Big Data Anwendungen

Recommender Systems

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation

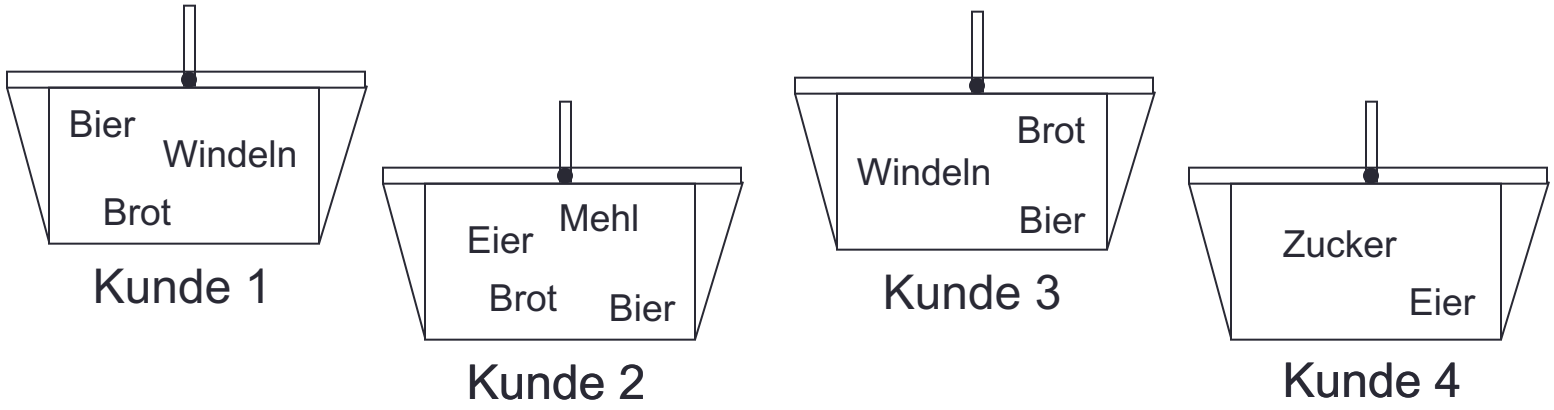
- Recommender Systems
  - Assoziationsregeln
  - Evaluation von Assoziationsregeln
  - Einsatzmöglichkeiten
  - Content-based Recommendations
  - Collaboratives Filtering

- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Motivation – Warenkorbanalyse I

- Gesucht: Einkaufsgewohnheiten
  - Höhere Kundenzufriedenheit durch günstige Anordnung
  - Höherer Absatz durch ungünstige Anordnung

- Warenkörbe (Beispiel)



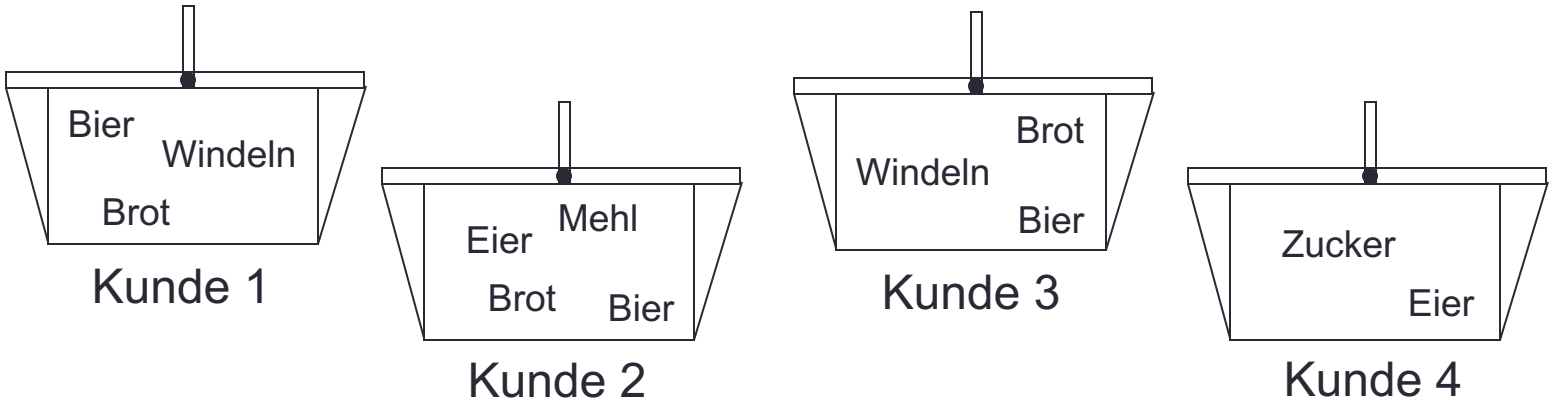
- Fragestellung: Welche Kombinationen werden häufig gekauft (Frequent Itemsets)?

Assoziationsregeln

- Darstellung von Assoziationsregeln
  - Antecedent X ⇨ Consequent Y
- Wahrscheinlichkeitsbasierter Charakter
  - Consequent Y ist mit der Wahrscheinlichkeit P wahr,
  - … wenn der Antecedent X wahr ist
  - Bedingte Wahrscheinlichkeit $P(Y|X)$!
- Zugelassene Wertebereiche
  - Besonders geeignet für kategorische Daten
  - Möglichkeit Grenzwerte für kontinuierliche Werte zu setzen

Motivation – Warenkorbanalyse II

- Warenkörbe



- Frequent Itemsets (mit mind. 2 Items)
  - {Brot, Bier}, {Brot, Bier, Windeln}, {Bier, Windeln}, {Brot, Windeln}

- Wie lassen sich aus Frequent Itemsets Assoziationsregeln ableiten?
  - Beispiel: Wer Windeln kauft, kauft auch Brot

Support

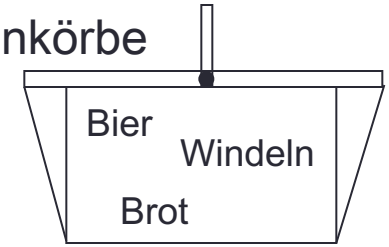
- Alternative Namen
  - Häufigkeit, Abdeckung
- Idee
  - Angabe bezüglich der Häufigkeit eines Portfolios
  - Anzahl bzw. Anteil der Transaktionen, die $X \cap Y$ enthalten
- Formal:
  - $supp(X \rightarrow Y) = P(X \cap Y) = \frac{|X \cap Y|}{|D|}$ mit D sind alle Elemente
- Beispiel:
  - Die Kombination Windeln, Bier tritt in 50% der Warenkörbe auf.
  - Support = 50%

Confidence

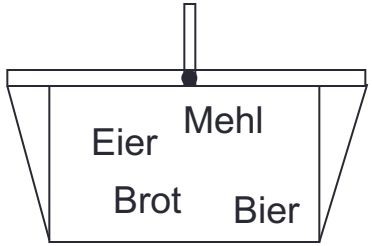
- Alternative Namen
  - Genauigkeit
  - „Überraschungsmass“
- Idee
  Wenn eine Transaktion X enthält, dann auch Y (mit gegebener Genauigkeit)
- Formal:
  $conf(X \rightarrow Y) = P(Y|X) = \frac{|X \cap Y|}{|X|} = \frac{supp(X \rightarrow Y)}{supp(X)}$
- Beispiel:
  - Wenn Windeln gekauft wurden, wurde in 100% aller Fälle auch Bier gekauft
  - Confidence = 100%
- Ziel: Finden von Regeln mit
  - … hohem Support (support > minSup) und …
  - … hoher Confidence (confidence > minConf)

Beispiel – Warenkorbanalyse

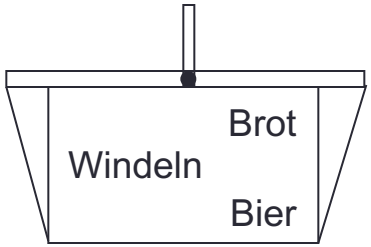
- Warenkörbe



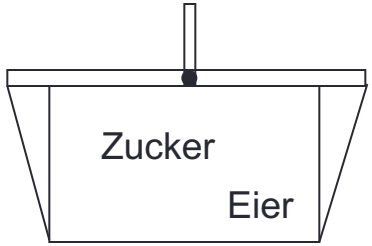
Kunde 1



Kunde 2



Kunde 3

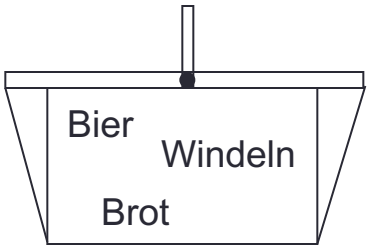


Kunde 4

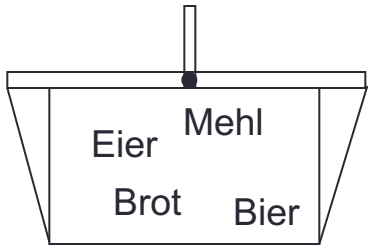
- Frequent Itemsets (minSup = $\frac{1}{2}$)
  - {Brot, Bier} (Support = $\frac{3}{4}$); {Brot, Bier, Windeln} (Support = $\frac{1}{2}$); {Bier, Windeln} (Support = $\frac{1}{2}$); {Brot, Windeln} (Support = $\frac{1}{2}$)
- Assoziationsregeln (Auswahl)
  - Brot $\Rightarrow$ Bier (Confidence = $\frac{3}{3}$)
  - Brot, Bier $\Rightarrow$ Windeln (Confidence = $\frac{2}{3}$)
  - Zucker $\Rightarrow$ Eier (Confidence = $\frac{1}{1}$)

Beispiel – Venn Diagramm

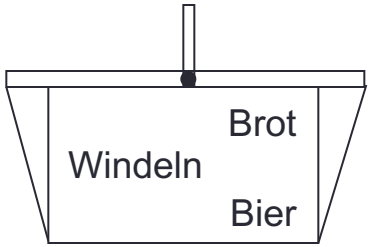
- Assoziation Rule: Brot, Bier ⇒ Windeln



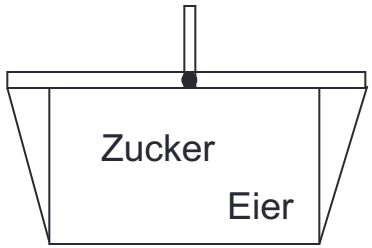
Kunde 1



Kunde 2

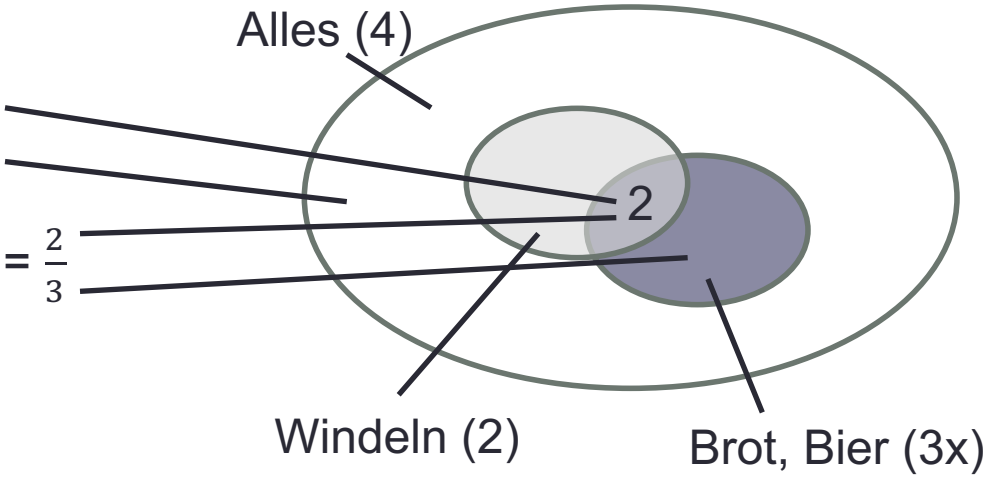


Kunde 3



Kunde 4

- Support (Brot, Bier ⇒ Windeln) = $\frac{2}{4}$
- Confidence (Brot, Bier ⇒ Windeln) = $\frac{2}{3}$



A-Priori Eigenschaft

- Herausforderung
  Datenbestände zum Finden von Assoziation Rules oft groß
  ⇒ effizienter Algorithmus notwendig

- Beobachtung
  - Itemset häufig, wenn Supermenge häufig
  - Itemset genau dann häufig, …
    … wenn alle Teilmengen häufig

- Beispiel:
  - {Bier, Windeln, Brot} häufig
  - {Bier, Windeln}, {Bier, Brot}, {Windeln, Brot} und
    {Bier}, {Windeln}, {Brot} häufig

- Damit mögliches Vorgehen
  Bestimmung von Frequent Itemsets mit n Elementen aus Frequent Itemsets mit $(n - 1)$ Elementen möglich

A-Priori Algorithmus – Frequent Itemsets

- Finden aller Itemsets mit ausreichendem Support:

- Beginn mit einelementigen Sets (1)-Sets:
  - einfaches Abzählen

- Berechnung der k-Sets aus den (k-1)-Sets:
  - Join-Step: Ermittlung von Kandidaten;
    Aus A-Priori Eigenschaft:
    Alle (k-1)-elementigen Teilmengen
    eines k-Sets sind (k-1)-Sets,
  - Prune-Step: Löschen aller Kandidaten,
    die eine „unzulässige“
    (k-1)-elementige Teilmenge haben.
  - Support Counting, d. h. Abzählen,
    wie häufig die Kandidaten wirklich sind.

A-Priori – Beispiel I

- Gegeben
  - Warenkörbe (rechts)
  - Mindest Support 2/4, d.h. 2 Warenkörbe

Warenkörbe			
Bier	Brot	Windeln	
Bier	Brot	Eier	Mehl
Bier	Brot	Windeln	
Eier	Zucker		

- Einelementige Warenkörbe

Warenkörbe	Anzahl
Bier	3
Brot	3
Eier	2
Windeln	2
Zucker	1
Mehl	1

- Streichen von Zucker und Mehl

A-priori – Beispiel II

- Zweielementige Warenkörbe

Warenkörbe		Anzahl
Bier	Brot	3
Bier	Eier	1
Bier	Windeln	2
Brot	Eier	1
Brot	Windeln	2
Eier	Windeln	0

Warenkörbe			
Bier	Brot	Windeln	
Bier	Brot	Eier	Mehl
Bier	Brot	Windeln	
Eier	Zucker		

Warenkörbe	Anzahl
Bier	3
Brot	3
Eier	2
Windeln	2

- Dreielementige Warenkörbe

Warenkörbe			Anzahl
Bier	Brot	Windeln	2

Frequent Pattern Growth (FP-Growth)

- Nachteile A priori Algorithmus
  - Anzahl möglicher Kandidaten kann sehr groß sein (insbesondere die mit ein und zwei Elementen)
  - Hohe Anzahl von kompletten „Datenscans“ (für Big Data also nur bedingt geeignet)

- Idee
  - Stufe 1: Ableiten des FP-Trees (Ableiten häufiger Itemsets in einem Baum)
  - Stufe 2: Ableiten der Frequent Itemsets (unter Einsatz des Baums anstelle von „Datenscans“)

- Vorteil
  - Datenbank muss nur zwei Mal komplett durchlaufen werden
  - Algorithmus ist bedeutend schneller als Apriori (bei identischen Ergebnissen)

FP-Growth – Beispiel I

- Durchführung 1. Scan der Daten...
  ... und zählen der Vorkommen der Elemente

- Hier:
  - 1: Mehl, Zucker
  - 2: Windeln, Eier
  - 3: Bier, Brot

Warenkörbe			
Bier	Brot	Windeln	
Bier	Brot	Eier	Mehl
Bier	Brot	Windeln	
Eier	Zucker		

- Entscheider gibt minimalen Support an

- Hier: minimaler Support ist 2
  Alle Elemente mit geringerem Support werden nicht mehr berücksichtigt!

	Relevante Objekte			
Index	0	1	2	3
Element	Brot	Bier	Eier	Windeln

Ohne Elemente mit geringem Support

Elemente sind absteigend nach Häufigkeit sortiert

FP-Growth – Beispiel II

	Relevante Objekte			
Index	0	1	2	3
Element	Brot	Bier	Eier	Windeln

Warenkörbe			
Bier	Brot	Windeln	
Bier	Brot	Eier	Mehl
Bier	Brot	Windeln	
Eier	Zucker		

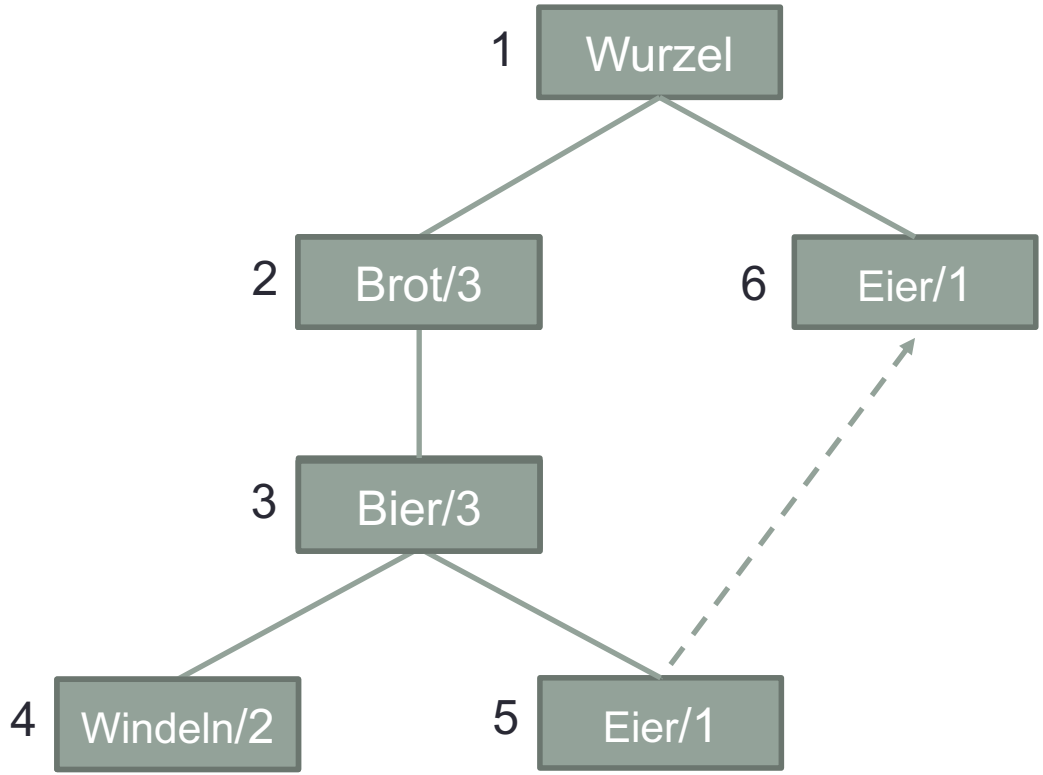
- Ableiten einer Tabelle der Warenkörbe...
  ... ohne seltene Artikel mit Sortierung

Warenkörbe		
Brot	Bier	Windeln
Brot	Bier	Eier
Brot	Bier	Windeln
Eier		

- Diese Tabelle wird direkt (also ohne Speicherung) in einen Baum überführt (siehe nächste Seite)

FP-Growth – Beispiel III

- Sequentielles Einfügen der Warenkörbe...
  ... in einen Baum

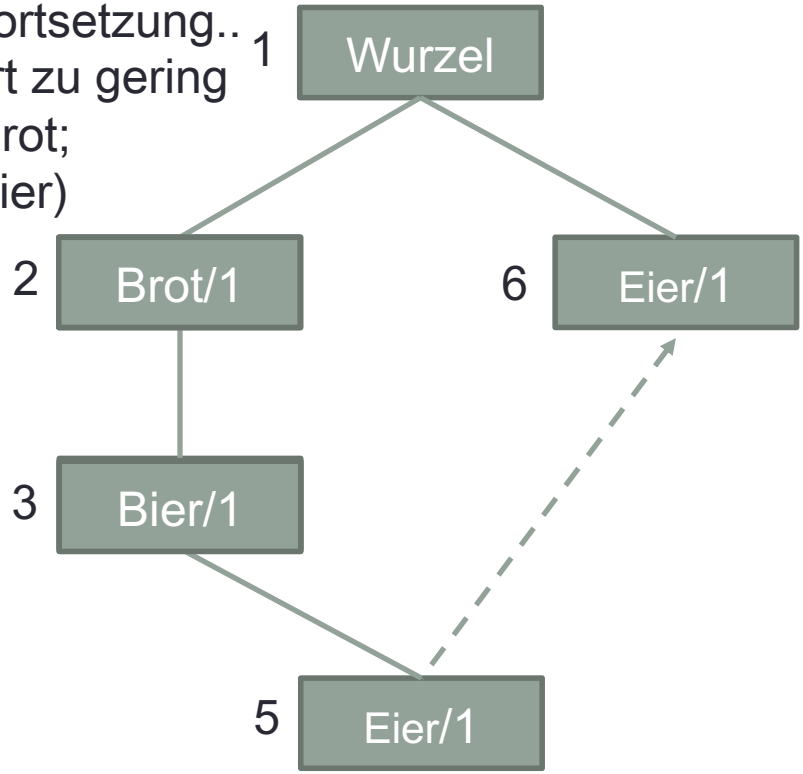


Warenkörbe		
Brot	Bier	Windeln
Brot	Bier	Eier
Brot	Bier	Windeln
Eier		

Anzahl	Elemente
3	Brot, Eier
2	Windeln, Eier

FP-Growth – Beispiel IV

- Finden der Frequent Itemsets
  - Entfernen aller Äste, die nicht auf Pfad...
    ... zwischen betrachtetem Knoten und Wurzel
  - Anpassung der Häufigkeiten
  - Rekursive Fortsetzung..
    .. bis Support zu gering
    (hier: Eier, Brot;
    Eier; Bier)

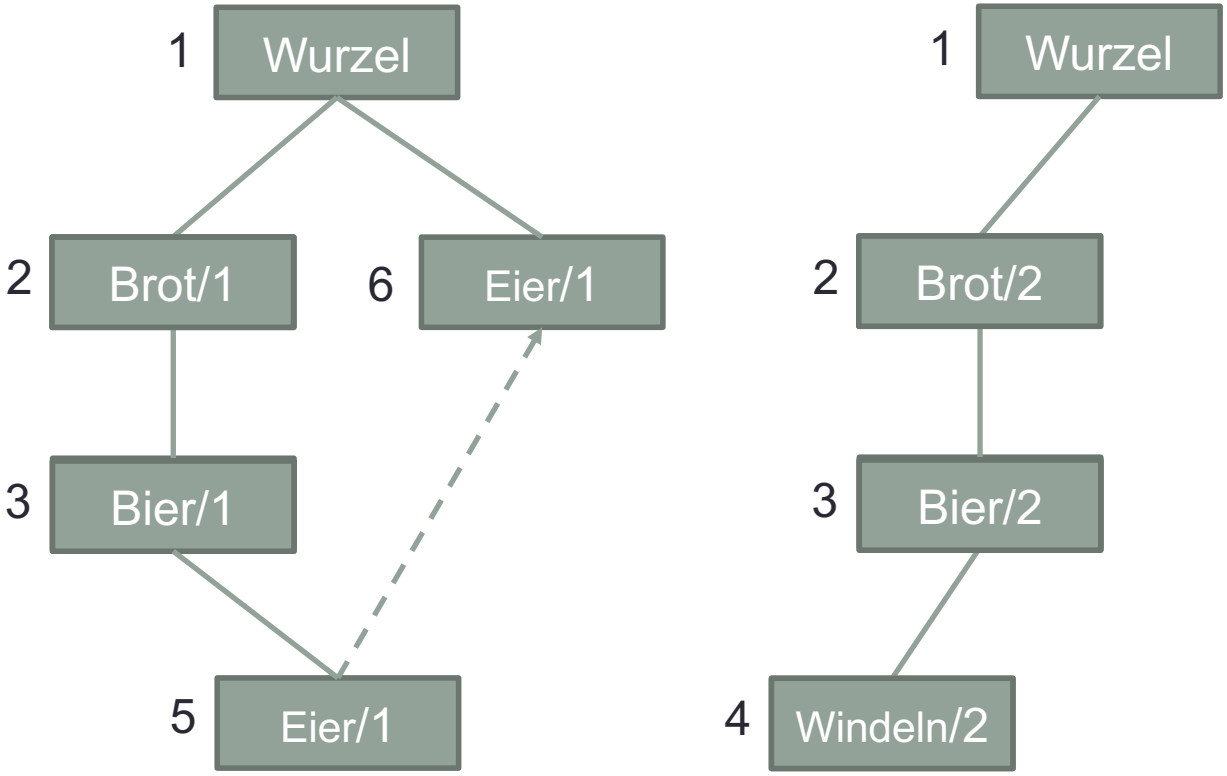


Warenkörbe		
Brot	Bier	Windeln
Brot	Bier	Eier
Brot	Bier	Windeln
Eier		

Anzahl	Elemente
3	Brot, Eier
2	Windeln, Eier

FP-Growth – Beispiel V

- Größte Teilbäume bevor Support zu gering



- Abgeleitete Frequent Itemsets
  {Eier}, {Windeln, Bier, Brot}

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation

- Recommender Systems
  - Assoziationsregeln
  - Evaluation von Assoziationsregeln
  - Einsatzmöglichkeiten
  - Content-based Recommendations
  - Collaboratives Filtering

- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Nachteile von Support und Confidence

- Bewertung Support
  - Wiederholung: $supp(X \rightarrow Y) = P(X \cap Y) = \frac{|X \cap Y|}{|D|}$
  - Probleme
    - Hoher Support bei häufigen Beobachtungen (meist wenig überraschend)
    - Geringer Support bei seltenen Beobachtungen (meist nicht relevant)
  - Support hilft kaum bei Bewertung der Güte von Assoziationsregeln

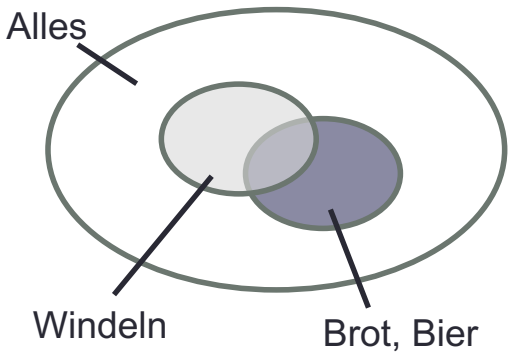
- Bewertung Confidence
  - Wiederholung: $conf(X \rightarrow Y) = P(Y|X) = \frac{|X \cap Y|}{|X|} = \frac{supp(X \rightarrow Y)}{supp(X)}$
  - Entspricht bedingter Wahrscheinlichkeit von Y gegeben X
  - Problem:
    - Häufigkeit von Y nicht beachtet
      In unserem Beispiel: Brot → Bier hat hohe Confidence
      (Bier ist aber in 3 der 4 Warenkörbe, damit kaum interessant)

Lift

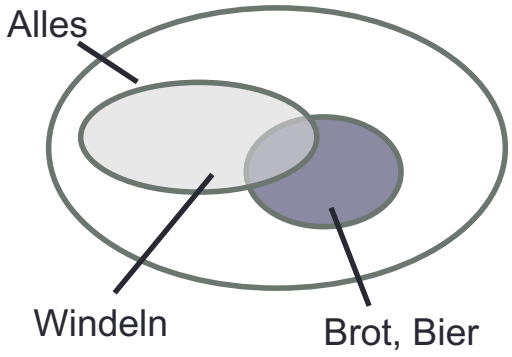
- Idee: Teilen der Confidence durch den Support von Y
- Bereits mit ähnlicher Intuition bekannt aus Kapitel Klassifikation
- Definition

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} = \frac{supp(X \cap Y)}{supp(X) \cdot supp(Y)} = \frac{|X \cap Y| \cdot |D|}{|X| \cdot |Y|}$$

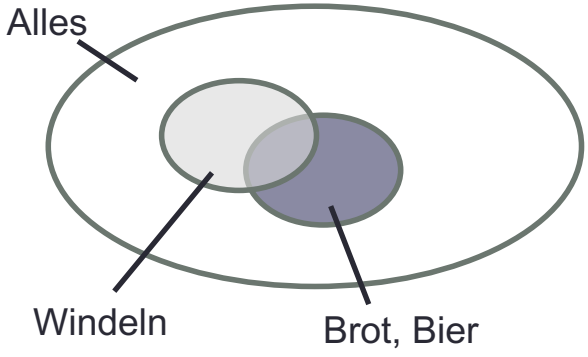
- Beispiel: Brot, Bier ⇒ Windeln



Referenz



Sinkt bei wachsendem Y



Steigt bei wachsendem D

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation

- Recommender Systems
  - Assoziationsregeln
  - Evaluation von Assoziationsregeln
  - Einsatzmöglichkeiten
  - Content-based Recommendations
  - Collaboratives Filtering

- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Assoziationsregeln zur Klassifikation – Vorgehen

- Idee
  - Antecedents sind Eingangsparameter
  - Consequent ist vorhergesagte Klasse

- Eine Regel passt:
  $\Rightarrow$ Klassifikation eindeutig (mit Konfidenz der Regel)

- Keine Regel passt:
  $\Rightarrow$ Mehrheitsklasse bzw. unklassifiziert

- Mehrere Regeln passen:
  - Berücksichtigung der Regel mit höchster Konfidenz
    - Regel entscheidet
  - Berücksichtigung der k Regeln mit höchster Konfidenz (oder auch aller Regeln)
    - Häufigste auftretende Klasse
    - Klasse mit höchster durchschnittlicher Konfidenz der Regeln
  - …

Assoziationsregeln zur Klassifikation - Beispiel

- Assoziationsregeln
  - Bier, Brot ⇒ Windeln (conf: 2/3)
  - Brot, Windeln ⇒ Bier (conf: 2/2)
  - Bier, Windeln ⇒ Brot (conf: 2/2)
  - Brot ⇒ Bier (conf: 3/3)
  - Windeln ⇒ Bier (conf: 2/2)

Warenkörbe			
Bier	Brot	Windeln	
Bier	Brot	Eier	Mehl
Bier	Brot	Windeln	
Eier	Zucker		

- Vorherzusagende Klasse
  (Kunde kauft) Bier

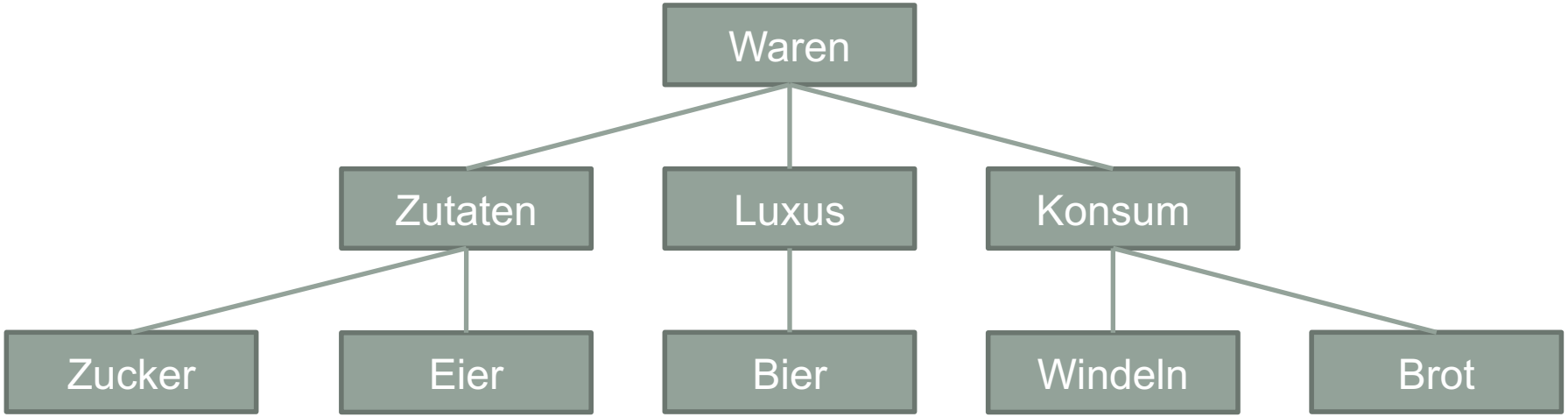
- Vorhersagen mit Hilfe der Assoziationsregeln
  - (Mehl, Eier) ⇒ kein Bier
  - (Zucker) ⇒ kein Bier
  - (Brot, Windeln) ⇒ Bier
  - (Brot, Zucker) ⇒ Bier

Sequenzanalysen

- Zentrale Änderung
  - Statt Items jetzt „Transaktionen“
  - Transaktionen: Item + Transaktionszeitpunkt
  - Oftmals zeitliche Abfolge von Transaktionen bekannt
- Beispiele für Sequenzen (Abfolgen von Transaktionen)
  - Warenkorbanalyse: Bezahlung erst mit Kreditkarte, dann mit Kundenkarte
  - Versandhandel: Erst Kauf von Hose, dann Bluse, dann Badeanzug
- Sequenzanalyse
  - Betrachtung eines Zeitraums
  - Frequent Itemsets nicht mehr beliebig in Reihenfolge, folgen Sequenz
- Beispiele typischer Sequenzen
  - Kunde besucht Onlineshop, stellt Warenkorb zusammen, verlässt Seite
  - Kunde kauft Fahrrad, kauft Auto, kauft Haus

Berücksichtigung von Taxonomien

- Taxonomie: Hierarchische Darstellung von „Abstraktion“



- Mögliche Anpassung der Bestimmung von Frequent Itemsets
  - Finden von ebenen-übergreifenden Warenkörben mit „Zutaten“
    Beispiel: Bier $\Rightarrow$ Zutaten
  - Finden von Warenkörben auf höherem Abstraktionsniveau

- Sinnvoll, bei
  - Geringer Beobachtungszahl auf unterer Taxonomieebene
  - Erweiterbar um zusätzliche Äste bspw. für „Sonderangebot“

Assoziationsregeln – Vor- und Nachteile

- Vorteile
  - Potentiell interessante Erkenntnisse
    - Bier $\Rightarrow$ Windeln
    - Warenkorbabbrüche
    - ...
  - Berechnung in überschaubarer Zeit möglich
  - Betrachtet andere Problemklasse als traditionelle statistische Verfahren

- Nachteile
  - Hoher manueller Aufwand
    (Separierung „interessanter“ von „uninteressanten“ Assoziationsregeln)
  - Integration weiterer Informationen nicht/kaum möglich
    (In Klassifikator kann alles eingefügt werden, hier nur Inhalt Warenkorb, ...)
  - Assoziationsregeln erfüllen an sie gestellte Anforderungen oft nicht
    - Unerklärbare Regeln (Ladeneröffnung führt zu anderem Verhalten)
    - Regel stark von Marketingaktionen beeinflusst
    - Keine Handlungsimplikationen

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation

- Recommender Systems
  - Assoziationsregeln
  - Evaluation von Assoziationsregeln
  - Einsatzmöglichkeiten
  - Content-based Recommendations
  - Collaboratives Filtering

- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Recommender Systems – Idee

- Assoziationsregeln revisited
  - Identifikation von Implikationen: „Wer A kauft, kauft auch B“
  - Assoziationsregeln entgegen aktuellem Trend in Big Data: Finden Regeln für alle, statt individuelle Empfehlungen (Generalisierung statt Individualisierung)

- Modernerer Ansatz: Recommender Systems
  - Identifikation ähnlicher Kunden
  - Empfehlung von Artikeln auf Basis ähnlicher Kunden

Content-Based Recommendations I

- Idee
  - Inhalte werden über Attribute beschrieben
  - Für User werden typische Attribute identifiziert (direkt abgefragt oder aus historischem Verhalten)
  - Empfehlungen basieren auf Ähnlichkeiten zwischen Artikel und User

- Matrix

	Krimi	Komödie	Thriller	Animation	Überlänge
Warrior	1				
The Body			1		1
Ted 2		1		1	

Enthält Informationen, die auch über User bekannt/ableitbar sind

Content-Based Recommendations II

- Methoden zum Ableiten des Uservektors
  - Direktes Feedback (z.B. likes / dislikes)
  - Klassifikatoren (zur Vorhersage einzelner Attribute)

- Für jeden Uservektor werden Ähnlichkeiten...
  ... zu allen Items ermittelt.

- Items werden nach Ähnlichkeit absteigend sortiert...
  ... und top Items ausgespielt

- Beispiele
  - Abspielen weiterer Songs auf Basis von Content-Based Recommendations (bei Pandora Radio)
  - Vorschläge Alternativer Kinofilme (bei Rotten Tomatoes)
  - Vorschlag ähnlicher Artikel bei Bestellungen im Call Center (bei Versandhändlern)

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation

- Recommender Systems
  - Assoziationsregeln
  - Evaluation von Assoziationsregeln
  - Einsatzmöglichkeiten
  - Content-based Recommendations
  - Collaboratives Filtering

- Clusteringverfahren
- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Collaborative Filtering

- Idee
  - Aufbau einer Matrix mit Benutzern in Zeilen,...
    ... Verhalten in Spalten
  - Ähnliche User werden identifiziert (auf Basis der Matrix)
  - Identifikation ähnlicher User mit Ähnlichkeitsmaß
  - Empfehlungen beruhen auf Items der top k ähnlichsten User

- Anmerkungen
  - Collaborative Filtering führt zu großen, aber dünnbesetzten Matritzen (vergleiche Filmbeispiel!)
  - Kann mit Content-Based Recommendations kombiniert werden
  - Anwendung von Collaborative Filtering häufiger...
    ... als Content-Based Recommendations

Nutzenmatrix

- Grundlage von Recommender Systems bildet eine „Nutzen“-Matrix

	Warrior	The Body	Ted 2	Minions	Triangle	It follows	Halo 4	Splice	Ungezähmt	Snatch	Born to Race	Sing Street
Kerstin				4		3			1			5
Stephan	1	3	5	1	2	3	5				1	
Julius				8								
Justus	5	2										
Niklas			5	3	5							
Silke												
Jesko		3		4		6		3		5		

Werte können auch auf 1 normiert sein
(Bspw. hat (nicht) gekauft)

Bestimmung der Ähnlichkeit von Usern

- Anwendung des Pearson Korrelationsmaßes
  - $Pearson(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^{S}(x_i - \hat{x}) \cdot (y_i - \hat{y})}{\sqrt{\sum_{i=1}^{S}(x_i - \hat{x})^2} \cdot \sqrt{\sum_{i=1}^{S}(y_i - \hat{y})^2}}$

    mit $\bar{X} = (x_1 \dots x_n)$ bzw. $\bar{Y} = (y_1 \dots y_n)$ sind Vektoren der User...
    ... und $\hat{x} = \sum_{i=1}^{S} \frac{x_i}{S}$ bzw $\hat{y} = \sum_{i=1}^{S} \frac{y_i}{S}$ die Mittelwerte der Vektoren

- Pearson Korrelationsmaß muss für betrachteten User...
  ... mit allen anderen Usern ermittelt werden

- Top k User mit höchstem Index bei Vergleich mit betrachteten User...
  ... bilden Grundlage für Empfehlungen

- Nachteil des Verfahrens
  - Bewertungen der Nutzer oft nicht gleich verteilt
    (Pessimisten vs. Optimisten)
  - Anpassung der Skalen notwendig

Beispiel

- Grundlage von Recommender Systems bildet eine „Nutzen“-Matrix

	Warrior	The Body	Ted 2	Minions	Triangle	It follows	Ha	Sp	Ur	Sn	Bo	Si
Kerstin				4		3			1			5
Stephan	1	3	5	1	2	3	5				1	
Julius				8								

Nur Betrachtung der besetzten Attribute (auch Betrachtung aller denkbar)

- $Pearson(\overline{X_{Kerstin}}, \overline{X_{Stephan}}) = \frac{(4-3.5)\cdot(1-2.0)+(3-3.5)\cdot(3-2.0)}{\sqrt{(4-3.5)^2+(3-3.5)^2}\cdot\sqrt{(1-2.0)^2+(3-2.0)^2}} = -1.00$
- $Pearson(\overline{X_{Kerstin}}, \overline{X_{Julius}}) = \frac{(4-4)\cdot(8-8)}{\sqrt{(4-4)^2}\cdot\sqrt{(8-8)^2}} = \frac{0}{0}$