TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Data Science: Advanced Analytics

Association Analysis

Dresden // 03.05.2023
Sommersemester 2023

Association Analysis

Overview

Association rules were first introduced in the area of **market basket analysis** with the aim of finding out which products are purchased together.

The resulting publications

"If you buy diapers on Friday night, you buy beer!"

then first drew the attention of the general public to the **possibilities of data mining** in the 1990s.

In the meantime, association rules are also increasingly used in other fields. Because they **do not perform disjoint data decomposition** as decision trees do, more **interesting associations** can be uncovered; however, a large number of uninteresting associations can also be uncovered.

Association Analysis

Idea and Examples

The **idea** can be described as follows:

- So-called **association rules** should be found that describe the correlation between items that occur together.
- Thus, the identification of items A that imply the occurrence of other items B within a transaction should succeed.
- A discovered **relationship** between two or more items can then be represented as a rule of the form **"If item(set) A, then item(set) B"** or **A → B**, respectively.

Examples include the following:

- **Point-of-sale** transactions
- **Financial** and banking services (current account, share deposit, credit card, building loan, etc.)
- Conclusion of various insurances (especially: sequence analyses)
- Combinations of insurance claims
- Interactions when taking different **medications** at the same time

Association Analysis

Basic Terms

Items are uninterpretable, discrete things or entities. The set of all items is denoted by **I**.

An **association rule A → B** consists of a set of items A in the rule body and a set of items B in the **rule head**, where A and B must be disjoint. (A can't be part of B)

Each **transaction t** consists of a set of items. The transactions considered in the association analysis are combined into a **database D**.

A transaction t satisfies a rule A → B if **$(A \cup B) \subseteq t$**, i.e., if all items contained in the rule also occur in the transaction.

Association Analysis
Support and Confidence

The **Support** of a rule is defined as the proportion of transactions from D that satisfy that rule:

$$\sup(A \rightarrow B) = \frac{\left|\{t \in D \mid (A \cup B) \subseteq t\}\right|}{|D|}$$

The **Confidence** of a rule is defined as the proportion of transactions involving A and B in the set of transactions satisfying the rule body A. The formal representation is:

$$\mathrm{conf}(A \rightarrow B) = \frac{\left|\{t \in D \mid (A \cup B) \subseteq t\}\right|}{\left|\{t \in D \mid A \subseteq t\}\right|} = \frac{\sup(A \rightarrow B)}{\sup(A)}$$

Support and Confidence

Example

Transaction table

- 1.000.000 transactions in total
- 200.000 chips
- 50.000 beer
- 20.000 chips and beer

Rule

When a customer buys chips, he/she has also bought beer in 10% of the cases.

Evaluation Criteria

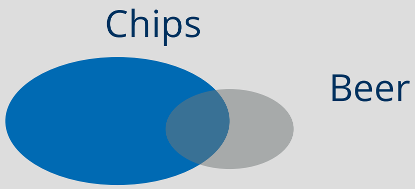
Support: 20.000/1.000.000 = 2%

Confidence: 20,000/200,000 = 10%

Expected Confidence = 50.000/1.000.000 = 5%

Lift = Confidence/Expected Confidence = 2

Venn Diagram



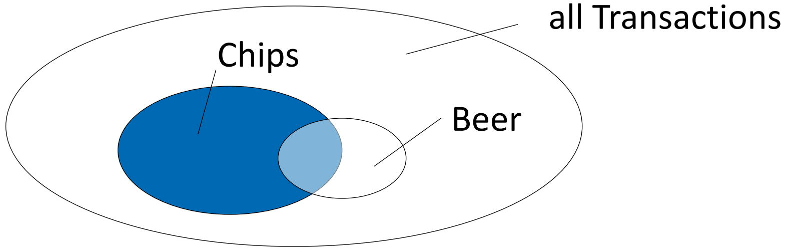
If a customer buys chips (**rule body/premise**), then he has also bought beer (**rule head/conclusion**) in 10% of all cases.

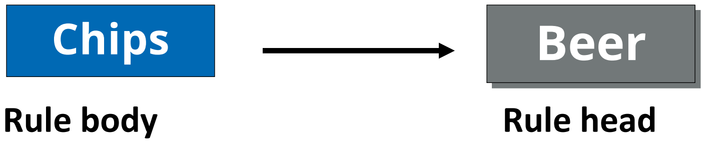
Confidence: information about the strength of the relationship between the rule body and the rule head: 10% (ratio of the proportion of transactions containing chips and beer to all transactions in which chips occur).

Support: simultaneous purchase of both products in 2% of all transactions

Support and Confidence

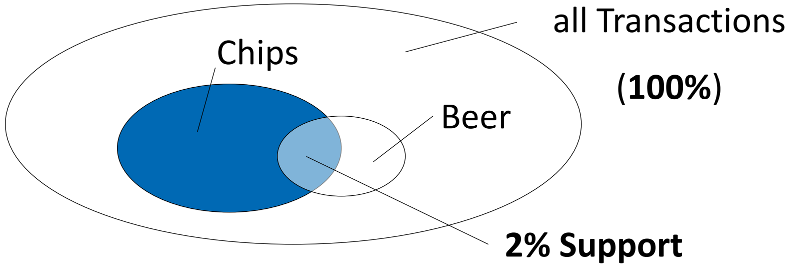
Interpretation

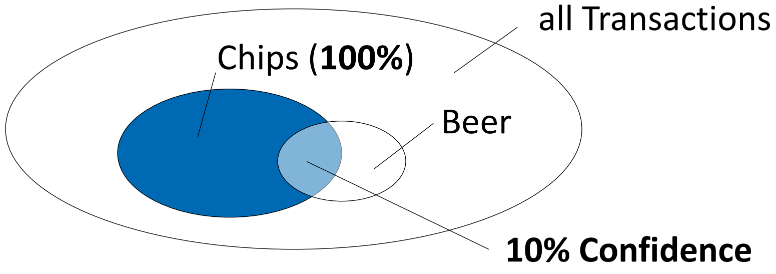




Interpretation:

In 10% of the cases where chips are purchased, beer is also purchased. Both products occur in 2% of all transactions.





Association Analysis

Algorithm overview

Since the data sets are typically very large, **efficient algorithms** are required.

The algorithms themselves generally search for **all association rules** that meet predefined minimum confidence and minimum support.

Depending on how this is done, different procedures are known. The best known methods are from Agrarwal:

Note: 100 items result in generally (100 over k) item combinations.

–	2-itemsets:	4,950 combinations
–	3-itemsets:	161,700 combinations
–	5-itemsets:	75.287.520 combinations

- **A-priori Algorithm**
- A-priori-Tid algorithm
- A-priori-Hybrid Algorithm
- AIS algorithm

The procedures have in common that they are often used as **starting procedures in the data mining**, since no assumptions are necessary to the connections of the data, which things could correlate e.g. (undirected data mining).

A-priori Algorithm

Idea

The **A-priori algorithm** is considered a classical method for the formation of association rules or for the efficient determination of all relevant relationships in a database.

The algorithm searches for all rules that have a high **confidence** value. It must now be taken into account that the important confidence of a rule is also expressed by its **support values**. Thus applies:

Let X be an item set from the population of all items I and X‘ $\subset$ X, then:

$$\mathrm{conf}((X - X') \rightarrow X') = \frac{\sup(X)}{\sup(X - X')}$$

Support values, in turn, represent frequencies of item sets in the population.

A-priori Algorithm

Process steps

The following steps are to be completed during rule generation using the A-priori algorithm.

must be completed:

Step 1: Compute the (frequent) item sets whose support values are at least equal to the minimum support (→ **frequent item set X**).

Step 2: For each subset X' of a frequent item set X holds:

$$\mathrm{conf}((X - X') \rightarrow X') = \frac{\sup(X)}{\sup(X - X')}$$

Formation of all rules whose Confidence is at least equal to the minimum Confidence.

A-priori Algorithm

Frequent sets of items – Part I

For X‘ $\subseteq$ X applies:

support (X‘) $\geq$ support (X) $\Rightarrow$ If X is frequent item set, then also X‘.

From the inversion of the implication follows immediately the **A priori property:**

If a k-itemset is not frequent, then a k+1 itemset containing it as a subset cannot be frequent either.

From this follows the procedure for the calculation of the frequent item sets:

1. Successive calculation of frequent item sets with i = 1,2,3,... Items
2. **Summary of the frequent item sets** with i items to the set I_i,
   d.h. I_i = {X : X is frequent item set, $|X|$ = i }

3.For X $\in$ I_{n+1} applies: All n-element subsets of X are often.

4.Calculation of I_{n+1} from I_n :

- N-element frequent sets are extended by one element.
- Check if (n +1)-elementary set is also frequent.If yes, take over the (n +1)-elementary set in I_{n+1}

A-priori Algorithm

Frequent sets of items – Part II

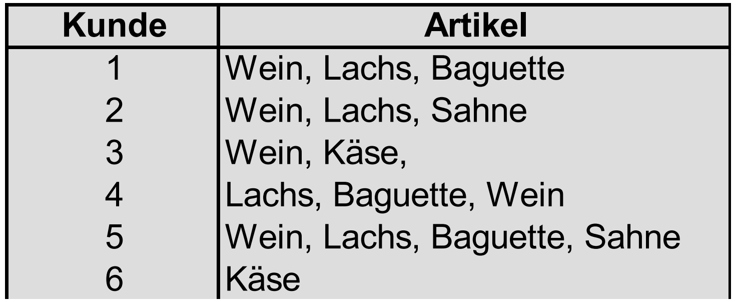
Note: If the algorithm ends, then for all elements of the sets of I1, ..., In, one computes the associated rules with their measures of interest and selects the relevant ones.

```
1. Initialization :
        $s_{min}$ = value for minimum support
        $n = 1$
        $I = \varnothing$
        $H_n = \{ \{i\} : i \text{ is an item}\}$

2. Determine in the database D for all $H \in H_n$ the Support

3. Determine $I_n = \{H \in H_n : \text{support}(H) \geq s_{min}\}$
        Set $I = I \cup I_n$

4. If $I_n = \varnothing$, then output set I as the result.

5. $H_{n+1} = \left\{ \{i_1,\dots,i_{n+1}\} : \forall j : 1 \leq j \leq n+1 : \{i_1,\dots,i_{n+1}\} - \{i_j\} \in I_n \right\}$

set $n = n + 1$
6.      Go to 2
```

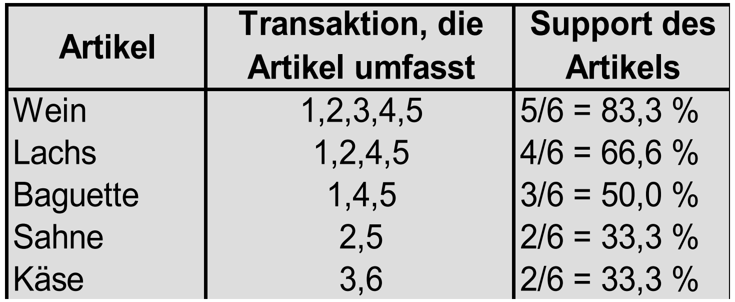
A-priori Algorithm

Example – Part I

Determination of **association rules** for transaction data of a **supermarket**

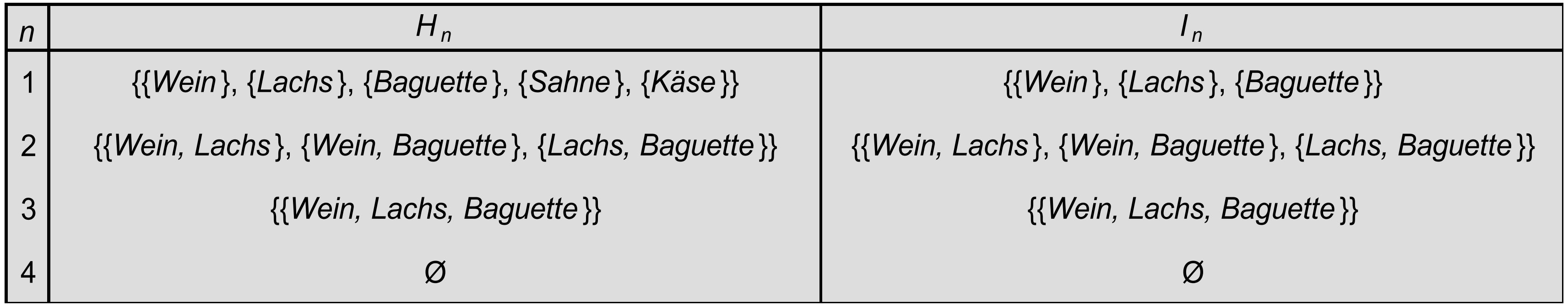


Kunde	Artikel
1	Wein, Lachs, Baguette
2	Wein, Lachs, Sahne
3	Wein, Käse,
4	Lachs, Baguette, Wein
5	Wein, Lachs, Baguette, Sahne
6	Käse



Artikel	Transaktion, die Artikel umfasst	Support des Artikels
Wein	1,2,3,4,5	5/6 = 83,3 %
Lachs	1,2,4,5	4/6 = 66,6 %
Baguette	1,4,5	3/6 = 50,0 %
Sahne	2,5	2/6 = 33,3 %
Käse	3,6	2/6 = 33,3 %

Note: min Support s_{min}=50%)



n	H_n	I_n
1	{{*Wein*}, {*Lachs*}, {*Baguette*}, {*Sahne*}, {*Käse*}}	{{*Wein*}, {*Lachs*}, {*Baguette*}}
2	{{*Wein, Lachs*}, {*Wein, Baguette*}, {*Lachs, Baguette*}}	{{*Wein, Lachs*}, {*Wein, Baguette*}, {*Lachs, Baguette*}}
3	{{*Wein, Lachs, Baguette*}}	{{*Wein, Lachs, Baguette*}}
4	Ø	Ø

A-priori Algorithm

Example – Part II

I_n
{{Wein}, {Lachs}, {Baguette}}
{{Wein, Lachs}, {Wein, Baguette}, {Lachs, Baguette}}
{{Wein, Lachs, Baguette}}
Ø

Regeln mit Support ≥ 50%	erfüllende Transaktionen	Support	Konfidenz
Wein → Lachs	1,2,4,5	66%	80%
Lachs → Wein	1,2,4,5	66%	100%
Lachs → Baguette	1,4,5	50%	75%
Baguette → Lachs	1,4,5	50%	100%
Wein → Baguette	1,4,5	50%	60%
Baguette → Wein	1,4,5	50%	100%
Wein, Baguette → Lachs	1,4,5	50%	100%
Lachs, Wein → Baguette	1,4,5	50%	75%
Baguette, Lachs → Wein	1,4,5	50%	100%
Lachs → Wein, Baguette	1,4,5	50%	75%
Baguette → Lachs, Wein	1,4,5	50%	100%
Wein → Lachs, Baguette	1,4,5	50%	60%

Based on the frequent item sets and the **minimal Confidence** c_{min} = 60 % values, such as e.g. {Wein, Lachs} $\in I_2$ can be calculated:

$$\text{conf}(\text{Wein} \rightarrow \text{Lachs}) = \frac{\sup(\{\text{Wein,Lachs}\})}{\sup(\{\text{Wein}\})} = 80\%$$

$$\text{conf}(\text{Lachs} \rightarrow \text{Wein}) = \frac{\sup(\{\text{Wein,Lachs}\})}{\sup(\{\text{Lachs}\})} = 100\%$$

Kunde	Artikel
1	Wein, Lachs, Baguette
2	Wein, Lachs, Sahne
3	Wein, Käse,
4	Lachs, Baguette, Wein
5	Wein, Lachs, Baguette, Sahne
6	Käse

Association Analysis

Classical measures of interest– Support

Support indicates the **percentage of transactions** that satisfy the rule.

The values are always between zero and one.

Formally, **support** is defined as follows:

$$\sup(A \rightarrow B) = \frac{\left|\left\{t \in D \middle| (A \cup B) \subseteq t\right\}\right|}{|D|}$$

Rules with high support values are often known facts.

Rules with very small support occur rather rarely.

There are **no generally valid statements** which support ranges should lead to a closer examination of the rules.

Association Analysis

Classical measures of interest– Confidence

Confidence describes the **conditional probability** of B given A (values from zero to one).

$$\mathrm{conf}(A \rightarrow B) = \frac{\left|\{t \in D \mid (A \cup B) \subseteq t\}\right|}{\left|\{t \in D \mid A \subseteq t\}\right|} = \frac{\sup(A \rightarrow B)}{\sup(A)}$$

Disadvantage: Probability of B is ignored;

Confidence of A → B can be equal to the probability of B in the population (i.e. B would be statistically independent of A) and still meet the minimum confidence.

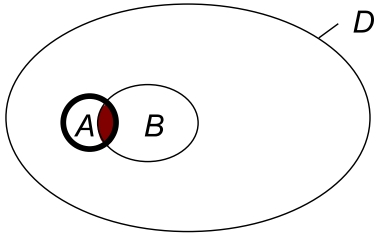
Example: If, for example, milk is purchased in 80% of all transactions and the purchase of milk would be completely uncorrelated with the purchase of salmon, the rule Lachs → Milch would still have a confidence of 80%. Rule is rated as "interesting" despite its habitual nature.

Association Analysis

Problems of the classical measures of interest

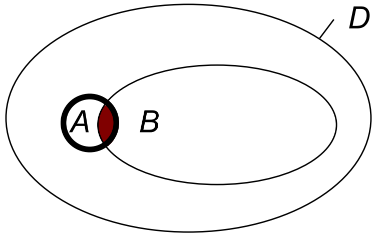
The classical evaluation or **selection measures** are fraught with drawbacks.

The following examples demonstrate this impressively:

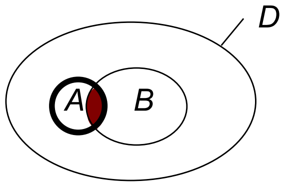


Referenzmodell

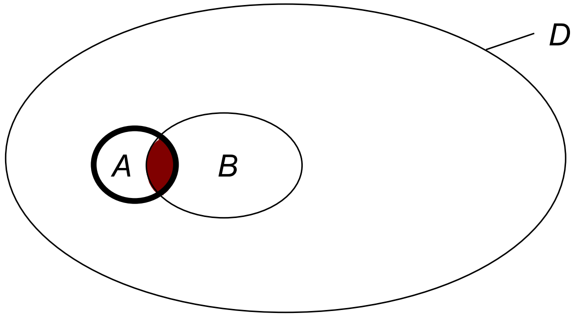
Disadvantages of Confidence:
Confidence does not consider the probability of B.
Confidence of a rule A B does not change if
if the population D increases or decreases,
if the frequency of B changes, or
if all frequencies increase at the same rate.



Häufigkeit von B steigt



Häufigkeit von *D* sinkt



Häufigkeiten steigen proportional an

Association Analysis

Measure of interest – Lift

The **lift** of a rule A B is calculated as the quotient of the confidence of the rule and the expected confidence of the rule, i.e. the support sup(B) of B. The lift indicates how much the distribution of certain items in a subset differs from the distribution in the population.

$$\mathrm{lift}(A \rightarrow B) = \frac{\mathrm{conf}(A \rightarrow B)}{\mathrm{sup}(B)} = \frac{\mathrm{sup}(A \cup B)}{\mathrm{sup}(A)\mathrm{sup}(B)}$$

The lift indicates how much the distribution of certain items in a subset differs from the distribution in the population.

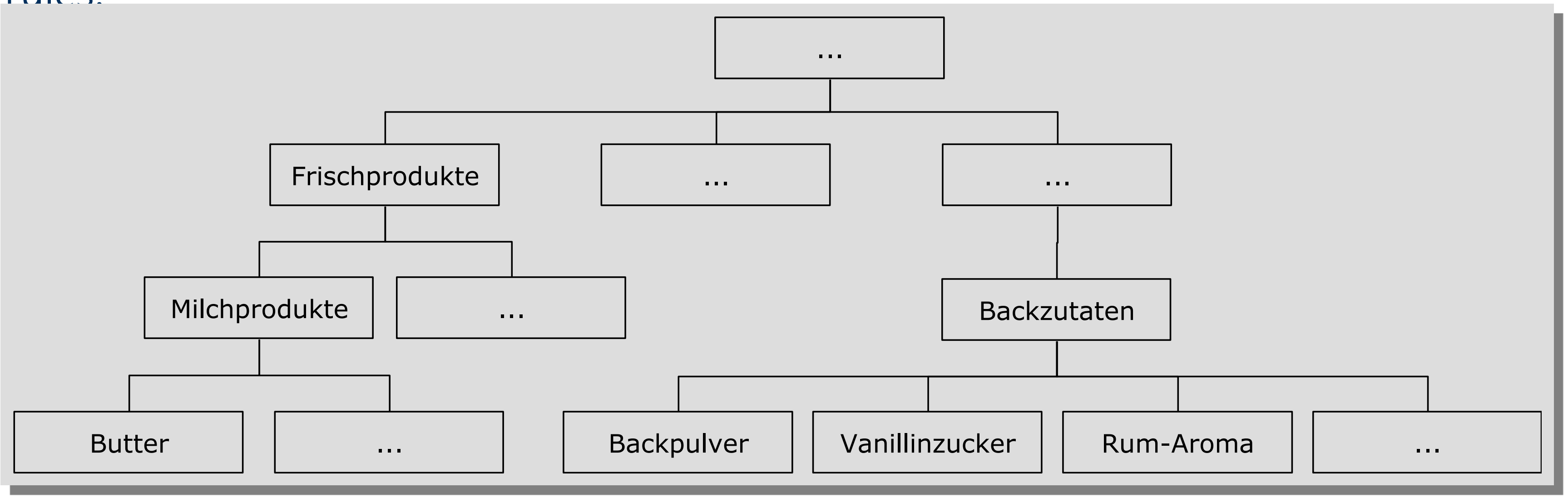
— Lift > 1: positive correlation
— Lift < 1: negative correlation
— Lift = 1: no correlation.

The lift of rule A → B and rule B → A are identical

Association Analysis

Taxonomies

The **integration of taxonomies**, i.e. systems, classes or orders over the items, can increase the usefulness of association rules.



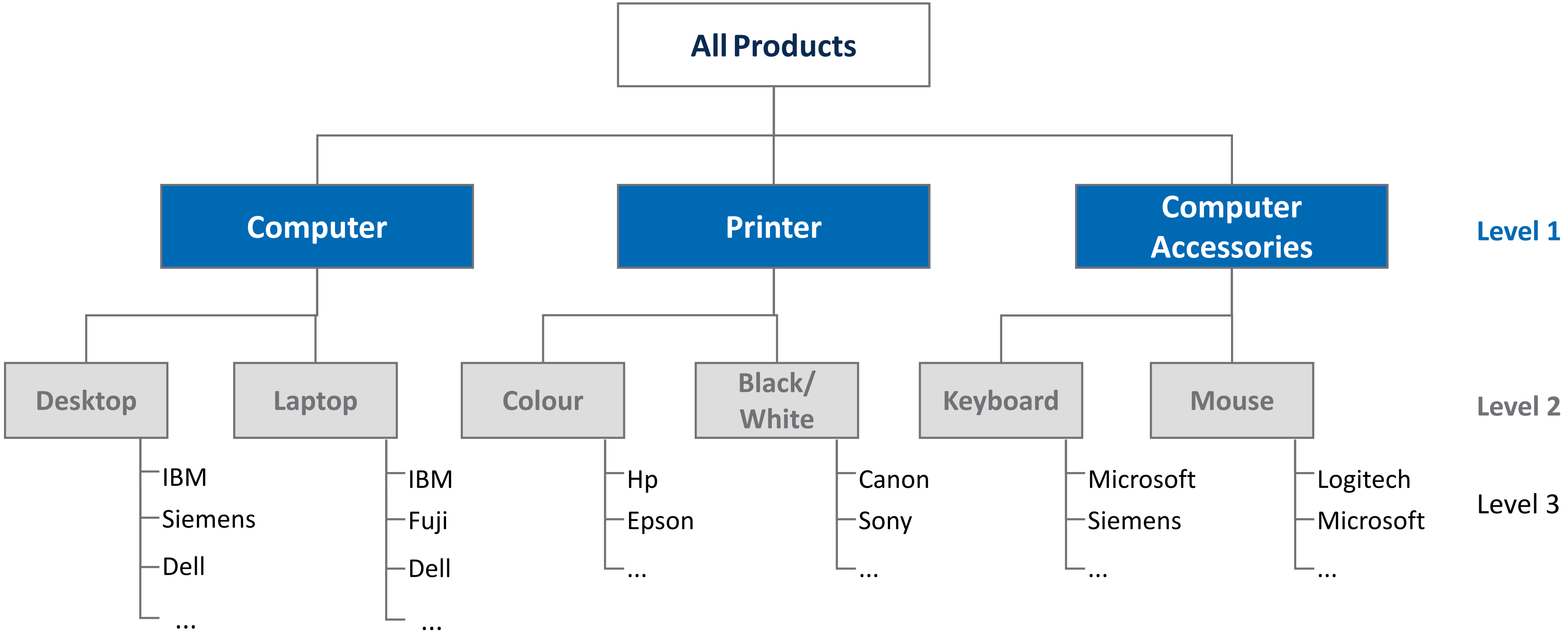
Benefit: Associations between commodity groups and products and commodity groups are also found; Support and Confidence tend to be higher as a result of grouping.

Simultaneous inclusion of **multiple independent taxonomies** in the analysis.

e.g.: classification of products into brand, category and commodity group and additionally assignment to the group “special offer” → in-depth examination of the rules containing the item special offer

Association Analysis

Taxonomies - Example Computer Products



Note:

The association analysis execution here can proceed at level 1, level 2 or level 3.

Association Analysis

Comments on the subject of transactions

The following applies to transactions in general:

Transactions consist of **item quantity and transaction time**.

Often, information on behavior over time is available in addition to the transactions.

- Market Basket analysis: Payment by credit card or customer card;
- Delivery Providers: collection of customer information e.g. based on orders received.

However, this **information is not used** in classical association analysis, although the inclusion of this additional information can significantly increase the meaningfulness of the rules (→ **sequence analysis**).

Association Analysis

Sequence Analysis – Beispiele

sportscheck.com

Transactions under consideration of temporal relations

Association Analysis

Sequence Analysis

Sequence analyses are used to analyze so-called sequences, i.e. **transactions** to which a **time specification** has been added.

Sequence analyses thus serve the investigation of transactions under consideration of temporal relations.

The differences to the classical association analysis are the following:

- Sequence analysis is not a time analysis, but a **time period analysis.**
- The objects of investigation are **sequences** of transactions that form a specific pattern.

Association Analysis

Real data and run times – Beispiele

Example 1 - Supermarket data

548 000 articles

Taxonomy with 4 levels, 118 roots

≈ 1.5 million transactions with an average of 9.6 items per transaction

→ optimized algorithm needs ≈ 90 minutes to generate all rules with support greater than 1%.

Beispiel 2 – Department-Store-Daten:

– 228 000 articles

– Taxonomy with 7 levels, 89 roots

– 570 000 transactions with an average of 4.4 articles per transaction

≈ optimized algorithm takes ≈ 5 minutes to generate all rules with support greater than 1%.

Association Analysis

Real data and applications– Data Mining

Example 3 - Toy store

"In 45% of the cases where a Barbie is purchased, candy is also purchased."

Recommendations:

Place the products close to each other.

Place products in different corners of the store.

Package Barbie and candy as one product as a bundle.

Package Barbie, candy, and a poor-selling product as a bundle.

Sell candy in the form of Barbie.

If the price of one of the products increases, lower the price of the other product.

Offer Barbie accessories as a loyalty reward for candy purchases.

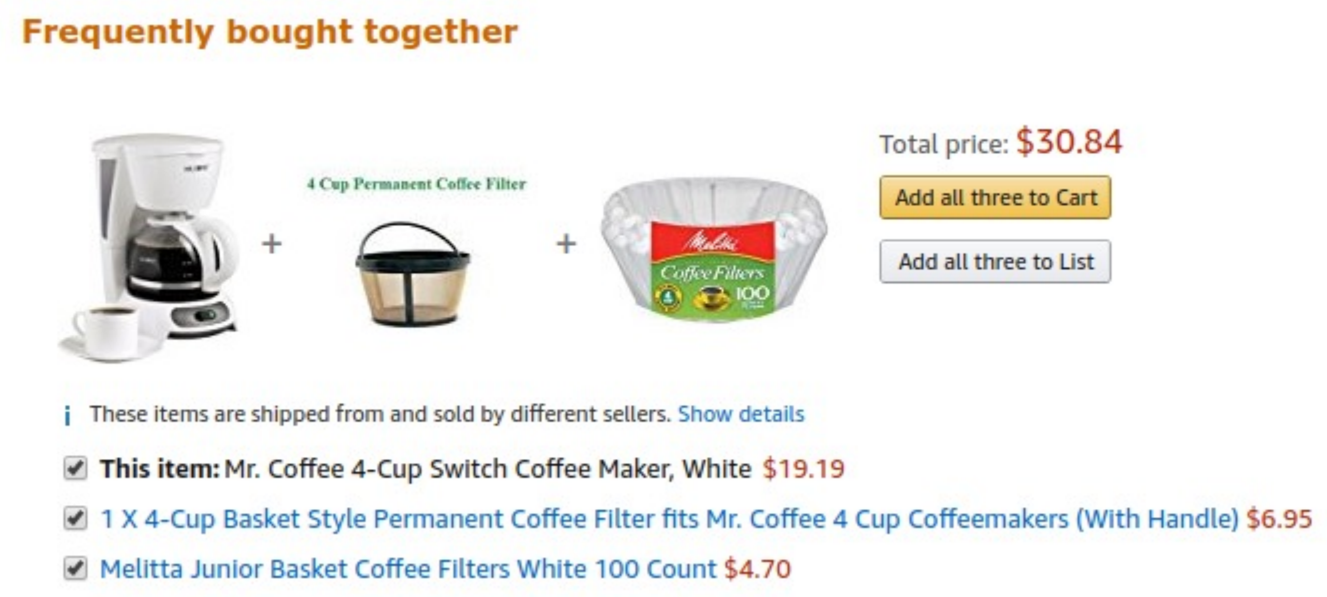
Do not promote both products together.

Association Analysis

Real data and applications– Data Mining

Example 4 - Amazon

"Customers who bought this product also bought the following products."



Frequently bought together

Total price: $30.84

i These items are shipped from and sold by different sellers. Show details

☑ **This item:** Mr. Coffee 4-Cup Switch Coffee Maker, White $19.19

☑ 1 X 4-Cup Basket Style Permanent Coffee Filter fits Mr. Coffee 4 Cup Coffeemakers (With Handle) $6.95

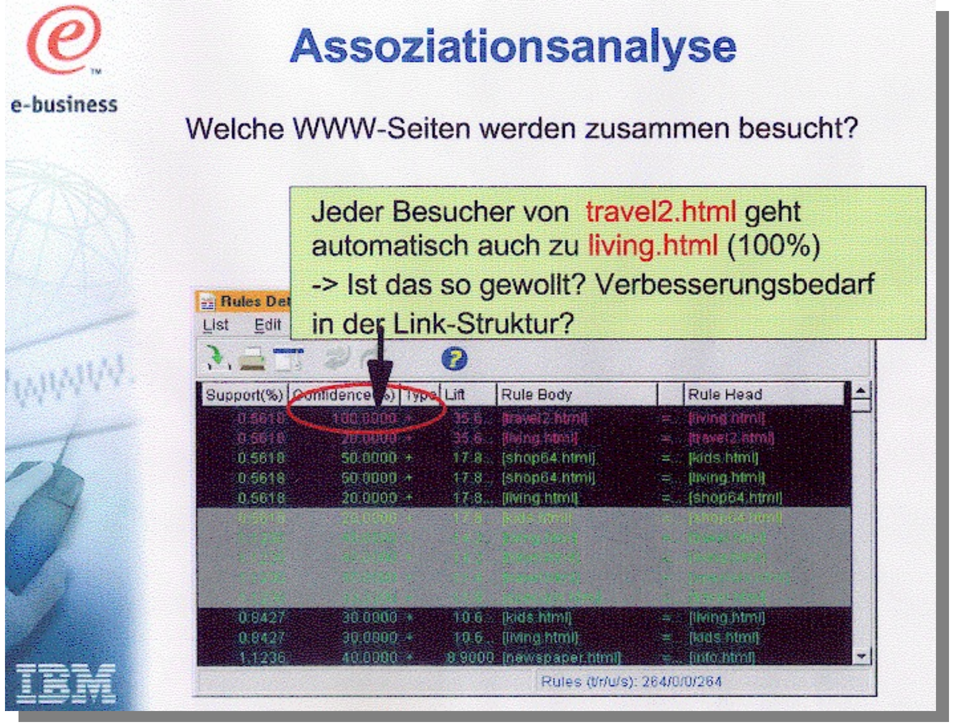
☑ Melitta Junior Basket Coffee Filters White 100 Count $4.70

Association Analysis

Real data and applications – Web Mining

Example 5 - Web log file analysis

Which links are frequently visited together?



Assoziationsanalyse

Welche WWW-Seiten werden zusammen besucht?

Jeder Besucher von travel2.html geht automatisch auch zu living.html (100%)
-> Ist das so gewollt? Verbesserungsbedarf in der Link-Struktur?

Support(%)	Confidence(%)	Type	Lift	Rule Body		Rule Head
0.5618	100.0000	+	35.6..	[travel2.html]	=..	[living.html]
0.5618	20.0000	+	35.6..	[living.html]	=..	[travel2.html]
0.5618	50.0000	+	17.8..	[shop64.html]	=..	[kids.html]
0.5618	50.0000	+	17.8..	[shop64.html]	=..	[living.html]
0.5618	20.0000	+	17.8..	[living.html]	=..	[shop64.html]
0.5618	20.0000	+	17.8	[kids.html]	=..	[shop64.html]
[illegible]	[illegible]	+	[illegible]	[illegible]	=..	[illegible]
[illegible]	[illegible]	+	[illegible]	[illegible]	=..	[illegible]
[illegible]	[illegible]	+	[illegible]	[illegible]	=..	[illegible]
[illegible]	[illegible]	+	[illegible]	[illegible]	=..	[illegible]
0.8427	30.0000	+	10.6..	[kids.html]	=..	[living.html]
0.8427	30.0000	+	10.6..	[living.html]	=..	[kids.html]
1.1236	40.0000	+	8.9000	[newspaper.html]	=..	[info.html]

Rules (t/r/u/s): 264/0/0/264

Association Analysis

Multidimensional Variant

Club Karstadt registration form

Special offer,
Postal code,
branch,
quantity,
price,
tag ...

Note: To account for the "environment" variables, they are included and treated in the same way as the previous items (products) in the analysis.

Association Analysis

Remarks & application recommendations – Teil I

Provided that frequent item sets with n elements exist, the A priori algorithm requires n iterations (1 MB shopping cart analyzable in seconds).

Algorithm discovers thousands of rules.

Visualization or **browsing** is necessary:

- Rules with very high support or confidence values are mostly already known.
- The so called **"medium" rules** are mainly interesting.

→ Thus, additional statistical key figures are necessary for the evaluation.

Many **external influencing factors** remain unconsidered, since transaction data do not include the following:

- socio-economic status of buyers
- time (time of day, day of week)
- Advertising campaigns

Association Analysis

Remarks & application recommendations – Teil II

Often association rules do not fulfill the expectations placed in them:

Often **trivial rules** are substituted:

- Computer purchase Completion Computer maintenance contract.
- Paint → brush
- …

Inexplicable rules lead to confusion

- New store opening causes changed buying behavior for many customers

Rules are the result of previous marketing action(s), but these were not mapped and are therefore not taken into account.

Rules do **not imply a (clear) action**

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Thank you for your attention