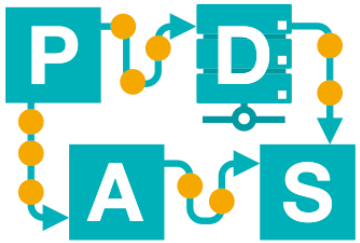
Association Rules and Sequence Mining

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:

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

2.Generate strong association rules 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.)**
- **Scan the database Repeatedly.**

So Apriori algorithm is a slow algorithm.

FP-Growth Algorithm

TID	Itemsets
1	{1,2,3,4,5,6}
2	{7,2,3,4,5,6}
3	{1,8,4,5}
4	{1,9,10,4,6}
5	{10,2,2,4,11,5}

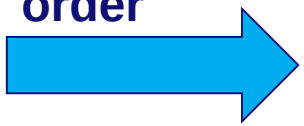
Item	Support count
1	3
2	3
3	2
4	5
5	4
6	3
7	1
8	1
9	1
10	2
11	1

FP-Growth Algorithm

- **Min-support =3**

Item	Support account
1	3
2	3
4	5
5	4
6	3

Write in descending order



L

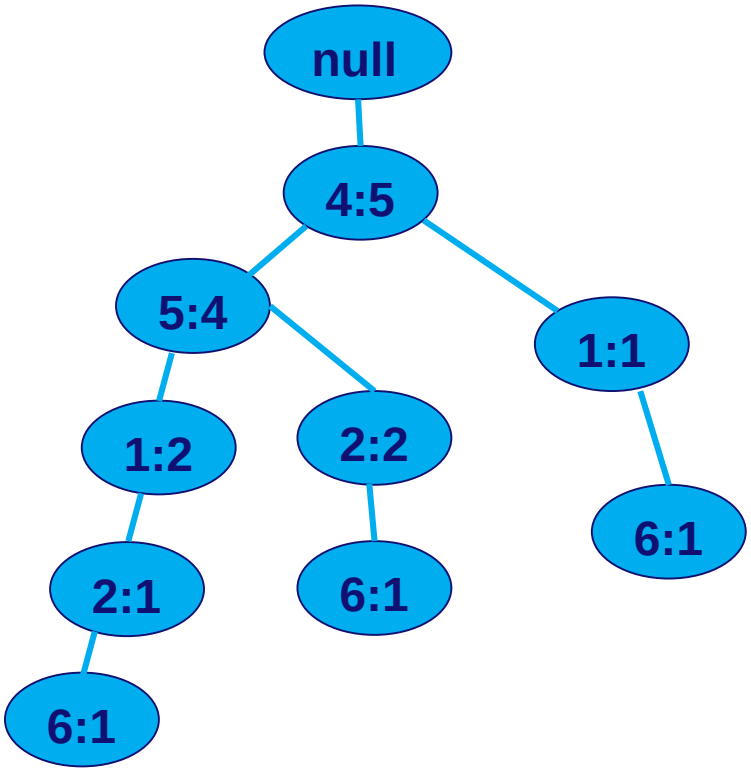
item	Support account
4	5
5	4
1	3
2	3
6	3

FP-Growth Algorithm

TID	Item sets	Ordered item set
1	{1,2,3,4,5,6}	{4,5,1,2,6}
2	{7,2,3,4,5,6}	{4,5,2,6}
3	{1,8,4,5}	{4,5,1}
4	{1,9,10,4,6}	{4,1,6}
5	{10,2,2,4,11,5}	{4,5,2}

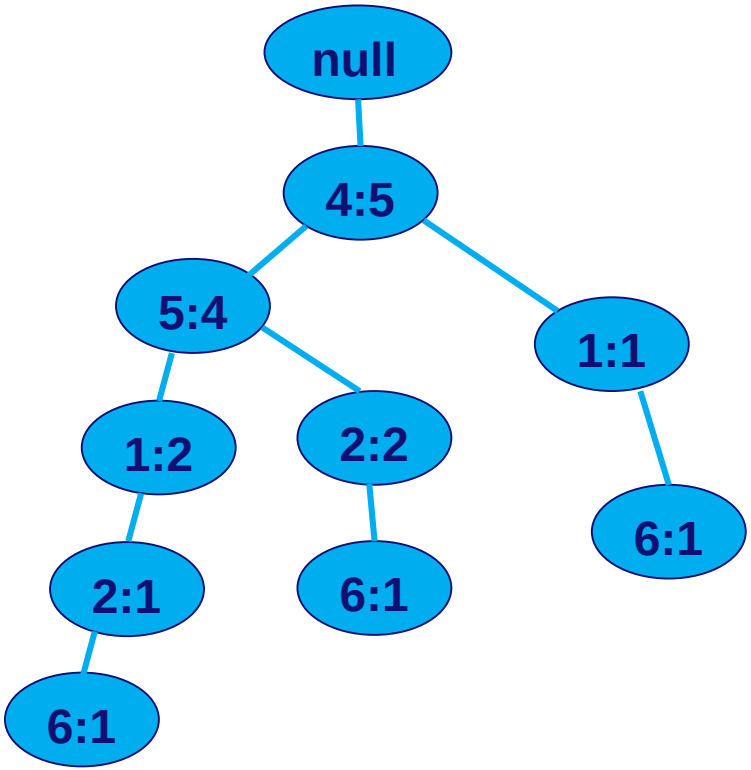
FP-Growth Algorithm

TID	Ordered item set
1	{4,5,1,2,6}
2	{4,5,2,6}
3	{4,5,1}
4	{4,1,6}
5	{4,5,2}



FP-Growth Algorithm

Item	Conditional pattern base
6	({4,5,1,2}:1, {4,5}:1,{4,1}:1)
2	({4,5,1}:1, {4,5}:2)
1	({4,5}:2, {4}:1)
5	({4}: 4)
4	-



FP-Growth Algorithm

In what ways we can reach each item in the tree?

What are the common items in the previous column?

Item	Conditional pattern base	Conditional FP tree	Frequent pattern generated
6	({4,5,1,2}:1, {4,5,2}:1,{4,1}:1)	[4:3]	<4,6: 3>
2	({4,5,1}:1, {4,5}:2)	[4,5:3]	<4,2:3> <5,2:3><2,5,4:3>
1	({4,5}:2, {4}:1)	[4:3]	<4,1:3>
5	({4}: 4)	[4:4]	<4,5:4>
4	-	-	-

Next step: association rules… what is confidence and support for 4 => 6 and 6 => 4 ?

Association Rules

- T is a 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

Association Rules

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

Association Rules

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

- 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

Association Rules

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

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

Set of transactions D

Association Rules

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

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

Set of transactions D

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

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

{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:

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

Set of transactions D

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

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.

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

Example data set S

Association Rules

- **Solution 3: 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, lift({A} => {B}) = 0.89, it is not a desired association rule**

- **support({A} => {C}) = 0.5, confidence({A} => {C}) = 0.67, lift({A} => {C}) = 0.89, it is not a desired association rule**

Support

A central concept in pattern 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)>

]

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)>

]

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)>

]

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)>

]