would like to discover the association rule A => B of which the support and confidence are above certain levels.

Association Rules

- $support(A \Rightarrow B) = support(A \cup B) = \dfrac{support_{count}(A \cup B)}{|T|}$
- $confidence(A \Rightarrow B) = \dfrac{support(A \cup B)}{support(A)} = \dfrac{support_{count}(A \cup B)}{support_{count}(A)}$

- **Min_sup represents minimum support and min_conf represents minimum confidence.**
- **A => B is a desired association rule if:**

$support(A \Rightarrow B)$ **>= min_sup and** $confidence(A \Rightarrow B)$ **>= min_conf**

Chair of Process and Data Science

Association Rules

- Set min_sup to 0.5 and min_conf to 0.7. Is {bread} => {meat} from D a desired association rule?

Set of transactions D

TID	Set of items
0	bread, meat, wine
1	bread, meat
2	pizza, wine
3	bread, meat, pizza, wine

Chair of Process and Data Science

Association Rules

- Set min_sup to 0.5 and min_conf to 0.7. Is {bread} => {meat} from D a desired association rule?

Set of transactions D

TID	Set of items
0	bread, meat, wine
1	bread, meat
2	pizza, wine
3	bread, meat, pizza, wine

$$support(\{\text{bread}\} \Rightarrow \{\text{meat}\}) = \frac{support_{count}(\{\text{bread,meat}\})}{|D|} = \frac{3}{4} = 0.75 > \text{min\_sup}$$

$$confidence(\{\text{bread}\} \Rightarrow \{\text{meat}\}) = \frac{support_{count}(\{\text{bread,meat}\})}{support_{count}(\{\text{bread}\})} = \frac{3}{3} = 1 > \text{min\_conf}$$

{bread} => {meat} is a desired association rule

Association Rules

We use lift to evaluate the quality of the discovered association rule A => B.

$$lift(A \Rightarrow B) = \frac{support(A \cup B)}{support(A) \cdot support(B)} = \frac{P(A \cup B)}{P(A) \cdot P(B)}$$

If $lift(A \Rightarrow B) \approx 1$ then A and B are independent

If $lift(A \Rightarrow B) \ll 1$ then A and B are negatively correlated

If $lift(A \Rightarrow B) \gg 1$ then A and B are positively correlated

Association Rules

Evaluate the quality of the association rule {bread} => {meat} by using lift:

Set of transactions D

TID	Set of items
0	bread, meat, wine
1	bread, meat
2	pizza, wine
3	bread, meat, pizza, wine

$$lift(\{bread\} \Rightarrow \{meat\}) = \frac{support(\{bread,meat\})}{support(\{bread\}) \cdot support(\{meat\})} = \frac{(3/4)}{(3/4) \cdot (3/4)} = 1.33$$

Association Rules

Exercise 3: Judge if {A, B} => {E}, {A} => {B} and {A} => {C} are the desired association rules under minimum support 0.5 and minimum confidence 0.75. Also evaluate the quality of the desired rules.

Example data set S

TID	Data items
1	A, B, E
2	C, A, D
3	C, B, D
4	C, A, B, E

Association Rules

Judge if {A, B} => {E}, {A} => {B} and {A} => {C}:

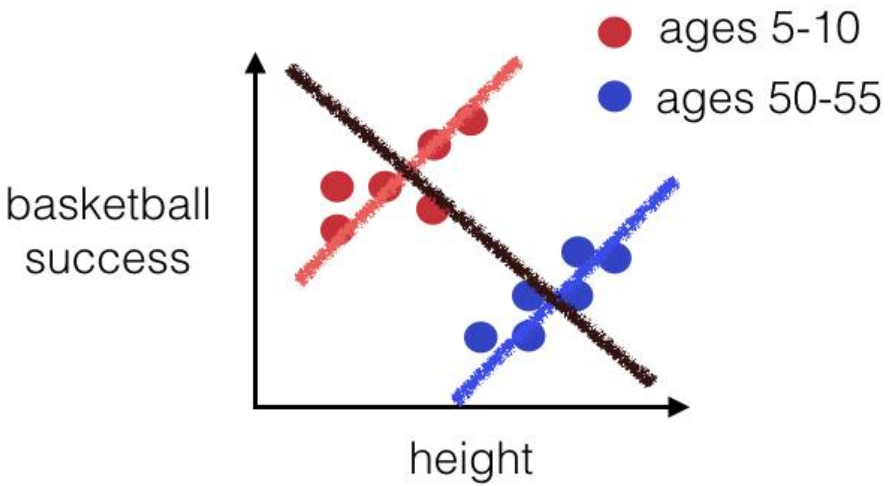
- **Support({A, B} => {E}) = 0.5, confidence({A, B} => {E}) = 1, lift({A, B} => {E}) = 2, it is a desired association rule, and lift is larger than 1.**
- **Support({A} => {B}) = 0.5, confidence({A} => {B}) = 0.67 it is not a desired association rule**
- **Support({A} => {C}) = 0.5, confidence({A} => {C}) = 0.67 it is not a desired association rule**

Simpson's Paradox

	Restaurant 1	Restaurant 2
Males	50\150= 33.3%	180\360= 50%
Females	200\250= 80%	36\40= 90%
General	250\400= 62.5%	216\400= 54%

Chair of Process and Data Science

Simpson's Paradox



Sequence Mining

Support

A central concept in sequence mining is *support* (and *support count*): the frequency of appearance (relative or absolute) of a certain pattern within the database.

Support

In this database, find the support count of:

(bc)(de)

b(de)

(bc)de

(ac)(bc)

(bc)(ac)

D = [

<a(bc)d(eb)>,

<(ac)(bc)de>,

<(ac)b(cd)>,

<ab(bc)(cde)>,

<(bc)(bd)(bde)>,

<(abc)(ac)(bc)de>,

<a(bd)c(de)>

]

Support: Solutions

In this database, find the support of:

(bc)(de): 2

b(de)

(bc)de

(ac)(bc)

(bc)(ac)

D = [

<a(bc)d(eb)>,

<(ac)(bc)de>,

<(ac)b(cd)>,

<ab(bc)(cde)>,

<(bc)(bd)(bde)>,

<(abc)(ac)(bc)de>,

<a(bd)c(de)>

]

Support: Solutions

In this database, find the support of:

(bc)(de): 2

b(de): 3

(bc)de

(ac)(bc)

(bc)(ac)

D = [

<a(bc)d(eb)>,

<(ac)(bc)de>,

<(ac)b(cd)>,

<ab(bc)(cde)>,

<(bc)(bd)(bde)>,

<(abc)(ac)(bc)de>,

<a(bd)c(de)>

]

Chair of Process and Data Science

Support: Solutions

In this database, find the support of:

(bc)(de): 2

b(de): 3

(bc)de: 4

(ac)(bc)

(bc)(ac)

D = [

<a(bc)d(eb)>,

<(ac)(bc)de>,

<(ac)b(cd)>,

<ab(bc)(cde)>,

<(bc)(bd)(bde)>,

<(abc)(ac)(bc)de>,

<a(bd)c(de)>

]

Chair of Process and Data Science

Support: Solutions

In this database, find the support of:

(bc)(de): 2

b(de): 3

(bc)de: 4

(ac)(bc): 2

(bc)(ac)

D = [

<a(bc)d(eb)>,

<(ac)(bc)de>,

<(ac)b(cd)>,

<ab(bc)(cde)>,

<(bc)(bd)(bde)>,

<(abc)(ac)(bc)de>,

<a(bd)c(de)>

]

Chair of Process and Data Science

Support: Solutions

In this database, find the support of:

(bc)(de): 2

b(de): 3

(bc)de: 4

(ac)(bc): 2

(bc)(ac): 1

D = [

<a(bc)d(eb)>,

<(ac)(bc)de>,

<(ac)b(cd)>,

<ab(bc)(cde)>,

<(bc)(bd)(bde)>,

<(abc)(ac)(bc)de>,

<a(bd)c(de)>

]

Chair of Process and Data Science