Probabilistische Ansätze in der Assoziationsanalyse

Habilitationsvortrag

Dr. Michael Hahsler

Institut für Informationswirtschaft

Wirtschaftsuniversität Wien

Wien, 19. Mai, 2006

Aufbau des Vortrags

1. ***Motivation***
2. ***Assoziationsanalyse mit Assoziationsregeln***
   - Assoziationsregeln (Support-Konfidenz-Framework)
   - Fragen aus betriebswirtschaftlicher Sicht
3. ***Probabilistische Interpretation, Schwächen und Weiterentwicklungen***
   - Probabilistische Interpretation von Assoziationsregeln
   - Schwächen von Assoziationsregeln
   - Lift und Chi-Quadrat-Unabhängigkeitstest
4. ***Probabilistisches Modell***
   - Das Unabhängigkeitsmodell
   - Anwendungen
     - Vergleich von simulierten- und Echtdaten
     - NB-Frequent Itemsets
     - Hyper-Konfidenz
5. ***Ausblick***

Motivation

Motivation

Enorme Datenmengen werden gesammelt. Z.B.:

- ***Transaktionsdaten*** im Einzelhandel (Scanner-Kassen) und E-Commerce
- ***Navigationsdaten*** im Web (Suchmaschinen, Digitale Bibliotheken, Wikis, etc.)

Typische Größe der Daten:

- Supermarkt: 10–500 Produktgruppen und 1000–10.000 Produkte
- Wikipedia (Engl.): ca. 1,1 Millionen Artikel (2006)
- Amazon: ca. 3 Millionen Bücher/CDs (1998)
- Google: ca. 8 Milliarden Seiten (ca. 70% des Webs) im Index (2005)

- Typischerweise 10.000–10 Millionen Transaktionen (Warenkörbe, Sessions, Beobachtungen, etc.)

Motivation

Durch Assoziationsanalyse sollen ***„interessante“ Beziehungen zwischen mehreren Items*** (Produkte, Dokumente, etc.) gefunden werden. Beispiel „Kaufverbund“:

Milch, Mehl und Eier werden häufig gemeinsam gekauft.
Oder
Wenn jemand ***Milch und Mehl*** kauft dann kauft die Person oft auch gleichzeitig ***Eier.***

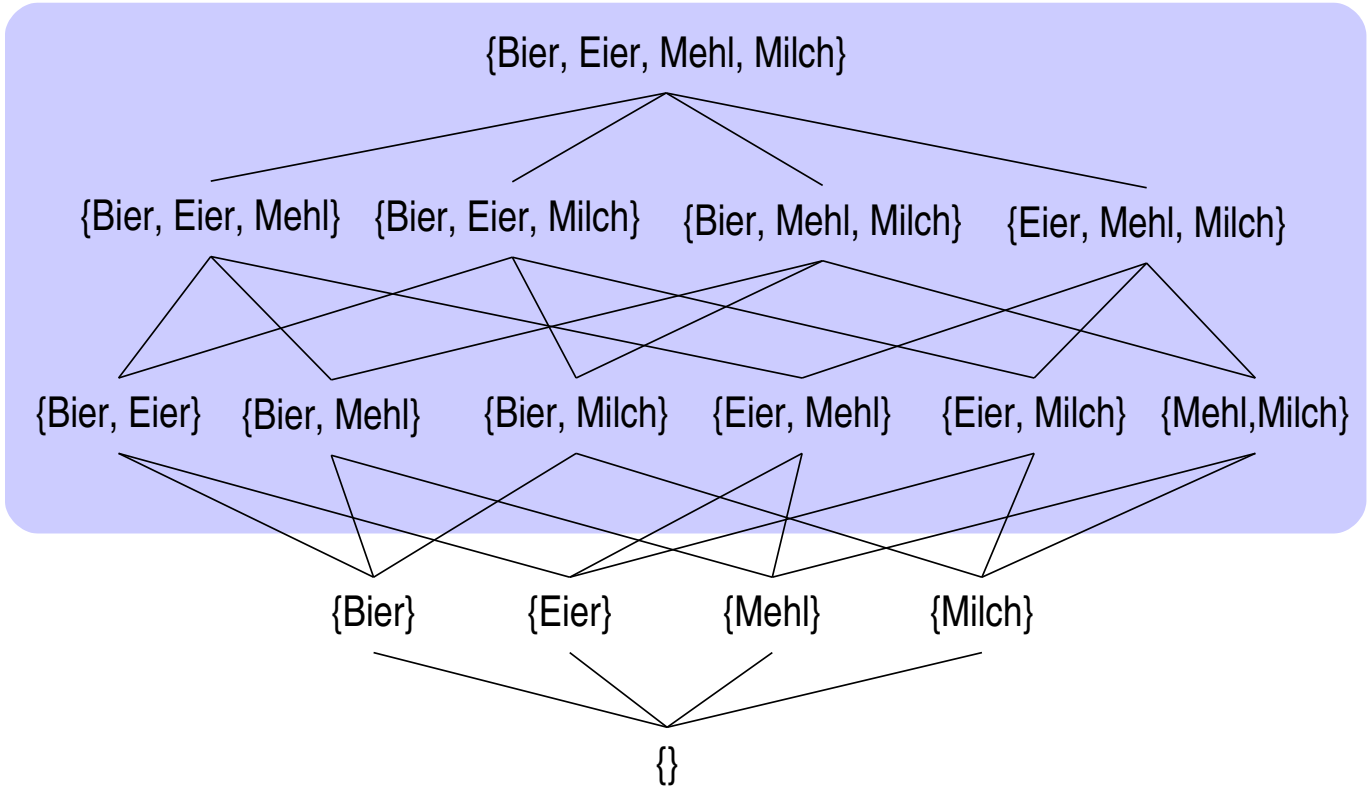
Anwendungsmöglichkeiten von gefundenen Assoziationen:

- Einzelhandel: Anordnung der Produkte im Geschäft, Planung von Aktionen, Sortimentsentscheidungen, etc.
  $\longrightarrow$ ***Explorative Warenkorbanalyse*** (Russell *et al.*, 1997; Berry & Linoff, 1997; Schnedlitz *et al.*, 2001).

- E-Commerce, Dig. Bibliotheken, Suchmaschinen: Personalisierung, autom. Generierung von Vorschlägen
  $\longrightarrow$ ***Recommender Systeme, Item-based Collaborative Filtering*** (Sarwar *et al.*, 2001; Linden *et al.*, 2003).

Motivation

Problem: bei k Items ergeben sich $2^k - k - 1$ mögliche Beziehungen zwischen Items.
Beispiel: Potenzmenge für $k = 4$ Items (dargestellt als Gitter).



Für $k = 100$ ergeben sich bereits mehr als 10^{30} mögliche Beziehungen!

$\rightarrow$ Data Mining: Suche von ***Frequent Itemsets*** und ***Assoziationsregeln***.

Assoziationsanalyse mittels Assoziationsregeln

Transaktionsdaten

Formale Definition:

$I = \{i_1, i_2, \ldots, i_k\}$ sei eine Menge von ***Items***.

$\mathcal{D} = \{Tr_1, Tr_2, \ldots, Tr_n\}$ sei eine Menge von ***Transaktionen***, genannt ***Datenbank***.

Jede Transaktionen in $\mathcal{D}$ hat eine eindeutige Transaktionsnummer und beinhaltet eine Teilmenge der Items in I.

Darstellung als binäre Kaufmatrix:

Transaktionsnummer	Bier	Eier	Mehl	Milch
1	0	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1

Assoziationsregeln

Eine ***Regel*** hat die Form $X \Rightarrow Y$ mit $X, Y \subseteq I$ und $X \cap Y = \emptyset$. Die Teilmengen der Items (abgekürzt ***Itemsets***) X und Y werden ***Antezedent*** (linke Seite) und ***Konsequent*** (rechte Seite) der Regel genannt.

Um ***„interessante“ Assoziationsregeln*** (Agrawal *et al.*, 1993) aus der Menge aller möglichen Regeln auszuwählen werden zwei Maße (*Measures of Interest*) verwendet:

1. Der ***Support*** eines Itemsets Z ist definiert als $\mathrm{supp}(Z) = n_Z/n$.
   $\rightarrow$ Relativer Anteil der Transaktionen in der Datenbank, die Z enthalten.
2. Die ***Konfidenz*** einer Regel $X \Rightarrow Y$ ist folgendermaßen definiert:
   $$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$
   $\rightarrow$ Anteil der Transaktionen die Y enthalten in den Transaktionen die X enthalten.

Jede Assoziationsregel $X \Rightarrow Y$ muss folgende benutzerdefinierte Grenzwerte erreichen:

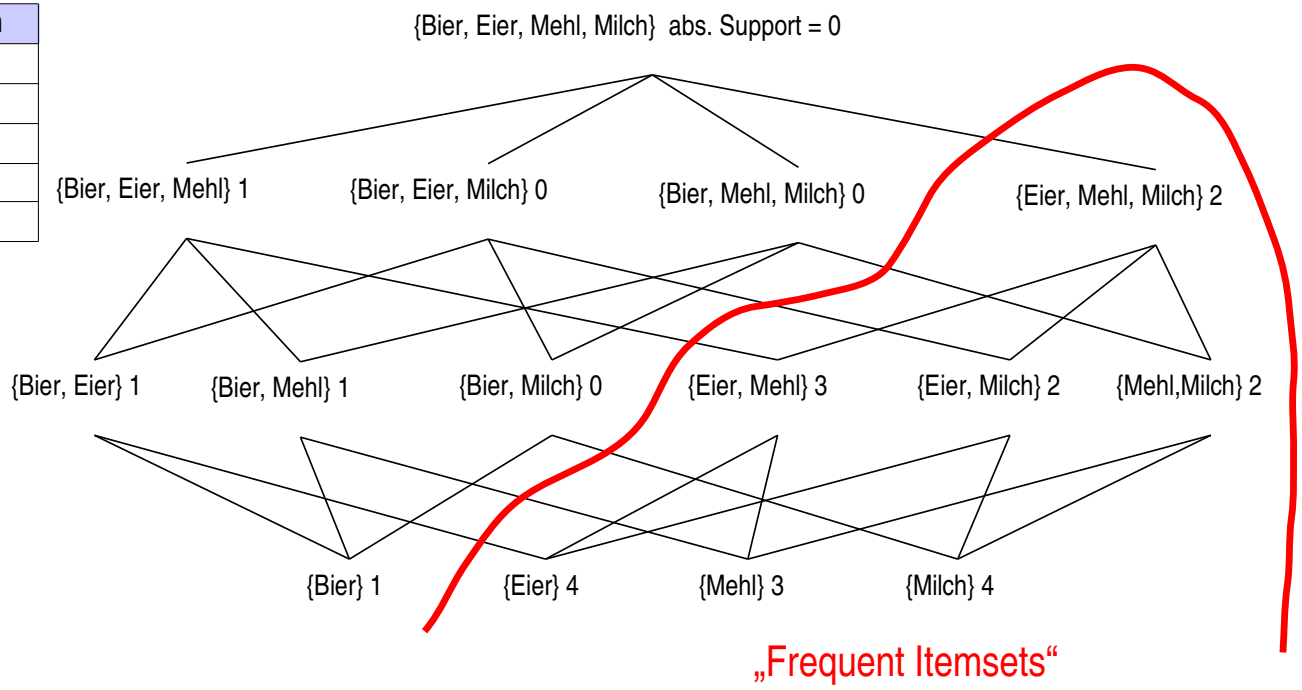
$$\mathrm{supp}(X \cup Y) \geq \sigma$$
$$\mathrm{conf}(X \Rightarrow Y) \geq \gamma$$

Minimum-Support

Idee: Setzen eines benutzerdefinierten Grenzwertes, da Itemsets, die öfter vorkommen, interessanter sind. Z.B. generieren Produkte, die oft gemeinsam gekauft werden mehr Umsatz.

Apriori-Eigenschaft von Support (Agrawal & Srikant, 1994): Support eines Itemsets kann durch Hinzufügen eines Items nicht steigen. Beispiel mit $\mathrm{supp}(Z) \geq 0,4$ (abs. Support ≥ 2):

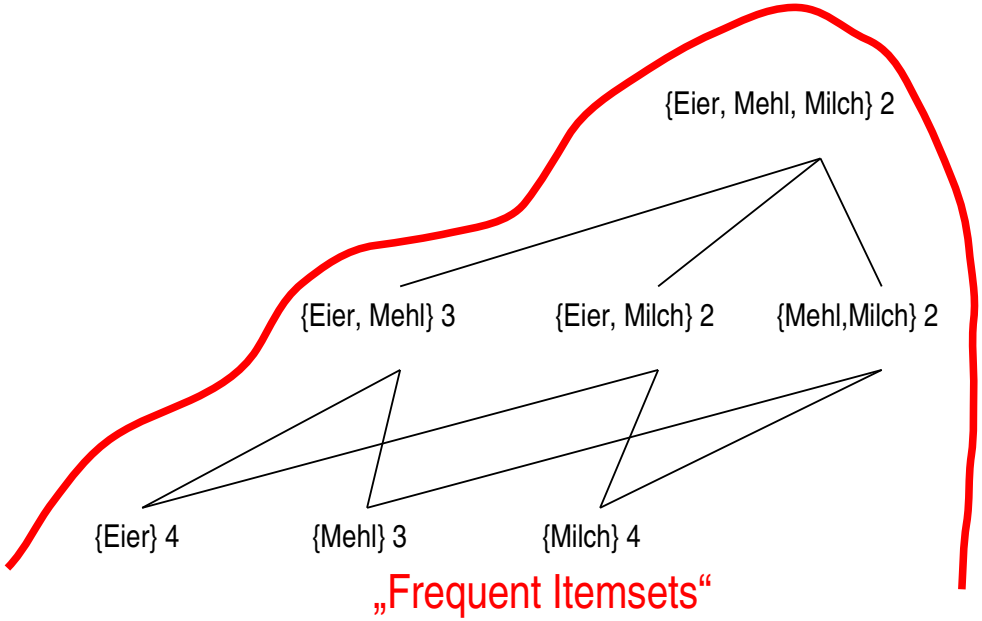
Transaktionsnummer	Bier	Eier	Mehl	Milch
1	0	1	1	1
2	1	1	1	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1



→ Grundlage für effiziente Algorithmen (Apriori, Eclat).

Minimum-Konfidenz

Aus den ***Frequent Itemsets*** werden alle Regeln erzeugt, die den Grenzwert für Konfidenz $\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \geq \gamma$ erreichen.



			Konfidenz
{Eier}	⇒	{Mehl}	$3/4 = 0,75$
{Mehl}	⇒	{Eier}	$3/3 = 1$
{Eier}	⇒	{Milch}	$2/4 = 0,5$
{Milch}	⇒	{Eier}	$2/4 = 0,5$
{Mehl}	⇒	{Milch}	$2/3 = 0,67$
{Milch}	⇒	{Mehl}	$2/4 = 0,5$
{Eier, Mehl}	⇒	{Milch}	$2/3 = 0,67$
{Eier, Milch}	⇒	{Mehl}	$2/2 = 1$
{Mehl, Milch}	⇒	{Eier}	$2/2 = 1$
{Eier}	⇒	{Mehl, Milch}	$2/4 = 0,5$
{Mehl}	⇒	{Eier, Milch}	$2/3 = 0,67$
{Milch}	⇒	{Eier, Mehl}	$2/4 = 0,5$

Bei $\gamma = 0,7$ werden folgende Regeln erzeugt:

			Support	Konfidenz
{Eier}	⇒	{Mehl}	$3/5 = 0,6$	$3/4 = 0,75$
{Mehl}	⇒	{Eier}	$3/5 = 0,6$	$3/3 = 1$
{Eier, Milch}	⇒	{Mehl}	$2/5 = 0,4$	$2/2 = 1$
{Mehl, Milch}	⇒	{Eier}	$2/5 = 0,4$	$2/2 = 1$

Fragen aus betriebswirtschaftlicher Sicht

1. Betriebswirtschaftlich sinnvolle Grenzwerte für Support und Konfidenz?
2. Interpretation der gefundenen Regeln?
3. Bewertung der gefundene Regeln?
4. Risiko durch „falsche“ Regeln?

Probabilistische Interpretation, Schwächen und Weiterentwicklungen

Probabilistische Interpretation von Support und Konfidenz

- Support

$$\text{supp}(Z) = n_Z/n$$

  entspricht dem Schätzer für die ***Auftretenswahrscheinlichkeit*** $P(E_Z)$, dem Ereignisses, dass Z in einer Transaktion enthalten ist.

- Konfidenz kann als Schätzer für die ***bedingte Wahrscheinlichkeit***

$$P(E_Y|E_X) = \frac{P(E_X \cap E_Y)}{P(E_X)}$$

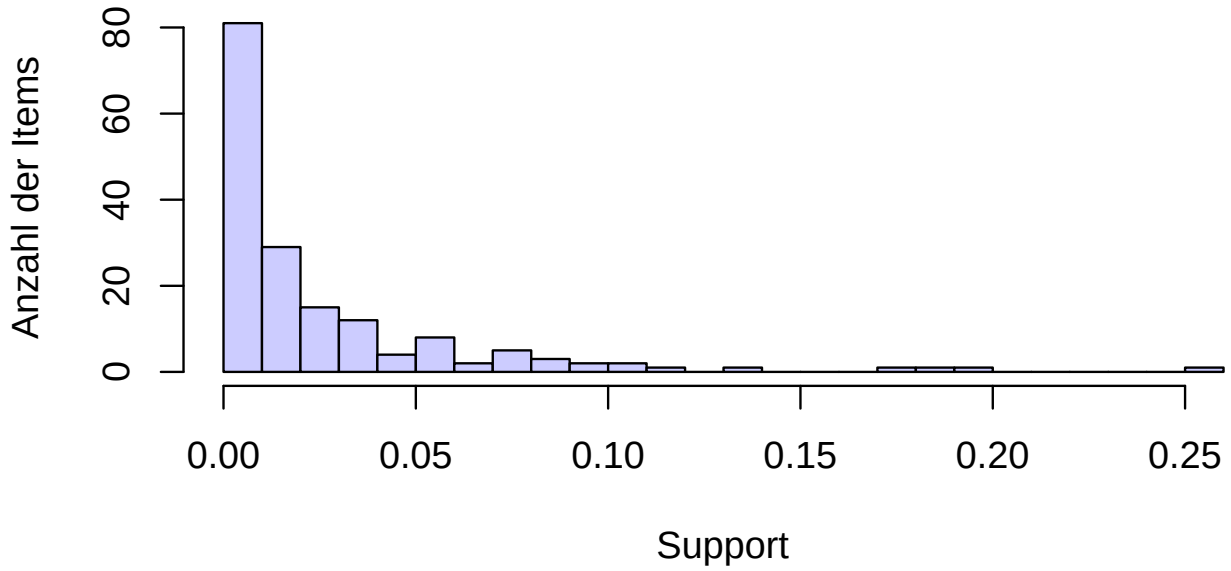
  interpretiert werden. Dies folgt direkt aus der Definition von Konfidenz:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{n_{X \cup Y}}{n_X}.$$

Schwächen von Support und Konfidenz

- Support unterliegt dem ***„Rare Item Problem“*** (Liu *et al.*, 1999a): Selten vorkommenden Items werden ignoriert. Problematisch wenn Produkte die seltener verkauft werden für einen Großteil des Gesamtumsatzes/-gewinns verantwortlich sind.

  Typische Support-Verteilung (Supermarkt POS-Daten mit 169 Items)



- Support nimmt mit der Länge der Itemsets schnell ab. Eine ***Minimum-Support-Schranke bevorzugt daher kurze Itemsets*** (Seno & Karypis, 2005).

Schwächen von Support und Konfidenz

- Konfidenz ignoriert die ***Häufigkeit von*** Y (Aggarwal & Yu, 1998; Silverstein *et al.*, 1998).

	X=0	X=1	Σ
Y=0	5	5	10
Y=1	70	**20**	**90**
Σ	75	**25**	100

$$\text{conf}(X \Rightarrow Y) = \frac{n_{X \cup Y}}{n_X} = \frac{20}{25} = 0,8 = \hat{P}(E_Y|E_X)$$

Konfidenz der Regel ist mit $0,8$ relativ hoch.
Die unbedingte Wahrscheinlichkeit $\hat{P}(E_Y) = n_Y/n = 90/100 = 0,9$ ist aber höher!

- Die ***Grenzwerte für Support und Konfidenz sind benutzerdefiniert.*** In der Praxis werden die Werte so gewählt, dass eine „vernünftige“ Anzahl von Itemsets bzw. Regeln gefunden wird.

  $\longrightarrow$ Aus betriebswirtschaftlicher Sicht möchte man Support anhand von Umsatz/Deckungsbeitrag festlegen oder das Risiko falscher Regeln kontrollieren.

Lift

Das Maß ***Lift*** (Interest Brin *et al.*, 1997) ist definiert als

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \cdot \text{supp}(Y)}$$

und kann als Schätzer für $P(E_X \cap E_Y)/(P(E_X) \cdot P(E_Y))$ interpretiert werden.

$\rightarrow$ Maß für die ***Abweichung von stochastischer Unabhängigkeit:***

$$P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)$$

Im Marketing wird Lift folgendermaßen interpretiert (Betancourt & Gautschi, 1990; Hruschka *et al.*, 1999):

- $\text{lift}(X \Rightarrow Y) > 1 \ldots$ Komplementäreffekte zwischen X und Y
- $\text{lift}(X \Rightarrow Y) < 1 \ldots$ Substitutionseffekte zwischen X und Y

Beispiel

	X=0	X=1	Σ
Y=0	5	5	10
Y=1	70	**20**	**90**
Σ	75	**25**	100

$$\text{lift}(X \Rightarrow Y) = \frac{0,2}{0,25 \cdot 0,9} = 0,89$$

Chi-Quadrat-Unabhängigkeitstest

Test auf Signifikanz der Abweichung von stochastischer Unabhängigkeit (Silverstein *et al.*, 1998; Liu *et al.*, 1999b).

Beispiel: Regel $X \Rightarrow Y$ – 2×2 Kontingenztafel ($l = 2$ Dimensionen)

	X=0	X=1	Σ
Y=0	5	5	10
Y=1	70	20	90
Σ	75	25	100

Nullhypothese: $P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)$
Die Teststatistik

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})} \quad \text{mit} \quad E(n_{ij}) = n_{i\cdot} \cdot n_{\cdot j}$$

ist annähernd χ^2-verteilt mit $2^l - l - 1$ Freiheitsgraden.

Ergebnis des Tests für die obige Kontingenztafel: $\chi^2 = 3.7037, \mathrm{df} = 1, \text{p-value} = 0.05429$
$\rightarrow$ Die Nullhypothese (Unabhängigkeit) kann bei $\alpha = 0.05$ gerade nicht verworfen werden.

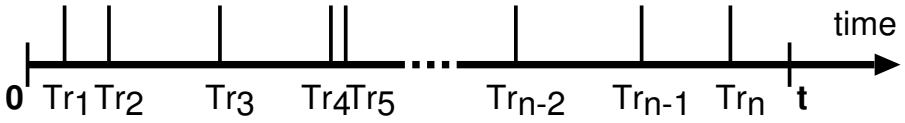
Auch für den Unabhängigkeitstest aller l Items in einem Itemset möglich – l-dimensionale Kontingenztafel.

Schwächen: Schlechte Approximation ($E(n_{ij}) < 5$); mehrfaches Testen.

Probabilistisches Modell

Das Unabhängigkeitsmodell

1. Das Auftreten von Transaktionen folgt einem homogenen Poisson-Prozess mit Parameter θ (Intensität).



$$P(N = n) = \frac{e^{-\theta t}(\theta t)^n}{n!}$$

2. Jedes Item hat eine bestimmte Auftretenswahrscheinlichkeit p_i und jede Transaktion ist das Ergebnis von k (Anzahl der Items) unabhängigen Bernoulli-Versuchen.

	i_1	i_2	i_3	...	i_k
p	0.0050	0.0100	0.0003	...	0.0250
Tr_1	0	1	0	...	1
Tr_2	0	1	0	...	1
Tr_3	0	1	0	...	0
Tr_4	0	0	0	...	0
...	.	.	.	...	.
Tr_{n-1}	1	0	0	...	1
Tr_n	0	0	1	...	1
n_i	99	201	7	...	411

$$P(N_i = n_i) = \sum_{m=n_i}^{\infty} P(N_i = n_i | N = n) \cdot P(N = n) = \frac{e^{-\lambda_i} \lambda_i^{n_i}}{n_i!} \quad \text{mit} \quad \lambda_i = p_i \theta t$$

Anwendung: Vergleich von Echtdaten mit simulierten Daten

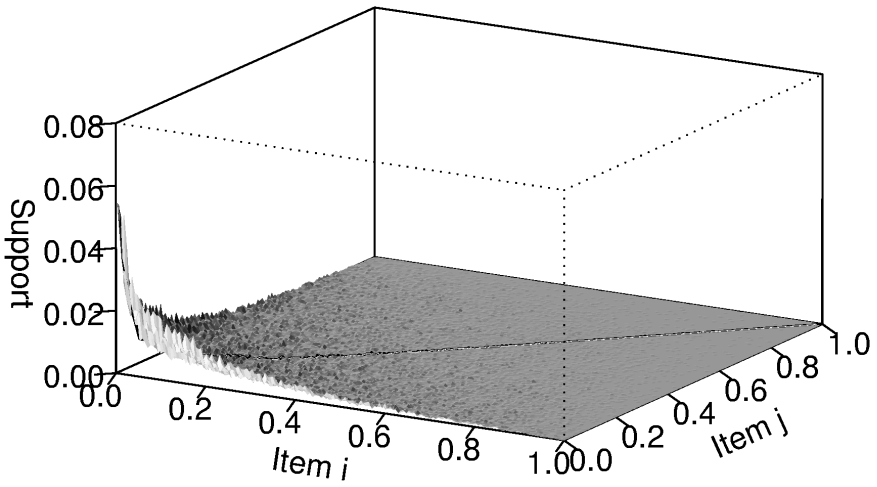
Bisher wurden in der Literatur immer nur eigens konstruierte Beispiel für Probleme von Support, Konfidenz und Lift angeführt (Brin *et al.*, 1997; Aggarwal & Yu, 1998; Silverstein *et al.*, 1998, und Andere)

Idee: Vergleich des Verhaltens der Maße auf Echtdaten und mittels des Unabhängigkeitsmodells simulierter Daten (Hahsler *et al.*, 2006).

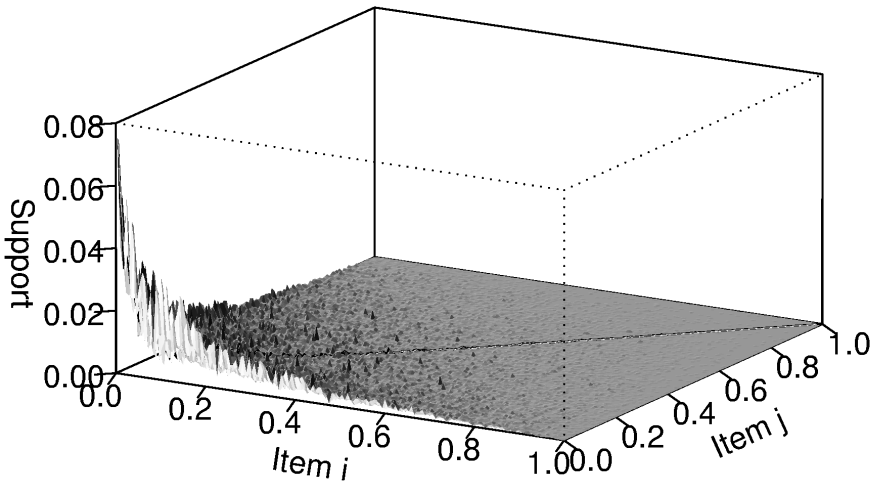
Charakteristiken der verwendeten Echtdaten: Typische Supermarktdaten.

- $t = 30$ Tage
- $k = 169$ Produktgruppen
- $n = 9835$ Transaktionen
- $\theta = n/t = 327,2$ Transaktionen/Tag
- Für p_i werden die beobachteten n_i/n verwendet.

Vergleich: Support



Simulierte Daten

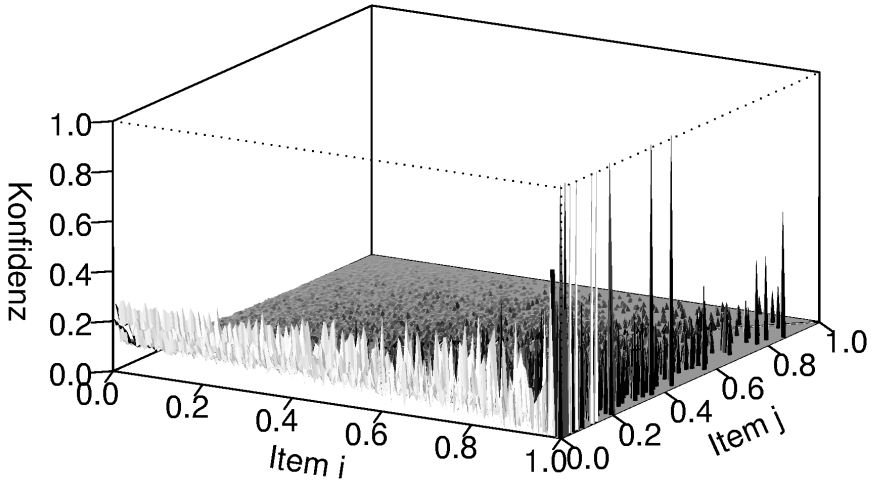


Supermarkt

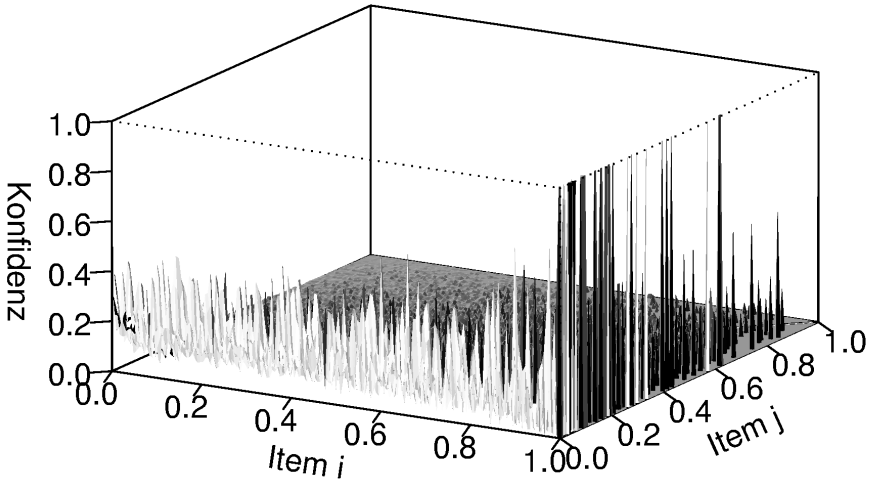
Betrachtet werden nur Regeln: $\{i_i\} \Rightarrow \{i_j\}$

X-Achse: Items i_i absteigend sortiert nach Support.
Y-Achse: Items i_j absteigend sortiert nach Support.
Z-Achse: Support der Regel.

Vergleich: Konfidenz



Simulierte Daten

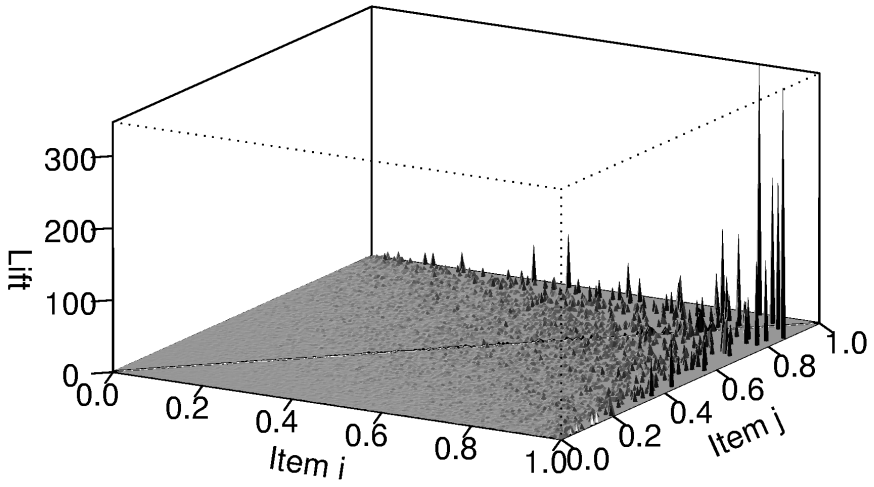


Supermarkt

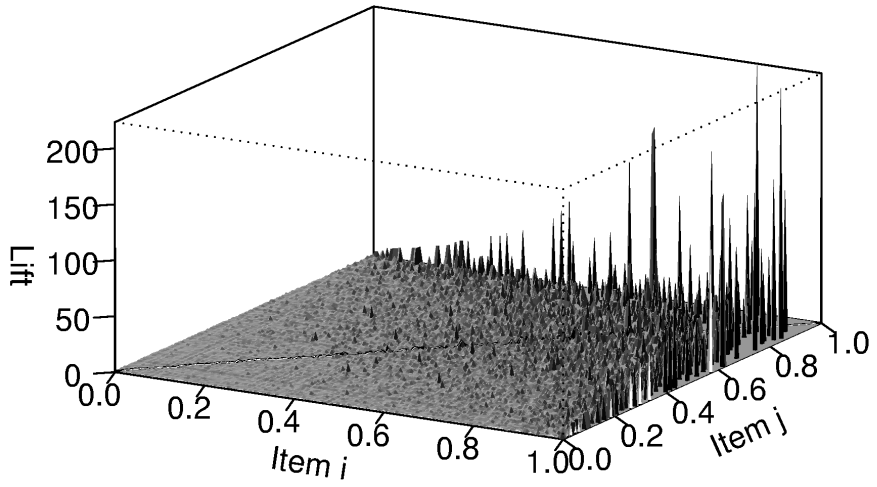
$$\text{conf}(\{i_i\} \Rightarrow \{i_j\}) = \frac{\text{supp}(\{i_i, i_j\})}{\text{supp}(\{i_i\})}$$

- Systematischer Einfluss von Support: Die Konfidenz nimmt mit dem Support des Items in der rechten Seite zu.

Vergleich: Lift



Simulierte Daten

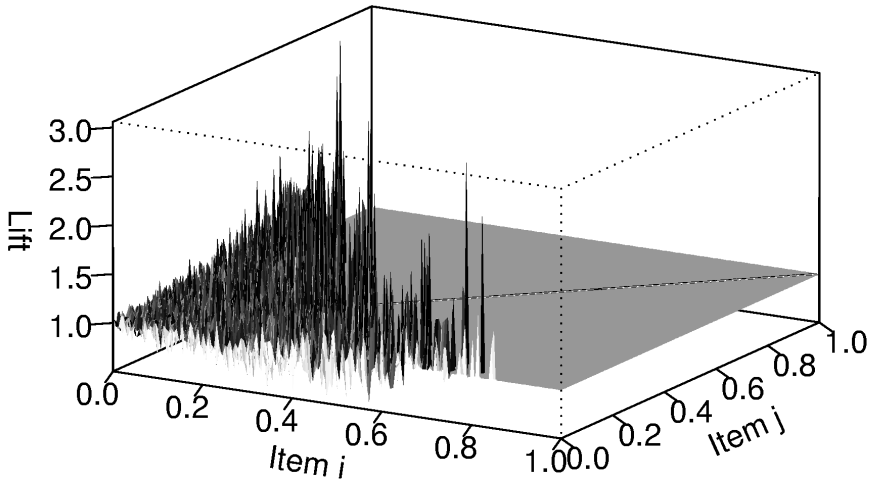


Supermarkt

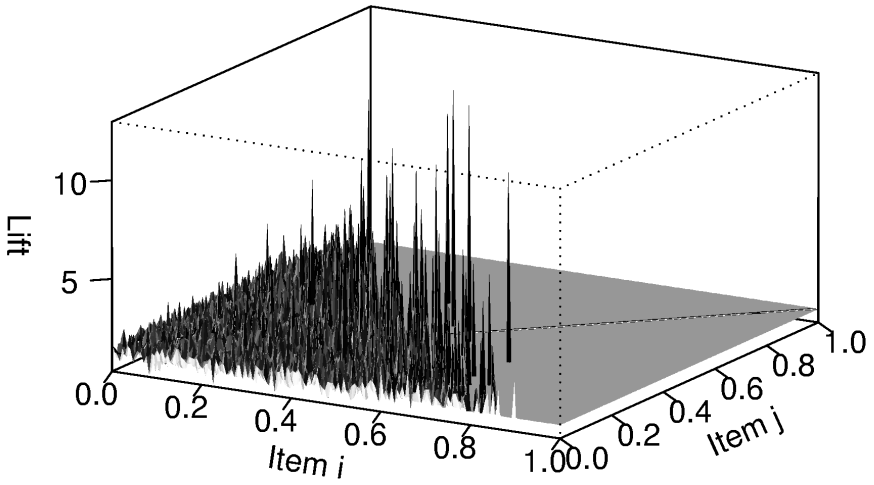
$$\text{lift}(\{i_i\} \Rightarrow \{i_j\}) = \frac{\text{supp}(\{i_i, i_j\})}{\text{supp}(\{i_i\}) \cdot \text{supp}(\{i_j\})}$$

- Ähnliche Verteilungen mit extremen Werten bei Items mit sehr geringem Support.

Vergleich: Lift + Minimum-Support



Simulierte Daten (Support: $\sigma = 0.1\%$)



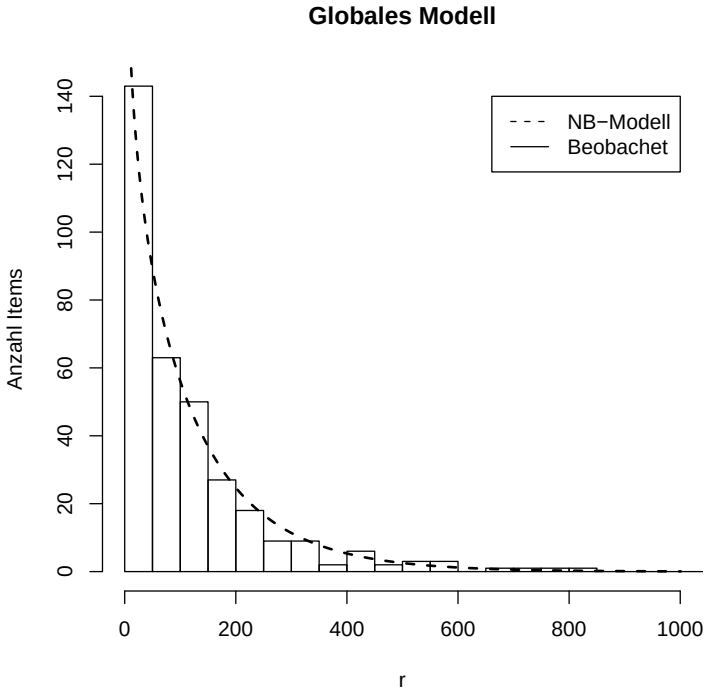
Supermarkt (Support: $\sigma = 0.1\%$)

- Deutlich höhere Werte in den Supermarktdaten (deuten auf Assoziationen hin).
- Starker systematischer Einfluss von Support.
- Höchste Werte an der Support-Konfidenz-Schranke (Bayardo Jr. & Agrawal, 1999). Falls Lift zur Sortierung der gefundenen Regeln verwendet wird, beeinflussen kleine Änderungen bei den Schranken direkt das Ergebnis.

Anwendung: NB-Frequent Itemsets

Idee: Identifikation von interessanten Assoziationen als Abweichungen vom Unabhängigkeitsmodell (Hahsler, 2004, 2006).

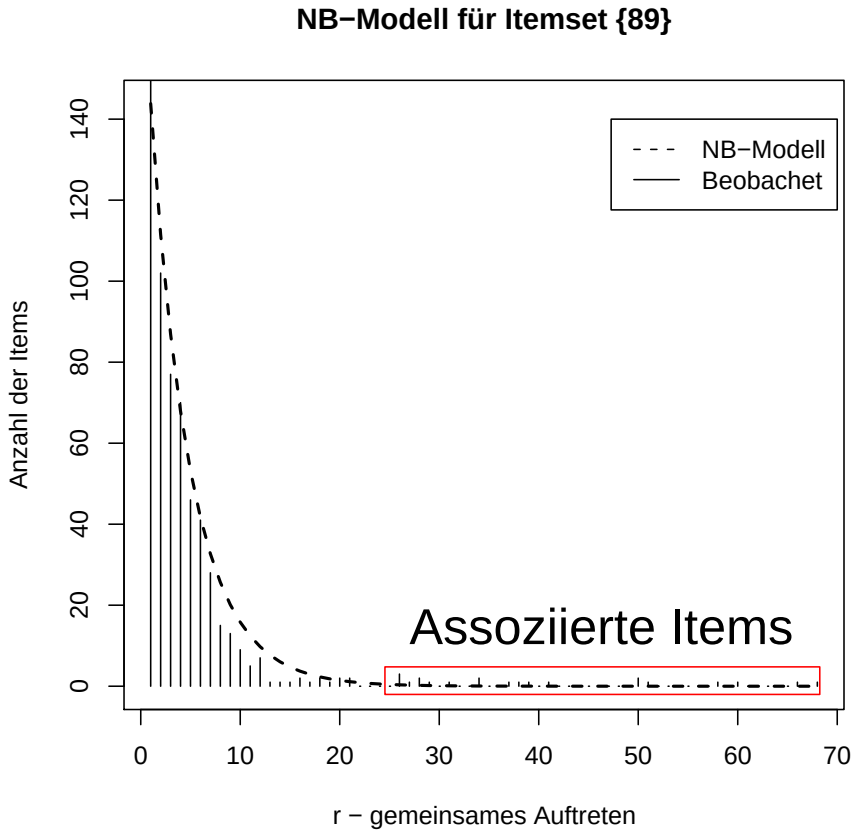
1. Schätzen eines ***globalen Unabhängigkeitsmodell*** aus den Häufigkeiten der Items in den Daten.
   Unabhängigkeitsmodell: k (Anzahl der Items) unabhängigen homogenen Poisson Prozesse. Parameter in der Population sind Γ-verteilt.



Anzahl der Items, die
$r = \{0, 1, \ldots, r_{max}\}$ mal in den Transaktionen vorkommen
$\longrightarrow$ ***Negative Binomialverteilung.***

NB-Frequent Itemsets

2. Transaktionen für ein Itemset Z werden ausgewählt. Alle von Z unabhängigen Items folgen in diesen Transaktionen weiter dem (reskalierten) globalen Unabhängigkeitsmodell. Assoziierte Items kommen „zu oft“ gemeinsam mit Z vor und können identifiziert werden.

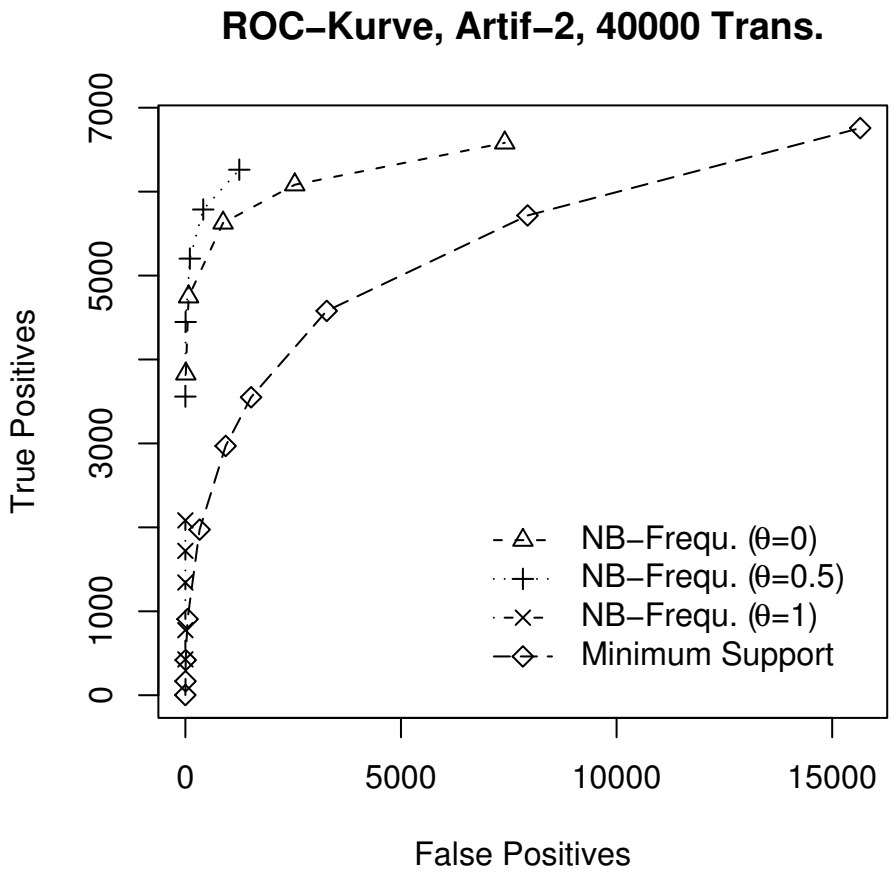


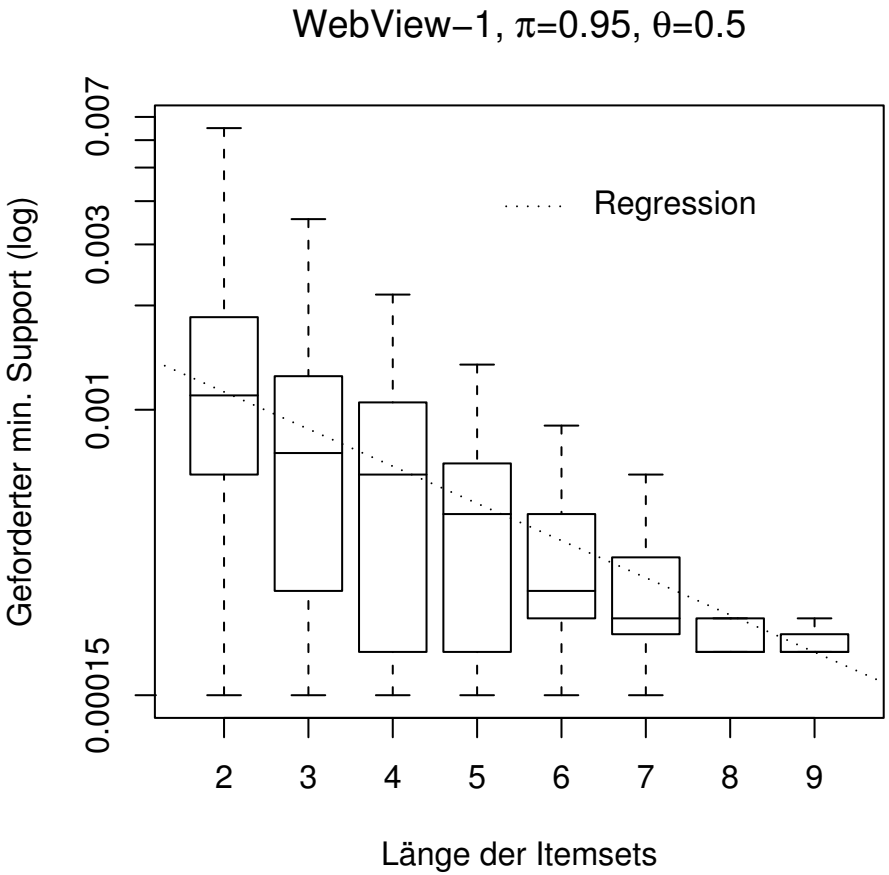
1. Verwendet einen ***benutzerdefinierten Grenzwert*** $1 - \pi$ ***für die max. zulässige Anzahl von „falschen Items“.***

2. ***Reskalierung des Modells*** für Z durch die Anzahl der Inzidenzen.

3. ***Einschränkung des Suchraums*** durch rekursive Definition und Parameter θ.

NB-Frequent Itemsets





Anwendung: Hyper-Konfidenz

Modellierung der Anzahl der Transaktionen, die die Regel $X \Rightarrow Y$ (X und Y) enthalten als Zufallsvariable N_{XY}. Gegeben den Häufigkeiten n_X und n_Y und Unabhängigkeit, hat N_{XY} eine ***Hypergeometrische Verteilung.***

> Die Hypergeometrische Verteilung kann durch das “Urnen Problem” erklärt werden: Ein Urne beinhaltet w weiße und b schwarze Bälle. Die Anzahl der weißen Bälle, die bei k Versuchen (ohne zurücklegen) gezogen wird ist hypergeometrisch verteilt.

Unter Unabhängigkeit kann die Datenbank als Urne mit n_X “guten” Transaktionen (enthalten X) und $N - n_X$ “schlechten” Transaktionen (enthalten nicht X) gesehen werden. Für die n_Y Transaktionen, die Y enthalten, wird nun n_Y mal aus der Datenbank gezogen. Die Anzahl der Transaktionen für Y, die auch X enthalten, ist damit hypergeometrische verteilt.

Die Wahrscheinlichkeit, dass X und Y unter Unabhängigkeit in genau r Transaktionen gemeinsam auftreten gegebenen n, n_X und n_Y, ist

$$P(N_{XY} = r) = \frac{\binom{n_Y}{r}\binom{n-n_Y}{n_X-r}}{\binom{n}{n_X}}.$$

Hyper-Konfidenz

$$\text{hyper-confidence}(X \Rightarrow Y) = P(N_{XY} < n_{XY}) = \sum_{i=0}^{n_{XY}-1} P(N_{XY} = i)$$

Ein sehr hoher Wert für Hyper-Konfidenz deutet darauf hin, dass die beobachtete Häufigkeit n_{XY} für die Unabhängigkeitsannahme zu hoch ist und dass zwischen X und Y ***komplementäre Effekte*** bestehen.

Wie für andere Maße kann ein Grenzwert gesetzt werden:

$$\text{hyper-confidence}(X \Rightarrow Y) \geq \gamma$$

Interpretation: Bei $\gamma = 0,99$ hat jede akzeptierte Regel max. eine 1% Chance, dass der Wert n_{XY} (gegeben n_X und n_Y) zufällig entstanden ist.

Hyper-Konfidenz

2×2 Kontingenztafel für $X \Rightarrow Y$

	$X = 0$	$X = 1$	
$Y = 0$	$n - n_Y - n_X - N_{XY}$	$n_X - N_{XY}$	$n - n_Y$
$Y = 1$	$n_Y - N_{XY}$	N_{XY}	n_Y
	$n - n_X$	n_X	n

Minimum-Hyper-Konfidenz (γ) ist äquivalent zu ***Fischer's exaktem Test*** mit Signifikanzniveau $\alpha = 1 - \gamma$.

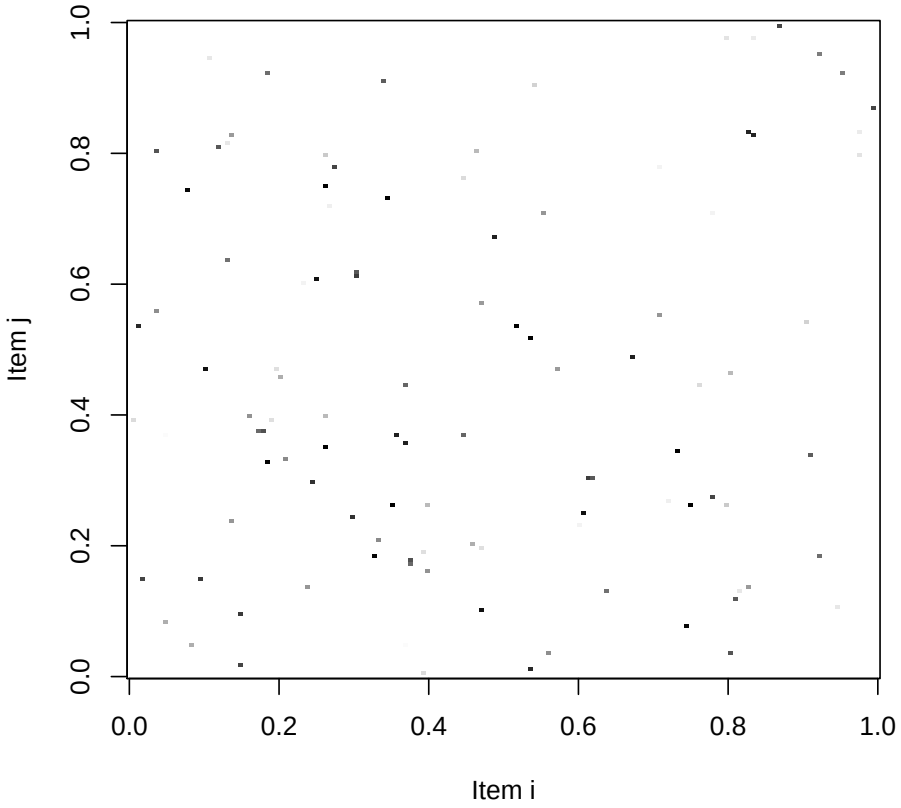
> Fischer's exakter Test ist ein Permutationstest bei dem unter der Annahme fixer Randhäufigkeiten die Wahrscheinlichkeit errechnet wird, eine noch extremere als die beobachtete Ungleichverteilung zu beobachten (einseitiger Test). Fischer zeigte, dass die Wahrscheinlichkeit eine bestimmte Konfiguration der Tabelle zu erreichen hypergeometrisch ist.

Damit ist der p-Wert des exakten Tests nach Fischer

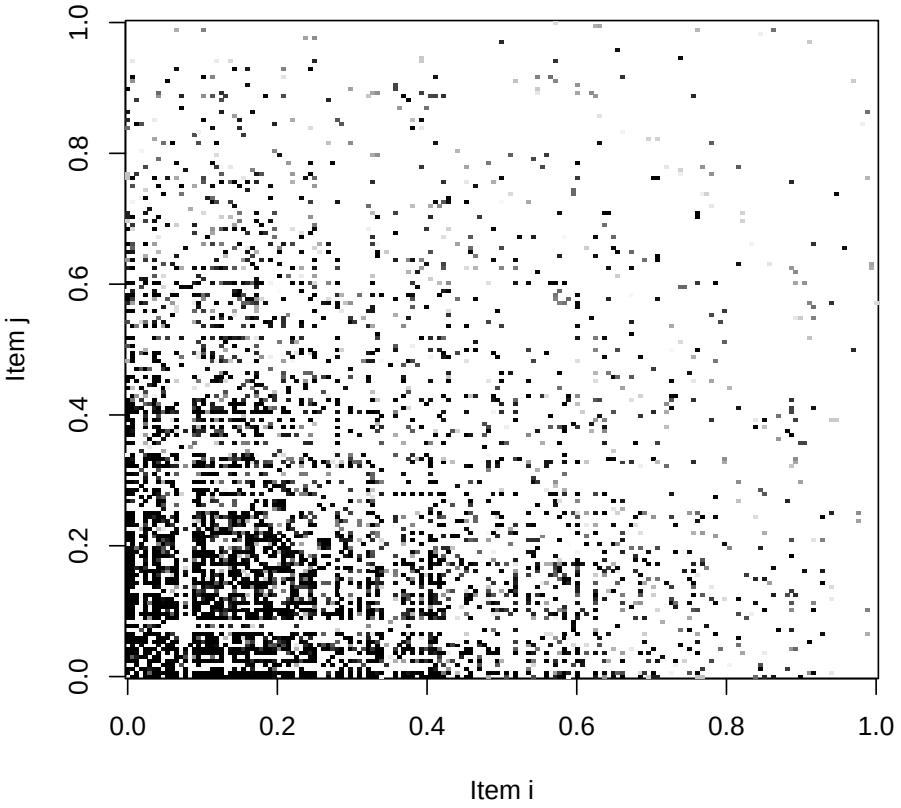
$$\text{p-Wert} = 1 - \text{hyper-confidence}(X \Rightarrow Y)$$

und das Signifikanzniveau $\alpha = 1 - \gamma$.

Hyper-Konfidenz: Komplementäreffekte



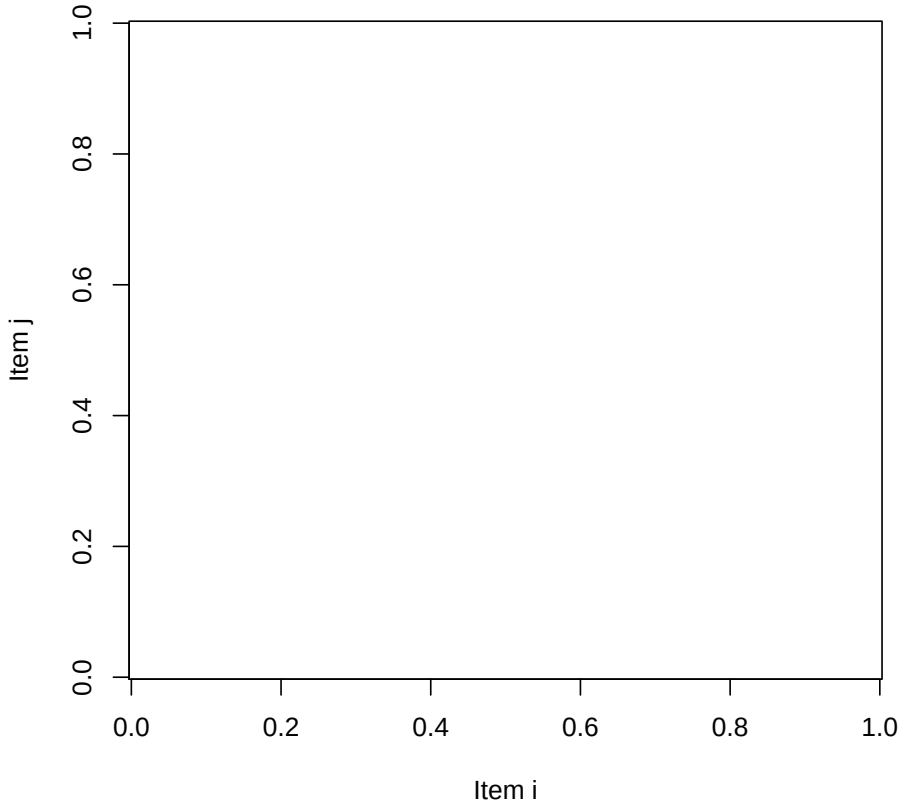
Simulierte Daten



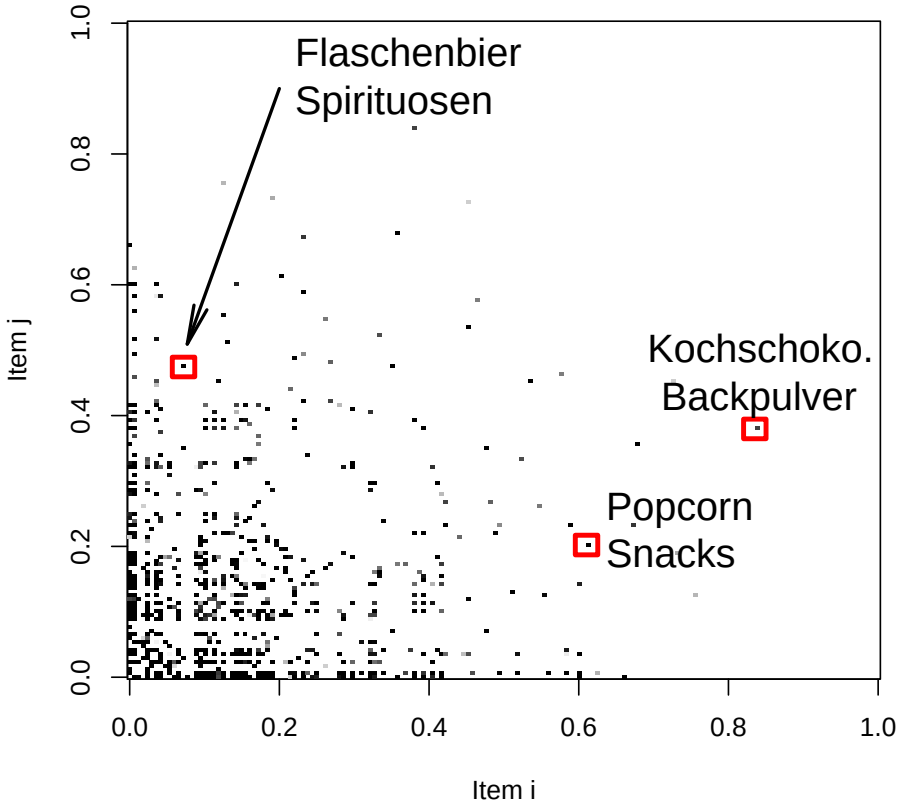
Supermarkt

$$\gamma = 0,99$$

Hyper-Konfidenz: Komplementäreffekte



Simulierte Daten



Supermarkt

$$\gamma = 0,9999993$$

Bonferroni Korrektur $\alpha = \frac{\alpha_i}{\binom{k}{2}}$

Hyper-Konfidenz: Substitutionseffekte

Hyper-Konfidenz findet Komplementäreffekte zwischen Items.

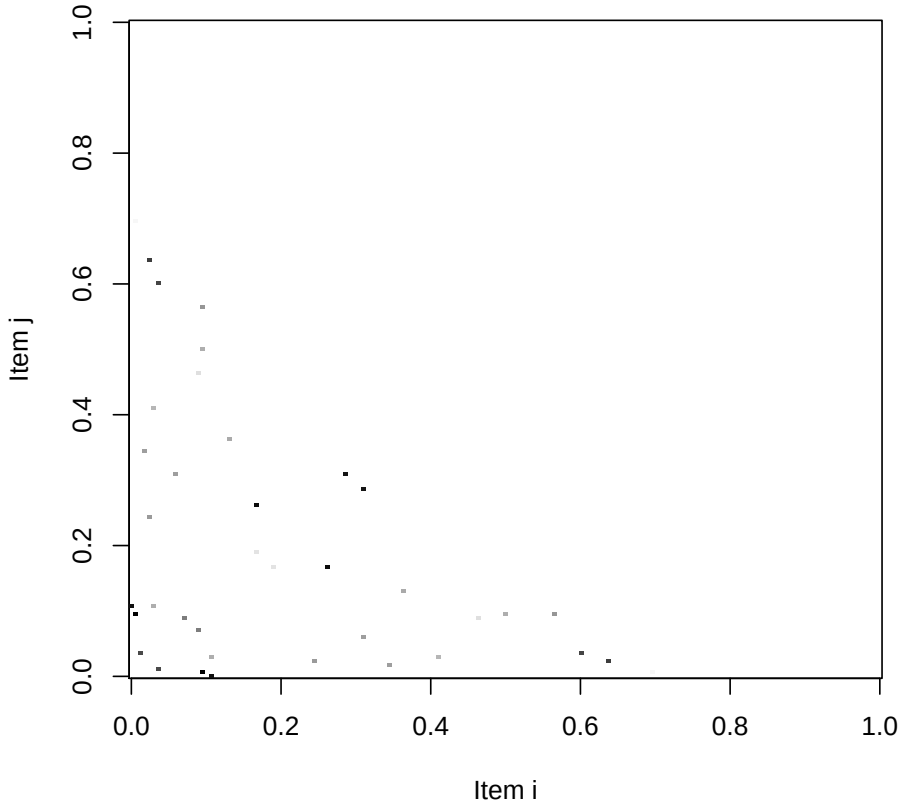
Um Substitutionseffekte aufzudecken, kann der Hyper-Konfidenz folgendermaßen angepasst werden:

$$\text{hyper-confidence}^{\text{sub}}(X \Rightarrow Y) = P(N_{XY} > n_{X,Y}) = 1 - \sum_{i=0}^{n_{XY}} P(N_{XY} = i)$$

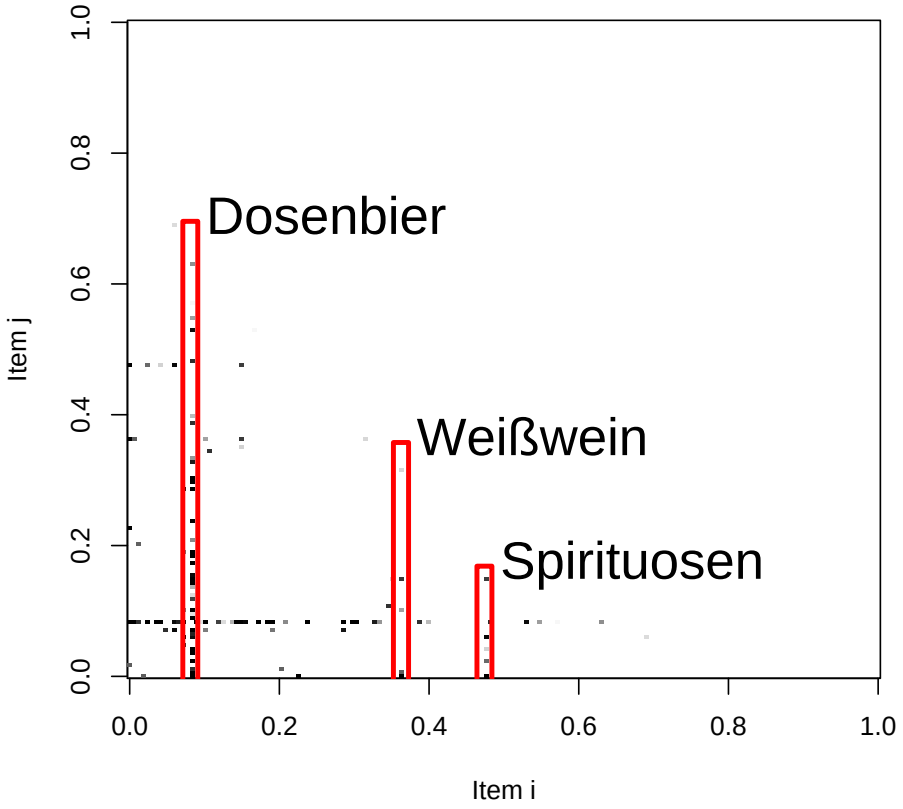
Und es wird verlangt:

$$\text{hyper-confidence}^{\text{sub}}(X \Rightarrow Y) \geq \gamma$$

Hyper-Konfidenz: Substitutionseffekte



Simulierte Daten

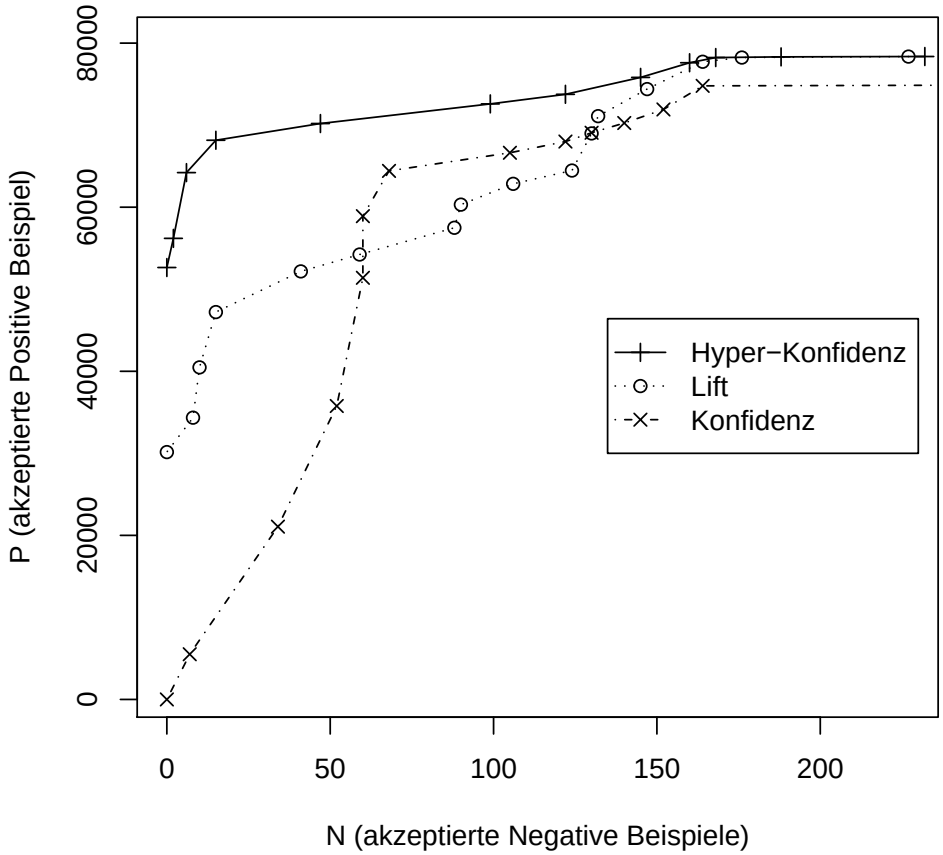


Supermarkt

$$\gamma = 0,99$$

Hyper-Konfidenz: Simulierte Daten

PN-Graph für den synthetischen Datensatz *T10I4D100K*
mit einer *Corruption-Rate* von 0,9.



Zusammenfassung und Ausblick

Für Assoziationsregeln (Support-Konfidenz-Framework) können für den betriebswirtschaftlichen Anwender wichtige Fragen nicht oder nur unzureichend beantworten werden:

- Sinnvolle Grenzwerte?
- Risiko durch „falsche“ Regeln?

$\longrightarrow$ ***Statistische Tests*** können helfen das Risiko einzugrenzen oder zumindest zu quantifizieren.

Probabilistische Modellierung der Daten kann verwendet werden um:

- ***Neue Maße*** zu entwickeln (NB-Frequent Itemsets mit Hilfe des Unabhängigkeitsmodells).
- ***Evaluierung und Vergleich*** von Maßen, Verfahren oder gesamten Data-Mining-Systemen mit Hilfe von synthetischen Daten aus Modellen mit Abhängigkeiten.

Ausblick für die Modellierung von anhängigen Daten:

- Modelle sollen die Erzeugung von Daten mit genau kontrollierbaren Abhängigkeiten ermöglichen.
- Betriebswirtschaftlich relevante Informationen sollen mit modelliert werden (Preise, Deckungsbeiträge,. . .).

Danke für die Aufmerksamkeit!

Literatur

C. C. Aggarwal & P. S. Yu. A new framework for itemset generation. In *PODS 98, Symposium on Principles of Database Systems*, Seiten 18–24, Seattle, WA, USA, 1998.

Rakesh Agrawal & Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, & Carlo Zaniolo, Hg., *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, Seiten 487–499, Santiago, Chile, September 1994.

R. Agrawal, T. Imielinski, & A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Seiten 207–216, Washington D.C., May 1993.

Robert J. Bayardo Jr. & Rakesh Agrawal. Mining the most interesting rules. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seiten 145–154. ACM Press, 1999.

M. J. Berry & G. Linoff. *Data Mining Techniques*. Wiley, New York, 1997.

R. Betancourt & D. Gautschi. Demand complementarities, household production and retail assortments. *Marketing Science*, 9(2):146–161, 1990.

Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, & Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, Seiten 255–264, Tucson, Arizona, USA, May 1997.

Michael Hahsler, Kurt Hornik, & Thomas Reutterer. Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, & W. Gaul, Hg., *From Data and Information Analysis to Knowledge Engineering*, Studies in Classification, Data Analysis, and Knowledge Organization, Seiten 598–605. Springer-Verlag, 2006.

Michael Hahsler. A model-based frequency constraint for mining associations from transaction data. Working Paper 07/2004, Working Papers on Information Processing and Information Management, Institut für Informationsverarbeitung und -wirtschaft, Wirtschaftsuniversität Wien, Augasse 2–6, 1090 Wien, Austria, November 2004.

Michael Hahsler. A model-based frequency constraint for mining associations from transaction data. *Data Mining and Knowledge Discovery*, 2006. Accepted for publication.

Harald Hruschka, Martin Lukanowicz, & Christian Buchta. Cross-category sales promotion effects. *Journal of Retailing and Consumer Services*, 6(2):99–105, 1999.

Greg Linden, Brent Smith, & Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, Jan/Feb 2003.

Bing Liu, Wynne Hsu, & Yiming Ma. Mining association rules with multiple minimum supports. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seiten 337–341. ACM Press, 1999.

Bing Liu, Wynne Hsu, & Yiming Ma. Pruning and summarizing the discovered associations. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seiten 125–134. ACM Press, 1999.

Gary J. Russell, David Bell, Anand Bodapati, Christina Brown, Joengwen Chiang, Gary Gaeth, Sunil Gupta, & Puneet Manchanda. Perspectives on multiple category choice. *Marketing Letters*, 8(3):297–305, 1997.

B. Sarwar, G. Karypis, J. Konstan, & J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the Tenth International World Wide Web Conference, Hong Kong, May 1-5*, 2001.

P. Schnedlitz, T. Reutterer, & W. Joos. Data-Mining und Sortimentsverbundanalyse im Einzelhandel. In H. Hippner, U. Müsters, M. Meyer, & K.D. Wilde, Hg., *Handbuch Data Mining im Marketing. Knowledge Discovery in Marketing Databases*, Seiten 951–970. Vieweg Verlag, Wiesbaden, 2001.

Masakazu Seno & George Karypis. Finding frequent itemsets using length-decreasing support constraint. *Data Mining and Knowledge Discovery*, 10:197–228, 2005.

Craig Silverstein, Sergey Brin, & Rajeev Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68, 1998.