GOETHE UNIVERSITÄT FRANKFURT AM MAIN

www.uni-frankfurt.de

SoSe 2014: M-TANI: Big Data Analytics

Lecture 5 – 28/05/2014

Sead Izberovic

Dr. Nikolaos Korfiatis

Agenda

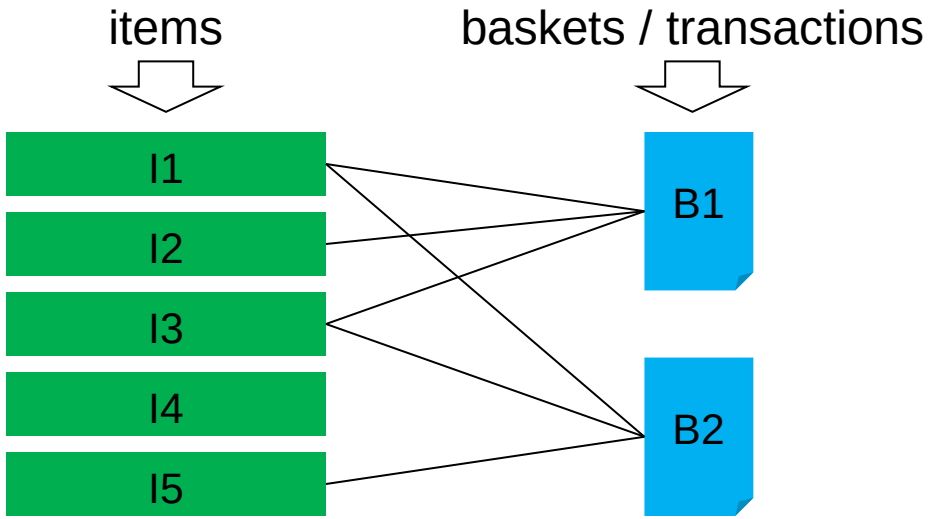
- **Recap from the previous session**
- **Frequent Itemsets Mining**
  - Market-Basket Model
  - Frequent Itemsets
  - Association Rules

Introduction

- “*The frequent-itemsets problem is that of finding sets of items that appear in (are related to) many of the same baskets*” [1]

Market-Basket Model

- **Describes a many to many mapping**



- **Looking for connections between items, not between baskets**

from [1]

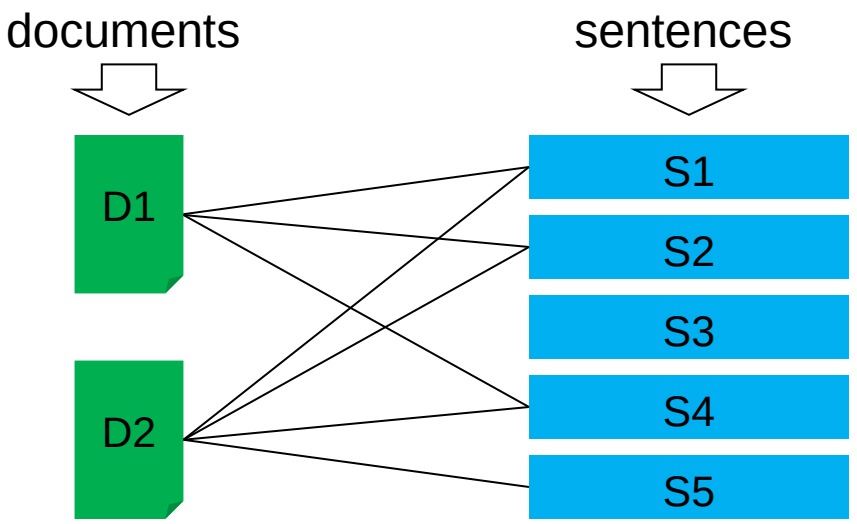
Applications

- **Frequent-Itemset applications**
  - Plagiarism
  - Biomarkers
  - Related concepts

from [1]

Frequent Itemsets

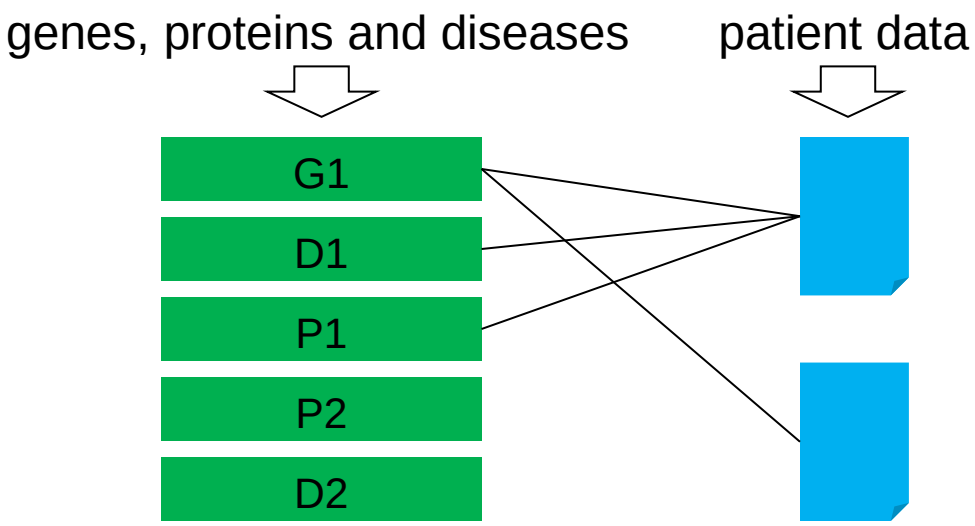
- **Frequent-Itemset applications**
  - Plagiarism
  - Biomarkers
  - Related concepts



from [1]

Frequent Itemsets

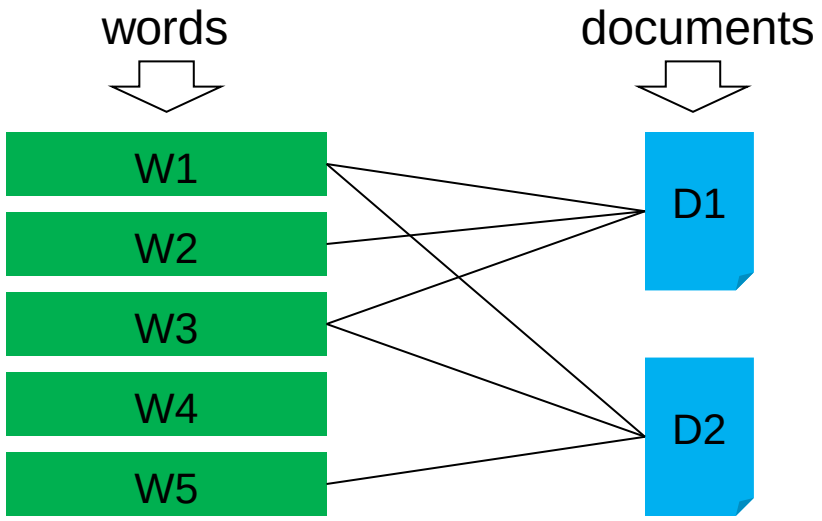
- **Frequent-Itemset applications**
  - Plagiarism
  - Biomarkers
  - Related concepts



from [1]

Frequent Itemsets

- **Frequent-Itemset applications**
  - Plagiarism
  - Biomarkers
  - Related concepts



from [1]

Frequent Itemsets

- **Definition**
  - I set of items
  - $b \in B$ is a basket
  - The $support$ for I is the number of baskets for which I is a subset
  - s is the support threshold
  - Sets are called **frequent Itemsets**, if $support \geq s$

from [1]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$

$\{\boldsymbol{m}\}$ is a frequent itemset. $support(\{\mathbf{m}\}) = 5 > s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$

$\{\boldsymbol{c}\}$ is a frequent itemset. $support(\{\boldsymbol{c}\}) = 5 > s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$

$\{\boldsymbol{b}\}$ is a frequent itemset. $support(\{\boldsymbol{b}\}) = 6 > s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$

$\{\boldsymbol{j}\}$ is a frequent itemset. $support(\{\boldsymbol{j}\}) = 4 > s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$

$\{\boldsymbol{m}, \boldsymbol{c}\}$ is not a frequent itemset! $support(\{\boldsymbol{m}, \boldsymbol{c}\}) = 2 < s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$

$\{\boldsymbol{m}, \boldsymbol{b}\}$ is a frequent itemset. $support(\{\boldsymbol{m}, \boldsymbol{b}\}) = 4 > s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$, $\{\boldsymbol{m}, \boldsymbol{b}\}$

$\{\boldsymbol{m}, \boldsymbol{j}\}$ is not a frequent itemset! $support(\{\boldsymbol{m}, \boldsymbol{j}\}) = 2 < s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$, $\{\boldsymbol{m}, \boldsymbol{b}\}$

$\{\boldsymbol{b}, \boldsymbol{c}\}$ is a frequent itemset! $support(\{\boldsymbol{b}, \boldsymbol{c}\}) = 4 > s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$, $\{\boldsymbol{m}, \boldsymbol{b}\}, \{\boldsymbol{b}, \boldsymbol{c}\}$

$\{\boldsymbol{c}, \boldsymbol{j}\}$ is a frequent itemset! $support(\{\boldsymbol{c}, \boldsymbol{j}\}) = 3 = s$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$, $\{\boldsymbol{m}, \boldsymbol{b}\}$, $\{\boldsymbol{b}, \boldsymbol{c}\}$, $\{\boldsymbol{c}, \boldsymbol{j}\}$

from [2]

Frequent Itemsets

Example

$I = \{\boldsymbol{m}ilk, \boldsymbol{c}oke, \boldsymbol{p}epsi, \boldsymbol{b}eer\ \boldsymbol{j}uice\}$
$s = 3$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\},\ \{\boldsymbol{c}\},\ \{\boldsymbol{b}\},\ \{\boldsymbol{j}\},\ \{\boldsymbol{m}, \boldsymbol{b}\},\ \{\boldsymbol{b}, \boldsymbol{c}\},\ \{\boldsymbol{c}, \boldsymbol{j}\}$

Why is $\{m, b, c\}$ not a frequent itemset?

from [2]

Association Rules

- **"Are implications that, if a basket contains a certain set of items $\{i_1, \dots, i_n\}$, then it is "*likely*" to contain another particular item j as well." [1]**

- **The $confidence$ of an association rule is the probability of j given $I = \{i_1, \dots, i_n\}$**

$$\boldsymbol{confidence(I \rightarrow j) = \frac{support(I \cup \{j\})}{support(I)}}$$

from [1] [2]

Association Rules

- **The $interest$ of an association rule is the difference between its confidence and the fraction of baskets that contain j**

$$\boldsymbol{interest(I \rightarrow j) = confidence(I \rightarrow j)} - \frac{|j \in B|}{|B|}$$

- **A association rule is intresting if:**
  - $\boldsymbol{interest(I \rightarrow j) > 0.5}$ or
  - $\boldsymbol{interest(I \rightarrow j) < -0.5}$

from [1] [2]

Association Rules

Example: Confidence and Interest

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$
$b_3 = \{\boldsymbol{m}, \boldsymbol{b}\}$ $b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$
$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$
$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{\boldsymbol{m}\}$, $\{\boldsymbol{c}\}$, $\{\boldsymbol{b}\}$, $\{\boldsymbol{j}\}$, $\{\boldsymbol{m}, \boldsymbol{b}\}$, $\{\boldsymbol{b}, \boldsymbol{c}\}$, $\{\boldsymbol{c}, \boldsymbol{j}\}$

$$confidence(\{m, b\} \rightarrow c) = \frac{support(\{m, b\} \cup \{c\})}{support(\{m, b\})} = \frac{2}{4} = \mathbf{\frac{1}{2}}$$

$$interest(\{m, b\} \rightarrow c) = confidence(\{m, b\} \rightarrow c) - \frac{|c \in B|}{|B|} = \frac{1}{2} - \frac{5}{8} = \mathbf{-\frac{1}{8}}$$

from [2]

Association Rule Mining

1. **Find all frequent itemsets J**
2. **Association Rule Generation**
   - For every $A \subset J$, generate a rule $A \rightarrow J \backslash A$
     - Since J is frequent $\rightarrow A$ is frequent
   - Compute the rule confidence
     - If a rule like $\{a, b, c\} \rightarrow d$ is below the confidence, so is $\{a, b\} \rightarrow \{c, d\}$
   - Output the rules $\mathrm{r} \in R$, with $confidence(r) \geq ct$
     - ct is a confidence threshold

from [2]

Association Rule Mining

Example with $s = 3$ and $confidence = \frac{3}{4}$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$

$b_3 = \{\boldsymbol{m}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{n}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$

$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$

$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{m\}, \{b\}, \{c\}, \{j\}$, $\{m, b\}, \{b, c\}, \{c, j\}, \{m, c, b\}$

Association Rules:

$$\{b\} \rightarrow \{m\};\ confidence = \frac{4}{6}$$

$$\{m\} \rightarrow \{b\};\ confidence = \frac{4}{5}$$

$$\{b\} \rightarrow \{c\};\ confidence = \frac{5}{6}$$

$$\{b, m\} \rightarrow \{c\};\ confidence = \frac{3}{4}$$

$$\vdots$$

from [2]

Association Rule Mining

Example with $s = 3$ and $confidence = \frac{3}{4}$

$b_1 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}\}$ $\quad b_2 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{j}\}$

$b_3 = \{\boldsymbol{m}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{n}\}$ $\quad b_4 = \{\boldsymbol{c}, \boldsymbol{j}\}$

$b_5 = \{\boldsymbol{m}, \boldsymbol{p}, \boldsymbol{b}\}$ $\quad b_6 = \{\boldsymbol{m}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$

$b_7 = \{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{j}\}$ $\quad b_8 = \{\boldsymbol{b}, \boldsymbol{c}\}$

Frequent Itemsets: $\{m\}, \{b\}, \{c\}, \{j\}$, $\{m, b\}, \{b, c\}, \{c, j\}, \{m, c, b\}$

Association Rules:

$\{b\} \rightarrow \{m\};\ confidence = \frac{4}{6}$

$\{m\} \rightarrow \{b\};\ confidence = \frac{4}{5}$

$\{b\} \rightarrow \{c\};\ confidence = \frac{5}{6}$

$\{b, m\} \rightarrow \{c\};\ confidence = \frac{3}{4}$

$\vdots$

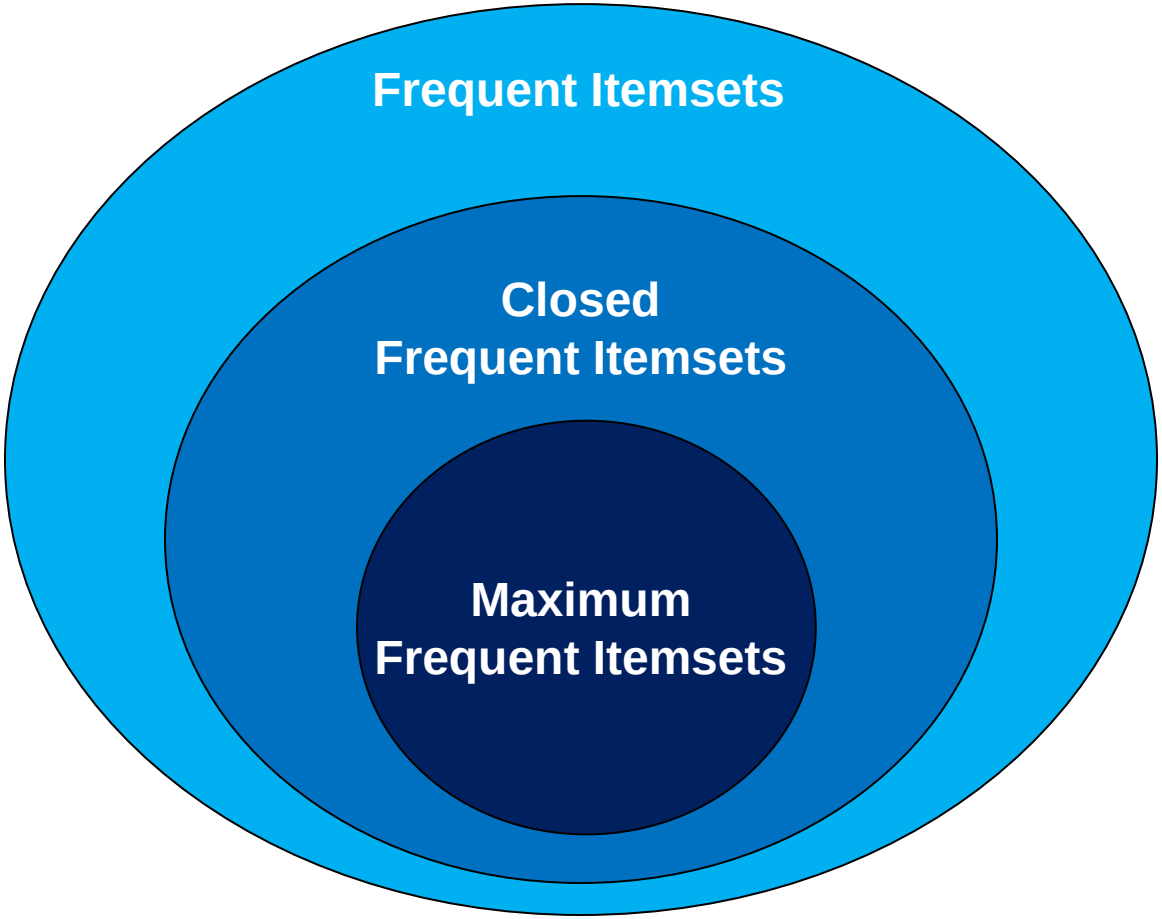
from [2]

Association Rule Mining

- **How to reduce the number of association rules?**
  1. Ajust s so that we do not get too many frequent itemsets
  2. Use maximum frequent itemsets
     - none of its immediate supersets is frequent
  3. Use closed frequent itemsets
     - none of its immediate supersets has the same support as the itemset

from [1] [2] [3]

Association Rule Mining



from [3]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4		
$\{B\}$	5		
$\{C\}$	3		
$\{A, B\}$	4		
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	
$\{B\}$	5		
$\{C\}$	3		
$\{A, B\}$	4		
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the immediate superset $\{A, B\}$ is frequent

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5		
$\{C\}$	3		
$\{A, B\}$	4		
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the support of the immediate superset $\{A, B\}$ is 4

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	
$\{C\}$	3		
$\{A, B\}$	4		
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the immediate superset $\{B, C\}$ is frequent

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3		
$\{A, B\}$	4		
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the immediate supersets $\{B, C\}$ and $\{A, B\}$ have a smaller support

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	
$\{A, B\}$	4		
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the immediate superset $\{B, C\}$ is frequent

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4		
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the support of immediate superset $\{B, C\}$ is 3

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the immediate superset $\{A, B, C\}$ is not frequent

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	Yes
$\{A, C\}$	2		
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the immediate superset $\{A, B, C\}$ has a smaller support

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	Yes
$\{A, C\}$	2	No	
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because not frequent!

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	Yes
$\{A, C\}$	2	No	No
$\{B, C\}$	3		
$\{A, B, C\}$	2		

Because the support of the immediate superset $\{A, B, C\}$ is 2

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	Yes
$\{A, C\}$	2	No	No
$\{B, C\}$	3	Yes	
$\{A, B, C\}$	2		

Because the immediate superset $\{A, B, C\}$ is not frequent

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	Yes
$\{A, C\}$	2	No	No
$\{B, C\}$	3	Yes	Yes
$\{A, B, C\}$	2		

Because the immediate superset $\{A, B, C\}$ has a smaller support

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	Yes
$\{A, C\}$	2	No	No
$\{B, C\}$	3	Yes	Yes
$\{A, B, C\}$	2	No	

Because not frequent!

from [2]

Association Rule Mining

Example: Maximum/Closed frequent itemsets

$s = 3$

	Support	Maximal	Closed
$\{A\}$	4	No	No
$\{B\}$	5	No	Yes
$\{C\}$	3	No	No
$\{A, B\}$	4	Yes	Yes
$\{A, C\}$	2	No	No
$\{B, C\}$	3	Yes	Yes
$\{A, B, C\}$	2	No	Yes

Because no immediate superset has the same support

from [2]

Literature

1. Anand Rajaraman, Jeffrey D. Ullman, Jure Leskovec. 2014
   Mining of Massive Datasets
   Cambridge University Press
2. Jure Leskovec. 2014
   Slides: **Mining Massive Data Sets**
   URL: http://www.stanford.edu/class/cs246/slides/02-assocrules.pdf
3. Pang-Ning Tan, Michael Steinbach, Vipin Kuma. 2004
   Slides: **Introduction to Data Mining**
   http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap6_basic_association_analysis.pdf