Warenkorbanalyse mit Hilfe der Statistik-Software R

Michael Hahsler, Kurt Hornik, Thomas Reutterer

Die Warenkorb- oder Sortimentsverbundanalyse bezeichnet eine Reihe von Methoden zur Untersuchung der bei einem Einkauf gemeinsam nachgefragten Produkte oder Kategorien aus einem Handelssortiment. In diesem Beitrag wird die explorative Warenkorbanalyse näher beleuchtet, welche eine Verdichtung und kompakte Darstellung der in (zumeist sehr umfangreichen) Transaktionsdaten des Einzelhandels auffindbaren Verbundbeziehungen beabsichtigt. Mit einer enormen Anzahl an verfügbaren Erweiterungspaketen bietet sich die frei verfügbare Statistik-Software R als ideale Basis für die Durchführung solcher Warenkorbanalysen an. Die im Erweiterungspaket *arules* vorhandene Infrastruktur für Transaktionsdaten stellt eine flexible Basis für die Warenkorbanalyse bereit. Unterstützt wird die effiziente Darstellung, Bearbeitung und Analyse von Warenkorbdaten mitsamt beliebigen Zusatzinformationen zu Produkten (zum Beispiel Sortimentshierarchie) und zu Transaktionen (zum Beispiel Umsatz oder Deckungsbeitrag). Das Paket ist nahtlos in R integriert und ermöglicht dadurch die direkte Anwendung von bereits vorhandenen modernsten Verfahren für Sampling, Clusterbildung und Visualisierung von Warenkorbdaten. Zusätzlich sind in *arules* gängige Algorithmen zum Auffinden von Assoziationsregeln und die notwendigen Datenstrukturen zur Analyse von Mustern vorhanden. Eine Auswahl der wichtigsten Funktionen wird anhand eines realen Transaktionsdatensatzes aus dem Lebensmitteleinzelhandel demonstriert.

Einführung

Nachdem eine Reihe deutschsprachiger Beiträge über längere Zeit hinweg die methodische Diskussion prägten (vergleiche Böcker 1975, 1978; Merkle 1979, 1981; Müller-Hagedorn 1978, Bordemann 1986, Hruschka 1985, 1991), erlebt die Analyse des Nachfrageverbunds zwischen Bestandteilen (Produkte, Warengruppen, et cetera) von Einzelhandelssortimenten in der internationalen Marketing-Forschung seit einigen Jahren eine gewisse Renaissance. Aktuelle Übersichtsbeiträge zur Sortimentsverbundanalyse auf Basis von Warenkorbdaten (Market Basket Analysis) stammen von Russell et al. (1999), Seetharaman et al. (2005) oder Boztug/Silberhorn (2006). Für das Handelsmanagement ist die Kenntnis von in Warenkörben verborgenen Verbundbeziehungen aus unterschiedlichen Gründen aufschlussreich. Traditionell interessiert eine Verwertung solcher Informationen mittels diverser marketingpolitischer Maßnahmen (zum Beispiel Platzierung, Preis- und Sonderangebotspolitik) im Rahmen des Category-Managements des Handels (vergleiche Müller-Hagedorn 2005). Auf kundenindividuellem oder segment-spezifischem Niveau stößt die Nutzung von Verbundrelationen auch im Zusammenhang mit der Gestaltung maßgeschneiderter Cross- und Upselling-Aktionen innerhalb von Loyalitätsprogrammen auf verstärktes Interesse (vergleiche Mild/Reutterer 2003, Boztug/Reutterer 2006, Reutterer et al. 2006).

Ausgangsbasis einer Warenkorbanalyse stellen regelmäßig die im Data Warehouse einer Handelsorganisation gesammelten Transaktionsdaten, die teilweise (zum Beispiel durch den Einsatz von Kundenkarten) auch in personalisierter Form vorliegen. Durch den heute fast flächendeckenden Einsatz von Scannerkassen fallen im Einzelhandel enorme Mengen solcher Transaktionsdaten an. Darüber hinaus ist im elektronischen Handel auch das Sortiment

besonders reichhaltig. Erklärende oder explanative Ansätze, die Auswirkungen (Kreuzeffekte) von Marketing-Aktionen in einer Warengruppe auf das Kaufverhalten von Kunden in einer anderen Warengruppe modellieren (Hruschka/Lukanowicz/Buchta 1999, Manchanda/Ansari/Gupta 1999, Russell/Petersen 2000, Boztug/Hildebrandt 2006), sind in der Regel wegen ihrer Komplexität nur auf sehr kleine Ausschnitte des Sortiments beschränkt (ausführlicher zu dieser Problematik vergleiche Boztug/Reutterer 2006).

Explorative Ansätze hingegen können dazu verwendet werden, große Datenmengen für die Analyse zu verdichten, Verbundeffekte effizient aufzufinden und für den/die Benutzer zu visualisieren (Berry/Linoff 2004, Schnedlitz/Reutterer/Joos 2001). Die vorliegende Arbeit behandelt solche explorative Ansätze der Warenkorbanalyse und illustriert deren praktischen Einsatz unter Verwendung des R-Erweiterungspakets[1] *arules* (Hahsler/Grün/Hornik 2005). Tabelle 1 liefert einen strukturierten Überblick über die in der einschlägigen Literatur bislang vorgestellten Verfahrensklassen der explorativen Verbundanalyse, die jeweils spezifische Eigenschaften und damit korrespondierende Einsatzgebiete aufweisen.

Als datenverarbeitungstechnische Infrastruktur stellt das Paket *arules* die benötigten Datenstrukturen für die Darstellung von Transaktionsdaten sowie für die durch Anwendung der jeweiligen Datenverdichtungstechnik gefundenen Sortimentsverbundmuster zur Verfügung. Alle Datenstrukturen sind für den effizienten Umgang mit typischen Transaktionsdaten geeignet.

Ansatz	*Ausgewählte Quellen*	*Methodische Kurzcharakteristik*	*Aggregations-niveau*
Affinitätsanalyse	Böcker (1975, 1978) Merkle (1979, 1981) Dickinson/Harris/Sircar (1992) Julander (1992) Schnedlitz/Kleinberg (1994)	Repräsentation einer Verbundmatrix bestehend aus paarweisen Assoziationsmaßen	Aggregiert
Prototypenbildende Clusterverfahren	Schnedlitz/Reutterer/Joos (2001) Decker/Monien (2003) Decker (2005) Reutterer et al. (2006)	Verdichtung von Verbundbeziehungen in Warenkörben zu prototypischen Warenkorbklassen	Disaggregiert (segment-spezifisch)
Assoziationsregeln	Agrawal/Srikant (1994) Hildermann et al. (1998) Decker/Schimmelpfennig (2002) Brin et al. (1997) Brijs et al. (2004) Hahsler/Hornik/Reutterer (2006)	Generierung von Verbundregeln als Implikationen des Kaufs einer Warengruppe A auf eine (oder mehrere) andere Warengruppe(n) B (C, D, …)	Aggregiert

Tab. 1: Überblick über alternative Verfahren der explorativen Verbundanalyse

Nachfolgend werden zunächst die derzeit verfügbaren Alternativen zur Darstellung von Transaktionsdaten in *arules* vorgestellt und deren Verwendung anhand einiger wichtiger Funktionen aufgezeigt. In weiterer Folge wird anhand eines durchgängig verwendeten realen Transaktionsdatensatzes aus dem Lebensmitteleinzelhandel vorgeführt, wie einfach das Paket für die in Tabelle 1 skizzierten Anwendungen der explorativen Warenkorbanalyse eingesetzt werden kann. Dabei liegt das Hauptaugenmerk in diesem Beitrag weniger auf der formalen Darstellung der verwendeten Methoden (diesbezüglich wird auf die weiterführende Literatur verwiesen), der Analyse oder den jeweils ermittelten Ergebnissen sondern vielmehr auf der Vorführung der speziell für diese Aufgaben entworfenen Software-Implementierungen in R. Im Rahmen einer Darstellung der „klassischen“ Affinitätsanalyse wird die Kaufverbundenheit zwischen Warengruppen aufgrund von paarweisen Ähnlichkeiten untersucht. Im Anschluss

daran beschäftigen wir uns mit der Verdichtung von Transaktionsdaten zu prototypischen Warenkörben mit Hilfe eines geeigneten Clusterverfahrens. Schließlich wird dargestellt, wie Verbundmuster mit Hilfe einer aus der Data-Mining-Literatur stammenden Methodik zur Suche nach Assoziationsregeln gefunden und analysiert werden können.

Darstellung von Transaktionsdaten

Im Einzelhandel fallen Transaktionsdaten über Scanningsysteme am Kassen-Check-Out („Point of Sale“, POS) in Form von Warenkörben an. Jeder Warenkorb (jede Transaktion) enthält alle Artikel, die während eines Einkaufsaktes gemeinsam nachgefragt wurden (Hruschka 1991, Berry/Linoff 2004). Transaktionsdaten werden typischerweise in Form von Tupeln, das heißt einer geordneten Zusammenstellung diverser Einträge, wie folgt gespeichert:

<Transaktionsnummer, Produktnummer, …>

Alle Tupel mit der gleichen Transaktionsnummer bilden eine Transaktion. Neben den Transaktionen selbst sind zusätzliche Informationen aus den Artikelstammdaten (Packungsgröße, Hersteller, et cetera) und zum Einkaufsvorgang selbst (Zeitpunkt, Filiale, Kassenplatz, et cetera) verfügbar. Typischerweise sind alle im Sortiment gelisteten Artikel in ein hierarchisches Klassifikationsschema von Warengruppen eingebunden. Durch diese Zusatzinformation wird es möglich, auch Zusammenhänge die eine gesamte Warengruppe betreffen zu analysieren. Durch geeignete Schnittstellen zum Warenwirtschafts- oder Marketing-Informationssystem des Einzelhändlers können auf Artikelebene weitere wichtige Marketing-Informationen wie zum Beispiel gekaufte Packungsmengen, Preise, Promotions und andere Aktionen (Flugblatt, Mehrfachplatzierungen, et cetera) oder Deckungsbeiträge einbezogen werden (vergleiche Gaul/Both 1998, Zentes 1998).

Gelegentlich sind für einen Teil der Transaktionen auch Hintergrundinformationen über die Kunden (zum Beispiel Identität und Kaufgeschichte, sozio-demographische Merkmale) bekannt. Diese stammen in der Regel aus den sich zunehmender Beliebtheit erfreuenden Loyalitätsprogrammen von Handelsunternehmen in Verbindung mit elektronisch lesbaren und direkt am POS präsentierten Kundenkarten (vergleiche Passingham 1998) oder anderen Quellen (zum Beispiel Authentifizierung der Kunden beim Login im Online-Retailing).

Um derartige Transaktionsdaten für die Warenkorbanalyse darzustellen, sind im Paket *arules* zwei Repräsentationsformen vorgesehen:

- Repräsentation als binäre Kaufinzidenzmatrix mit Transaktionen in den Zeilen und Artikeln beziehungsweise Warengruppen in den Spalten. Die Einträge stellen den Kauf (1) beziehungsweise den Nichtkauf (0) einer Warengruppe in einer Transaktion dar[2]. Im Kontext von Assoziationsregeln wird diese Darstellung oft als horizontales Datenbanklayout bezeichnet (Zaki 2000).
- Repräsentation als Transaktionsnummernliste, wobei für jede spaltenweise angeordnete Warengruppe eine Liste jener Transaktionsnummern, welche die betreffende Warengruppe enthält, gespeichert wird. Diese Darstellung wird auch als vertikales Datenbanklayout bezeichnet (Zaki 2000).

Ein Beispiel der beiden Darstellungsformen ist in Abbildung 1 dargestellt.

Produkte

Transaktionen	Milch	Brot	Butter	Bier	...
1	1	1	0	0	
2	0	1	1	0	
3	0	0	0	1	
4	1	1	1	0	
5	0	1	1	0	

(a)

Artikel	Transaktionsnummern
Milch	1, 4
Brot	1, 2, 4, 5
Butter	2, 4
Bier	3
⋮	

(b)

Abb. 1: Beispiel von Transaktionsdaten als (a) Kaufinzidenzmatrix und (b) Transaktionsnummernliste

Zusätzlich zu den Transaktionen können die oben erwähnten Zusatzinformationen gespeichert werden, die dann für Manipulationen zur Verfügung stehen. Neben einfachen Attributen der gekauften Artikel oder Informationen zum Einkaufsvorgang kann eine Sortimentshierarchie, wie ausschnittsweise in Abbildung 2 dargestellt, mit den Transaktionsdaten assoziiert werden. Gespeichert wird die hierarchische Struktur dadurch, dass zu jedem Produkt die Zugehörigkeit zu den Kategorien der einzelnen Ebenen aufgezeichnet wird. Dadurch wird es möglich, aus den Transaktionsdaten all jene Transaktionen zu selektieren die beispielsweise Molkereiprodukte beinhalten. Ebenso einfach können dadurch gewisse Warengruppen (zum Beispiel Brot und Gebäck) aus Analysen ausgeschlossen werden.

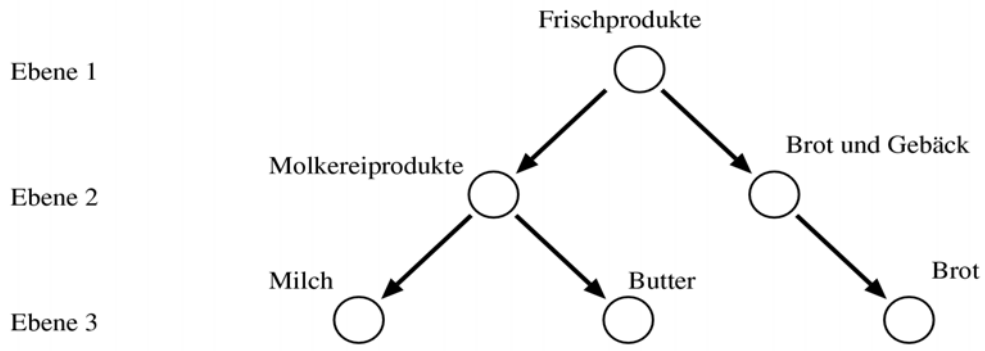
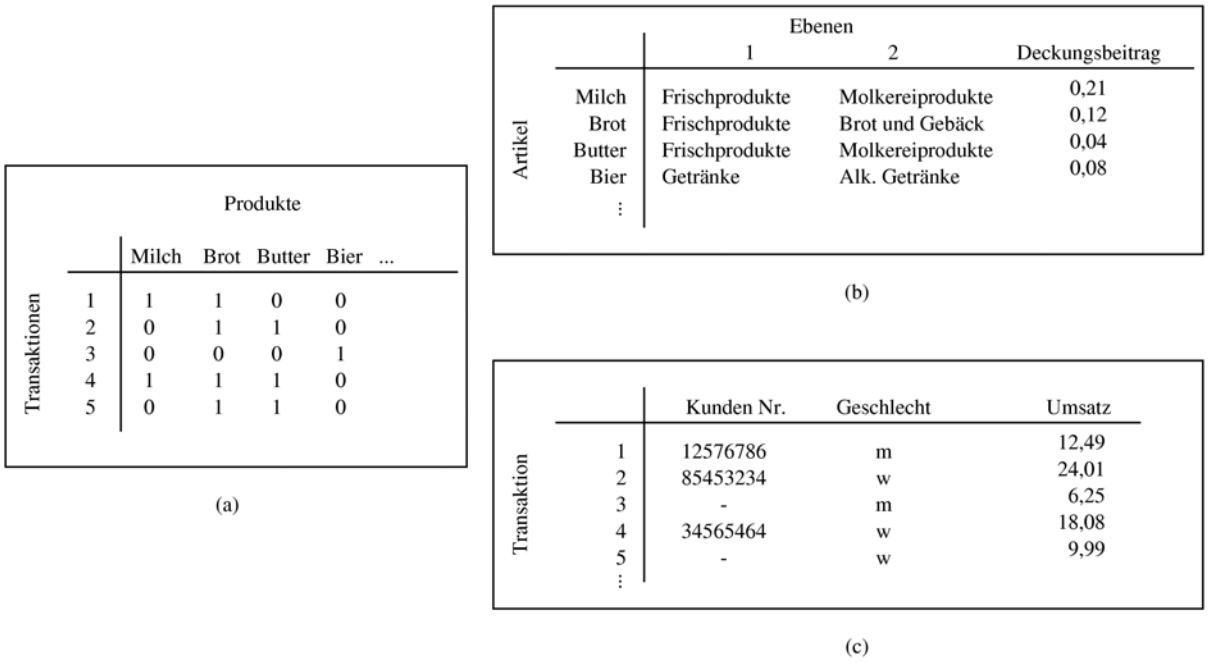


Abb. 2: Ausschnitt aus einer dreistufigen Sortimentshierarchie

Transaktionsbezogene Informationen können verwendet werden um bestimmte Transaktionen auszuwählen, etwa nur Transaktionen die einen Umsatz größer als EUR 20 generieren. Abbildung 3 zeigt den typischen Aufbau von Transaktionsdaten im Paket *arules*.

Für Illustrationszwecke ist ein realer Einzelhandelsdatensatz bereits in *arules* enthalten. Der Datensatz enthält die Transaktionsdaten einer Supermarktfiliale eines Lebensmitteleinzelhändlers für den Zeitraum von 30 Tagen. Die Daten sind bereits zu 169 Warengruppen aggregiert. Nachdem das Paket geladen wurde, kann der Datensatz mit dem Namen „Groceries“ wie folgt aktiviert werden[3]:

```
> library("arules")

> data("Groceries")
```



Produkte

Transaktionen	Milch	Brot	Butter	Bier	...
1	1	1	0	0	
2	0	1	1	0	
3	0	0	0	1	
4	1	1	1	0	
5	0	1	1	0	

(a)

Ebenen

Artikel	1	2	Deckungsbeitrag
Milch	Frischprodukte	Molkereiprodukte	0,21
Brot	Frischprodukte	Brot und Gebäck	0,12
Butter	Frischprodukte	Molkereiprodukte	0,04
Bier	Getränke	Alk. Getränke	0,08
⋮			

(b)

Transaktion	Kunden Nr.	Geschlecht	Umsatz
1	12576786	m	12,49
2	85453234	w	24,01
3	-	m	6,25
4	34565464	w	18,08
5	-	w	9,99
⋮			

(c)

Abb. 3: Teile von typischen Transaktionsdaten in arules bestehend aus (a) Transaktionen, (b) Warengruppeninformationen und (c) Transaktionsinformationen

Nachfolgend führen wir eine Reihe elementarer in *arules* verfügbarer Funktionen vor, die es ermöglichen, den Datensatz auf einfache Art und Weise näher zu charakterisieren. Eine erste kurze Zusammenfassung der grundlegenden Eigenschaften des Datensatzes kann zunächst durch die Funktion `summary()` erstellt werden.

```
> summary(Groceries)


transactions as itemMatrix in sparse format with

  9835 rows (elements/itemsets/transactions) and

  169 columns (items)


most frequent items:

      whole milk     other vegetables       rolls/buns         soda

            2513              1903              1809             1715

          yogurt          (Other)

            1372             34055


element (itemset/transaction) length distribution:

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  1.000   2.000   3.000   4.409   6.000  32.000
```

```
includes extended item information - examples:

       labels    level2           level1

1 frankfurter    sausage       meet and sausage

2     sausage    sausage       meet and sausage
```

Aus der Zusammenfassung kann man entnehmen, dass es sich bei dem Datensatz um Transaktionsdaten mit 9835 Transaktionen und 169 Warengruppen (`items`) handelt. Aus Effizienzgründen werden für die schwach besetzte („sparse“) 9835 × 169 Inzidenzmatrix nur die Einsen gespeichert. Als nächstes werden die fünf am häufigsten gekauften Warengruppen mit der Anzahl der Transaktionen, in denen sie vorkommen, aufgelistet. Am häufigsten werden Artikel aus der Warengruppe Vollmilch („whole milk“) gekauft (in 2513 aus 9835 Transaktionen). Die Verteilung der „Länge“ der Transaktionen wird als nächstes angegeben: Im Durchschnitt enthält ein Warenkorb 4,409 unterschiedliche Warengruppen. Ein Median kleiner als der Mittelwert deutet auf eine schiefe Verteilung mit sehr vielen „kurzen“ Transaktionen, die nur wenige Warengruppen enthalten (also klassische Bagatell-, oder Kleinsteinkäufe), hin. Der letzte Teil der Zusammenfassung weist darauf hin, dass zusätzlich zu den Bezeichnungen der Warengruppen (`labels`) auch Informationen zur Sortimentshierarchie vorhanden sind (`level1` und `level2`). Informationen zu den Warengruppen können durch die Funktion `itemInfo()` abgefragt werden, Informationen zu den Transaktionen durch `transactionInfo()`. Beispielsweise können alle Warengruppenbezeichnungen, die auf Ebene 2 (`level2`) der Hauptwarengruppe Molkereiprodukte (dairy products) angehören, angezeigt werden.

```
> itemLabels(Groceries)[itemInfo(Groceries)$level2 == "dairy products"]

[1] "whole milk"         "butter"             "curd"

[4] "dessert"            "butter milk"        "yogurt"

[7] "whipped/sour cream" "beverages"
```

Im Beispiel gibt die Funktion `itemLabels()` die Bezeichnungen der Warengruppen, die im Datensatz verwendet werden, zurück. Die acht Warengruppen, die zur Hauptwarengruppe Molkereiprodukte gehören, werden dann mithilfe der Zusatzinformationen gefunden.

Die Warengruppen in einzelnen Transaktionen können mittels `inspect()` angezeigt werden. Beispielsweise enthalten die ersten drei Transaktionen folgende Warengruppen:

```
> inspect(Groceries[1:3])

  items

1 {citrus fruit, semi-finished bread, margarine, ready soups}

2 {tropical fruit, yogurt, coffee}

3 {whole milk}
```

Die Länge der Transaktionen kann mittels der Funktion `size()` ermittelt werden. Dadurch können beispielsweise besonders „lange“ Transaktionen gefunden werden oder es kann ein Histogramm der Transaktionslängen erzeugt werden.

```
> which(size(Groceries) > 25)

[1] 1092 1217 2939 2974 4431 5611 9002

> hist(size(Groceries))
```

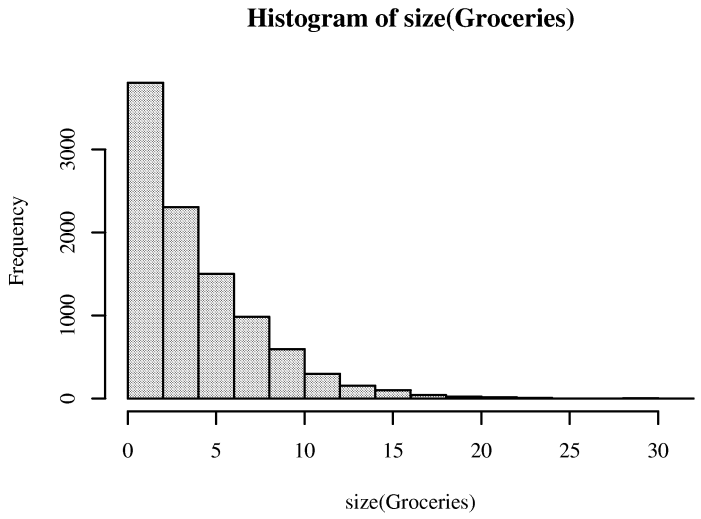


Abb. 4: Verteilung der Transaktionslängen

Der erste Befehl findet die Transaktionsnummern jener 7 Transaktionen (Großeinkäufe), die mehr als 25 Warengruppen beinhalten. Das Histogramm, welches der zweite Befehl erzeugt, ist in Abbildung 4 dargestellt. Im Histogramm ist die typische Verteilung mit sehr vielen „kurzen“ und sehr wenigen „langen“ Transaktionen klar erkennbar.

Als weitere grundlegende Eigenschaft von Transaktionsdaten interessiert meist die Häufigkeit, mit der einzelne Warengruppen in den Daten vorkommen. Diese Häufigkeiten können mittels der Funktionen `itemFrequency()` beziehungsweise `itemFrequencyPlot()` angezeigt und grafisch dargestellt werden. Der folgende Befehl stellt die Häufigkeiten der Warengruppen im Datensatz dar. Aus Gründen der Übersichtlichkeit werden nicht alle 169 Warengruppen dargestellt, sondern nur jene, die in mindestens 5 % der Transaktionen vorkommen (siehe Abbildung 5).

```
> itemFrequencyPlot(Groceries, support = 0.05)
```

Die einfache Handhabung und Manipulation von Transaktionsdaten wie sie im Paket *arules* unterstützt wird, ist die Grundlage für effizientes Analysieren und Arbeiten mit Transaktionsdaten, wie es in den folgenden Abschnitten dargestellt wird.

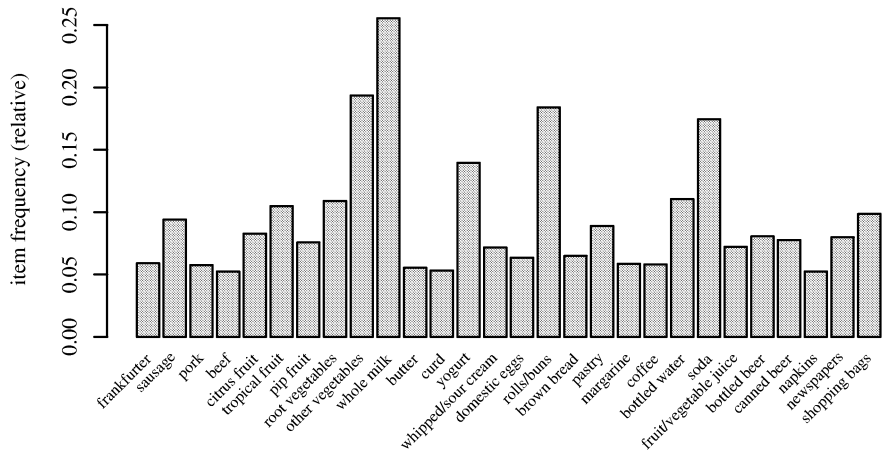


Abb. 5: Kaufhäufigkeiten der Warengruppen (Minimum 5 %)

Affinitätsanalyse auf Basis paarweiser Assoziationsmaße

Die insbesondere von den Proponenten der frühen deutschsprachigen Literatur (Böcker 1978, Merkle 1981) vorgestellten Ansätze der Affinitätsanalyse ermitteln zunächst ein zweidimensionale „Frequenzmatrix“, welche die Häufigkeiten der gemeinsamen Kaufhäufigkeiten für Paare von Warengruppen enthält (vergleiche Hruschka 1991). Diese Matrix lässt sich bequem wie folgt bestimmen:

```
> ct <- crossTable(Groceries)

> ct[1:5, 1:5]

             frankfurter sausage liver loaf ham meat

frankfurter         580      99          7  25   32

sausage              99     924         10  49   52

liver loaf            7      10         50   3    0

ham                  25      49          3 256    9

meat                 32      52          0   9  254
```

Die Warengruppen werden in Zeilen und Spalten angeordnet und die einzelnen Zellen der symmetrischen Matrix enthalten jeweils die Anzahl der Transaktionen, in denen beide Warengruppen enthalten sind. Hier wurden nur die gemeinsamen Kaufhäufigkeiten für die ersten fünf Warengruppen dargestellt. Auf diese Frequenzmatrix können Assoziationskoeffizienten zwischen Paaren von Warengruppen definiert und die Frequenzmatrix in eine sogenannte „Verbundmatrix“ überführt werden. Da letztere mit zunehmender Warengruppenanzahl sehr schnell unüberschaubar wird (im vorliegenden Fall handelt es sich bereits um eine 169 × 169 Matrix), ist regelmäßig eine geeignete Verdichtung von Interesse. Üblicherweise gelangen dabei Projektionsmethoden wie Verfahren der mehrdimensionale Skalierung (MDS) oder hierarchische Clusteranalysemethoden zum Einsatz (vergleiche Merkle 1979, Bordemann 1986, Decker/Schimmelpfennig 2002). Die Visualisierung der Verbundbeziehungen erfolgt dann entweder in einem niedrigdimensionalen geometrischen Raum oder über eine Baumstruktur, auf deren Basis in weiterer Folge eine Typologie der Warengruppen erstellt werden kann.

Einen Überblick über in der Verbundanalyse bewährte Assoziationskoeffizienten für binäre Transaktionsdaten findet man bei Böcker (1978) oder Hruschka (1991). Unter den als besonders geeignet geltenden Koeffizienten befindet sich der Tanimoto-Koeffizient, dessen Pendant als Unähnlichkeitsmaß der Jaccard-Koeffizient (Sneath 1957) darstellt. Der Tanimoto-Koeffizient ist ein asymmetrisches Ähnlichkeitsmaß, in das übereinstimmende Nullen nicht eingehen. Dies verhindert, dass Warengruppen, die sehr selten frequentiert werden und daher in vielen Transaktionen übereinstimmende Nullen aufweisen als ähnlicher eingestuft werden, als häufiger vorkommende Warengruppen. Im Kontext der Verbundanalyse spricht man in diesem Zusammenhang auch von der „negativen Verbundenheit“ von Warengruppen (vergleiche ausführlicher dazu Böcker 1978, Bordemann 1986, experimentelle Befunde liefern Mild/Reutterer 2001). Zwecks Verdichtung der Verbundmatrix wird wie bei Schnedlitz und Kleinberg (1994) im nachfolgenden Illustrationsbeispiel ein hierarchisches Clusterverfahren, und zwar die Minimum-Varianz-Methode nach Ward, verwendet. Damit die Baumstruktur mit den Bezeichnungen der

Warengruppen dargestellt werden kann, werden nur Warengruppen verwendet, die in mindestens 2 % der Transaktionen vorkommen.

```
> diss <- dissimilarity(Groceries[, itemFrequency(Groceries) > 0.02],
      method = "Jaccard", which = "items")
> hc <- hclust(diss, method = "ward")
> plot(hc)
```

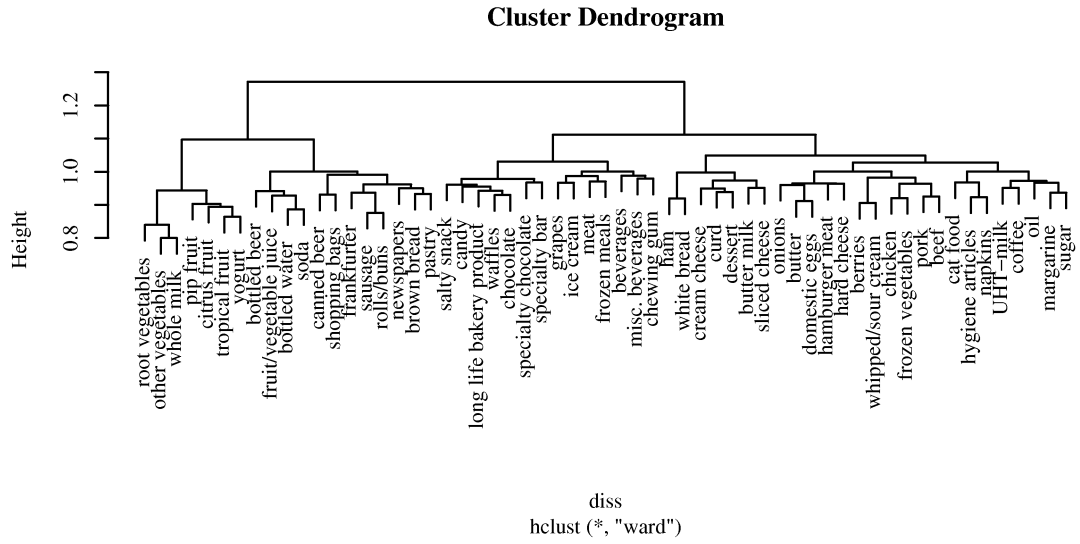


Abb. 6: Darstellung der hierarchischen Clusterlösung als Dendrogramm

Die gefundene typologische Struktur der Warengruppen ist in Abbildung 6 in Form eines Dendrogramms dargestellt. Das Dendrogramm ist die Darstellung einer hierarchischen Clusterstruktur als Baum mit den einzelnen Objekten (Warengruppen) als Blätter. Je weiter man sich entlang der Äste in Richtung Blattspitzen bewegt, desto stärker die Kaufverbundenheit der beteiligten Warengruppen. Die einzelnen Verästelungen der Baumstruktur können nun verwendet werden, um Gruppen von Warengruppen, innerhalb derer höhere Kaufverbundenheit herrscht, zu identifizieren. Beispielsweise kann eine Gruppe von Frischesortimentsteilen bestehend aus Gemüse, Früchte, Milch und Yoghurt (vegetables, fruits, milk, yogurt) am linken Rand des Dendrogramms in Abbildung 6 gefunden werden. Die absatzpolitische Relevanz solcher auf faktischem Kaufverhalten begründeten hierarchischen Warengruppentypologien wird primär in der Nutzung für Verbundplatzierungen im Verkaufsraum oder Werbemedien (zum Beispiel Flugblätter, Kataloge) sowie in der nachfragerorientierten Bildung sogenannte Categories im Rahmen des Category-Management-Prozesses gesehen (vergleiche dazu etwa Bordemann 1986, Zielke 2002, Müller-Hagedorn 2005).

Verdichtung von Transaktionsdaten zu Prototypen

Die oben beschriebene Konstruktion einer Verbundmatrix enthält nach Anwendung eines geeigneten Assoziationsmaßes die paarweisen Verbundkoeffizienten für den gesamten (gepoolten) Transaktionsdatensatz. Abgesehen vom häufig vorgebrachten Einwand, dass bei einer solchen Analyse lediglich auf Verbundrelationen zwischen Paaren von Warengruppen abgestellt wird (vergleiche Hruschka 1991), wird unter anderem von Schnedlitz, Reutterer und Joos (2001) die dabei erfolgende a-priori Aggregation der Transaktionsdaten

problematisiert. Tatsächlich ist damit eine nachträgliche Untersuchung des Sortimentsverbunds auf disaggregiertem Niveau einzelner oder Gruppen von Transaktionen nicht mehr möglich. Gelegentlich ist dies allerdings erwünscht, man denke etwa an die Untersuchung der Dynamik von Verbundbeziehungen (zum Beispiel über Tageszeiten, Wochentagen, Saisonen hinweg) oder die Identifikation von Filialgruppen oder Kundensegmenten, die durch ähnliche Verbundbeziehungen zwischen den Warengruppen eines Sortiments gekennzeichnet sind.

In solchen Fällen erweist es sich als hilfreich, die einzelnen Transaktionen einer (kleinen) Gruppe von prototypischen Warenkörben zuzuordnen und damit eine Warenkorb-Typologie zu schaffen (vergleiche Schnedlitz/Reutterer/Joos 2001). Die Verdichtung zu sogenannten Warenkorb-Protoypen kann durch Clusterbildung basierend auf den binären Transaktionsdaten realisiert werden. Im Folgenden wird ein Beispiel für die Verdichtung des Groceries Datensatzes gezeigt.

Für die Analyse sollen nur Transaktionen verwendet werden, die mindestens zwei verschiedene Warengruppen beinhalten. Des Weiteren soll die Warengruppe Tragtaschen (shopping bags) aus der Analyse ausgeschlossen werden, da diese in Experimenten die Clusterbildung negativ beeinflusst haben. Dies kann folgendermaßen realisiert werden:

```
> groc <- Groceries[size(Groceries)>1,

      which(itemLabels(Groceries) == "shopping bags")]
```

Nach der Auswahl stehen 7676 Transaktionen und 168 Warengruppen für die Analyse zur Verfügung.

Schnedlitz, Reutterer und Joos (2001), Decker und Monien (2003) sowie Decker (2005) schlagen für diesen Analysezweck die Verwendung von Verfahren der Vektorquantisierung oder geeignete neurale Netzwerkmethoden vor. Allen gemeinsam ist, dass eine Partitionierung der Daten vorgenommen und die gefundenen Centroide (Mittelpunkte der Cluster) zur Beschreibung von prototypischen Warenkorbklassen herangezogen werden. Für das Beispiel hier wird eine andere Clustermethode, das von Kaufman und Rousseeuw (1990) entwickelte Verfahren Partitioning Around Medians (PAM), eingesetzt. PAM arbeitet ähnlich wie der bekannte K-Means Algorithmus mit dem Hauptunterschied, dass die Cluster nicht durch deren Centroide sondern durch „Medoide“ repräsentiert wird. Als Medoid eines Clusters wird dabei jene Transaktion (oder genauer: jener Warenkorb) verstanden, deren durchschnittlicher Abstand zu allen anderen Transaktionen im selben Cluster minimal ist. Durch die iterative Suche nach einer repräsentativen Transaktion pro Cluster erweist sich PAM als vergleichsweise unempfindlich gegenüber Ausreißern (für eine Marketing-Anwendung in einem anderen Kontext vergleiche Larson/Eric/Fader 2005).

Um die Größe der in PAM verwendeten Unähnlichkeitsmatrix zu reduzieren, werden zufällig 2000 Transaktionen aus dem Datensatz gezogen. Danach werden Unähnlichkeiten zwischen den gezogenen Transaktionen wieder mittels des Jaccard-Koeffizienten errechnet.

```
> samp <- sample(groc, 2000)

> diss <- dissimilarity(samp, method = "Jaccard")
```

Die Funktionen `sample()` und `dissimilarity()` sind für Transaktionsdaten im Paket *arules* definiert. Der Clusteralgorithmus PAM befindet sich im Erweiterungspaket *cluster* und wird mit der Unähnlichkeitsmatrix *diss* und der gewünschten Anzahl von Clustern *k* aufgerufen. Als Ergebnis liefert der Algorithmus unter anderem die Zuordnung der Transaktionen zu den Clustern, sowie Informationen zu den Clustern. Im vorliegenden Illustrationsbeispiel werden alle Clusterlösungen mit $k = 1$ bis 8 bestimmt.

```
> library("cluster")
```

```
> clust <- lapply(1:8, function(x) pam(diss, k = x))
```

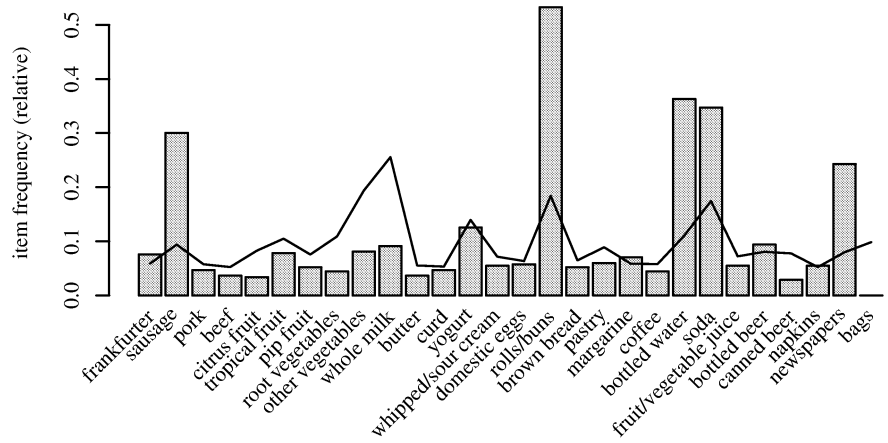
Das Objekt `clust` enthält nun Informationen zu den gefundenen Clusterlösungen für eine aufsteigende Anzahl von Klassen *k* = 1, …, 8. Für die Auswahl einer „geeigneten" Clusteranzahl steht eine reichhaltige Auswahl an internen Validitätsmaßen zur Verfügung (einen Überblick dazu findet man bei Milligan/Cooper 1985), die Großteils auch in R verfügbar sind. Im vorliegenden Fall wird aus den generierten Clusterlösungen mit Hilfe der durchschnittlichen Breite der sogenannten „Silhouette" (vergleiche Kaufman/Rousseeuw 1990, S. 83ff.) eine Klassenanzahl k = 5 ausgewählt, die sich wie folgt charakterisieren lässt:

```
> clust[[5]]$clusinfo

     size max_diss av_diss diameter separation

[1,]  513   0.9474  0.8113        1     0.2857

[2,]  383   0.9286  0.7700        1     0.2500

[3,]  578   1.0000  0.8427        1     0.2500

[4,]  285   0.9091  0.7294        1     0.2500

[5,]  241   0.9091  0.7146        1     0.2500
```

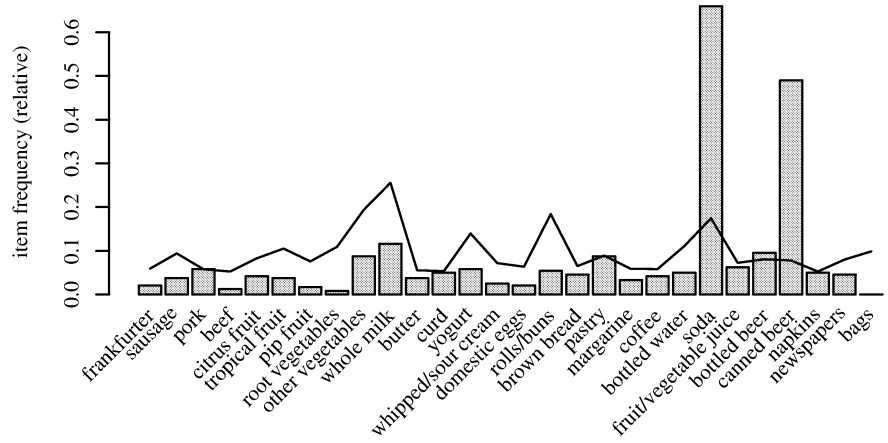
Aus den Informationen zu den einzelnen Warenkorb-Clustern erkennt man neben der absoluten Anzahl der den einzelnen Cluster zugewiesenen Transaktionen (`size`), dass die durchschnittliche Unähnlichkeit innerhalb der Cluster (`av_diss`) sehr hoch ist. Sehr geringe klasseninterne Unähnlichkeiten sind jedoch für hochdimensionale Daten (n=168) typisch und deuten daher nicht unbedingt auf eine schlechte Qualität der Clusterlösung hin. Von den fünf Warenkorb-Cluster können mit Hilfe der folgenden Anweisungen exemplarisch die Prototypenprofile für die Cluster mit der Nummerierung 2 und 5 dargestellt werden:

```
> itemFrequencyPlot(samp[clust[[5]]$clustering == 2],

       population = Groceries, support = 0.05)

> itemFrequencyPlot(samp[clust[[5]]$clustering == 5],

       population = Groceries, support = 0.05)
```

In Abbildung 7 werden die beiden ausgewählten Prototypenprofile visualisiert. Aus Darstellungsgründen werden nur jene Warengruppen herangezogen, die in mindestens 5 % der Transaktionen vorkommen (Parameter `support`). Die Linie kennzeichnet jeweils die relative Frequenzverteilung im gesamten Datensatz (Parameter `population`) und die Balken zeigen die Verteilung der warengruppenspezifischen Einkaufsfrequenzen innerhalb der Cluster. Da die Warenkörbe als Binärdaten codiert sind, entsprechen die Prototypenprofile den bedingten Kaufwahrscheinlichkeiten der Warengruppen innerhalb eines Clusters. Für die Verbundanalyse sind insbesondere Warengruppenkombinationen mit starken Abweichungen vom Profil der unbedingten Kaufwahrscheinlichkeiten (dargestellt durch die Linie) interessant. Deutlich positive (negative) Abweichungen signalisieren komplementäre (substitutive) Verbundbeziehungen zwischen den betreffenden Warengruppen. So stellt Cluster 2 beispielsweise ein prototypisches Einkaufsmuster mit besonders hohen Anteilen in den Warengruppen Wurst, Semmel, Getränke (bottled water und soda) und Zeitungen dar. Cluster 5 repräsentiert hingegen ein typisches Getränke-Cluster mit hohen Kaufanteilen in den Warengruppen Limonade (soda) und Dosenbier (canned beer).



(a)



(b)

Abb. 7: Prototypenprofile für Transaktionen in (a) Cluster-Nummer 2 und (b) Cluster-Nummer 5.

Neben den Prototypenprofilen der Transaktions-Cluster können auch die in den Medoiden der einzelnen Cluster enthaltenen Warengruppen, Artikel, und so weiter, näher analysiert werden. Auf Warengruppen-Niveau entsprechen diese natürlich den überdurchschnittlich häufig nachgefragten Warengruppen, wie sie bereits in Abbildung 7 dargestellt sind.

```
> inspect(samp[clust[[5]]$medoids[2]])

  items

1 {sausage, rolls/buns, bottled water, soda, newspapers}

> inspect(samp[clust[[5]]$medoids[5]])

  items

1 {soda, canned beer}
```

Wie bereits zu Beginn dieses Abschnittes angedeutet, kann als Hauptvorzug einer Prototypen-basierten Verbundanalyse die Möglichkeit einer disaggregierten beziehungsweise segmentspezifischen Betrachtungsweise des Phänomens Sortimentsverbund gesehen werden (vergleiche Schnedlitz/Reutterer/Joos 2001, Decker/Monien 2003). In der gegenwärtigen Marketing-Praxis ist dies insbesondere dann von Interesse, wenn personalisierte Transaktionshistorien (wie dies zum Beispiel bei Loyalitätsprogrammen in Kombination mit elektronisch lesbaren Kundenkarten der Fall ist) verfügbar sind und segmentspezifisch disaggregierte Verbundanalysen für eine effektivere und effizientere Kundenansprache genutzt werden sollen. Eine Erweiterung in Richtung dynamischer Kundensegmentierung sowie ein Anwendungsbeispiel aus der Direktmarketing-Praxis findet sich bei Reutterer et al. (2006). Eine ähnliche Vorgehensweise zur zielgruppengerechten Segmentansprache findet auch im prominenten Tesco Clubcard-Programm Anwendung (vergleiche Humby/Hunt 1003, S. 143 ff.). Gelegentlich wird ein solcher auf realen Transaktionsdaten basierender Segmentierungsansatz in der einschlägigen Literatur auch als „Subsegmentation“ bezeichnet (vergleiche Malthouse 2003). Wie Boztug und Reutterer (2006) zeigen, bietet sich eine prototypenbasierte Warenkorbanalyse auch als geeignete Methode zur Vorverdichtung und Warenkorbselektion an, um nachfolgend segmentspezifisch maßgeschneiderte Erklärungsmodelle für Kreuzeffekte zwischen Warengruppen in Abhängigkeit von diversen Marketing-Variablen zu schätzen.

Generierung von bedeutsamen Assoziationsregeln

Ähnlich wie die soeben diskutierte Vorgehensweise zielen auch neuere aus der Data-Mining-Literatur stammende Ansätze auf die Analyse der gemeinsamen Kaufhäufigkeiten für eine (typischerweise sehr große) Auswahl von Warengruppen oder einzelnen Artikeln ab. Es handelt sich hierbei um Methoden zur Konstruktion und Beurteilung sogenannter Assoziationsregeln (Association Rules, Agrawal/Srikant 1994), die über die Einschränkung der Affinitätsanalyse auf paarweise Verbundbeziehungen hinaus gehen und die in den beobachteten Transaktionsdaten verborgenen Interdependenzstrukturen über einen probabilistischen Messansatz in Form von Regeln zwischen beliebigen Mengen von Artikeln beziehungsweise Warengruppen abbilden.

Gehen wir zunächst von folgender einfachen (und in mancherlei Hinsicht auch plausibel erscheinenden) Regel aus: „Wenn ein Kunde Brot und Milch kauft, wird er auch Butter kaufen“. Eine solche Regel beschreibt ein Muster im Aufbau von Warenkörben und wird aus Marketingsicht dann interessant, wenn sie ausgestattet mit gewissen Wahrscheinlichkeitsaussagen auch Rückschlüsse auf beobachtbares Kaufverhalten zulässt (Berry/Linoff 2004). Assoziationsregeln (Agrawal/Imielinski/Swami 1993) formalisieren nun derartige Regeln als Implikationen der Art {Brot, Milch} $\Rightarrow$ {Butter}. Ausgehend von einem gegebenen Transaktionsdatensatz, beinhalten Assoziationsregeln zunächst alle möglichen Regeln, die sich aus dem Sortiment bilden lassen und die einen Mindestwert des Signifikanzmaßes Support und des Qualitätsmaßes Konfidenz übersteigen. Der Support ist dabei als relativer Anteil an allen Transaktionen definiert, in welchen sich die in einer Regel befindlichen Warengruppen oder Artikel (Items) befinden. Weist die oben erwähnte Regel beispielsweise einen Support von 1 % auf, ist dies gleichbedeutend damit, dass die auch als Itemmenge bezeichnete Kombination von Warengruppen {Brot, Milch, Butter} gemeinsam in 1 % aller Transaktionen als Nachfragemuster gefunden wurde. Die Konfidenz einer Regel gibt hingegen an, wie oft die von einer Regel behauptete Implikation zutrifft. Weist die erwähnte Regel etwa eine Konfidenz von 50 % auf, ist dies gleichbedeutend damit, dass in der Hälfte der Warenkörbe, die Brot und Milch enthalten, auch Butter nachgefragt wird, also die Regel {Brot, Milch} $\Rightarrow$ {Butter} zutrifft. Dies muss allerdings nicht notwendigerweise auf all jene Regeln gleichermaßen zutreffen, die sich aus Permutationen der in einem Itemset

enthaltenen Warengruppen beziehungsweise Artikel erzeugen lassen. Mit anderen Worten, können die Regeln {Brot, Butter} ⇒ {Milch} oder {Butter, Milch} ⇒ {Brot} trotz identischen Supports durchaus mit anderen Konfidenzen ausgestattet sein. Dies macht deutlich, dass Support ein symmetrisches und Konfidenz ein asymmetrisches Verbundmaß darstellt (vergleiche dazu auch das Illustrationsbeispiel bei Decker/Schimmelpfennig 2002).

Für Regeln der oben beschriebenen Art existieren eine Reihe weiterer Qualitätsmaße. Ein wichtiges Maß ist Interest (Brin et al. 1997), das im Anschluss an die einschlägige Literatur auch als Lift bezeichnet wird. Das Lift-Maß gibt das Verhältnis der gemeinsamen Vorkommenshäufigkeit der linken (dem sogenannten Rumpf oder Prämisse) und der rechten Seite (dem sogenannten Kopf oder Konklusion) einer Regel von der unter Annahme stochastischer Unabhängigkeit erwarteten Vorkommenshäufigkeit an[4]. Lift-Werte größer als 1 weisen somit auf Komplementäreffekte zwischen den im Regelrumpf enthaltenen Items und dem Regelkopf hin, während Werte kleiner als 1 Substitutionseffekte signalisieren.

Als Engpassfaktor bei der Identifikation von bedeutsamen Assoziationsregeln erweist sich die mit zunehmender Sortimentsgröße explodierende Menge aller möglichen Itemsets (einschließlich der daraus ableitbaren Regeln). Zur Bewältigung des damit einhergehenden Komplexitätsproblems wurde in der einschlägigen Literatur zum Association-Rule-Mining folglich eine Reihe effizienter Suchstrategien vorgeschlagen. Im Packet *arules* ist eine Variante des weit verbreiteten APRIORI-Algorithms (Agrawal/Srikant 1994) implementiert, der für ein vorgegebenes minimales Support- und Konfidenzkriterium alle zulässigen Regeln findet. Um auf das Anwendungsbeispiel des „Groceries"-Datensatzes zurückzukehren, werden Assoziationsregeln mit minimalem Support von 0,1 % und einer Konfidenz von mindestens 20 % gesucht (um die Ausgabe des Algorithmus zu unterdrücken wird der Kontrollparameter `verbose` auf Falsch gesetzt).

```
> rules <- apriori(Groceries, parameter = list(support = 0.001,

      confidence = 0.2), control = list(verbose = FALSE))
```

Eine Übersicht zur gefundenen Regelmenge kann nun mittels der Funktion `summary()` erzeugt werden:

```
> summary(rules)

set of 21574 rules

rule length distribution (lhs + rhs):

  1    2    3    4    5

  1  620 9337 9824 1792

   Min.  1st Qu.  Median    Mean  3rd Qu.   Max.

  1.000   3.000   4.000   3.593   4.000   5.000

summary of quality measures:

    support           confidence         lift

 Min.   :0.00102   Min.   :0.200   Min.   : 0.803

 1st Qu.:0.00112   1st Qu.:0.263   1st Qu.: 2.115

 Median :0.00132   Median :0.355   Median : 2.756

 Mean   :0.00196   Mean   :0.396   Mean   : 3.016

 3rd Qu.:0.00193   3rd Qu.:0.500   3rd Qu.: 3.612
```

```
 Max.   :0.25552   Max.   :1.000   Max.   :35.716
```

Für die hier gewählten Minimalvorgaben hinsichtlich Support und Konfidenz wurden 21574 Regeln gefunden. Die Verteilung der Anzahl der Warengruppen, die in den Regeln (und zwar unabhängig ob im Regelrumpf oder –kopf) vorkommen wird als nächstes angegeben. Im Durchschnitt enthalten die gefundenen Regeln 3,593 unterschiedliche Warengruppen. Der letzte Teil der Ausgabe betrifft die Qualitätsmaße der Regeln. Es werden Statistiken zur Verteilung von Support, Konfidenz und Lift angegeben.

Selbstverständlich wird dem Benutzer nicht die Inspektion der gesamten Regelliste abverlangt, sondern eine Reihe von Selektions- und Auswahlhilfen zur Aufdeckung „interessanter“ Regeln angeboten. Einzelne Regeln können mit der Funktion `inspect()` betrachtet werden. Im Folgenden werden die Regeln mit der Funktion `SORT()` nach ihrem Wert des Lift-Maßes absteigend sortiert und die ersten drei Regeln ausgegeben.

```
> inspect(SORT(rules, by = "lift")[1:3])

   lhs                   rhs                        support confidence  lift

1 {bottled beer,

   red/blush wine} => {liquor}                   0.001932     0.3958 35.72

2 {hamburger meat,

   soda}           => {Instant food products}   0.001220     0.2105 26.21

3 {ham,

   white bread}    => {processed cheese}        0.001932     0.3800 22.93
```

Das Ergebnis liefert erwartungsgemäß typische Beziehungen zwischen Warengruppen, wie zum Beispiel: Schinken, Weißbrot und Käse (Regel 3). Die Interpretation des Liftmaßes impliziert, dass falls Schinken und Weißbrot im Warenkorb enthalten sind, Käse um einen Faktor von 22,93 Mal häufiger nachgefragt wird als im gesamten Datensatz (vergleiche Decker/Schimmelpfennig 2002 und die dort angeführte weiterführende Literatur). Analoges gilt für die anderen Regeln.

Die Darstellung von Regeln kann aber auch auf ganz bestimmte Warengruppen eingeschränkt werden. Interessieren beispielsweise nur jene Regeln, die auf den Kauf von Rindfleisch (beef) hinweisen, kann die rechte Seite der Regeln (`rhs`) mittels der Funktion `subset()` entsprechend ausgewählt werden. Im vorliegenden Beispiel sollen weiters nur die drei Regeln mit der höchsten Konfidenz dargestellt werden:

```
> rulesBeef <- subset(rules, rhs %in% "beef")

> inspect(SORT(rulesBeef, by = "conf")[1:3])

  lhs                   rhs     support confidence  lift

1 {root vegetables,

   whole milk,

   butter,

   rolls/buns}      => {beef}   0.001118     0.4783 9.116

2 {sausage,

   root vegetables,
```

```
   butter}          => {beef}   0.001017     0.4545 8.664

3 {root vegetables,

   butter,

   yogurt}          => {beef}   0.001525     0.3947 7.524
```

Aufgrund der vorhandenen effizienten Suchalgorithmen ist das Finden von Assoziationsregeln auch für große Datenmengen und umfangreiche Sortimente selbst auf Artikelebene möglich und kann unter anderem zur Entscheidungsunterstützung in den Bereichen Sortiments-, Platzierungs- und Werbeplanung hilfreich sein (vergleiche Brijs et al. 2004, Van den Poel/De Schamphelaere/Wets 2004). Allerdings muss auf Probleme mit den Maßen Support, Konfidenz und Lift hingewiesen werden. Wie Hahsler, Hornik und Reutterer (2006) mittels Experimenten zeigen, werden Support, Konfidenz und auch Lift systematisch durch die Kaufhäufigkeit der Produkte beeinflusst, was zu Problemen beim Vergleich von Regeln anhand der Maße führen kann.

Zusammenfassung und Ausblick

Die mit dem R-Erweiterungspaket *arules* zur Verfügung gestellte Infrastruktur erleichtert die explorative Warenkorbanalyse dadurch, dass eine Vielzahl von Analysen schnell und einfach durchgeführt werden können. Die Grundlage dafür wird durch die spezifischen Darstellungsstrukturen für Transaktionsdaten geschaffen, die neben den Kaufinzidenzmatrizen auch zusätzliche Informationen zu einzelnen Artikeln und Transaktionen beinhalten können. Beispielhaft wurden im vorliegenden Beitrag anhand eines realen Supermarktdatensatzes paarweise Ähnlichkeiten zwischen Warengruppen analysiert, es wurden Transaktionen zu prototypischen Warenkörben verdichtet und Assoziationsregeln gesucht.

Die *arules* Infrastruktur ist flexibel und ermöglicht unkomplizierte und rasche Erweiterungen durch die Verwendung weiterer R-Pakete. Beispielsweise können für das Problem der effizienten Verdichtung von hochdimensionalen binären Transaktionsdaten zu prototypischen Warenkörben moderne Kompressionsverfahren wie beispielsweise PROXIMUS (Koyuturk/Grama/Ramakrishnan 2005) oder nachbarschaftsbasierte Clusteralgorithmen wie ROCK (Guha/Rastogi/Shim 2000) eingesetzt werden. Beide Verfahren sind bereits im R-Paket *cba* implementiert und können auch mit *arules* verwendet werden.

Für das Suchen von interessanten Regeln bieten sich neben der Verwendung von den auch in dieser Arbeit verwendeten Qualitätsmaßen für Assoziationsregeln auch probabilistische Verfahren an, die explizit ein Wahrscheinlichkeitsmodell der Daten in die Auswahl von Regeln integrieren. Ein erster Schritt in diese Richtung wurden von Hahsler, Hornik und Reutterer (2005) mit der Entwicklung des Konzeptes Hyperlift bereits getan, welches ebenfalls in *arules* zur Verfügung steht.

Literatur

Agrawal, R. /Srikant, R. (1994): Fast Algorithms for Mining Association Rules in Large Databases. In: Bocca, J. B./Jarke, M./Zaniolo, C. (eds.): Proceedings of the 20th International Conference on Very Large Data Bases (VLDB). Santiago, Chile, pp. 487-499.

Agrawal, R./Imielinski, T./Swami, A. (1993): Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Washington D.C., pp. 207-216.

Berry, M./Linoff, G. (2004): Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. New York.

Böcker, F. (1975): Die Analyse des Kaufverbundes. Ein Ansatz zur bedarfsorientierten Warentypologie. In: zfbf, 27. Jhg., S. 290-306.

Böcker, F. (1978): Die Bestimmung der Kaufverbundenheit von Produkten. Berlin.

Bordemannn, H.-G. (1986): Analyse von Verbundbeziehungen zwischen Sortimentsteilen im Einzelhandel. Duisburg.

Boztug, Y./Hildebrandt, L. (2006): A Market Basket Analysis Conducted with a Multivariate Logit Model. In: Spiliopoulou, M./Kruse, R./Borgelt, Ch./Nürnberger, A./Gaul, W. (eds.): Studies in Classification, Data Analysis, and Knowledge Organization. From Data and Information Analysis to Knowledge Engineering. Berlin-New York, pp. 558-565.

Boztug, Y./Reutterer, T. (2006): A Combined Approach for Segment-Specific Analysis of Market Basket Data. Humboldt-Universität zu Berlin: SFB 649 Discussion Paper, No. 06. Berlin.

Boztug, Y./Silberhorn, N. (2006): Modellierungsansätze in der Warenkorbanalyse im Überblick. In: Journal für Betriebswirtschaft (im Erscheinen).

Brijs, T./Swinnen, G./Vanhoof, K./Wets, G. (2004): Building an Association Rules Framework to Improve Product Assortment Decisions. In: Knowledge Discovery and Data Mining, Vol. 8, 1, pp. 7-23.

Brin, S./Motwani, R./Ullman, J. D./Tsur, S. (1997): Dynamic Itemset Counting and Implication Rules for Market Basket Data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Tucson, AZ, pp. 255-264.

Decker, R. (2005): Market Basket Analysis by Means of a Growing Neural Network, The International Review of Retail. In: Distribution and Consumer Research, Vol. 15, 2, pp. 151-169.

Decker. R./Monien, K. (2003): Market Basket Analysis with Neural Gas Networks and Self-organising Maps. In: Journal of Targeting, Measurement and Analysis for Marketing, Vol. 11, 4, pp. 373-386.

Decker, R./Schimmelpfennig, H. (2002): Alternative Ansätze zur datengestützten Verbundmessung im Electronic Retailing. In: Ahlert, D./Olbrich, R./Schröder, H. (Hrsg.): Jahrbuch Handelsmanagement 2002, Electronic Retailing. Frankfurt am Main, S. 193-212.

Dickinson, R./Harris, F./Sircar, S. (1992): Merchandise Compatibility: An Exploratory Study of its Measurement and Effect on Department Store Performance. In: International Review of Retail, Distribution and Consumer Research, Vol. 2, 4, pp. 351-379.

Gaul, W./Both, M. (1998): Computergestütztes Marketing. Berlin.

Guha, S./Rastogi, R./Shim, K., (2000): ROCK: A Robust Clustering Algorithm for Categorical Attributes. In: Information Systems, Vol. 25, 5, pp. 345-366.

Hahsler, M./Grün, B./Hornik, K. (2005): Arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. In: Journal of Statistical Software, Vol. 14, 15, pp. 1-25.

Hahsler, M./Hornik, K./Reutterer, T. (2005): Implications of Probabilistic Data Modeling for Rule Mining. Report 14. Wien: Research Report Series, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien.

Hahsler, M./Hornik, K./Reutterer, T. (2006): Implications of Probabilistic Data Modeling for Mining Association Rules. In: Spiliopoulou, M.,/Kruse, R./Borgelt, Ch./Nürnberger, A./Gaul, W. (eds.): Studies in Classification, Data Analysis, and Knowledge Organization. From Data and Information Analysis to Knowledge Engineering, pp. 598-605, Berlin-New York.

Hilderman, R. J./Hamilton, H. J./Carter, C. L./Cercone, N. (1998): Mining Association Rules from Market Basket Data Using Share Measures and Characterized Itemsets. In: International Journal on Artificial Intelligence Tools, Vol. 7, 3, pp. 189-220.

Hruschka, H. (1985): Der Zusammenhang zwischen Verbundbeziehungen und Kaufaktbeziehungsweise Käuferstrukturmerkmalen. In: zfbf, 37. Jhg., S. 218-231.

Hruschka, H. (1991): Bestimmung der Kaufverbundenheit mit Hilfe eines probabilistischen Messmodells. In: zfbf, 43. Jhg., S. 418-434.

Hruschka, H./Lukanowicz, M./Buchta, C. (1999): Cross-Category Sales Promotion Effects. In: Journal of Retailing and Consumer Services, Vol. 6, 2, pp. 99-105.

Humby, C./Hunt, T. (2003): Scoring Points. How Tesco Is Winning Customer Loyalty. London.

Julander, C.-R. (1992): Basket Analysis. A New Way of Analyzing Scanner Data. In: International Journal of Retail and Distribution Management, Vol. 20, 1, pp. 10-18.

Kaufman, L./Rousseeuw, P. J. (1990): Finding Groups in Data: An Introduction to Cluster Analysis. New York.

Koyuturk, M./Grama, A./Ramakrishnan, N. (2005): Compression, Clustering and Pattern Discovery in Very High Dimensional Discrete-attribute Datasets. In: IEEE Transactions on Knowledge and Data Engineering, Vol. 17, 4, pp. 447-461.

Larson, J. S./Eric, T. B./Fader, P. S. (2005): An Exploratory Look at Supermarket Shopping Paths. In: International Journal of Research in Marketing, Vol. 22, 4, pp. 395-414.

Malthouse, E. C. (2003): Database Sub-segmentation. In: Iacobucci, D./Calder, B. (eds.): Kellogg on Integrated Marketing. New York, pp. 162-188.

Manchanda, P./Ansari, A./Gupta, S. (1999): The „Shopping Basket“: A Model for Multi-Category Purchase Incidence Decisions, In: Marketing Science, Vol. 18, 2, pp. 95-114.

Merkle, E. (1979): Die Analyse des Sortimentsverbundes. In: Dichtl, E./Schobert, R. (Hrsg.): Mehrdimensionale Skalierung. Methodische Grundlagen und betriebswirtschaftliche Anwendungen. Berlin, S. 75-88.

Merkle, E. (1981): Die Erfassung und Nutzung von Informationen über den Sortimentsverbund in Handelsbetrieben. Berlin.

Mild, A./Reutterer, T. (2001): Collaborative Filtering Methods for Binary Market Basket Data Analysis. In: Lecture Notes in Computer Science, Vol. 2252, pp. 302-313.

Mild, A./Reutterer, T. (2003): An Improved Collaborative Filtering Approach for Predicting Cross-category Purchases Based on Binary Market Basket Data. In: Journal of Retailing and Consumer Services, Vol. 10, 3, pp. 123-133.

Milligan, G. W./Cooper, M. C. (1985): An Examination of Procedures for Determining the Number of Clusters in a Data Set. In: Psychometrika, Vol. 50, 2, pp. 159-179.

Müller-Hagedorn, L. (2005): Handelsmarketing. Stuttgart.

Müller-Hagedorn, L. (1978): Das Problem des Nachfrageberbundes in erweiterter Sicht. In: zfbf, 30. Jg., S. 181-193.

Passingham, J. (1998): Grocery Retailing and the Loyalty Card. In: Journal of Market Research Society, Vol. 40, Jan., pp. 55-63.

Reutterer, T./Mild, A./Natter, M./Taudes, A. (2006): A Dynamic Segmentation Approach for Targeting Direct Marketing Efforts. In: Journal of Interactive Marketing (forthcoming).

Russell, G. J./Petersen, A. (2000): Analysis of Cross Category Dependence in Market Basket Selection. In: Journal of Retailing, Vol. 76, 3, pp. 367-392.

Russell, G. J./Ratneshwar, S./Shocker, A. D./Bell, D./Bodapati, A./Degeratu, A./Hildebrandt, L./Kim, N./Ramaswami, S./Shankar, V. H. (1999): Multiple-Category Decision-Making: Review and Synthesis. In: Marketing Letters, Vol. 10, 3, pp. 319-332.

Schnedlitz, P./Kleinberg, M. (1994): Einsatzmöglichkeiten der Verbundanalyse im Lebensmittelhandel. In: Der Markt, 33. Jhg., S. 31-39.

Schnedlitz, P./Reutterer, T./Joos, W. (2001): Data-Mining und Sortimentsverbundanalyse im Einzelhandel. In: Hippner, H./Küsters, U./Meyer, M./Wilde, K. D. (Hrsg.): Handbuch Data Mining im Marketing. Knowledge Discovery in Marketing Databases. Wiesbaden, S. 951-970.

Seetharaman, P. B./Chib, S./Ainslie, A./Boatwright, P./Chan, T./Gupta, S./Mehta, N./Rao, V./Strijnev, A. (2005): Models of Multi-Category Choice Behavior. In: Marketing Letters, Vol. 16, 3-4, pp. 239-254.

Sneath, P. H. (1957): Some Thoughts on Bacterial Classification. In: Journal of General Microbiology, Vol. 17, 2, pp. 184-200.

Van den Poel, D./De Schamphelaere, J./Wets, G. (2004): Direct and Indirect Effects of Retail Promotions on Sales and Profits in the Do-it-yourself Market. In: Expert Systems with Applications, Vol. 27, 1, pp. 53-62.

Zaki, M. J. (2000): Scalable Algorithms for Association Mining. In: IEEE Transactions on Knowledge and Data Engineering, Vol. 12, 3, pp. 372-390.

Zentes, J. (1998): EDV-gestütztes Marketing. Ein informations- und kommunikationstheoretischer Ansatz. Berlin.

Zielke, S. (2002): Kundenorientierte Warenplatzierung. Modelle und Methoden für das Category Management. Stuttgart.

---

[1] R ist eine freie Programmierumgebung für statistische Anwendungen. Aktuelle Versionen des R Basissystems und eine umfangreiche Kollektion an Erweiterungspaketen wie **arules** können vom Comprehensive R Archive Network (CRAN) unter http://cran.r-project.org/ bezogen werden. Dort findet man auch eine ausführliche Dokumentation von Installations- und Download-Anweisungen.

[2] Wir schließen uns damit der in der einschlägigen Literatur gebräuchlichen Konvention an, die numerische Repräsentation der Transaktionen als Realisationen sogenannte „Pick-Any-Daten" aufzufassen (vergleiche Manchanda/Ansari/Gupta 1999, Russel/Petersen 2000). Für Illustrationszwecke und vor dem Hintergrund der nachfolgenden Anwendungsbeispiele beschränken wir uns in der weiteren Darstellung auf Warengruppen anstelle von Produkten beziehungsweise Artikeln.

[3] Für das Einlesen eigener Datensätzen steht in **arules** die Funktion `read.transactions()` zur Verfügung. Diese Funktion kann vorbearbeitete Daten in verschiedenen Formaten von der Festplatte einlesen. Nähere Informationen dazu findet man in der Dokumentation zum **arules** Paket unter http://cran.r-project.org/src/contrib/Descriptions/arules.html.

[4] Vor dem Hintergrund der Beuteilung von komplementären beziehungsweise substitutionalen Verbundeffekten ist diese Annahme in der Sortimentsverbundanalyse keineswegs neu und gelangt auch bereits in der „klassischen" Affinitätsanalyse zur Anwendung (vergleiche dazu bereits bei Böcker 1978, S. 20f. und S. 80f. oder Hruschka 1991).