Lecture 04: Frequent pattern mining

Meelis Kull
meelis.kull@ut.ee
Autumn 2019
Lecture 04 – Frequent pattern mining

• Frequent patterns
• Frequent itemsets and association rules
• Frequent itemset mining and Apriori algorithm
• Compact representations of itemsets
• Association rule mining
• Frequent itemsets in non-transactional data
• Pattern interestingness measures
Lecture 04 – Frequent pattern mining

- **Frequent patterns**
- Frequent itemsets and association rules
- Frequent itemset mining and Apriori algorithm
- Compact representations of itemsets
- Association rule mining
- Frequent itemsets in non-transactional data
- Pattern interestingness measures
Frequent patterns

Humpty Dumpty sat on a wall,
Humpty Dumpty had a great fall;
All the King's horses and all the King's men
Couldn't put Humpty together again

Adapted from slides by Konstantin Tretyakov
Frequent patterns

Humpty Dumpty sat on a wall,
Humpty Dumpty had a great fall;
All the King's horses and all the King's men
Couldn't put Humpty together again

Adapted from slides by Konstantin Tretyakov
Frequent patterns

Humpty Dumpty sat on a wall,
Humpty Dumpty had a great fall;
All the King's horses and all the King's men
Couldn't put Humpty together again

Adapted from slides by Konstantin Tretyakov
Frequent patterns

Humpty Dumpty sat on a wall,
Humpty Dumpty had a great fall;
All the King's horses and all the King's men
Couldn't put Humpty together again

[H|D]umpty

Adapted from slides by Konstantin Tretyakov
Frequent patterns

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Bread, Milk</strong></td>
</tr>
<tr>
<td>2</td>
<td><strong>Bread, Diaper, Beer, Eggs</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>Milk, Diaper, Beer, Coke</strong></td>
</tr>
<tr>
<td>4</td>
<td><strong>Bread, Milk, Diaper, Beer</strong></td>
</tr>
<tr>
<td>5</td>
<td><strong>Bread, Milk, Diaper, Coke</strong></td>
</tr>
</tbody>
</table>

Adapted from slides by Konstantin Tretyakov
Frequent patterns

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

Adapted from slides by Konstantin Tretyakov
Frequent patterns

Adapted from slides by Konstantin Tretyakov
Frequent patterns

Adapted from slides by Konstantin Tretyakov
Frequent patterns

[Diagram showing various paths through a website, including:
- Front page → Search → Product A → Product B → Product C
- Front page → Ad → Product B → Purchase → Product C
- Front page → Front page → Feedback
- Front page → Search → Product B → Search → Product C → Purchase
- Front page → Ad → Product C → Search → Product B → Purchase
- Front page → Front page → Front page → Front page]

Adapted from slides by Konstantin Tretyakov
Why frequent patterns

• Repetition in data always indicates **structure**.

• This structure may be due **known processes**.

• Otherwise, we want to **know about it and explain it**.

• Consequently, **searching for frequent patterns** in data is one of the most basic procedures used in descriptive data analysis.

Adapted from slides by Konstantin Tretyakov
Is a frequent pattern always interesting?

A. Yes

B. No

C. Not sure
Is a frequent pattern always interesting?

• No
  – e.g. “Front page” is frequent but uninteresting
Is an interesting pattern always frequent?

A. Yes  
B. No  
C. Not sure
Is an interesting pattern always frequent?

• No
  – e.g. “Diapers & Beer” may be a rare but a very valuable pattern
Frequent pattern mining

• Finding frequent patterns:
  – Primarily an algorithmic problem

• Finding “interesting” patterns:
  – Primarily a statistical problem
Types of frequent patterns

• In general, we may be interested in different types of data and types of patterns:

  – Natural language / Word sequences
  – Text / Regular expression patterns
  – Web log / Event sequences (various models)
  – Purchases / Item sets
  – …

Today

(however, many notions and ideas apply to all data / pattern types).

Adapted from slides by Konstantin Tretyakov
Frequent patterns

- **Frequent itemsets and association rules**
- Frequent itemset mining and Apriori algorithm
- Compact representations of itemsets
- Association rule mining
- Frequent itemsets in non-transactional data
- Pattern interestingness measures
Definitions

Let the data consist of transactions (i.e. sets of items)

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

- The patterns we are interested are *itemsets*.
  - E.g. \{Milk, Eggs\} is an itemset.
  - NB! Here we are looking at sets, therefore *order does not matter* and \{Milk, Eggs\} is the same as \{Eggs, Milk\}

Adapted from slides by Konstantin Tretyakov
Definitions

The **support count** of an itemset is the number of transactions where it occurs

- \( \text{support\_count}(\{\text{Milk, Eggs}\}) = 0 \)
- \( \text{support\_count}(\{\text{Bread}\}) = 4 \)
- \( \text{support\_count}(\{\text{Milk, Beer}\}) = 2 \)

Adapted from slides by Konstantin Tretyakov
Definitions

The **support** of an itemset is the proportion of transactions where it occurs.

- \[ \text{support}\{\{\text{Milk, Eggs}\}\} = 0 \]
- \[ \text{support}\{\{\text{Bread}\}\} = \frac{4}{5} = 0.8 \]
- \[ \text{support}\{\{\text{Milk, Beer}\}\} = \frac{2}{5} = 0.4 \]

Adapted from slides by Konstantin Tretyakov
Support count of Milk is

A. 1
B. 2
C. 3
D. 4  ✔
E. 5

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>
Definitions

An itemset is **frequent** if its support is greater than or equal to some predefined threshold $s_{\text{min}}$.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

Let $s_{\text{min}} = 3$. Find all frequent itemsets.

{Bread}, {Milk}, {Diaper}, {Beer}, {Bread, Milk}, {Bread, Diaper}, {Milk, Diaper}, {Diaper, Beer}, {}.

Adapted from slides by Konstantin Tretyakov
Explanations for itemsets

- Suppose we found that \{Beer, Milk, Diapers\} is a frequent itemset.
- How do we explain it?
- Perhaps it is frequent due to the fact that there is a causal relationship
  \{Diapers, Milk\} => Beer
- We cannot claim causality based on transaction data, but we can find association rules:
  \{Diapers, Milk\} -> Beer
  - Often when people purchased Diapers and Milk they also purchased Beer

Adapted from slides by Konstantin Tretyakov
Association rules

• **An association rule** is an implication of the form $X \rightarrow Y$, where $X$ and $Y$ are itemsets. $X$ – antecedent; $Y$ - consequent

• The **support of a rule** is just the support of $X \cup Y$.
  \[
  \text{support}(X \rightarrow Y) = \text{support}(X \cup Y)
  \]

• The **confidence** of a rule is the proportion of transactions satisfying the right part among the transactions, which satisfy the left.
  \[
  \text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}
  \]

Adapted from slides by Konstantin Tretyakov
Confidence of Milk→Bread is

A. 1/5
B. 1/3
C. 3/4
D. 4/5
E. 1

\[
\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}
\]

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>
Association rule mining

Given a set of transactions, find all rules $X \rightarrow Y$, such that

- $\text{support}(X \rightarrow Y) \geq s_{\text{min}}$
- $\text{confidence}(X \rightarrow Y) \geq c_{\text{min}}$

Solution:

- Find frequent itemset $A$ with support $\geq s_{\text{min}}$
- Then find a partitioning of $A$ into left and right part, so that the resulting rule has high confidence.

- The algorithmic approach is the same in both steps. We’ll start with the first one.

Adapted from slides by Konstantin Tretyakov
Lecture 04 – Frequent pattern mining

✓ Frequent patterns
✓ Frequent itemsets and association rules
  • **Frequent itemset mining and Apriori algorithm**
  • Compact representations of itemsets
  • Association rule mining
  • Frequent itemsets in non-transactional data
  • Pattern interestingness measures
Frequent itemset mining

• Brute force approach:
  – Enumerate all possible sets of items
    • For each set compute its support in the database
    • Output sets with support $\geq s_{\text{min}}$

Adapted from slides by Konstantin Tretyakov
Frequent itemset mining

• Brute force approach:
  – Enumerate all possible sets of items
    • For each set compute its support in the database
    • Output sets with support $\geq s_{\text{min}}$

  $2^d$ iterations

• Let there be $d$ different items, $n$ transactions, average transaction size $w$.
What is the complexity of this algorithm?

$O(2^d nw)$

Adapted from slides by Konstantin Tretyakov
Faster itemset mining

- **Apriori**: Avoid scanning through all \(2^d\) itemsets.

Adapted from slides by Konstantin Tretyakov
The Apriori idea

• Suppose that
  \( \text{support}(\{A, C\}) = 42/100 \)

• It follows that
  \( \text{support}(\{A, B, C\}) \leq 42/100 \)

Adapted from slides by Konstantin Tretyakov
Anti-monotonicity of support

In general,

\[ X \subseteq Y \Rightarrow \text{support}(X) \geq \text{support}(Y) \]

it follows that:

- If an itemset is **not frequent**, all of its **supersets** are also not frequent.
- If an itemset is **frequent**, all of its **subsets** are **also frequent**.

Adapted from slides by Konstantin Tretyakov
Apriori principle

The origin of the name ‘Apriori’: if a set is infrequent than the algorithm concludes *a priori* (from the earlier) that its supersets are also infrequent.

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

• First generate frequent 1-sets,

• Next, generate frequent 2-sets from 1-sets,

• … then generate frequent 3-sets from 2-sets,

• … etc, until there are no frequent \( k \)-sets

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

• First generate frequent 1-sets,
  – Simply count the frequency of each item and leave only the frequent ones.

<table>
<thead>
<tr>
<th>Item</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A}</td>
<td>10</td>
</tr>
<tr>
<td>{B}</td>
<td>15</td>
</tr>
<tr>
<td>{C}</td>
<td>3</td>
</tr>
<tr>
<td>{D}</td>
<td>4</td>
</tr>
<tr>
<td>{E}</td>
<td>6</td>
</tr>
<tr>
<td>{F}</td>
<td>10</td>
</tr>
<tr>
<td>{G}</td>
<td>4</td>
</tr>
</tbody>
</table>

Let min support count = 5

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

• Next, generate frequent 2-sets. This is done in several steps.

1. Generate candidate 2-sets

\{A\}: 10
\{B\}: 15
\{E\}: 6
\{F\}: 10

\{A, B\}
\{A, E\}
\{A, F\}
\{B, E\}
\{B, F\}
\{E, F\}

All subsets of a candidate set must be frequent. For 2-sets it simply means that both elements are from frequent 1-sets.

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

- Next, generate frequent 2-sets. This is done in several steps.

2. Count actual support count for each candidate

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>15</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
</tr>
<tr>
<td>A,B</td>
<td>10</td>
</tr>
<tr>
<td>A,E</td>
<td>3</td>
</tr>
<tr>
<td>A,F</td>
<td>5</td>
</tr>
<tr>
<td>B,E</td>
<td>6</td>
</tr>
<tr>
<td>B,F</td>
<td>10</td>
</tr>
<tr>
<td>E,F</td>
<td>5</td>
</tr>
</tbody>
</table>

(Requires a full pass over the transaction database for each candidate)

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

• Next, generate frequent 2-sets. This is done in several steps.

3. … and leave only actually frequent ones

\[
\begin{array}{c|c}
\{A\} & 10 \\
\{B\} & 15 \\
\{E\} & 6 \\
\{F\} & 10 \\
\{A, B\} & 10 \\
\{A, E\} & 3 \\
\{A, F\} & 5 \\
\{B, E\} & 6 \\
\{B, F\} & 10 \\
\{E, F\} & 5 \\
\end{array}
\]

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

• Now we have frequent 1- and 2-sets.

In many practical situations the algorithm stops here.

1. Because there are so many items that enumerating beyond 2-sets is impractical.
2. Because knowing frequent 2-sets is already useful enough (think of the “Beer/Diapers” example)

\[
\begin{align*}
\{A\} & : 10 \\
\{B\} & : 15 \\
\{E\} & : 6 \\
\{F\} & : 10 \\
\{A,B\} & : 10 \\
\{A,F\} & : 5 \\
\{B,E\} & : 6 \\
\{B,F\} & : 10 \\
\{E,F\} & : 5 \\
\end{align*}
\]

Adapted from slides by Konstantin Tretyakov
But let’s try generating frequent 3-sets. We proceed as before.

1. Generate candidate 3-sets

We augment each 2-set with an additional element and check that all 2-subsets of the resulting set are frequent.

Adapted from slides by Konstantin Tretyakov

This can be optimized somewhat, see, e.g:
http://www.dais.unive.it/~orlando/PAPERS/dawak01.pdf
Basic Apriori algorithm

- But let’s try generating frequent 3-sets. We proceed as before.

2. Count actual support count

\[
\begin{align*}
\{A\} &: 10 \\
\{B\} &: 15 \\
\{E\} &: 6 \\
\{F\} &: 10 \\
\{A,B\} &: 10 \\
\{A,F\} &: 5 \\
\{B,E\} &: 6 \\
\{B,F\} &: 10 \\
\{E,F\} &: 5 \\
\{A,B,F\} &: 4 \\
\{B,E,F\} &: 5 \\
\end{align*}
\]

(Again, a pass over the whole DB)

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

• But let’s try generating frequent 3-sets. We proceed as before.

\{A\}: 10
\{B\}: 15
\{E\}: 6
\{F\}: 10

3. … and throw away the non-frequent ones

\{A,B\}: 10
\{A,F\}: 5
\{B,E\}: 6
\{B,F\}: 10
\{E,F\}: 5

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

In this particular example, can there be frequent 4-sets?

\[
\begin{align*}
\{A\} &: 10 \\
\{B\} &: 15 \\
\{E\} &: 6 \\
\{F\} &: 10 \\
\{A, B\} &: 10 \\
\{A, F\} &: 5 \\
\{B, E\} &: 6 \\
\{B, F\} &: 10 \\
\{E, F\} &: 5 \\
\{B, E, F\} &: 5
\end{align*}
\]

Adapted from slides by Konstantin Tretyakov
In this particular example, can there be frequent 4-sets?

A. Yes

✓ B. No

C. Not sure

All frequent sets up to size 3:

{A}: 10
{B}: 15
{E}: 6
{F}: 10
{B,E,F}: 5
In this particular example, can there be frequent 4-sets?

No, because if, say, \{B, E, F, X\} is frequent, then \{B, E, X\} must be frequent too!

\[
\begin{align*}
\{A\} & : 10 \\
\{B\} & : 15 \\
\{E\} & : 6 \\
\{F\} & : 10 \\
\{A, B\} & : 10 \\
\{A, F\} & : 5 \\
\{B, E\} & : 6 \\
\{B, F\} & : 10 \\
\{E, F\} & : 5 \\
\{B, E, F\} & : 5 \\
\end{align*}
\]

Adapted from slides by Konstantin Tretyakov
Basic Apriori algorithm

• Generate all candidate 1-sets
• Discard 1-sets with support below $s_{\text{min}}$
• For $k = 2, 3, ...$
  – Generate all candidate $k$-sets from frequent $(k - 1)$-sets
  – Discard $k$-sets with an infrequent $(k - 1)$-subset
  – Discard $k$-sets with support below $s_{\text{min}}$
Compare to naïve algorithm

- For \( k=1,2,3,\ldots \)
  - Generate all candidate \( k \)-sets
  - Discard \( k \)-sets with support below \( S_{\text{min}} \)
Many optimizations are possible

• Avoid generating the separate candidate set explicitly (do it on-the-fly while counting).
• Store \(k\)-sets in a hash tree data structure, (speeds up the counting/generation process).
• Use only a part of the whole transaction database (sample or partition)
• Use Bloom-filter like data structures to reduce candidate set.
  – Bloom filter: probabilistic data structure to represent sets; replies to queries either “possibly in set” or “definitely not in set”

Adapted from slides by Konstantin Tretyakov (see, e.g. http://i.stanford.edu/~ullman/mmds/ch6a.pdf)
Lecture 04 – Frequent pattern mining

✓ Frequent patterns
✓ Frequent itemsets and association rules
✓ Frequent itemset mining and Apriori algorithm

• **Compact representations of itemsets**
• Association rule mining
• Frequent itemsets in non-transactional data
• Pattern interestingness measures
Compact representation of frequent itemsets

If \( \{A, B, C, D, E\} \) is frequent, then also those sets are frequent:

\[
\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \\
\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \\
\{B, D\}, \{B, E\}, \{C, D\}, \{C, E\}, \{D, E\}, \\
\{A, B, C\}, \{A, B, D\}, \{A, B, E\}, \{A, C, D\}, \{A, C, E\}, \\
\{A, D, E\}, \{B, C, D\}, \{B, C, E\}, \{B, D, E\}, \{C, D, E\}, \\
\{A, B, C, D\}, \{A, B, C, E\}, \{A, B, D, E\}, \\
\{A, C, D, E\}, \{B, C, D, E\}
\]

Do we really want our algorithm to report those?

Adapted from slides by Konstantin Tretyakov
Maximal frequent itemsets

An itemset is maximal frequent if none of its immediate supersets is frequent.

© Tan, Steinbach, Kumar
Introduction to Data Mining
Maximal frequent itemsets

• Why is maximality useful?
• Frequent itemsets = Subsets of maximal frequent itemset

• However, if $s_{\text{min}}$ is changed, then need to find maximal frequent itemsets again
Closed itemsets

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

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{A,B}</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>{B,C,D}</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>{A,B,C,D}</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>{A,B,D}</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>{A,B,C,D}</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A}</td>
<td>4</td>
</tr>
<tr>
<td>{B}</td>
<td>5</td>
</tr>
<tr>
<td>{C}</td>
<td>3</td>
</tr>
<tr>
<td>{D}</td>
<td>4</td>
</tr>
<tr>
<td>{A,B}</td>
<td>4</td>
</tr>
<tr>
<td>{A,C}</td>
<td>2</td>
</tr>
<tr>
<td>{A,D}</td>
<td>3</td>
</tr>
<tr>
<td>{B,C}</td>
<td>3</td>
</tr>
<tr>
<td>{B,D}</td>
<td>4</td>
</tr>
<tr>
<td>{C,D}</td>
<td>3</td>
</tr>
</tbody>
</table>

Closed itemsets highlighted with yellow

© Tan, Steinbach, Kumar

Introduction to Data Mining
Note that Tan, Steinbach, Kumar have used 'support' to stand for 'support count'. We will still use 'support count' here.

Closed itemsets highlighted with yellow

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
<th>Support count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{A,B}</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>{B,C,D}</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>{A,B,C,D}</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>{A,B,D}</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>{A,B,C,D}</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support count</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A,B,C}</td>
<td>2</td>
</tr>
<tr>
<td>{A,B,D}</td>
<td>3</td>
</tr>
<tr>
<td>{A,C,D}</td>
<td>2</td>
</tr>
<tr>
<td>{B,C,D}</td>
<td>3</td>
</tr>
<tr>
<td>{A,B,C,D}</td>
<td>2</td>
</tr>
</tbody>
</table>
Closed itemsets

• Closed itemsets = Maximal frequent itemsets across all $s_{min}$

• Why is closeness useful?

• Finding all frequent itemsets for a fixed $s_{min}$:
  – Frequent itemsets = All closed itemsets with support at least $s_{min}$ and all their subsets
Can an itemset be closed and not maximal

A. Yes
B. No
C. Not sure
Quiz

• Can an itemset be:
  – **Closed and not Maximal? Yes.**
    • E.g. adding any item will reduce support, but adding some items will still make a frequent set.
Can an itemset be not closed and maximal

A. Yes
B. No
C. Not sure

Response Counter
Quiz

• Can an itemset be:
  – Closed and not Maximal? Yes.
    • E.g. adding any item will reduce support, but adding some items will still make a frequent set.
  – Not closed and Maximal? No.
    • Not closed, hence we can add some other item without reducing support.
Can an itemset be closed and maximal

A. Yes
B. No
C. Not sure
Quiz

• Can an itemset be:
  – Closed and not Maximal? Yes.
    • E.g. adding any item will reduce support, but adding some items will still make a frequent set.
  – Not closed and Maximal? No.
    • Not closed, hence we can add some other item without reducing support.
  – **Closed and Maximal? Yes.**
    • Adding any item will reduce support and make the set infrequent.

Adapted from slides by Konstantin Tretyakov
Closed vs maximal frequent itemsets

- Frequent Itemsets
  - Closed Frequent Itemsets
    - Maximal Frequent Itemsets
Intermediate summary

• **Frequent itemsets** are interesting because those correspond to **structure in the data**.

• **Association rules** identify predictive relationships within frequent itemsets

• **Apriori-like algorithms** search for frequent sets better than brute-force.

• It is sufficient to find only **maximal itemsets** (or **closed itemsets**, if some flexibility in changing cutoff later is needed).

Adapted from slides by Konstantin Tretyakov
Lecture 04 – Frequent pattern mining

✓ Frequent patterns
✓ Frequent itemsets and association rules
✓ Frequent itemset mining and Apriori algorithm
✓ Compact representations of itemsets
  • **Association rule mining**
  • Frequent itemsets in non-transactional data
  • Pattern interestingness measures
Recall association rule mining:

- **Problem:** find association rules $X \rightarrow Y$ with
  - Support at least $S_{\text{min}}$
  - Confidence at least $C_{\text{min}}$
- **Solution:**
  - Find frequent itemsets with support at least $S_{\text{min}}$
  - For each itemset find a split into $X \rightarrow Y$, which ensures required confidence.

Adapted from slides by Konstantin Tretyakov
Rule generation

• Suppose we found that \( \{A, B, C, D\} \) is a frequent itemset with the necessary support.

• We can make a variety of rules from it:
  - \( \{A\} \rightarrow \{B, C, D\} \)
  - \( \{B\} \rightarrow \{A, C, D\} \)
  - ...
  - \( \{A, B\} \rightarrow \{C, D\} \)
  - ...
  - \( \{B, C, D\} \rightarrow \{A\} \)

• How to efficiently find those which satisfy the confidence threshold?

Adapted from slides by Konstantin Tretyakov
Rule generation

• Recall that

\[
\text{confidence} \left( \{A\} \rightarrow \{B, C, D\} \right) = \frac{\text{support} \left( \{A, B, C, D\} \right)}{\text{support} \left( \{A\} \right)}
\]

\[
\text{confidence} \left( \{A, B\} \rightarrow \{C, D\} \right) = \frac{\text{support} \left( \{A, B, C, D\} \right)}{\text{support} \left( \{A, B\} \right)}
\]

\[
\text{confidence} \left( \{A, C, D\} \rightarrow \{B\} \right) = \frac{\text{support} \left( \{A, B, C, D\} \right)}{\text{support} \left( \{A, C, D\} \right)}
\]
Rule generation

• Recall that

\[ \text{confidence}(\{A\} \rightarrow \{B, C, D\}) = \frac{\text{support}(\{A, B, C, D\})}{\text{support}(\{A\})} \]

• Consequently, among all rules built on the set \( \{A, B, C, D\} \), confidence is inverse proportional to the support of antecedent (the left part of the rule).

• I.e. confidence is monotonic with respect to antecedent (and anti-monotonic with respect to consequent).

Adapted from slides by Konstantin Tretyakov
Rule generation

• Suppose we are converting an Itemset into a Rule = Antecedent → Consequent

• We want to find rules with confidence(Rule) ≥ \( c_{\text{min}} \)

• This is equivalent to \( \frac{\text{support(Itemset)}}{\text{support(Antecedent)}} \) ≥ \( c_{\text{min}} \)

• This is equivalent to \( \text{support(Antecedent)} \leq \frac{\text{support(Itemset)}}{c_{\text{min}}} \)
Rule generation

In other words, you can use Apriori for rule generation from a found frequent set.
Rule generation

• When generating rules from an itemset:
  – Find all subsets with support below a threshold
  – This is inverse to the frequent itemset task

• Apply apriori algorithm (in reverse):
  – Start from the full subset as antecedent and empty consequent
  – Gradually move items from antecedent to consequent
Basic Apriori for generating all high-confidence rules from an itemset

- Generate all 1-rules with consequent size 1
- Discard 1-rules with confidence below $c_{\text{min}}$
- For $k = 2, 3, ...$
  - Generate all candidate $k$-rules from confident $(k - 1)$-rules by combining consequents
  - Discard $k$-rules with a non-confident $(k - 1)$-sub-rule (with respect to consequent sets)
  - Discard $k$-rules with confidence below $c_{\text{min}}$
Lecture 04 – Frequent pattern mining

✓ Frequent patterns
✓ Frequent itemsets and association rules
✓ Frequent itemset mining and Apriori algorithm
✓ Compact representations of itemsets
✓ Association rule mining

• **Frequent itemsets in non-transactional data**
• Pattern interestingness measures
Frequent itemsets in non-transactional data

- Many non-transactional datasets can be converted to transactional

- For example:

<table>
<thead>
<tr>
<th>Person ID</th>
<th>Age</th>
<th>Gender</th>
<th>Number of children</th>
</tr>
</thead>
<tbody>
<tr>
<td>48472</td>
<td>30-40</td>
<td>Female</td>
<td>3</td>
</tr>
<tr>
<td>93719</td>
<td>10-20</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td>40184</td>
<td>0-10</td>
<td>Female</td>
<td>0</td>
</tr>
</tbody>
</table>

- Can be converted into:

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>48472</td>
<td>Age_30_40, Female, Has_3_children</td>
</tr>
<tr>
<td>93719</td>
<td>Age_10_20, Male, Has_1_child</td>
</tr>
<tr>
<td>40184</td>
<td>Age_below_10, Female, Has_no_children</td>
</tr>
</tbody>
</table>

- Example rule: \( \{\text{Age\_below\_10}\} \rightarrow \text{Has\_no\_children} \)
Lecture 04 – Frequent pattern mining

✓ Frequent patterns
✓ Frequent itemsets and association rules
✓ Frequent itemset mining and Apriori algorithm
✓ Compact representations of itemsets
✓ Association rule mining
✓ Frequent itemsets in non-transactional data

• Pattern interestingness measures
Found the frequent patterns, what next?

• Are the discovered patterns useful for the business?
  Often: “Alas! Too many patterns!”

• Need to rank patterns by interestingness

• How to measure interestingness?
  – Depends on the business
    (depends on the goal in the application domain)
Rank patterns by interestingness
Pattern interestingness

• How to measure interestingness?
  – Depends on the business
    (depends on the goal in the application domain)
Example application scenarios for association rules

- Client has product A (e.g. cereals) in the basket of an online grocery store

  - Scenario 1 (Prediction):
    - Show 3 products that are most likely to be added

  - Scenario 2 (Recommendation):
    - Recommend 3 products commonly purchased together with product A

  - Scenario 3 (Combined discounts):
    - Offer 3 products with a discount as the client might consider these if discounted
Example dataset

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Butter, Cereals, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Apple, Bread, Cereals, Milk</td>
</tr>
<tr>
<td>3</td>
<td>Bread, Cereals</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Butter, Jam</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Jam</td>
</tr>
</tbody>
</table>

Suppose this is the database of our past client transactions. Now we have a client with Cereals in the shopping basket.

Let us convert into a different format:

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Scenario 1: Prediction

- Show 3 products that are most likely to be added

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- \{Cereals\} -> \{Bread\} \hspace{2cm} \text{confidence: 3/3}
- \{Cereals\} -> \{Milk\} \hspace{2cm} \text{confidence: 2/3}
- \{Cereals\} -> \{Apple\} \hspace{2cm} \text{confidence: 1/3}
Scenario 2: Recommendation

- Recommend 3 products commonly purchased together with Cereals

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- \{Cereals\} \rightarrow \{Bread\} ?
- **No point in recommending Bread, because everyone buys Bread**
Scenario 2: Recommendation

- Recommend 3 products commonly purchased together with Cereals

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- Should predict products which become more likely given that Cereals are in the basket
## Scenario 2: Recommendation

- Recommend 3 products commonly purchased together with Cereals

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### confidence({}->…)
- 1.0
- 0.4
- 0.2
- 0.4
- 0.4

- Should predict products which become more likely given that Cereals are in the basket
Scenario 2: Recommendation

- Recommend 3 products commonly purchased together with Cereals

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

confidence({})-→...

|   | confidence({})-→...
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

confidence({Cereals})-→...

|   | confidence({Cereals})-→...
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0.667</td>
</tr>
<tr>
<td>3</td>
<td>0.333</td>
</tr>
<tr>
<td>4</td>
<td>0.333</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- Should predict products which become more likely given that Cereals are in the basket
New notation and new measure

- $P(X)$ – support of itemset $X$
- Confidence of rule $X \rightarrow Y$:
  $$\text{confidence}(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X)}$$
- New measure ‘lift’:
  $$\text{lift}(X \rightarrow Y) = \frac{\text{confidence}(X \rightarrow Y)}{\text{confidence}(\emptyset \rightarrow Y)} = \frac{P(X \cup Y)}{P(X)P(Y)}$$
- Lift is related to statistical independence:
  - $\text{lift}(X \rightarrow Y) > 1$  $X, Y$ positively correlated
  - $\text{lift}(X \rightarrow Y) = 1$  $X, Y$ statistically independent
  - $\text{lift}(X \rightarrow Y) < 1$  $X, Y$ negatively correlated
Scenario 2: Recommendation

- Recommend 3 products commonly purchased together with Cereals

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>confidence({}-{})</th>
<th>confidence({Cereals}-{})</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>confidence({}-{})</td>
<td>1.0</td>
<td>0.4</td>
<td>1.0</td>
</tr>
<tr>
<td>confidence({Cereals}-{})</td>
<td>1.0</td>
<td>0.667</td>
<td>1.667</td>
</tr>
<tr>
<td>Lift</td>
<td>1.0</td>
<td>1.667</td>
<td>0.833</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.333</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.333</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Scenario 2: Recommendation

• Recommend 3 products commonly purchased together with Cereals

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- \{Cereals\} \rightarrow \{Milk\} \quad \text{lift: } \frac{2/3}{2/5} = \frac{5}{3}
- \{Cereals\} \rightarrow \{Apple\} \quad \text{lift: } \frac{1/3}{1/5} = \frac{5}{3}
- \{Cereals\} \rightarrow \{Bread\} \quad \text{lift: } \frac{3/3}{5/5} = \frac{1}{1}
Scenario 3: Combined discounts

- Offer 3 products with a discount as the client might consider these if discounted

<table>
<thead>
<tr>
<th></th>
<th>Cereals</th>
<th>Bread</th>
<th>Milk</th>
<th>Apple</th>
<th>Butter</th>
<th>Jam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- \{Cereals\} $\rightarrow$ \{Milk\} perhaps people buy milk anyways?
- \{Cereals\} $\rightarrow$ \{Apple\} is this a better rule?
- Not clear what measure to use here
- Need more domain knowledge
Take-home message

• Pattern evaluation measure should be derived based on the domain knowledge

• Next slide has some known measures
# Interestness measures for association patterns

<table>
<thead>
<tr>
<th>#</th>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\phi$-coefficient</td>
<td>$\frac{P(A, B) - P(A)P(B)}{\sqrt{(P(A)P(B)(1-P(A))(1-P(B)))}}$</td>
</tr>
<tr>
<td>2</td>
<td>Goodman-Kruskal's ($\lambda$)</td>
<td>$\frac{\sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$</td>
</tr>
<tr>
<td>3</td>
<td>Odds ratio ($\alpha$)</td>
<td>$\frac{P(A, B)P(A, B)}{P(A)P(B)}$</td>
</tr>
<tr>
<td>4</td>
<td>Yule's $Q$</td>
<td>$\frac{P(A, B)P(\overline{A}B) - P(A, \overline{B})P(\overline{A}, B)}{P(A, B) + P(A, \overline{B})P(\overline{A}, B)} = \frac{\alpha - 1}{\alpha + 1}$</td>
</tr>
<tr>
<td>5</td>
<td>Yule's $Y$</td>
<td>$\frac{\sqrt{P(A, B)P(\overline{A}, B)} - \sqrt{P(A, \overline{B})P(\overline{A}, B)}}{\sqrt{P(A, B)P(\overline{A}, B)} + \sqrt{P(A, \overline{B})P(\overline{A}, B)}} = \frac{\sqrt{\alpha - 1}}{\sqrt{\alpha + 1}}$</td>
</tr>
<tr>
<td>6</td>
<td>Kappa ($\kappa$)</td>
<td>$\frac{1 - P(A)P(B) - P(\overline{A})P(\overline{B})}{\sum_i \sum_j P(A_i, B_j)\log P(A_i)P(B_j)}$</td>
</tr>
<tr>
<td>7</td>
<td>Mutual Information ($I(M)$)</td>
<td>$\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))$</td>
</tr>
<tr>
<td>8</td>
<td>J-Measure ($J$)</td>
<td>$\max(P(A, B)\log\left(\frac{P(B</td>
</tr>
<tr>
<td>9</td>
<td>Gini index ($G$)</td>
<td>$\max\left(\frac{P(A)[P(B</td>
</tr>
<tr>
<td>10</td>
<td>Support ($s$)</td>
<td>$P(A, B)$</td>
</tr>
<tr>
<td>11</td>
<td>Confidence ($c$)</td>
<td>$\max(P(B</td>
</tr>
<tr>
<td>12</td>
<td>Laplace ($L$)</td>
<td>$\max\left(\frac{NP(A, B) + 1}{NP(A) + 2}, \frac{NP(A, B) + 1}{NP(B) + 2}\right)$</td>
</tr>
<tr>
<td>13</td>
<td>Conviction ($V$)</td>
<td>$\max\left(\frac{P(A)P(B)}{P(AB)}, \frac{P(B)P(A)}{P(BA)}\right)$</td>
</tr>
<tr>
<td>14</td>
<td>Interest ($I$)</td>
<td>$\frac{P(A, B)^2}{P(AB)}$</td>
</tr>
<tr>
<td>15</td>
<td>cosine ($IS$)</td>
<td>$\sqrt{P(A)P(B)}$</td>
</tr>
<tr>
<td>16</td>
<td>Piatetsky-Shapiro's ($PS$)</td>
<td>$P(A, B) - P(A)P(B)$</td>
</tr>
<tr>
<td>17</td>
<td>Certainty factor ($F$)</td>
<td>$\max\left(\frac{P(B</td>
</tr>
<tr>
<td>18</td>
<td>Added Value ($AV$)</td>
<td>$\max(P(B</td>
</tr>
<tr>
<td>19</td>
<td>Collective strength ($S$)</td>
<td>$\frac{P(A, B) + P(\overline{A})P(\overline{B})}{P(A)P(B)} \times \frac{1 - P(A)P(B) - P(\overline{A})P(\overline{B})}{P(AB)}$</td>
</tr>
<tr>
<td>20</td>
<td>Jaccard ($\zeta$)</td>
<td>$\frac{P(A) + P(B) - P(A, B)}{\sqrt{P(A, B)\max(P(B</td>
</tr>
<tr>
<td>21</td>
<td>Klosgen ($K$)</td>
<td>$\sqrt{P(A, B)\max(P(B</td>
</tr>
</tbody>
</table>

**Source:**
Tan PN, Kumar V, Srivastava J. Selecting the right interestness measure for association patterns. ACM SIGKDD International Conference on Knowledge discovery and data mining, 2002.
## Interestingness measures for association patterns

**Support** and **Confidence** are the same as we know them except that confidence is here maximum of confidence \((A \rightarrow B)\) and confidence \((B \rightarrow A)\)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>(P(A, B))</td>
</tr>
<tr>
<td>Confidence</td>
<td>(\max(P(B</td>
</tr>
<tr>
<td>Laplace</td>
<td>(\frac{P(A,B) + 1}{NP(A) + N</td>
</tr>
<tr>
<td>Conviction</td>
<td>(\frac{P(A,B)}{P(A)p(B)})</td>
</tr>
<tr>
<td>Interest</td>
<td>(\frac{P(A,B)}{P(A)p(B)})</td>
</tr>
<tr>
<td>cosine</td>
<td>(\frac{\sqrt{P(A)p(B)}}{P(A,B) - P(A)p(B)})</td>
</tr>
<tr>
<td>Klosgen</td>
<td>(\sqrt{P(A,B) + P(A)p(B)})</td>
</tr>
</tbody>
</table>

**Interest** is an old name for **Lift**

Source:
Tan PN, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. ACM SIGKDD International Conference on Knowledge discovery and data mining, 2002.
Useful properties of measures

- Some properties that an interestingness measure might have or not:
  - **Symmetricity**: value does not change if $X$ and $Y$ are swapped: $\text{measure}(X \rightarrow Y) = \text{measure}(Y \rightarrow X)$
  - **Scaling invariance**: value does not change if all transactions with $X$ (or $Y$) are duplicated
  - **Inversion invariance**: value does not change if $X$ (or $Y$) is inverted (removed from where it was present and added to where it was absent)
  - **Invariant to null additions**: value does not change if transactions without $X$ and $Y$ are added
Lecture 04 – Frequent pattern mining

✓ Frequent patterns
✓ Frequent itemsets and association rules
✓ Frequent itemset mining and Apriori algorithm
✓ Compact representations of itemsets
✓ Association rule mining
✓ Frequent itemsets in non-transactional data
✓ Pattern interestingness measures
Quotes

• “Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.”
  – Richard P. Feynman

• “Nobody knows the future with certainty. We can, however, identify ongoing patterns of change.”
  – Alvin Toffler

• “I'm a student of patterns. At heart, I'm a physicist. I look at everything in my life as trying to find the single equation, the theory of everything.”
  – Will Smith

• Source: https://www.brainyquote.com