DM583

Data Mining

Frequent Itemsets and Association Rules

Outline

Recommended Reading

Recommended Reading:

- *Tan et al. [2006], Chapter 6; Tan et al. [2020], Ch. 4.*
- *Han et al. [2011], Chapter 6.*
- *Zaki and Meira Jr. [2020], Chapters 8+9.*
- *Witten et al. [2011], Chapter 4.5.*
- *Advanced topics: Aggarwal and Han [2014].*

Outline

Definition of Sets

A set is a collection of objects (e.g., $S = \{1,4,9\}$).

- The objects are said to be elements of the set (e.g., $1 \in S, 4 \in S, 9 \in S$). Each element is *unique*.

We can define sets

extensionally: by enumerating the elements that define the set (e.g., $S = \{1,4,9\}$).

intensionally: by characterizing the elements of the set,

- describing what condition (the "*characteristic function*" of the set) holds for all the elements and only for the elements of the set (e.g. $S = \{x | \sqrt{x} \in \mathbb{N}$ *AND* $x < 15\}$ — read '|' as 'for which holds' or 'such that').
- The intensional definition typically resorts to a domain over which the set is defined (here: $\mathbb{N}$).

Properties of Sets

- A set can be finite (e.g., $S = \{1, 2, 3, \ldots, n\}$: $1 \in S, 2 \in S, 3 \in S, \ldots, n \in S$). In this case, the set has a finite size (*cardinality*), that is the number of elements of the set (e.g., $|S| = n$).
- A set can be countably infinite. Example:

$$\begin{aligned} S &= \{1, -1, 3, -3, 5, -5, \ldots\} \\ &= \{x | x = 2k + 1 \textit{ OR } x = -2k + 1, k \in \mathbb{N}_0\} \end{aligned}$$

- A set can be uncountable (e.g., $\mathbb{R}$).
- A set can be empty: $S = \{\} = \emptyset$. $|S| = 0$.

Notations

- logical operators:
  - $\vee$: or
  - $\wedge$: and
  - $\neg$: not
  - $\not x$: negation of x (e.g., $\neq$, $\notin$)
  - $A \Rightarrow B$: if A, then B (short for $\neg(A \wedge \neg B)$)
  - $A \Leftarrow B$: if B, then A (short for $\neg(B \wedge \neg A)$)
  - $A \iff B$: $A \Rightarrow B \wedge A \Leftarrow B$ ("iff", read: "if and only if", also: $\equiv$, read: "is equivalent to")
- $\exists x : p(x)$, means: there exists some x such that $p(x)$
- $\forall x : p(x)$, means: for all x holds $p(x)$
- subset: $T \subseteq S \equiv \forall x \in T : x \in S$
- proper subset: $T \subset S \equiv (\forall x \in T : x \in S) \wedge (\exists x \in S : x \notin T)$

Algebra of Sets

An algebra is defined over a base set Ω, all sets involved in the algebra are subsets of Ω.

basic operations for $S, T \subseteq \Omega$:

- **union** $S \cup T \equiv \{x | x \in S \vee x \in T\}$
- **intersection** $S \cap T \equiv \{x | x \in S \wedge x \in T\}$
- **complement** $\bar{S} \equiv S^C \equiv \{x | x \notin S\}$
- **difference** $S \setminus T \equiv \{x | x \in S \wedge x \notin T\}$
- **product** $S \times T \equiv \{(x, y) | x \in S \wedge y \in T\}$

Powerset $\mathcal{P}(S) \equiv \wp(S) \equiv 2^S \equiv \{S' | S' \subseteq S\}$

example Let $\Omega = \mathbb{N}$, $S = \{1, 2, 3\}$ and $T = \{2, 3, 4\}$ – what are the values of all these expressions?

Tuple and Relations

n-tuple element of the (Cartesian) product of n sets:

$$S_1 \times S_2 \times S_3 \times \ldots \times S_n =$$
$$\{(a_1, a_2, a_3, \ldots, a_n) | a_1 \in S_1, a_2 \in S_2, a_3 \in S_3, \ldots, a_n \in S_n\}$$

If $S_1 = S_2 = S_3 = \ldots = S_n$, we write:

$$S^n \equiv S_1 \times S_2 \times S_3 \times \ldots \times S_n$$

n-ary relation R is a set of n-tuples:

$$R(x_1, \ldots, x_n) \equiv (x_1, \ldots, x_n) \in R$$

characteristic function of $R \subseteq S_1 \times \ldots \times S_n$ (can also be seen as “predicate”: “R is true/false”)

$$S_1 \times \ldots \times S_n \rightarrow \{\text{true}, \text{false}\}$$

$$t \mapsto \begin{cases} \text{true} & \text{if } t \in R \\ \text{false} & \text{otherwise} \end{cases}$$

Properties of Homogeneous Relations

Relation $R \subseteq S_1 \times S_2 \times S_3 \times \ldots \times S_n$ is homogeneous iff $S_1 = S_2 = S_3 = \ldots = S_n$.

A binary homogeneous relation $R \subseteq S \times S$ is

reflexive iff $\forall x \in S : (x, x) \in R$

symmetric iff $\forall x, y \in S : (x, y) \in R \Rightarrow (y, x) \in R$

antisymmetric iff $\forall x, y \in S : (x, y) \in R \wedge (y, x) \in R \Rightarrow x = y$

transitive iff $\forall x, y, z \in S : (x, y) \in R \wedge (y, z) \in R \Rightarrow (x, z) \in R$

total iff $\forall x, y \in S : (x, y) \in R \vee (y, x) \in R$

Examples:

- ‘$<$’ $\subseteq \mathbb{N} \times \mathbb{N}$: $(1, 2) \in$ ‘$<$’
  (more customary is the infix notation $1 < 2$)
- ‘$=$’ $\subseteq \mathbb{N} \times \mathbb{N}$: $(361, 361) \in$ ‘$=$’ or $361 = 361$
- ‘$\subset$’ $\subseteq \wp(S)^2$

Partial and Total Order

A relation $R \in S \times S$ can be an order:

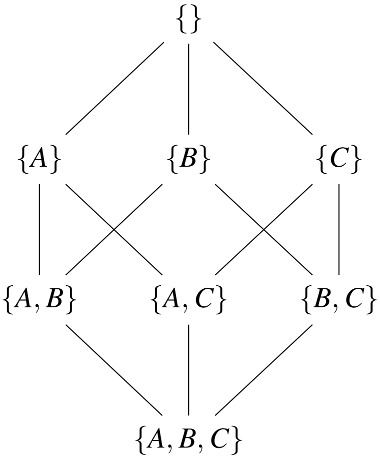
R is a partial order if R is antisymmetric and transitive.

R is a total order if R is a partial order and is total.

Are these relations orders? Which kind of order?

- alphanumeric sorting of strings?
- '$<$' $\subseteq \mathbb{N} \times \mathbb{N}$
- '$\leq$' $\subseteq \mathbb{N} \times \mathbb{N}$
- '$\subset$' $\subseteq \wp(S)^2$
- '$\subseteq$' $\subseteq \wp(S)^2$

Lattice of Subsets



Functions

- A *function* is a mapping from some set (the domain) to some set (the image).
- We can see functions as relations with particular properties: a *univalent* (or *right-unique*) relation over domain and image.
- Formally, a function f is a binary relation over $D \times I$: $f \subseteq D \times I$, for which holds:

$$(x,y) \in f \land (x,z) \in f \Rightarrow y = z$$

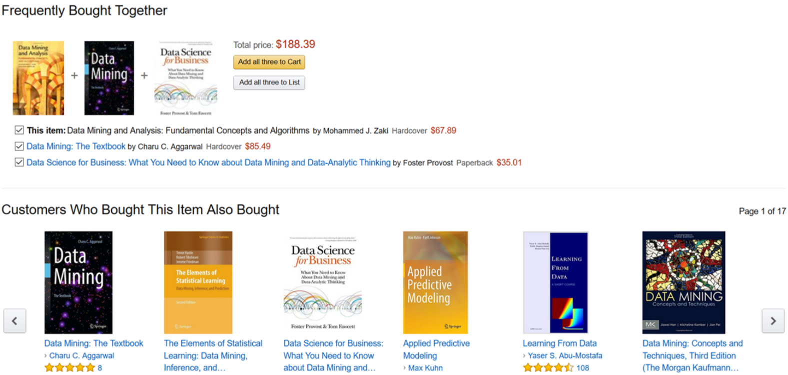
  i.e., for each $d \in D$, f maps to at most one $i \in I$.
- Notation:

$$(x,y) \in f \iff y = f(x) \iff f(x) = y \iff f : x \mapsto y$$

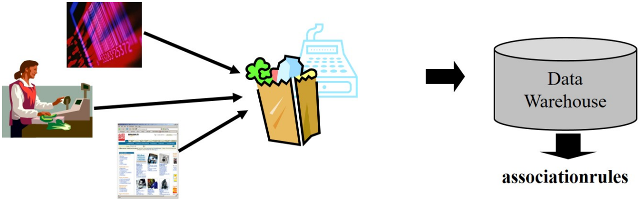
Outline

Motivation

on the webpage of some online bookseller:



Market Basket Analysis



application examples:

- layout of supermarket: optimize the arrangement of items bought together
- online seller: recommend related items
- cross-marketing, add-on sales, targeted attached mailings

read about the infamous example of beer and diapers:

`http://www.theregister.co.uk/2006/08/15/beer_diapers/`

Frequent Patterns

Transaction database:

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
⋮
⋮
```

If we observe patterns, can we conclude on associations between items?

Associations

a,**b,c,d**,e
b,c,d
a,**b,c,d**
c,d,f
a,**b,c,d**,e
a,c,d
a,c,e,f
c,d,e,f
a,**b,c,d**,f
a,b,e,f

In 5 out of 10 (50%) cases,

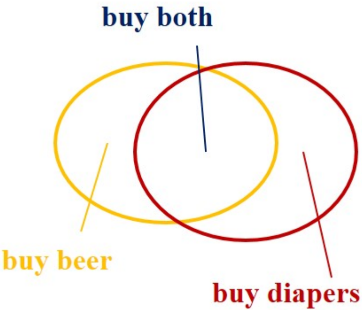
b,c,d

occur together.

In 5 cases we have **b,c**, and in all those 5 cases we also have **d**:

Rule with 100% confidence:

If **b,c** *are in the set, then also* **d** *is in the set.*



Outline

Association Rules

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

items bread, butter, eggs, milk etc.

transaction a database entry as a set of items

rule $L \Rightarrow R$, $L \cap R = \emptyset$ (disjoint sets of items)

- L *left-hand-side* or *antecedent*
- R *right-hand-side* or *consequent*

- {butter, flour} $\Rightarrow$ {milk}
- {sugar} $\Rightarrow$ {butter}

Definition: Frequent Itemsets

items: $I = \{i_1, \ldots, i_m\}$ a set of literals (e.g., items in a shop)

itemset: $X \subseteq I$ (e.g., the items in a basket)

transaction: $T = (tid, X_{tid})$ designates a specific itemset

transaction database $\mathcal{D}$: a set of transactions

order: items in an itemset are ordered by some strict total order (e.g., alphabetical order of the literals), i.e.:
$X = (x_1, x_2, \ldots, x_k) \Rightarrow x_1 < x_2 < \ldots < x_k$

length of an itemset: number of elements contained in the itemset

k-itemset: an itemset of length k (e.g., T1 is a 4-itemset, T4 is a 1-itemset)

Definition: Cover, Support, Frequency

cover of an itemset: set of all transactions that contain the itemset: $\text{cover}(X) = \{(tid, X_{tid}) | (tid, X_{tid}) \in \mathcal{D} \wedge X \subseteq X_{tid}\}$

support of an itemset: the support s of an itemset X ($s(X)$) is the number of transactions containing X (i.e., the size of the cover set): $s(X) = |\text{cover}(X)|$

frequency of an itemset: the frequency of an itemset X is its support relative to the database size $f(X) = \frac{s(X)}{|\mathcal{D}|}$

frequent itemset: given some support threshold σ, an itemset X is frequent (w.r.t. σ) iff: $s(X) \geq \sigma$ or equivalently $f(X) \geq \frac{\sigma}{|\mathcal{D}|}$

Problem 1: Frequent Itemset Mining

Given:

- a set of items I
- a transaction database $\mathcal{D}$ over I
- a support threshold σ

Find all frequent itemsets in $\mathcal{D}$, i.e., $\{X | X \subseteq I \wedge s(X) \geq \sigma\}$

example: which itemsets are frequent with $\sigma = 3$ in $\mathcal{D}$:

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Definition: Association Rule

association rule: expresses an implication of the form $X \Rightarrow Y$, where X and Y are itemsets, $X \cap Y = \emptyset$

implication: describes a co-occurrence, not a causality

An association rule does not necessarily need to hold in all cases. We can describe its strength (or weakness), based on the observed cases:

support: The support of an association rule in $\mathcal{D}$ is the support of the union of its components:
$s(X \Rightarrow Y) = s(X \cup Y)$

frequency: Analogously, $f(X \Rightarrow Y) = f(X \cup Y)$

confidence: $\mathrm{conf}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$

Problem 2: Association Rule Mining

Given:

- a set of items I
- a transaction database $\mathcal{D}$ over I
- a support threshold σ and a confidence threshold c

Find all association rules $X \Rightarrow Y$ in $\mathcal{D}$ with a support of at least σ and a confidence of at least c, i.e.:

$$\{X \Rightarrow Y | s(X \Rightarrow Y) \geq \sigma \wedge \text{conf}(X \Rightarrow Y) \geq c\}$$

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Problem 1 $\subset$ Problem 2

Problem 1 is part of Problem 2:

- Itemset A is frequent w.r.t. σ
- Given
  - $A = X \cup Y$
  - $X \cap Y = \emptyset$
  - $X = A \setminus Y$
  - $Y = A \setminus X$
- '$X \Rightarrow Y$' is frequent w.r.t. σ

Two-step approach:

1. find all frequent itemsets w.r.t. σ
2. generate rules with confidence $\geq c$ from each frequent itemset, where each rule is a binary partition of the itemset

Find Frequent Itemsets: Naïve Algorithm (Brute Force)

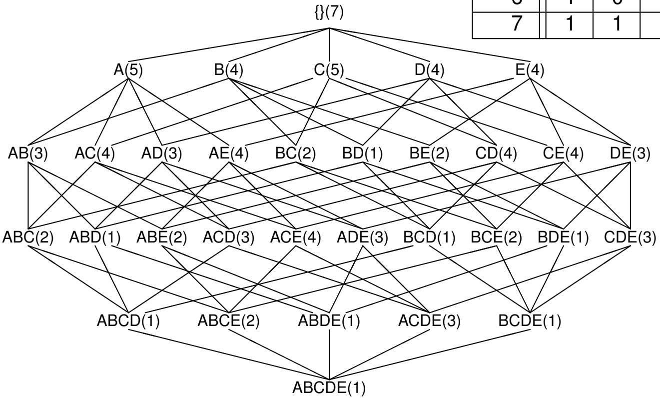
- Each possible itemset is a *candidate* for being a *frequent* itemset.
- We need to check the database to count if the itemset is actually frequent.
- Complexity roughly amounts to:
  number of candidates $\times$ number of transactions
- How many candidates do we have, given n items?

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Outline

Lattice of Itemsets

TID	A	B	C	D	E
1	0	1	0	0	0
2	1	0	1	1	1
3	1	1	1	0	1
4	0	0	1	1	0
5	1	1	1	1	1
6	1	0	1	1	1
7	1	1	0	0	0

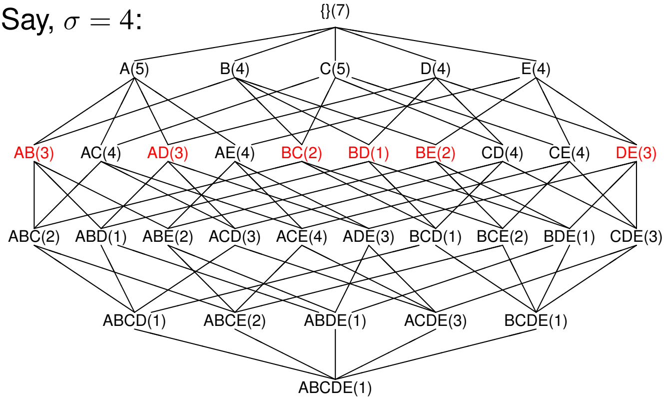


Monotonicity and Anti-Monotonicity

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 4$:

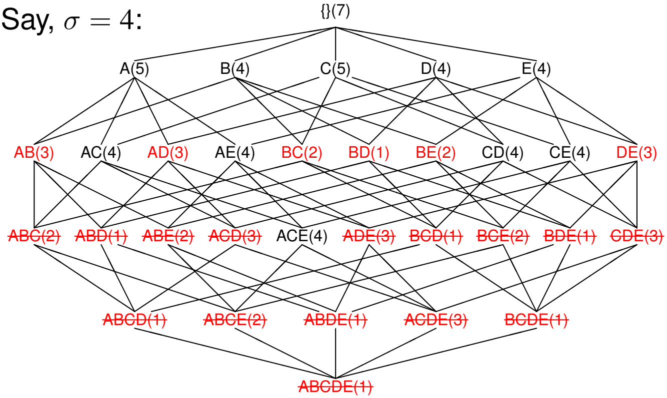


Monotonicity and Anti-Monotonicity

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 4$:

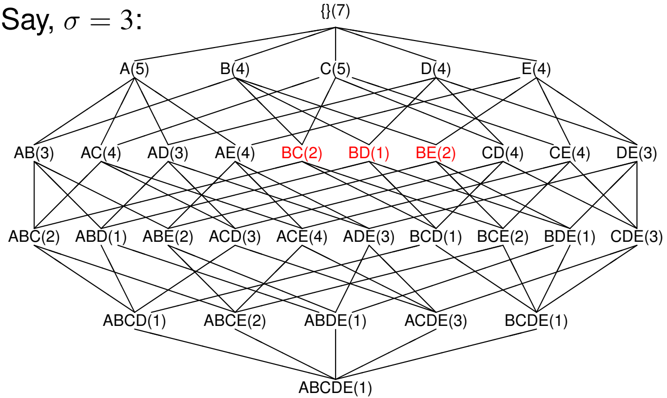


Monotonicity and Anti-Monotonicity

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 3$:

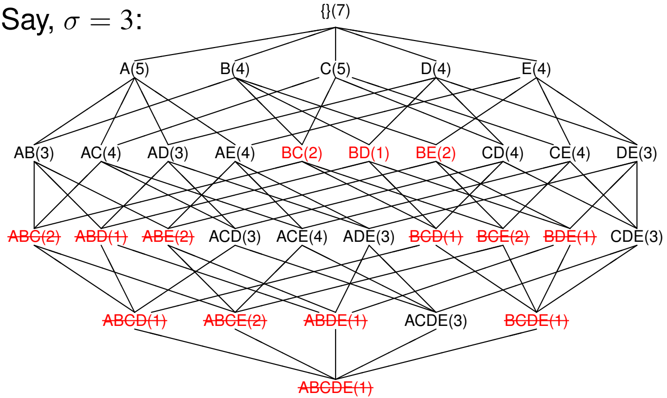


Monotonicity and Anti-Monotonicity

Observation:

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supseteq X$ can be frequent.

Say, $\sigma = 3$:



Maximal Frequent Itemsets

Due to the anti-monotonicity, we can summarize solutions by their *border* in the lattice:

- An itemset X belongs to the border (w.r.t. some σ), if:
  - $\forall Y \subset X : Y$ is frequent (w.r.t. σ)
  - $\forall Z \supset X : Z$ is not frequent (w.r.t. σ)
- positive border: X itself is frequent
- such frequent itemsets are called **maximal frequent itemsets**
- they are frequent itemsets with no frequent supersets

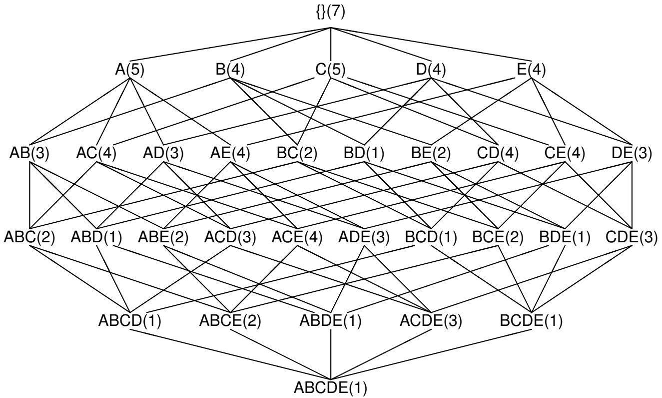
Maximal frequent itemsets can be used as a *condensed representation of a solution*, as all frequent itemsets can be derived from the maximal frequent itemsets.

Note that:

The anti-monotonicity property of support is also called downward-closure property.

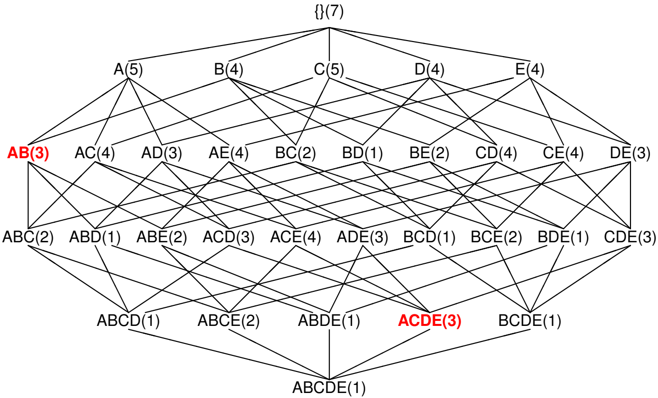
Maximal Frequent Itemsets

Example: find the positive border (maximal frequent itemsets) for $\sigma = 3$



Maximal Frequent Itemsets

Example: find the positive border (maximal frequent itemsets) for $\sigma = 3$



Closed (Frequent) Itemsets

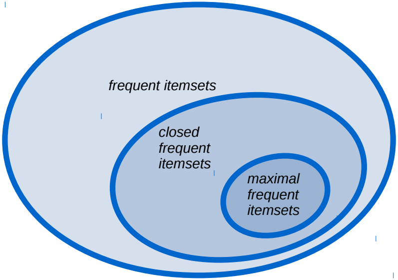
closed itemset: An itemset X is *closed* if none of its immediate supersets has exactly the same support as X.

closed frequent itemset: An itemset is a closed frequent itemset (w.r.t. some σ) if it is a closed itemset and is frequent (w.r.t. σ).

Closed frequent itemsets (including their support) represent *a solution* (all frequent itemsets *and* their support).

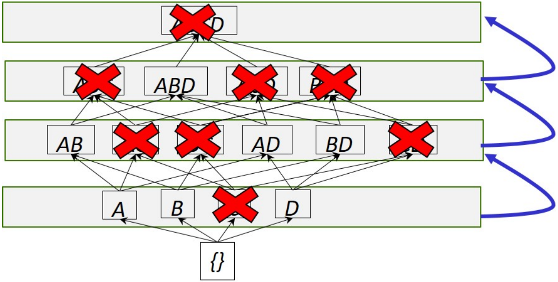
CFIs can also generate all frequent itemsets since they necessarily **contain** the MFIs

The supports of all frequent itemsets can also be retrieved from the CFIs' supports



Outline

Apriori Algorithm [Srikant and Agrawal, 1996]: Idea



1. find frequent 1-itemsets first, then 2-itemsets, 3-itemsets etc. (breadth-first search in the lattice)
2. for finding $(k+1)$-itemsets C_{k+1}: consider only those as candidates, where *all* k-itemsets $C_k \subset C_{k+1}$ are frequent
3. count frequency of all k-itemset candidates in a single database scan (hashing of the candidate itemsets)

Apriori Algorithm [Srikant and Agrawal, 1996]: Pseudo Code

C_k: k-itemset candidates, S_k: frequent k-itemsets (solution)

Algorithm 2.1 (Apriori [Srikant and Agrawal, 1996])

Apriori(I, $\mathcal{D}$, σ)
  $S_1 =$ {frequent 1-itemsets};
  $k = 2$;
  while $S_{k-1} \neq \emptyset$ do
    $C_k =$ AprioriGenerateCandidates(S_{k-1});
    for each transaction $T \in \mathcal{D}$ do
      $C_T = \{c \in C_k | c \subseteq T\}$;
      for each $c \in C_T$ do
        c.count++;
    $S_k = \{c \in C_k | c.count \geq \sigma\}$;
    k++;
  return $\cup_k S_k$;

Apriori Candidate Generation

Algorithm 2.2 (AprioriGenerateCandidates(S_{k-1}))

1. *join:*
   two frequent $(k-1)$-itemsets $p, q \in S_{k-1}$ are joined if they are identical in the ***first*** *(order!) $k-2$ items:*

$$\begin{array}{rl} p \in S_{k-1}: & (\underline{A,B},C) \\ q \in S_{k-1}: & (\underline{A,B},D) \\ \hline \Rightarrow & (A,B,C,D) \quad \in C_k \end{array}$$

2. *pruning:*
   remove all k-itemsets from C_k that contain any $(k-1)$-itemset $\notin S_{k-1}$

example: $S_3 = \{(1,2,3),(1,2,4),(1,3,4),(1,3,5),(2,3,4)\}$

1. join: $C_4 = \{(1,2,3,4),(1,3,4,5)\}$
2. pruning: remove $(1,3,4,5)$

result: $C_4 = \{(1,2,3,4)\}$

Outline

Apriori Example: Find Frequent Itemsets X with $f(X) \geq 0.3$

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	
B	
C	
D	
E	
F	
G	
H	
I	
J	
K	
L	

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	
B	1
C	
D	
E	1
F	
G	1
H	1
I	
J	
K	
L	

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	1
B	2
C	1
D	
E	2
F	
G	2
H	2
I	
J	
K	
L	

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	2
B	3
C	2
D	
E	3
F	1
G	2
H	3
I	
J	
K	
L	

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	2
B	4
C	3
D	1
E	4
F	2
G	3
H	4
I	
J	
K	
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	3
B	5
C	3
D	1
E	5
F	2
G	3
H	5
I	
J	
K	1
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	3
B	6
C	3
D	1
E	6
F	3
G	4
H	6
I	1
J	
K	2
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	4
B	7
C	3
D	2
E	6
F	3
G	5
H	7
I	1
J	
K	2
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	5
B	8
C	3
D	3
E	6
F	3
G	6
H	7
I	1
J	
K	2
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	5
B	9
C	3
D	4
E	6
F	4
G	7
H	7
I	1
J	
K	2
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	5
B	9
C	4
D	4
E	7
F	5
G	7
H	7
I	1
J	
K	2
L	1

SDU
SYDDANSK UNIVERSITET

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	6
B	9
C	5
D	4
E	8
F	6
G	7
H	8
I	1
J	
K	2
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8
I	1
J	
K	2
L	1

C1 and S1

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8
I	$\cancel{1}$
J	$\cancel{0}$
K	$\cancel{2}$
L	$\cancel{1}$

$\sigma = 30\% \Leftrightarrow$ support ≥ 4

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB		CE	
AC		CF	
AD		CG	
AE		CH	
AF		DE	
AG		DF	
AH		DG	
BC		DH	
BD		EF	
BE		EG	
BF		EH	
BG		FG	
BH		FH	
CD		GH	

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB		CE	
AC		CF	
AD		CG	
AE		CH	
AF		DE	
AG		DF	
AH		DG	
BC		DH	
BD		EF	
BE	1	EG	1
BF		EH	1
BG	1	FG	
BH	1	FH	
CD		GH	1

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	1	CE	1
AC	1	CF	
AD		CG	1
AE	1	CH	1
AF		DE	
AG	1	DF	
AH	1	DG	
BC	1	DH	
BD		EF	
BE	2	EG	2
BF		EH	2
BG	2	FG	
BH	2	FH	
CD		GH	2

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	2	CE	2
AC	2	CF	1
AD		CG	1
AE	2	CH	2
AF	1	DE	
AG	1	DF	
AH	2	DG	
BC	2	DH	
BD		EF	1
BE	3	EG	2
BF	1	EH	3
BG	2	FG	
BH	3	FH	1
CD		GH	2

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	2	CE	3
AC	2	CF	2
AD		CG	2
AE	2	CH	3
AF	1	DE	1
AG	1	DF	1
AH	2	DG	1
BC	3	DH	1
BD	1	EF	2
BE	4	EG	3
BF	2	EH	4
BG	3	FG	1
BH	4	FH	2
CD	1	GH	3

SDU SYDDANSK UNIVERSITET

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	3	CE	3
AC	2	CF	2
AD		CG	2
AE	3	CH	3
AF	1	DE	1
AG	1	DF	1
AH	3	DG	1
BC	3	DH	1
BD	1	EF	2
BE	5	EG	3
BF	2	EH	5
BG	3	FG	1
BH	5	FH	2
CD	1	GH	3

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	3	CE	3
AC	2	CF	2
AD		CG	2
AE	3	CH	3
AF	1	DE	1
AG	1	DF	1
AH	3	DG	1
BC	3	DH	1
BD	1	EF	3
BE	6	EG	4
BF	3	EH	6
BG	4	FG	2
BH	6	FH	3
CD	1	GH	4

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	4	CE	3
AC	2	CF	2
AD	1	CG	2
AE	3	CH	3
AF	1	DE	1
AG	2	DF	1
AH	4	DG	2
BC	3	DH	2
BD	2	EF	3
BE	6	EG	4
BF	3	EH	6
BG	5	FG	2
BH	7	FH	3
CD	1	GH	5

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	3
AC	2	CF	2
AD	2	CG	2
AE	3	CH	3
AF	1	DE	1
AG	3	DF	1
AH	4	DG	3
BC	3	DH	2
BD	3	EF	3
BE	6	EG	4
BF	3	EH	6
BG	6	FG	2
BH	7	FH	3
CD	1	GH	5

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	3
AC	2	CF	2
AD	2	CG	2
AE	3	CH	3
AF	1	DE	1
AG	3	DF	2
AH	4	DG	4
BC	3	DH	2
BD	4	EF	3
BE	6	EG	4
BF	4	EH	6
BG	7	FG	3
BH	7	FH	3
CD	1	GH	5

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
→ 10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	4
AC	2	CF	3
AD	2	CG	2
AE	3	CH	3
AF	1	DE	1
AG	3	DF	2
AH	4	DG	4
BC	3	DH	2
BD	4	EF	4
BE	6	EG	4
BF	4	EH	6
BG	7	FG	3
BH	7	FH	3
CD	1	GH	5

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
→ 11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	5	CE	5
AC	3	CF	4
AD	2	CG	2
AE	4	CH	4
AF	2	DE	1
AG	3	DF	2
AH	5	DG	4
BC	3	DH	2
BD	4	EF	5
BE	6	EG	4
BF	4	EH	7
BG	7	FG	3
BH	7	FH	4
CD	1	GH	5

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
→ 12	A B E G

S1

A
B
C
D
E
F
G
H

C2

AB	6	CE	5
AC	3	CF	4
AD	2	CG	2
AE	5	CH	4
AF	2	DE	1
AG	4	DF	2
AH	5	DG	4
BC	3	DH	2
BD	4	EF	5
BE	7	EG	5
BF	4	EH	7
BG	8	FG	3
BH	7	FH	4
CD	1	GH	5

C2 and S2

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S1

A
B
C
D
E
F
G
H

S2

AB	6	CE	5
AC	~~3~~	CF	4
AD	~~2~~	CG	~~2~~
AE	5	CH	4
AF	~~2~~	DE	~~1~~
AG	4	DF	~~2~~
AH	5	DG	4
BC	~~3~~	DH	~~2~~
BD	4	EF	5
BE	7	EG	5
BF	4	EH	7
BG	8	FG	~~3~~
BH	7	FH	4
CD	~~1~~	GH	5

S2 to C3

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

S2 to C3

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

→

C3

ABE	
ABG	
ABH	
AEG	
AEH	
AGH	

S2 to C3

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

→

C3

ABE	BEG
ABG	BEH
ABH	BFG
AEG	BFH
AEH	BGH
AGH	
BDE	
BDF	
BDG	
BDH	
BEF	

S2 to C3

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

ABE	BEG
ABG	BEH
ABH	BFG
AEG	BFH
AEH	BGH
AGH	CEF
BDE	CEH
BDF	CFH
BDG	
BDH	
BEF	

S2 to C3

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

$\longrightarrow$

C3

ABE	BEG
ABG	BEH
ABH	BFG
AEG	BFH
AEH	BGH
AGH	CEF
BDE	CEH
BDF	CFH
BDG	EFG
BDH	EFH
BEF	EGH

S2 to C3

S2

AB	CE
AE	CF
AG	CH
AH	DG
BD	EF
BE	EG
BF	EH
BG	FH
BH	GH

→

C3

ABE	BEG
ABG	BEH
ABH	~~BFG~~
AEG	BFH
AEH	BGH
AGH	CEF
~~BDE~~	CEH
~~BDF~~	CFH
BDG	~~EFG~~
~~BDH~~	EFH
BEF	EGH

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE		BEH	
ABG		BFH	
ABH		BGH	
AEG		CEF	
AEH		CEH	
AGH		CFH	
BDG		EFH	
BEF		EGH	
BEG			

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE		BEH	1
ABG		BFH	
ABH		BGH	1
AEG		CEF	
AEH		CEH	
AGH		CFH	
BDG		EFH	
BEF		EGH	1
BEG	1		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	1	BEH	2
ABG	1	BFH	
ABH	1	BGH	2
AEG	1	CEF	
AEH	1	CEH	1
AGH	1	CFH	
BDG		EFH	
BEF		EGH	2
BEG	2		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	2	BEH	3
ABG	1	BFH	1
ABH	2	BGH	2
AEG	1	CEF	1
AEH	2	CEH	2
AGH	1	CFH	1
BDG		EFH	1
BEF	1	EGH	2
BEG	2		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	2	BEH	4
ABG	1	BFH	2
ABH	2	BGH	3
AEG	1	CEF	2
AEH	2	CEH	3
AGH	1	CFH	2
BDG	1	EFH	2
BEF	2	EGH	3
BEG	3		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	5
ABG	1	BFH	2
ABH	3	BGH	3
AEG	1	CEF	2
AEH	3	CEH	3
AGH	1	CFH	2
BDG	1	EFH	2
BEF	2	EGH	3
BEG	3		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	1	BFH	3
ABH	3	BGH	4
AEG	1	CEF	2
AEH	3	CEH	3
AGH	1	CFH	2
BDG	1	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	2	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	2	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	3	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	4	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	3
AEH	3	CEH	3
AGH	2	CFH	2
BDG	4	EFH	3
BEF	3	EGH	4
BEG	4		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C3

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	4
AEH	4	CEH	4
AGH	2	CFH	3
BDG	4	EFH	4
BEF	3	EGH	4
BEG	4		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
→ 12	A B E G

C3

ABE	4	BEH	6
ABG	4	BFH	3
ABH	4	BGH	5
AEG	2	CEF	4
AEH	4	CEH	4
AGH	2	CFH	3
BDG	4	EFH	4
BEF	3	EGH	4
BEG	5		

C3 and S3

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S3

ABE	4	BEH	6
ABG	4	BFH	~~3~~
ABH	4	BGH	5
AEG	~~2~~	CEF	4
AEH	4	CEH	4
AGH	~~2~~	CFH	~~3~~
BDG	4	EFH	4
BEF	~~3~~	EGH	4
BEG	5		

S3 to C4

S3
ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

$\longrightarrow$

C4

S3 to C4

S3
ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

$\longrightarrow$

C4
ABEG
ABEH
ABGH

S3 to C4

S3
ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

→

C4
ABEG
ABEH
ABGH
BEGH

S3 to C4

S3
ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

$\longrightarrow$

C4
ABEG
ABEH
ABGH
BEGH
CEFH

S3 to C4

S3
ABE
ABG
ABH
AEH
BDG
BEG
BEH
BGH
CEF
CEH
EFH
EGH

→

C4
~~ABEG~~
ABEH
~~ABGH~~
BEGH
~~CEFH~~

AEG, AGH and CFH not *frequent*!

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	
BEGH	

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	
BEGH	1

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	1
BEGH	2

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	2
BEGH	2

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	2
BEGH	3

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	3
BEGH	3

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	3
BEGH	4

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

C4

ABEH	3
BEGH	4

C4 and S4

Database

1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

S4

ABEH	~~3~~
BEGH	4

Only one frequent 4-Itemset remaining.

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI
CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI
CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI
CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH	4

MFI and CFI
CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets

S1

A	7
B	10
C	5
D	4
E	9
F	6
G	8
H	8

S2

AB	6
AE	5
AG	4
AH	5
BD	4
BE	7
BF	4
BG	8
BH	7
CE	5
CF	4
CH	4
DG	4
EF	5
EG	5
EH	7
FH	4
GH	5

S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

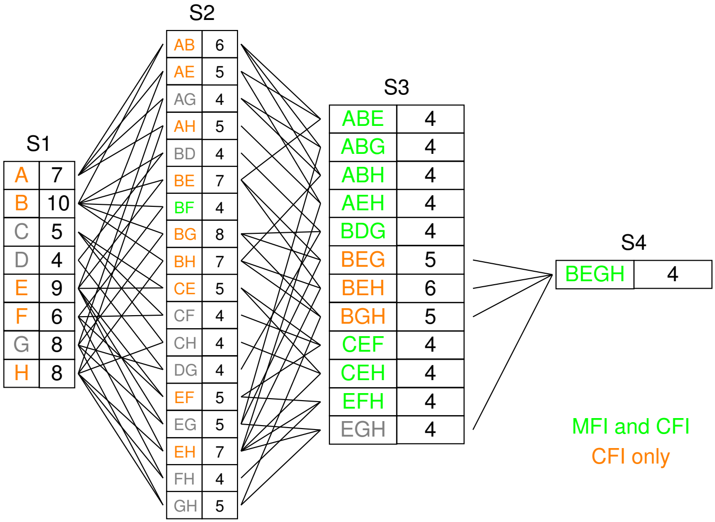
S4

BEGH	4

MFI and CFI

CFI only

Maximal Frequent Itemsets and Closed Frequent Itemsets



MFI and CFI

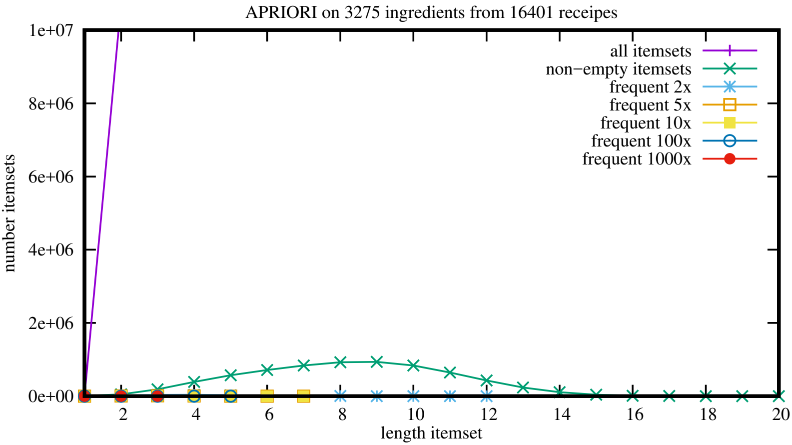
CFI only

Outline

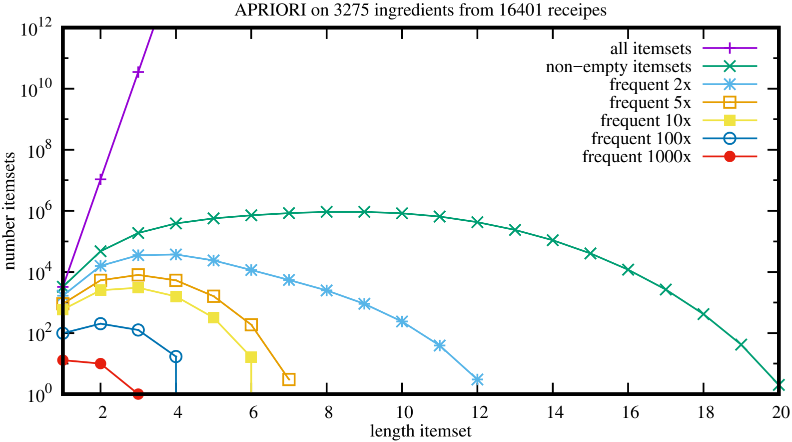
In Theory. . .

- naïve algorithm: count frequency of all k-itemsets, for all $\binom{I}{k}$ k-itemsets, for all k
- number of possible itemsets $0 \leq k \leq |I|$?
- Apriori: one database scan for all frequent k-itemset *candidates* of a given k
- reduction of number of candidates by the anti-monotonicity principle of frequency: generate only candidates that have a chance to be frequent (join of frequent $(k-1)$-itemsets and pruning)

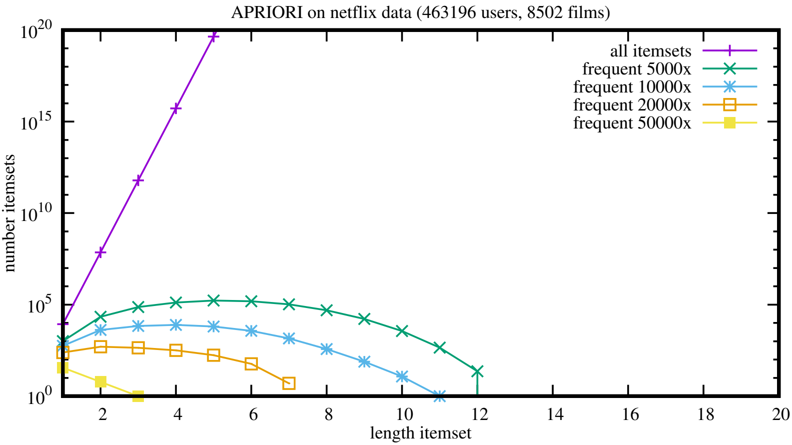
Experiments



Experiments



Experiments



Outline

Definition: Association Rule

association rule: expresses an implication of the form $X \Rightarrow Y$, where X and Y are itemsets, $X \cap Y = \emptyset$

implication: describes a co-occurrence, not a causality

An association rule does not necessarily need to hold in all cases. We can describe its strength (or weakness), based on the observed cases:

support: The support of an association rule in $\mathcal{D}$ is the support of the union of its components:
$s(X \Rightarrow Y) = s(X \cup Y)$

frequency: Analogously, $f(X \Rightarrow Y) = f(X \cup Y)$

confidence: $\mathrm{conf}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$

Problem 2: Association Rule Mining

Given:

- a set of items I
- a transaction database $\mathcal{D}$ over I
- a support threshold σ and a confidence threshold c

Find all association rules $X \Rightarrow Y$ in $\mathcal{D}$ with a support of at least σ and a confidence of at least c, i.e.:

$$\{X \Rightarrow Y | s(X \Rightarrow Y) \geq \sigma \wedge \mathrm{conf}(X \Rightarrow Y) \geq c\}$$

```
T1:  {bread, butter, milk, sugar}
T2:  {butter, flour, milk, sugar}
T3:  {butter, eggs, milk, salt}
T4:  {eggs}
T5:  {butter, flour, milk, salt, sugar}
```

Find Association Rules

This is part of the Apriori algorithm [Srikant and Agrawal, 1996].

for frequent itemset X:

- for each $Y \subset X$, $Y \neq \emptyset$, build the rule $Y \Rightarrow (X \setminus Y)$
- $\mathrm{conf}(Y \Rightarrow (X \setminus Y)) = \frac{s(X)}{s(Y)}$
- delete rules with confidence below a given threshold c

Note that:

For all involved itemsets (X, Y, $(X \setminus Y)$), we have the support from the solution of Problem 1 (stored or reconstructable from closed frequent itemsets). Thus we don't need a single database scan here.

Monotonicity of Confidence

Theorem 2.1

Given:

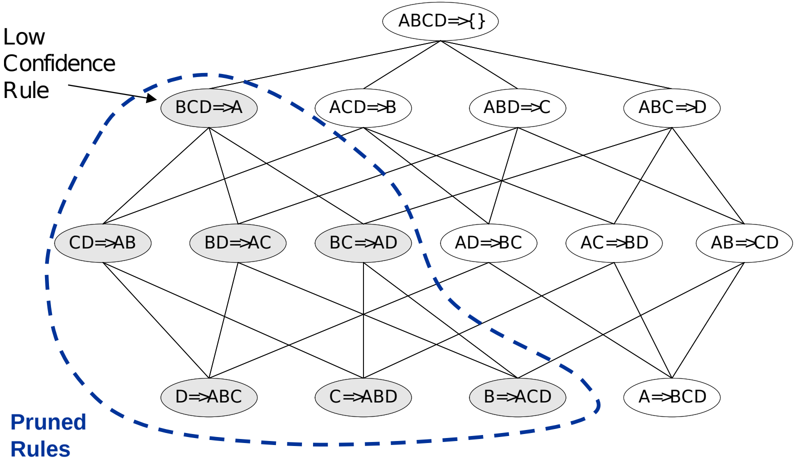
- *itemset* X
- $Y \subset X, Y \neq \emptyset$

If $\operatorname{conf}(Y \Rightarrow (X \setminus Y)) < c$, *then* $\forall Y' \subset Y$*:*

$$\operatorname{conf}(Y' \Rightarrow (X \setminus Y')) < c.$$

This property allows the construction of all association rules satisfying some confidence threshold from all frequent itemsets with a procedure similar to the Apriori construction of frequent itemsets, but without database scan. [Srikant and Agrawal, 1996]

Pruning of Association Rules



Adapted from Tan et al. [2006], Fig. 6.15.

Example: Association Rules

Association rules for BEGH (see slide 88), confidence $\geq 60\%$, support 4:

Antecedent	Consequent	S(Antecedent)	Confidence		Rule
BEGH	$\emptyset$	4		1.000	
BEG	H	5	$4/5 =$	**0.800**	BEG $\Rightarrow$ H
BE	GH	7	$4/7 \approx$	0.571	-
BG	EH	8	$4/8 =$	0.500	-
EG	BH	5	$4/5 =$	**0.800**	EG $\Rightarrow$ BH
BEH	G	6	$4/6 \approx$	**0.667**	BEH $\Rightarrow$ G
BH	EG	7	$4/7 \approx$	0.571	-
EH	BG	7	$4/7 \approx$	0.571	-
BGH	E	5	$4/5 =$	**0.800**	BGH $\Rightarrow$ E
GH	BE	5	$4/5 =$	**0.800**	GH $\Rightarrow$ BE
EGH	B	4	$4/4 =$	**1.000**	EGH $\Rightarrow$ B

Interpretation of Support and Confidence

support: measures the frequency of the item set

- rules with very low support may occur simply by chance
- rules with low support are uninteresting from a business perspective

confidence: measures the reliability of the rule

- $X \Rightarrow Y$ – the higher the confidence, the more likely Y is present in transactions that contain X
- estimate of the conditional probability of Y given X

Limitations of Support and Confidence as Measures

	coffee	no coffee	
tea	150	50	200
no tea	650	150	800
	800	200	1000

$\{tea\} \Rightarrow \{coffee\}$

support?

confidence?

$\{\} \Rightarrow \{coffee\}$

support?

confidence?

Conclusion?

(Discussed by Tan et al. [2006], page 372f., example 6.3.)

Interestingness

Other measures to assess the interestingness of a rule include:

Lift: $\mathit{Lift}(A \Rightarrow B) = \frac{\mathrm{conf}(A \Rightarrow B)}{f(B)}$

Jaccard: $\mathit{Jaccard}(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)+s(B)-s(A \cup B)}$

conviction: $\mathit{conviction}(A \Rightarrow B) = \frac{1-f(B)}{1-\mathrm{conf}(A \Rightarrow B)}$

Recommended Reading:

Advanced reading:

- *Vreeken and Tatti [2014]*

Example of Lift:

- Lift:
  - Lift(I $\rightarrow$ Z) = Conf(I $\rightarrow$ Z) / (Sup(Z) / N)
- Example:
  - Lift(q $\rightarrow$ p) = Conf(q $\rightarrow$ p) / (Sup(p) / N) =

    = 1 / (25/30) = **1.2**
  - Lift(q $\rightarrow$ r) = Conf(q $\rightarrow$ r) / (Sup(r) / N) =

    = 1 / (5/30) = **6**

p	q	r	...
1	1	1	...
1	1	1	...
1	1	1	...
1	1	1	...
1	1	1	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
1	0	0	...
0	0	0	...
0	0	0	...
0	0	0	...
0	0	0	...
0	0	0	...

Frequent Itemsets and Association Rules in R

- There are useful packages to perform frequent itemset and association rule mining in R
- These packages contain a variety of off-the-shelf algorithms and sophisticated analysis tools, including visualization-based tools
- One such package is the `arules` package
- You can learn about this package, e.g., at:
  - CRAN Package Documentation
  - Arules Tutorial by Michael Hahsler

Frequent Itemsets and Association Rules in R

- **Optional Exercise**: Consider the following data set with 10 transactions, represented as a binary matrix where each column (except for the 1st, which contains the transaction ID) stands for an item and the presence or absence of that item in each transaction is denoted by a value 1 or 0, respectively.

ID	Milk	Coffee	Tea	Bread	Butter	Rice	Beans
1	0	1	0	1	1	0	0
2	1	0	1	1	1	0	0
3	0	1	0	1	1	0	0
4	1	1	0	1	1	0	0
5	0	0	1	0	0	0	0
6	0	0	0	0	1	0	0
7	0	0	0	1	0	0	0
8	0	0	0	0	0	0	1
9	0	0	0	0	0	1	1
10	0	0	0	0	0	1	0

Frequent Itemsets and Association Rules in R

- **Optional Exercise**:

This data set is available in a CSV file `10_Groceries_Transactions.csv`. This file has a header with the variable names in the 1st line, the other 10 lines represent the transactions. Each line contains 8 values (columns) separated by comma (","). In this exercise you are asked to:

1. Read the file into a data.frame called `TR_10_Frame` using `read.table()`;
2. Remove the 1st column (Transaction IDs) and represent the remaining 7 columns as a binary matrix called `TR_Matrix`;
3. Convert this matrix into an object of class `transactions` using function `as()`, and name this object `TR_obj`;
4. Inspect and visualise this object using functions `summary()`, `inspect()` and `image()`;
5. Use functions `apriori()` and `inspect()` to generate the rules with minimum relative support of 3/10 and minimum confidence of 9/10.

Outlook: Other Algorithms

- Frequent pattern mining is a large research field, for an overview see the collection of topics edited by Aggarwal and Han [2014].
- Another important algorithm for frequent pattern mining is FP-Growth [Han et al., 2000].
- Frequent patterns have been defined in other application scenarios such as, e.g., graphs, spatiotemporal data, sequential data (for example protein sequences, as in the work of Birzele and Kramer [2006]).
- The principle of anti-monotonicity for pruning has been applied in many other application areas (e.g., in subspace clustering [Zimek et al., 2014]).

References

- M. J. Zaki and W. Meira Jr., “Data Mining and Machine Learning. Fundamental Concepts and Algorithms”. Cambridge University Press, 2nd edition, 2020
- P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining", Addison-Wesley, 2006