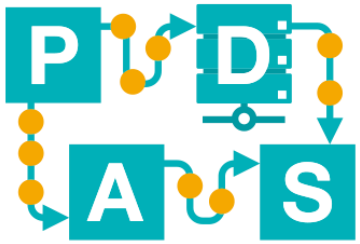
Association Rules and Sequence Mining

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	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	{10,2,4,11,5}

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

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.

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 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

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

- **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**

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)>

]