

Chapter 05

Association Rule Mining

Dr. Steffen Herbold

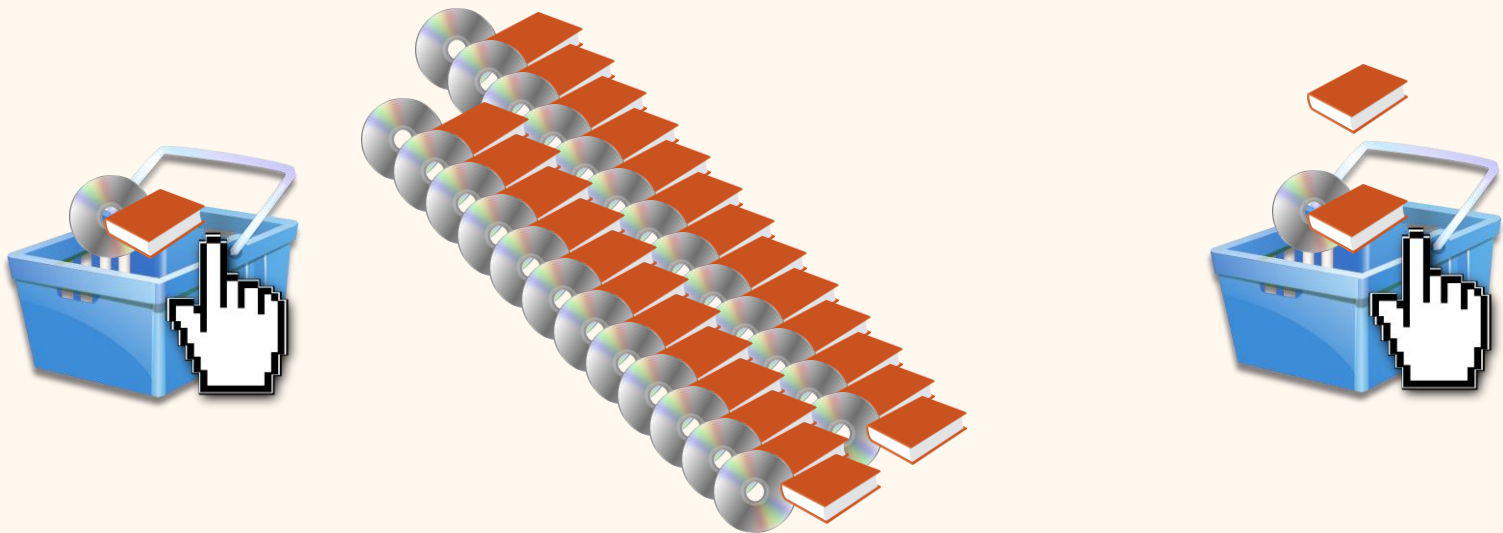
herbold@cs.uni-goettingen.de

Outline

- Overview
- The Apriori Algorithm
- Summary

Example of Association Rules

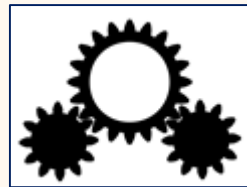
Items already in basket + Available items → Item likely to be added



The General Problem

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- ...



Association
Rule Mining

Association Rules

- item2 \rightarrow item3
- item2 \rightarrow item4
- item2,item3 \rightarrow item4
- ...



Rules describe
„interesting relationships“

The Formal Problem

- Items

- $I = \{i_1, i_2, \dots, i_m\}$

- Transactions

- $T = \{t_1, \dots, t_n\}$ with $t_i \subseteq I$

- Rules

- $X \Rightarrow Y$ such that $X, Y \subseteq I$ and $X \cap Y = \emptyset$

- X is also called antecedent or left-hand-side

- Y is also called consequent or right-hand-side



Defining „Interesting Relationships“

Set of Transactions

- item1, item2, item3
- item2, item4
- item1, item5
- item6, item7
- item2, item3, item4, item7
- item2, item3, item4, item8
- item2, item4, item5
- item2, item3, item4
- item4, item5
- ...

item2, item3,
item4 occur often
together



Interesting == often together

Outline

- Overview
- **The Apriori Algorithm**
- Summary

Support and Frequent Item Sets

- Support
 - Percentage of occurrences of an itemset IS
 - $support(IS) = \frac{|\{t \in T : IS \subseteq t\}|}{|T|}$
- Frequent item set
 - Itemsets that appear together „often enough“
 - Defined using a threshold
 - $IS \subseteq I$ is frequent if $support(IS) \geq minsupp$
- Rules can be generated by splitting itemsets
 - $IS = X \cup Y$

Example for Generating Rules

$minsupp = 0.3$

- $support(\{item2, item3, item4\}) = \frac{3}{10} \geq 0.3$

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

- Eight possible rules:
 - $\emptyset \Rightarrow \{item2, item3, item4\}$
 - $\{item2\} \Rightarrow \{item3, item4\}$
 - $\{item3\} \Rightarrow \{item2, item4\}$
 - $\{item4\} \Rightarrow \{item2, item3\}$
 - $\{item2, item3\} \Rightarrow \{item4\}$
 - $\{item2, item4\} \Rightarrow \{item3\}$
 - $\{item3, item4\} \Rightarrow \{item2\}$
 - $\{item2, item3, item4\} \Rightarrow \emptyset$

Are all rules interesting?



Confidence, Lift, and Leverage

- Confidence

- Percentage of transactions that contain the antecedent, which also contain consequent

- $confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} = \frac{|\{t \in T: X \cup Y \subseteq T\}|}{|\{t \in T: X \subseteq T\}|}$

- Lift

- Ratio of the probability of X and Y together and independently

- $lift(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X) \cdot support(Y)}$

- Leverage

- Difference in the probability of X and Y together and independently

- $leverage(X \Rightarrow Y) = support(X \cup Y) - support(x) \cdot support(Y)$



Usually, lift favors itemsets with lower support, leverage with higher support

Confidence for the Example Rules

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

- $confidence(\emptyset \Rightarrow \{item2, item3, item4\}) = \frac{0.3}{1} = 0.3$
- $confidence(\{item2\} \Rightarrow \{item3, item4\}) = \frac{0.3}{0.6} = 0.5$
- $confidence(\{item3\} \Rightarrow \{item2, item4\}) = \frac{0.3}{0.4} = 0.75$
- $confidence(\{item4\} \Rightarrow \{item2, item3\}) = \frac{0.3}{0.6} = 0.5$
- $confidence(\{item2, item3\} \Rightarrow \{item4\}) = \frac{0.3}{0.4} = 0.75$
- $confidence(\{item2, item4\} \Rightarrow \{item3\}) = \frac{0.3}{0.5} = 0.6$
- $confidence(\{item3, item4\} \Rightarrow \{item2\}) = \frac{0.3}{0.3} = 1$
- $confidence(\{item2, item3, item4\} \Rightarrow \emptyset) = \frac{0.3}{0.3} = 1$

Lift for the Example Rules

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

- $lift(\emptyset \Rightarrow \{item2, item3, item4\}) = \frac{0.3}{1 \cdot 0.3} = 1$
- $lift(\{item2\} \Rightarrow \{item3, item4\}) = \frac{0.3}{0.6 \cdot 0.3} = 1.66$
- $lift(\{item3\} \Rightarrow \{item2, item4\}) = \frac{0.3}{0.4 \cdot 0.5} = 1.5$
- $lift(\{item4\} \Rightarrow \{item2, item3\}) = \frac{0.3}{0.6 \cdot 0.4} = 1.25$
- $lift(\{item2, item3\} \Rightarrow \{item4\}) = \frac{0.3}{0.4 \cdot 0.6} = 1.25$
- $lift(\{item2, item4\} \Rightarrow \{item3\}) = \frac{0.3}{0.5 \cdot 0.4} = 1.5$
- $lift(\{item3, item4\} \Rightarrow \{item2\}) = \frac{0.3}{0.3 \cdot 0.6} = 1.66$
- $lift(\{item2, item3, item4\} \Rightarrow \emptyset) = \frac{0.3}{0.3 \cdot 1} = 1$

Overview of Scores for Example

Rule	Confidence	Lift	Leverage
$\emptyset \Rightarrow \{item2, item3, item4\}$	0.30	1.00	0.00
$\{item2\} \Rightarrow \{item3, item4\}$	0.50	1.66	0.12
$\{item3\} \Rightarrow \{item2, item4\}$	0.75	1.50	0.10
$\{item4\} \Rightarrow \{item2, item3\}$	0.50	1.25	0.06
$\{item2, item3\} \Rightarrow \{item4\}$	0.75	1.25	0.06
$\{item2, item4\} \Rightarrow \{item3\}$	0.60	1.50	0.10
$\{item3, item4\} \Rightarrow \{item2\}$	1.00	1.66	0.12
$\{item2, item3, item4\} \Rightarrow \emptyset$	1.00	1.00	0.00

Perfect confidence, but
no gain over randomness

Perfect confidence, and
1.66 times more likely
than randomness

Itemsets and Rules = Exponential

- Number of itemset is exponential
 - All possible itemsets are the powerset \mathcal{P} of I
 - $|\mathcal{P}(I)| = 2^{|I|}$
 - Still exponential if we restrict the size
 - $|I|$ itemsets with $k = 1$ items
 - $\frac{|I| \cdot (|I|-1)}{2}$ itemsets with $k = 2$ items
 - $\binom{|I|}{k} = \frac{|I|!}{(|I|-k)!k!}$ itemsets with k items
- Number of rules per itemset is exponential
 - Possible antecedents of itemset i are the powerset \mathcal{P} of i
 - $|\mathcal{P}(i)| = 2^{|i|}$
- Example: $|I| = 100, k = 3$
 - 161,700 possible itemsets
 - 1,293,600 possible rules

How do we
restrict the
search space?



Pruning the Search Space

- Apriori Property
 - **All subsets of frequent itemsets are also frequent**
 - $support(i') \geq support(i)$ for all $i' \subseteq i, i \in I$
- „Grow“ itemsets and prune search space by applying Apriori property
 - Start with itemsets of size $k = 1$
 - Drop all itemsets that do not have minimal support
 - Build all combinations of size $k + 1$
 - Repeat until
 - No itemsets with minimal support are found
 - A threshold for k is reached
- Still has exponential complexity, but better bounded

Example for Growing Itemsets (k=1)

$minsupp = 0.3$

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

Rule	Support
{item1}	0.2
{item2}	0.6
{item3}	0.4
{item4}	0.5
{item5}	0.3
{item6}	0.2
{item7}	0.3
{item8}	0.1

← Drop

← Drop

← Drop

Example for Growing Itemsets (k=2)

$minsupp = 0.3$

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

Rule	Support
{item2, item3}	0.4
{item2, item4}	0.5
{item2, item5}	0.1
{item2, item7}	0.1
{item3, item4}	0.3
{item3, item5}	0.0
{item3, item7}	0.1
...	...

← Drop

← Drop

← Drop

← Drop

← Drop

Example for Growing Itemsets (k=3)

$minsupp = 0.3$

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- item6,item7

Rule	Support
{item2, item3, item4}	0.3

Only itemset remaining, growing terminates.

- Found the following frequent itemsets with at least two items:
 - {item2, item3}, {item2, item4}, {item3, item4}
 - {item2, item3, item4}

Candidates for Rules

- Usually, not all possible rules are considered
- Two common restrictions:
 - No empty antecedents and consequents
 - $X \neq \emptyset$ and $Y \neq \emptyset$
 - Only one item as consequent
 - $|Y| = 1$
- Example:

~~$\emptyset \Rightarrow \{item2, item3, item4\}$~~
 ~~$\{item2\} \Rightarrow \{item3, item4\}$~~
 ~~$\{item3\} \Rightarrow \{item2, item4\}$~~
 ~~$\{item4\} \Rightarrow \{item2, item3\}$~~
 $\{item2, item3\} \Rightarrow \{item4\}$
 $\{item2, item4\} \Rightarrow \{item3\}$
 $\{item3, item4\} \Rightarrow \{item2\}$
 ~~$\{item2, item3, item4\} \Rightarrow \emptyset$~~

Evaluating Association Rules

- Use different criteria, not just support and confidence
 - Lift and leverage can tell you if rules are coincidental
- Validate if rules hold on test data
 - Check if the rules would also be found on the test data
- For basket prediction, use incomplete itemsets on test data
 - Example: remove item4 from all itemsets and see if the rules would correctly predict where it is associated
- Manually inspect rules
 - Do they make sense?
 - Ask domain experts

Outline

- Overview
- The Apriori Algorithm
- **Summary**

Summary

- Associations are interesting relationships between items
- Interesting usually means appear together and not coincidental
- Number of possible itemsets/rules exponential
 - Apriori property for bounding itemsets
 - Restrictions on rule structures
- Test data and manual inspections for validation