Data Mining und Marketing am Beispiel der explorativen Warenkorbanalyse

Von Thomas Reutterer, Michael Hahsler und Kurt Hornik

Techniken des Data Mining stellen für die Marketingforschung und -praxis eine zunehmend bedeutsamere Bereicherung des herkömmlichen Methodenarsenals dar. Mit dem Einsatz solcher primär datengetriebener Analysewerkzeuge wird das Ziel verfolgt, marketingrelevante Informationen „intelligent" aus großen Datenbanken (sog. Data Warehouses) zu extrahieren und für die weitere Entscheidungsvorbereitung in geeigneter Form aufzubereiten. Im vorliegenden Beitrag werden Berührungspunkte zwischen Data Mining und Marketing diskutiert und der konkrete Einsatz ausgewählter Data-Mining-Methoden am Beispiel der explorativen Warenkorb- bzw. Sortimentsverbundanalyse für einen Transaktionsdatensatz aus dem Lebensmitteleinzelhandel demonstriert. Zur Anwendung gelangen dabei Techniken aus dem Bereich der klassischen Affinitätsanalyse, ein *K*-Medoid-Verfahren der Clusteranalyse sowie Werkzeuge zur Generierung und anschließenden Beurteilung von Assoziationsregeln zwischen im Sortiment enthaltenen Warengruppen. Die Vorgehensweise wird dabei anhand des mit der Statistik-Software R frei verfügbaren Erweiterungspakets **arules** illustriert.

1. Einführung

Die modellgestützte Entscheidungsvorbereitung blickt in der Marketingwissenschaft auf eine langjährige Forschungstradition zurück und kann auch in der Praxis beachtliche Implementierungserfolge verzeichnen (vgl. *Leeflang et al.* 2000). Hinsichtlich der Verfügbarkeit von potenziell marketingentscheidungsrelevanten Datenbeständen brachte die von *Blattberg/Glazer/Little* (1994) treffend als *Information Revolution* bezeichnete Entwicklung der beiden letzten Jahrzehnte enorme Fortschritte mit sich: Sowohl Anfallsvolumina und -frequenzen, aber auch die Zugriffsgeschwindigkeiten haben enorme Steigerungen erfahren. Im Gegensatz zur *Datenarmut* aus vergangenen Tagen sehen sich die Modellkonstrukteure heute mit einer wahren *Datenexplosion* konfrontiert (vgl. *McCann/Gallagher* 1990; *Leeflang et al.* 2000, S. 301 ff.). Nicht selten hat letztere zur Konsequenz, dass eine aussagekräftige Interpretation der in der modernen Marketingentscheidungspraxis von unterschiedlichsten unternehmensinternen wie -externen Herkunftsquellen bezogenen und meist in einem sog. *Data Warehouse* organisierten Daten ohne eine gewisse Automatisierung des Modellierungsprozesses kaum mehr mit vertretbarem Aufwand möglich ist (*Bucklin/Lehmann/Little* 1998).

Vor diesem Hintergrund scheint sich immer mehr die Einsicht durchzusetzen, dass die weitere erfolgreiche Anwendung leistungsfähiger Entscheidungsmodelle eine Anreicherung bestehender Ansätze um fortschrittliche computergestützte Analysekonzepte benötigt. Letztere

Thomas Reutterer ist außerordentlicher Professor am Institut für Handel und Marketing an der Wirtschaftsuniversität Wien, Augasse 2–6, A-1090 Wien, Tel.: +43/1/3 13 36-4619, E-Mail: Thomas.Reutterer@wu-wien.ac.at

Michael Hahsler ist Privatdozent am Institut für Informationswirtschaft an der Wirtschaftsuniversität Wien, Augasse 2–6, A-1090 Wien, Tel.: +43/1/3 13 36-6081, E-Mail: Michael.Hahsler@wu-wien.ac.at

Kurt Hornik ist Universitätsprofessor am Department für Statistik und Mathematik an der Wirtschaftsuniversität Wien, Augasse 2–6, A-1090 Wien, Tel.: +43/1/3 13 36-4756, E-Mail: Kurt.Hornik@wu-wien.ac.at

sollen dabei eine gehaltvolle Verdichtung und eine – im Sinne von adaptiv und wissensbasiert verstandene – „intelligente“ (Vor-) Selektion interessanter Zusammenhangs- und/oder Abhängigkeitsstrukturen in den Daten erlauben (vgl. etwa *Rangaswamy* 1993; *Wierenga/Van Bruggen* 2000, S. 119 ff.; *Matsatsinis/Siskos* 2003). Die meisten in jüngerer Zeit für die Bewältigung solcher Aufgaben eingesetzten Verfahren finden ihre methodologischen Wurzeln im Bereich des maschinellen Lernens, wo sie auch als sog. *Data-Mining-Techniken* bezeichnet werden (vgl. *Berry/Linoff* 1997; *Hastie/Tibshirani/Friedman* 2001).

Wie die Vergangenheit gezeigt hat, ist für die Marketingwissenschaft der Methodenimport aus diversen formal- wie substanzwissenschaftlichen Disziplinen freilich nicht ungewöhnlich (vgl. *Hildebrandt* 2000). Abgesehen von den im Gefolge der Informationsrevolution erwähnten Zweckmäßigkeitsüberlegungen setzt ein erfolgreicher Transfer „neuer“ methodischer Ansätze wie den hier interessierenden Verfahren des Data Mining unter anderem die nachweisliche Überlegenheit gegenüber prinzipiell vergleichbaren Alternativen aus dem etablierten Methodenrepertoire, nicht zuletzt aber auch die Verfügbarkeit geeigneter Programmierumgebungen bzw. Softwarelösungen voraus. Letzteres dürfte insbesondere im Umgang mit Massendaten, wie es bei Data-Mining-Anwendungen typischerweise der Fall ist, von besonderer Bedeutung sein.

In diesem Beitrag werden zum einen wesentliche Charakteristika von Techniken des Data Mining im Vergleich zu herkömmlichen statistischen Modellierungsansätzen erörtert und dabei Berührungspunkte mit konkreten marketingrelevanten Datenanalyseproblemen näher beleuchtet. Den Ausführungen von *Ravi/Raman/Mantrala* (2006) folgend, dass sich die soeben skizzierten Entwicklungen im Handel besonders bemerkbar machen, wird zum zweiten am Beispiel der explorativen Analyse von Warenkorbdaten des Handels der Einsatz ausgewählter Data-Mining-Methoden näher erläutert und auch vorgeführt. Für Demonstrationszwecke werden dabei unter Zuhilfenahme einer geeigneten Programmierumgebung ein Datensatz aus dem Lebensmitteleinzelhandel herangezogen und die einzelnen Analyseschritte ausführlich kommentiert.

2. Zum Verhältnis von Data Mining und Marketing

Technisch betrachtet kann Data Mining zunächst als Teil eines Prozesses aufgefasst werden, der in der einschlägigen Literatur als *Knowledge Discovery in Databases* (KDD, *Fayyad/Piatetsky-Shapiro/Smith* 1996) bezeichnet wird. Wie in *Abb. 1* dargestellt, besteht dieser Prozess aus mehreren Schritten mit dem Ziel, aus umfangreichen und schlecht strukturierten Rohdaten assoziative Muster zu extrahieren, die für substanzwissenschaftliche Fragestellungen zu neuer Erkenntnis führen (können). Dabei liefern die durch Data Mining gefundenen Muster per se keinerlei Erkenntniszuwachs, sondern erst durch deren fachkundige Interpretation im Lichte der interessierenden Problemstellung (vgl. *Brachman/Anand* 1996).

Der sinnvolle Einsatz von Data Mining kommt somit vor allem dann in Betracht, wenn lediglich rudimentär vorhandene aber kaum hinreichend abgesicherte A-Priori-Vermutungen über diverse im Problemkontext als relevant erachtete Zusammenhangsstrukturen vorliegen. Charakteristisch sind auch die im Unternehmen zwar grundsätzlich verfügbaren aber selten direkt verwertbaren Daten sowie der in der Regel voll- oder teilautomatisierte Ablauf des Mining-Prozesses (vgl. *Hand/Mannila/Smyth* 2001; *Tan/Steinbach/Kumar* 2006). Die Entwicklung von diesbezüglichen Problemlösungen und deren Implementierung in Informations- und Entscheidungsunterstützungssystemen war und ist eine Domäne der Informationstechnologie- (IT) und Informationsmanagement-Experten. Es ist daher auch nahe liegend, dass sich gewisse Zweige der Wirtschaftsinformatik in Richtung KDD entwickelten und im Zuge dessen auch „klassische“ marketingrelevante Forschungsfelder wie Kundensegmentierung oder Warenkorbanalyse das Forschungsinteresse von Informatikern bzw. Vertretern der angewandten Statistik finden (vgl. dazu etwa *Berry/Linoff* 1997 oder der Sammelband von *Hippner et al.* 2001).

Trotz einer inzwischen reichhaltigen Palette an kommerziellen (vgl. den Überblick von *Haughton et. al* 2003) und frei verfügbaren (z. B. das Projekt *Weka*; vgl. *Witten/Frank* 2005) Softwarepaketen ist bislang ein Aufbrechen disziplinärer Grenzen nur sehr zögerlich zu beobachten. Tatsächlich finden sich hinter den bislang im Marketing breiter diskutierten Data-Mining-Anwendungen meist sehr spezifische, typischerweise mit der Verarbeitung von Massendaten verbundene Problemstellungen im Di-

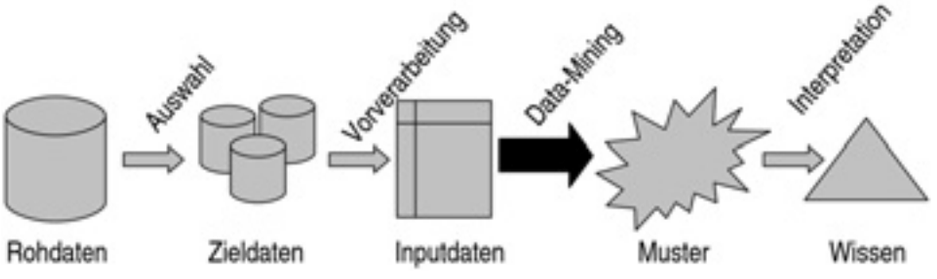


Abb. 1: Der KDD-Prozess

rektmarketing oder des Customer-Relationship-Managements (CRM) wieder (*Levin/Zahavi* 2001; *Shaw et al.* 2001; *Rygielski/Wang/Yen* 2002; *Verhoef et al.* 2002; *Neslin et al.* 2006). Dieser Befund wird durch eine Reihe spezialisierter Konferenzen (vgl. z. B. *Winer/Tuzhilin* 2005) oder vereinzelte Beiträge in renommierten Zeitschriften wie *Marketing Science*, *Management Science* oder dem *Journal of Interactive Marketing*, belegt. Eine gewisse bislang beobachtbare Skepsis der akademischen Marketingforschung gegenüber der Adoption von Data-Mining-Techniken mag aber auch daher rühren, dass eine datengetriebene Komplexitätsreduktion sowie robuste (und eventuell sogar automatisierte) Prognosen primär für den zeitkritischen Routinebetrieb von Modellanwendungen in der tagtäglichen Marketingpraxis benötigt werden (vgl. *Cui/Curry* 2005, S. 597 f.; *Levin/Zahavi* 1998). Dies sollte unseres Erachtens die Marketingwissenschaft jedoch nicht dazu veranlassen, die Verantwortlichkeit für die Entwicklung angemessener Problemlösungen gänzlich in den IT-Kontext abzuschieben.

Was die innerhalb des KDD-Prozesses eingesetzten Data-Mining-Methoden anbelangt, bezeichnen diese als Sammelbegriff eine heterogene Menge von Datenmanipulationstechniken, die traditionelle statistische Analysemethoden mit moderneren Algorithmen zur Verarbeitung von großen Datenmengen (aus den Fachgebieten der künstlichen Intelligenz, des maschinellen Lernens und der Mustererkennung) vereint. Methodologisch betrachtet unterscheiden sich erstere von letzteren fundamental hinsichtlich ihrer Annahmen über die Daten generierenden Prozesse: Während traditionelle Modellierungsansätze der parametrischen Statistik ein rigides stochastisches Datenmodell unterstellen, betrachten Ansätze des „algorithmischen Modellierens“ (*Data Mining im engeren Sinne*) die Beschaffenheit der Daten als komplexe Unbekannte und versuchen, diese als flexible, nichtparametrische Funktion zu approximieren (vgl. *Breiman* 2001; *Hastie/Tibshirani/Friedman* 2001, S. 9 ff.). Die Maschine „lernt“ dabei die den Daten zugrunde liegende unbekannte Verteilungsfunktion als algorithmisches Modell in Form von neuronalen Netzen, Support Vector Machines, Entscheidungsbäumen (erzeugt mittels AID, CHAID, CART, etc.), Random Forests oder dergleichen.

Wie *Cooper/Giuffrida* (2000) beispielhaft vorführen, schließen sich diese beiden Welten aber nicht gegenseitig aus und können durchaus sinnvoll miteinander kombiniert werden. So schätzen die Autoren in einem zweistufigen Prozess zunächst herkömmliche Marktresponsemodelle und zeigen, dass deren prognostische Güte verbessert werden kann, wenn ein einfaches Verfahren der regelbasierten Induktion auf die Residuen angewendet wird. Weitere Marketing-Beispiele für einen solchen „Tandem-Zutritt“ findet man auch bei *Elsner/Krafft/Huchzermeier* (2004), *Natter* et al. (2006) oder *Van den Poel/De Schamphelaere/Wets* (2004).

Ein gebräuchliches Einteilungsschema für Data-Mining-Methoden unterscheidet zwischen den zugrunde liegenden Analyseproblemen, die (1) *prädiktiver* oder (2) *explorativer* Natur sein können (vgl. *Fayyad/Piatetsky-Shapiro/Smyth* 1996; *Tan/Steinbach/Kumar* 2006). Insbesondere hinsichtlich (1) dürfte nicht nur im Marketing das von *Leo Breiman* (2001, S. 199) zwar etwas überzeichnend formulierte „98:2-Verhältnis“ zugunsten der traditionellen statistischen Modellierung gegenüber „algorithmisch“ geprägten Modellierungsansätzen nicht allzu weit von der Realität entfernt liegen. Dennoch stößt man bei Durchsicht der einschlägigen Literatur immer häufiger auf empirische Leistungsvergleiche, die für eine Überlegenheit moderner Data-Mining-Techniken sprechen. Letztere äußert sich insbesondere in Bezug auf die Robustheit und Prognosegüte von Modellen, die komplexe und nicht-lineare Beziehungsstrukturen zwischen Ziel- und einer Vielzahl an (unterschiedlich skalierten) erklärenden Variablen erfassen wollen.

Besonders augenscheinlich wird dieser Befund bei einer Gegenüberstellung von herkömmlichen statischen Modellen und mehrschichtigen neuronalen Feedforward-Netzen, die jede beliebige funktionale Relation mit hinreichender Exaktheit abzubilden imstande sind (vgl. *Hornik/Stinchcombe/White* 1989). *Hruschka* und Kollegen zeigen für eine Reihe von Marketing-Anwendungen (z. B. *Hruschka* 1993, 2001; *Hruschka/Natter* 1993; *Hruschka et al.* 2002; *Hruschka/Fettes/Probst* 2004), dass bestimmte neuronale Netzwerkmodelle als Verallgemeinerungen von traditionelleren statistischen Modellen, wie etwa logistische Regressions- oder Logit-Modelle interpretiert werden können (vgl. dazu auch *Kumar/Rao/Soni* 1995). Wie von *West/Brockett/Golden* (1997) an einem Beispiel der Einkaufsstättenwahl von Konsumenten vorgeführt wird, sind geeignete neuronale Netzwerkarchitekturen auch in der Lage, nichtkompensatorische Entscheidungsheuristiken darzustellen.

Eine Reihe von Anwendungen diverser Data-Mining-Methoden (i.e.S.) im prädiktiven Analysekontext des Database-Marketings findet man in Beiträgen des Sammelbandes von *Hippner et al.* (2001). Darüber hinaus sei auf den Einsatz Bayesianischer Netze zur Responseanalyse im Direktmarketing bei *Cui/Wong/Liu* (2006) und *Baesens et al*. (2002), von Support Vector Machines zur Neukundenklassifikation bei *Decker/Monien* (2003b) und *Cui/Curry* (2005) oder von Random Forests und neuronalen Netzen zur Kundenabwanderungsanalyse („Churn Prediction“) bei *Buckinx/Van den Poel* (2005), *Neslin et al.* (2006) und *Buckinx/Verstraeten/Van den Poel* (2006) hingewiesen.

Probleme der explorativen Datenanalyse treten im Marketing insbesondere im Rahmen der A-Posteriori-Marktsegmentierung, der Ableitung kompetitiver Marktstrukturen („Competitive Market Structure Analysis“; vgl. *Elrod* 1991) sowie bei der nachfolgend noch ausführlicher behandelten Warenkorbanalyse in Erscheinung. Bei den beiden zuerst genannten – und nicht selten miteinander kombinierten – Analyseproblemen stehen parametrische Verfahren der statistischen Modellierung (z. B. Mischmodelle) mit „modellfreien“ algorithmischen Zutritten

(z. B. Clusteralgorithmen) bekanntlich traditionell im Methodenwettbewerb (für einen Überblick vgl. *Wedel/Kamakura* 1999; *DeSarbo/Manrai/Manrai* 1993). Auch hier findet man wiederum eine Reihe von Beiträgen, die sich geeigneter neuronaler Netzwerkarchitekturen bedienen (*Hruschka/Natter* 1995; *Balakrishnan et al.* 1996; *Mazanec* 1999, 2000; *Reutterer/Natter* 2000). Für diese gelten prinzipiell analoge Überlegungen wie die bereits im Zusammenhang mit der prädiktiven Analyse genannten Vorzüge.

3. Explorative Warenkorbanalyse

Nachdem die deutschsprachige Marketing-Literatur über längere Zeit hinweg die methodische Diskussion prägte (vgl. *Böcker* 1978; *Merkle* 1981; *Müller-Hagedorn* 1978; *Bordemann* 1985; *Hruschka* 1985, 1991), erlebt die Analyse des Nachfrage- bzw. Kaufverbunds zwischen Bestandteilen (Produkten, Warengruppen, etc.) von Einzelhandelssortimenten in der internationalen Marketing-Forschung, aber auch in der einschlägigen Literatur zum Data Mining (vgl. stellvertretend dazu *Berry/Linoff* 1997) 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 mittels diverser marketingpolitischer Maßnahmen (z. B. Platzierung, Preis- und Sonderangebotspolitik, etc.) im Rahmen des Category-Managements des Handels (vgl. *Müller-Hagedorn* 2005). Auf kundenindividuellem oder segmentspezifischem Niveau stößt die Nutzung von Verbundrelationen auch im Rahmen maßgeschneiderter Cross-/Upselling-Aktionen innerhalb von Loyalitätsprogrammen auf verstärktes Interesse (vgl. *Mild/Reutterer* 2003; *Reutterer et al.* 2006).

Ausgangsbasis einer Warenkorbanalyse stellen regelmäßig die im Data Warehouse einer Handelsorganisation gesammelten Transaktionsdaten, die teilweise (z. B. 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 permanent an. Darüber hinaus ist im elektronischen Handel auch das Sortiment besonders reichhaltig. Die prädiktive Analyse besteht hierbei in der Schätzung von Modellen für Auswirkungen (Kreuzeffekte) von Marketing-Aktionen in einer Warengruppe auf das Kaufverhalten in einer anderen Warengruppe (*Hruschka/Lukanowicz/Buchta* 1999, *Manchanda/Ansari/Gupta* 1999; *Russell/Petersen* 2000; *Boztug/Hildebrandt* 2006). Der Anwendungsspielraum solcher Modelle ist jedoch aufgrund der sehr rasch steigenden Modellkomplexität meist auf verhältnismäßig kleine Ausschnitte des Sortiments beschränkt (ausführlicher zu dieser Problematik: *Boztug/Reutterer* 2007).

Explorative Ansätze beabsichtigen demgegenüber, die in großen Mengen von Transaktionsdaten beobachtbaren interdependenten Nachfragemuster mit Hinblick auf die Aufdeckung bedeutsamer Verbundbeziehungen angemessen zu verdichten und für eine nachfolgende Verwertung (z. B. im Rahmen einer prädiktiven Analyse) komprimiert darzustellen bzw. geeignet zu visualisieren (*Berry/Linoff* 1997, *Schnedlitz/Reutterer/Joos* 2001). Da neben konventionellen Zutritten diverse Data-Mining-Algorithmen für eine derartige Aufgabenstellung geradezu prädestiniert sind, liegen in der einschlägigen Literatur auch bereits eine Reihe viel versprechender Anwendungen vor. *Tab. 1* liefert dazu einen strukturierten Überblick über die bislang vorgestellten Verfahrensklassen der explorativen Kaufverbundanalyse, die jeweils spezifische Vorzüge und damit korrespondierende Einsatzgebiete aufweisen. In weiterer Folge wird auf diese Ansätze näher eingegangen.

Softwareseitig sind derzeit einige Systeme verfügbar, die ähnliche Analysemöglichkeiten aus dem Repertoire der

Ansatz	Ausgewählte Quellen	Methodische Kurzcharakteristik	Aggregationsniveau
(1) Affinitätsanalyse	*Böcker* (1978); *Merkle* (1981); *Dickinson/Harris/Sircar* (1992); *Julander* (1992); *Schnedlitz/Kleinberg* (1994)	Repräsentation einer Verbundmatrix bestehend aus paarweisen Assoziationsmaßen	Aggregiert
(2) Prototypenbildende Clusterverfahren	*Schnedlitz/Reutterer/Joos* (2001); *Decker/Monien* (2003a); *Decker* (2005); *Reutterer et al.* (2006)	Verdichtung von Verbundbeziehungen in Warenkörben zu prototypischen Warenkorbklassen	Disaggregiert (segmentspezifisch)
(3) Generierung von 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 zur explorativen Kaufverbundanalyse

explorativen Warenkorbanalyse anbieten. Die führenden kommerziellen Systeme sind zweifelsohne *SPSS Clementine* und der *SAS Enterprise Miner* (vgl. *Herschel* 2006). Eine ausführliche Gegenüberstellung der von den bekanntesten einschlägigen kommerziellen Software-Paketen angebotenen Data-Mining-Ressourcen findet man bei *Haughton et al.* (2003). In regelmäßig durchgeführten Benutzerbefragungen (beispielsweise durch *KDnuggets*[1]) hat sich allerdings herausgestellt, dass frei verfügbare Software und insbesondere R (vgl. *R Development Core Team* 2007)[2] unter anderem auch für typische Data-Mining-Anwendungen immer öfter eingesetzt wird.

Ein direkter Vergleich zwischen *Open-Source-Software* und den naturgemäß wenig transparenten proprietären Softwaresystemen ist (vor allem mit Hinblick auf die jeweils implementierten Algorithmen) freilich nur sehr eingeschränkt möglich. Generell dürften jedenfalls die folgenden Entscheidungskriterien für die Auswahl einer geeigneten Data-Mining-Software von hoher Relevanz sein: Benutzerfreundlichkeit, Verarbeitungsgeschwindigkeit, Skalierbarkeit für sehr große Datenmengen, Verfügbarkeit fortschrittlicher Analysemethoden und laufzeiteffizienter Algorithmen, angebotene Hilfestellung beim Umgang mit technischen Problemen und nicht zuletzt die Lizenzkosten.

Vor diesem Hintergrund weisen die führenden *kommerziellen Pakete* gegenüber frei verfügbarer Software wie R insbesondere im Zusammenhang mit der Skalierbarkeit der Anwendungen, hinsichtlich der für eine benutzerfreundliche Steuerung des Mining-Prozesses verfügbaren interaktiven grafischen Benutzeroberfläche und durch den angebotenen technischen Support Vorteile auf. Bei R wurde bislang die Skalierung für Datenmengen, die weit jenseits der bei PCs gebräuchlichen Hauptspeicherkapazitäten liegen, weitestgehend vernachlässigt. An der Beseitigung dieser Schwäche wird erst seit Kurzem gearbeitet. Zwar sind auch unter R bereits diverse grafische Schnittstellen verfügbar, die den Benutzerkomfort beträchtlich zu steigern imstande sind[3]. Diese sind derzeit jedoch noch deutlich weniger ausgereift als es bei kommerziellen Paketen üblich ist.

Was die Beurteilung der Verarbeitungsgeschwindigkeit der implementierten Algorithmen anbelangt, sind die Leistungsunterschiede differenziert zu sehen. Während bei einigen R Paketen nicht optimierter Programmcode verwendet wird, basiert beispielsweise die Suchstrategie nach Assoziationsregeln in dem nachfolgend näher vorgestellten Paket **arules** (*Hahsler/Grün/Hornik* 2005, 2006) auf einer sehr effizienten Implementierung von *Borgelt* (2003), welche auch in kommerziellen Paketen zur Anwendung gelangt. Zu den Stärken von R zählen neben den entfallenden Lizenzkosten die schnelle Weiterentwicklung der Software und damit die Verfügbarkeit modernster Analysemethoden, die nicht selten erst Jahre später Ihren Weg in kommerzielle Produkte finden. Außerdem hat sich R in den letzten Jahren zur *Lingua Franca* der akademischen Statistik entwickelt und man hat damit (freien) Zugriff auf die jeweils aktuellsten Methodeninnovationen. Schließlich dürften marketingspezifische Erweiterungspakete, insbesondere das Paket **bayesm** mit leistungsfähigen Implementierungen empirischer Bayes-Methoden (vgl. *Rossi/Allenby/McCulloch* 2005), den Verbreitungsgrad von R innerhalb der Marketing-Disziplin weiter begünstigen.

Diesen Entwicklungen folgend wird der praktische Einsatz der in *Tab. 1* skizzierten Ansätze zur explorativen Kaufverbundanalyse unter Zuhilfenahme der im R-Erweiterungspaket **arules** implementierten Data-Mining-Software dargestellt. Für Illustrationszwecke wird dabei ein durchgängig verwendeter, für die Warenkorbanalyse typischer Datensatz aus dem Lebensmitteleinzelhandel herangezogen, dessen datentechnische Repräsentation sowie grundlegenden Eigenschaften nachfolgend kurz erörtert werden.

4. Darstellung von Transaktionsdaten

Jede Transaktion (jeder Warenkorb) enthält alle während eines Einkaufsaktes gemeinsam nachgefragten Artikel. Transaktionsdaten werden im Data Warehouse typischerweise in Form sog. Tupel, d.h. einer geordneten Zusammenstellung von Einträgen, wie folgt abgespeichert:

<Transaktionsnummer, Produktnummer, ...>

Alle Tupel mit der gleichen Transaktionsnummer bilden eine Transaktion, wobei neben den Artikeln zumindest Informationen aus den Artikelstammdaten (z. B. Packungsgröße, Hersteller, etc.) und zum Einkaufsvorgang (z. B. Zeitpunkt, Filiale, Kassenplatz, etc.) verfügbar sind. In der Handelspraxis sind die im Sortiment gelisteten Artikel meist in ein hierarchisches Klassifikationsschema von Warengruppen eingebunden. Eine solche Sortimentshierarchie, wie sie in *Abb. 2* dargestellt ist, kann dann mit in den Transaktionen enthaltenen Artikeln assoziiert werden. Dadurch wird es möglich, die verfügbaren Transaktionen auf jeder beliebigen Hierarchieebene einer Vorselektion zu unterziehen. Das spezifische Untersuchungsinteresse könnte es beispielsweise erfordern, all jene Transaktionen auszuwählen, die Artikel aus gewissen Warengruppen (z. B. Brot und Gebäck) beinhalten. Analog dazu können gewisse Warengruppen (z. B. Molkereiprodukte) aus der weiteren Analyse ausgeschlossen werden. Letzteres erweist sich insbesondere dann als zweckmäßig, wenn es sich um Warengruppen handelt, die besonders häufig nachgefragt werden und daher per se einen starken Ausstrahlungseffekt auf das Restsortiment ausüben.

Für die Warenkorbanalyse werden derartige Transaktionsdaten typischerweise in eine der folgenden zwei *Repräsentationsformen* transformiert (vgl. *Abb. 3* für ein Beispiel):

- Eine binäre Kaufinzidenzmatrix mit Transaktionen in den Zeilen und Artikeln bzw. Warengruppen in den

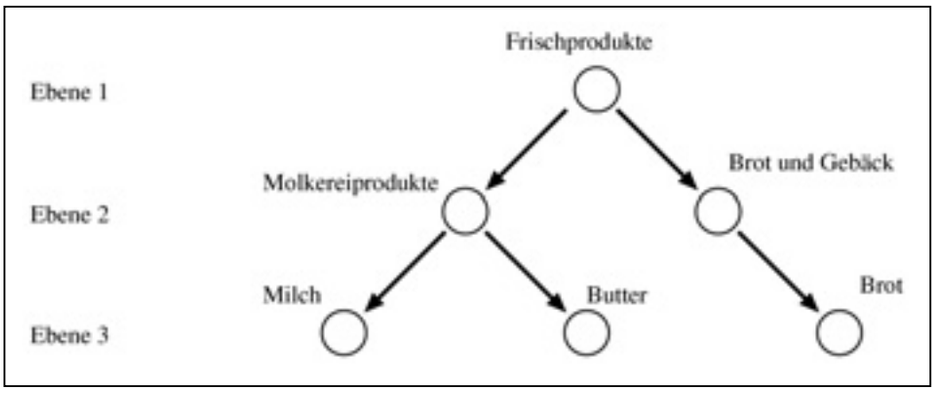


Abb. 2: Exemplarischer Ausschnitt aus einer dreistufigen Sortimentshierarchie

Spalten. Die Einträge stellen den Kauf (1) bzw. Nichtkauf (0) einer Warengruppe in einer Transaktion dar[4]. Diese Darstellung wird oft als *horizontales Datenbanklayout* bezeichnet (vgl. *Zaki* 2000).

- Eine Transaktionsnummernliste, wobei für jede der zeilenweise angeordneten Warengruppen eine Liste jener Transaktionsnummern, welche die betreffende Warengruppe enthält, gespeichert wird. Diese Darstellung wird auch als *vertikales Datenbanklayout* bezeichnet (*Zaki* 2000).

Nachfolgend wird auf die Eigenschaften des Transaktionsdatensatzes näher eingegangen, der im Software-Paket **arules** integriert ist und daher für Testzwecke von Data-Mining-Anwendungen im Rahmen der explorativen Warenkorbanalyse zur Verfügung steht[5]. Es handelt sich dabei um die in einer Supermarktfiliale über einen Kalendermonat hinweg angefallenen Transaktionen, wobei die Warenkörbe bereits zu 169 Warengruppen aggregiert wurden. Nach dem Start von R wird zunächst das Paket **arules** mittels `library("arules")` geladen und der als „Groceries" bezeichnete Datensatz mit der Anweisung `data("Groceries")` eingelesen.

Der Datensatz enthält $N = 9835$ Transaktionen und $J = 169$ Warengruppen, die als J-dimensionale binäre Warenkorbvektoren $\mathbf{x}_n \in \{0,1\}^J$ in einer $N \times J$ Kaufinzidenzmatrix $\boldsymbol{X}^T = [\boldsymbol{x}_n]_{n=1,\ldots,N}$ dargestellt werden können. Die Anweisung summary (`Groceries`) kann verwendet werden, um eine Reihe elementarer Eigenschaften des Datensatzes zusammenzufassen. Die in *Abb. 4* wiedergegebenen relativen Kaufhäufigkeiten der in zumindest 5 % aller Transaktionen vorkommenden Warengruppen verdeutlichen, dass am häufigsten Artikel aus klassischen Nahrungsmittelwarengruppen, wie Vollmilch („whole milk"), Gemüse („other vegetables"), Brötchen („rolls/buns") etc. nachgefragt werden.

Als weitere grundlegende Eigenschaften von Transaktionsdaten interessieren meist Lage- und Streumaße der Warenkorbgröße, wobei letztere als Anzahl der in einem Warenkorb miteinander kombinierten Warengruppen verstanden wird: Im Durchschnitt enthält ein Warenkorb des vorliegenden Datensatzes 4,409 unterschiedliche Warengruppen. Der Median der Warenkorbgröße ist mit 3 Warengruppen deutlich kleiner als der Mittelwert. Dies weist auf eine schiefe Verteilung mit sehr vielen „kurzen" Transaktionen, die nur wenige Warengruppen enthalten (also klassische „Bagatell-", oder Kleinsteinkäufe), hin. Die Verteilung der Warenkorbgrößen wird typischerweise mittels eines Histogramms dargestellt. Das Histogramm für den Datensatz (inklusive der notwendigen Anweisung) ist in *Abb. 5* dargestellt, in welchem eine für Warenkorbdaten typische Verteilung mit sehr vielen „kurzen" und nur wenigen „langen" Transaktionen klar erkennbar ist.

Nachfolgend werden die soeben skizzierten Transaktionsdaten unter Anwendung der in *Tab. 1* erwähnten Verfahren einer explorativen Kaufverbundanalyse unterzogen.

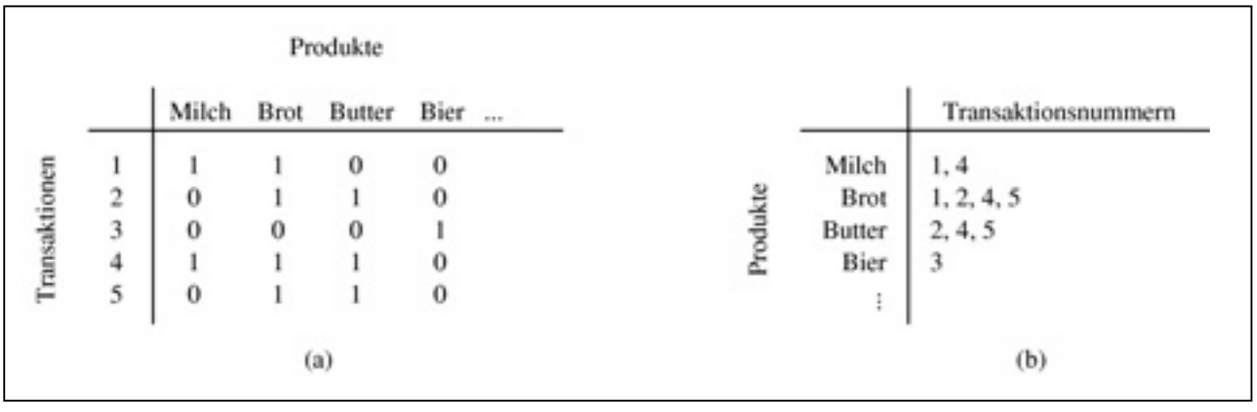


Abb. 3: Transaktionsdaten repräsentiert als (a) Kaufinzidenzmatrix und (b) Transaktionsnummernliste

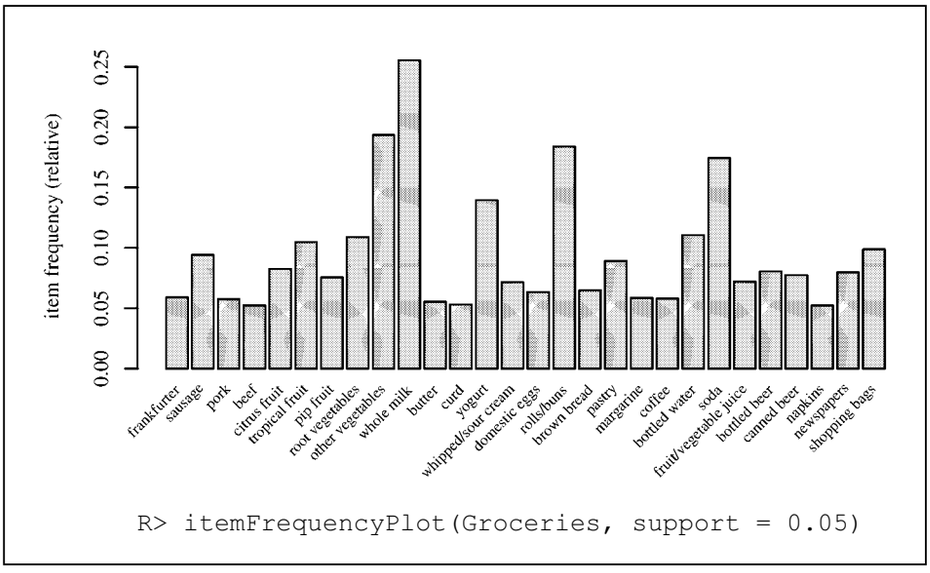


Abb. 4: Relative Kaufhäufigkeiten von Warengruppen (Minimum 5 %)

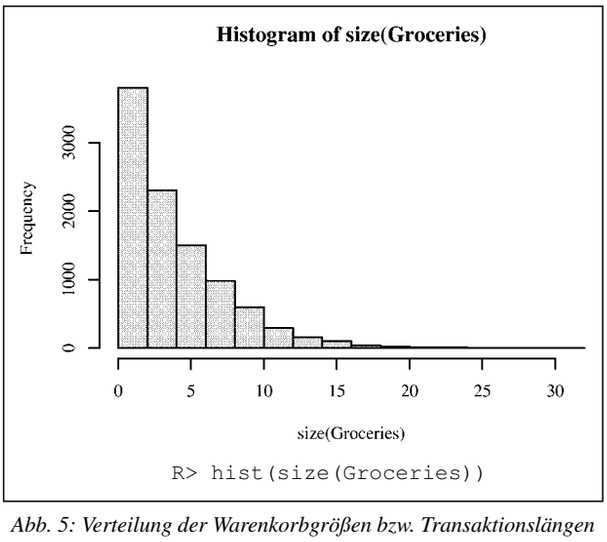


Abb. 5: Verteilung der Warenkorbgrößen bzw. Transaktionslängen

5. Konventionelle Affinitätsanalyse auf Basis paarweiser Assoziationsmaße

Die insbesondere von den Proponenten der früheren deutschsprachigen Literatur (*Böcker* 1978; *Merkle* 1981) vorgestellten Ansätze ermitteln auf Basis der Kaufinzidenzen zunächst eine zweidimensionale *Frequenzmatrix*

$$\boldsymbol{X}^T \cdot \boldsymbol{X} = C = [c_{ij}], \tag{1}$$

welche die absoluten gemeinsamen Kaufhäufigkeiten für jedes Warengruppenpaar $i,j = 1,...,J$ enthält (vgl. *Hruschka* 1991). Diese Matrix **C** stellt den Ausgangspunkt für die nachfolgende Verbundanalyse dar und lässt sich bequem mit Hilfe der Anweisung `crossTable(Groceries)` erzeugen[6]. Aus Übersichtlichkeitsgründen werden in *Tab. 2* nur die ersten 5 Warengruppen dargestellt. Die beiden Dreieckshälften der symmetrischen $J \times J$ Matrix **C** beinhalten jeweils die absoluten (Co-) Häufigkeiten, mit denen die Warengruppen i und j gemeinsam gekauft wurden; die Diagonalelemente c_{ii} enthalten die Anzahl der Transaktionen, in denen die Warengruppen i vorkommen.

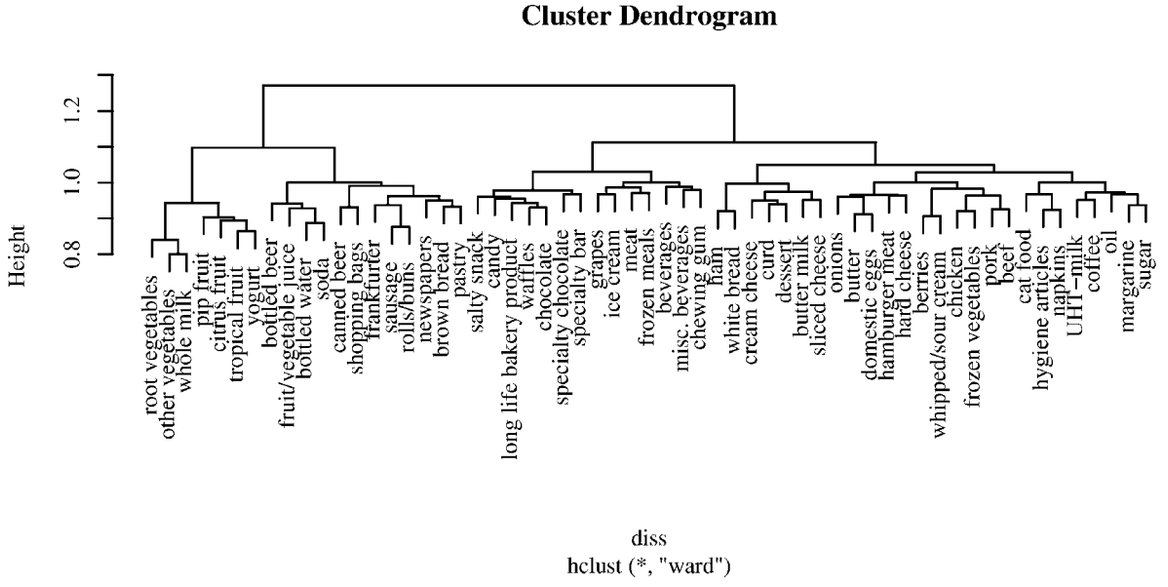
	frankfurter	sausage	liver loaf	ham	meat	…
frankfurter	580	99	7	25	32	…
sausage	99	924	10	49	52	…
liver loaf	7	10	50	3	0	…
ham	25	49	3	256	9	…
meat	32	52	0	9	254	…
...	…	…	…	…	…	…

Tab. 2: Beispiel einer Frequenzmatrix (Ausschnitt)

Zur Messung der Kaufverbundenheit zwischen allen möglichen Warengruppenpaaren können nun geeignete Assoziationskoeffizienten angewendet und die Frequenzmatrix **C** in eine sog. „Verbundmatrix" $\mathbf{D} = [d_{ij}]$ überführt werden. Da diese mit zunehmender Warengruppenanzahl J rasch sehr unübersichtlich wird und sich daher einer unmittelbaren Verwertung seitens des Marketing-Managements entzieht, ist regelmäßig eine angemessene Verdichtung und Visualisierung von Interesse. Traditionell gelangen hierfür diverse Projektionsmethoden wie Verfahren der mehrdimensionale Skalierung (MDS) oder hierarchische Clusteranalysemethoden zum Einsatz (vgl. *Merkle* 1979; *Bordemann* 1985; *Decker/Schimmelpfennig* 2002). Die Visualisierung der Verbundbeziehungen erfolgt 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 „klassischen" explorativen Verbundanalyse bewährte Assoziationskoeffizienten für binäre Transaktionsdaten findet man bei *Böcker* (1978) oder *Hruschka* (1991). Unter den als besonders geeignet geltenden Koeffizienten findet sich der *Tanimoto-Koeffizient*, dessen Pendant als Unähnlichkeitsmaß der *Jaccard-Koeffizient* (*Sneath* 1957) darstellt:

$$d_{ij} = 1 - \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}}, \quad \forall\ i,j = 1,...,J \tag{2}$$




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

Abb. 6: Verdichtung der Verbundmatrix mittels hierarchischer Clusteranalyse und Darstellung der Lösung als Dendrogramm

Der als Subtrahend in (2) direkt aus der Frequenzmatrix ableitbare *Tanimoto-Koeffizient* ist als ein asymmetrisches Ähnlichkeitsmaß definiert, das gegenüber übereinstimmenden Nullen (also gemeinsamen Nichtkäufen) unsensitiv ist. Diese Eigenschaft verhindert, dass sehr selten frequentierte – und daher in vielen Transaktionen übereinstimmende Nichtkäufe aufweisende – Warengruppen als miteinander stärker verbunden eingestuft werden, als häufiger in Kombination nachgefragte Warengruppen. Im Kontext der Verbundanalyse spricht man in diesem Zusammenhang auch vom Problem der „negativen Verbundenheit“ (vgl. ausführlicher dazu *Böcker* 1978 oder *Bordemann* 1985).

Im vorliegenden Illustrationsbeispiel wird für die Verdichtung der – wie soeben besprochenen unter Anwendung des Jaccard-Koeffizienten erzeugten – Verbundmatrix **D** gemäß *Abb. 6* ein nicht nur in der Verbundanalyse durchaus gebräuchliches hierarchisches Clusterverfahren (vgl. *Schnedlitz/Kleinberg* 1994), nämlich die Minimum-Varianz-Methode nach *Ward*, verwendet. Um eine übersichtliche Darstellung der Baumstruktur mit den Bezeichnungen der Warengruppen zu ermöglichen, werden dabei nur Warengruppen verwendet, die in mindestens 2 % aller Transaktionen vorkommen.

Die gefundene typologische Kaufverbundenheitsstruktur zwischen den Warengruppen ist in *Abb. 6* in Form eines Dendrogramms dargestellt, wobei die einzelnen Warengruppen die Endpunkte (Blätter) des Baumes repräsentieren. Je weiter man sich entlang der Äste in Richtung Blattspitzen bewegt, desto stärker ist der zwischen den beteiligten Warengruppen gemessene Kaufverbund. Die einzelnen Verästelungen der Baumstruktur können nun verwendet werden, um Warengruppen-Cluster mit intern höherer Kaufverbundenheit zu identifizieren. Beispielsweise kann eine Gruppe von Teilen des Frischesortiments bestehend aus Gemüse, Früchten, Milch und Joghurt („vegetables“, „fruits“, „milk“, „yogurt“) am linken Rand des Dendrogramms in *Abb. 6* identifiziert werden. Die absatzpolitische Relevanz solcher auf faktischem Kaufverhalten begründeten hierarchischen Warengruppentypologien wird primär in der Nutzung für Verbundplatzierungen im Verkaufsraum oder Werbemedien (z. B. Flugblätter, Kataloge) sowie in der nachfragerorientierten Bildung sog. „Categories“ im Rahmen des Category-Management-Prozesses gesehen (vgl. dazu etwa *Bordemann* 1985; *Zielke* 2002; *Müller-Hagedorn* 2005).

Der traditionellen explorativen Verbundanalyse werden folgende Haupteinwände entgegengehalten: (1) Die Einschränkung der Analyse auf Verbundrelationen zwischen Paaren von Warengruppen (vgl. z. B. *Hruschka* 1991) und (2) die bei Konstruktion der Verbundmatrix **D** vorgenommene *A-Priori-Aggregation* der Transaktionsdaten (vgl. z. B. *Schnedlitz/Reutterer/Joos* 2001). Gelegentlich ist es aber erwünscht, den Sortimentsverbund zwischen mehr als zwei Warengruppen und/oder auf disaggregiertem Niveau zu untersuchen. Man denke etwa an das Studium der Dynamik von Verbundbeziehungen (z. B. über

Tageszeiten, Wochentagen, Saisonen, usw. hinweg) oder die im Rahmen von CRM-Programmen besonders interessierende Identifikation von Kundensegmenten, die durch ähnliche Verbundbeziehungen gekennzeichnet sind. Wie sich anhand der nachfolgend vorgestellten Ansätze zeigt, können die Defizite der herkömmlichen Verbundanalyse mit Hilfe geeigneter Data-Mining-Techniken vermieden werden.

6. Verdichtung von Transaktionsdaten zu Prototypen

Die oben kritisierte Konstruktion einer homogenen Verbundmatrix für den gesamten (gepoolten) Transaktionsdatensatz wird bei der prototypenbasierten Verbundanalyse durch eine disaggregierte bzw. segmentspezifische Betrachtungsweise des Phänomens Sortimentsverbund ersetzt (vgl. *Schnedlitz/Reutterer/Joos* 2001; *Decker/Monien* 2003a). In der gegenwärtigen Marketing-Praxis ist dies insbesondere dann von Interesse, wenn personalisierte Transaktionshistorien (z. B. bei Loyalitätsprogrammen in Kombination mit elektronisch lesbaren Kundenkarten) verfügbar sind und *segmentspezifisch disaggregierte Verbundanalysen* für eine effektivere und effizientere Kundenansprache genutzt werden sollen. Von zentraler Bedeutung ist dabei eine Menge sog. *Prototypen*, die jeweils eine bestimmte Klasse von Warenkörben mit intern besonders prägnanter Kaufverbundstruktur repräsentieren sollen.

Obwohl für die Verdichtung der binären Kaufinzidenzen zu Warenkorb-Prototypen grundsätzlich eine Vielzahl partitionierender Verfahren der Clusteranalyse in Frage kommt, wird angesichts der Datenmengen gelegentlich für den Einsatz adaptiver Methoden plädiert. *Schnedlitz et al.* (2001), *Decker/Monien* (2003a) sowie *Decker* (2005) schlagen hierfür die Verwendung von Verfahren der Vektorquantisierung oder geeignete neurale Netzwerkmethoden vor. Als zumeist auf die Besonderheiten der Warenkorbanalyse entsprechend adaptierte Online-Variationen des bekannten *K*-Means-Alorithmus (*Mac Queen* 1967) ist diesen Methoden gemeinsam, dass sie durch Lösung des sog. Principal-Point-Problems die inneren Varianzen einer Partition $C = \{C_1,...,C_K\}$ zu minimieren versuchen (vgl. *Bock* 1999). Unter stationären Optimalitätsbedingungen entsprechen dabei die mittels stochastischer Approximation optimierten Prototypen $P = (p_1,..., p_K)$ mit $p_k \in \mathbb{R}^J \forall k$ den Zentroiden (Clustermittelpunkten) der durch sie erzeugten *K*-Partition und werden dann zur näheren inhaltlichen Charakterisierung der prototypischen Warenkorbcluster herangezogen.

Als Alternative dazu verwenden wir im vorliegenden Anwendungsbeispiel ein Verfahren zur Lösung des folgenden *K*-Medoid-Problems (*Hastie/Tibshirani/Friedman* 2001, S. 468 f.):

$$\sum_{k=1}^{K} \sum_{n \in C_k} d(x_n, m_k) \rightarrow \min_{C, \{m_k\}_1^K}, \tag{3}$$

wobei als Medoid m_k jener Warenkorb verstanden wird, dessen mittlerer Abstand zu allen anderen Transaktionen im selben Cluster C_k minimal ist. In Kombination mit der Minimalvorschrift (3) wird für den Medoiden eines Warenkorbclusters also folgende Eigenschaft gefordert:

$$m_k \in \{x_n\}_{n \in C_k} \forall k \tag{4}$$

Der grundlegende methodologische Unterschied zwischen Principal-Point- und *K*-Medoid-Partitionen ergibt sich aus der Restriktion (4). Während ein *K*-Means-Zentroid als hypothetischer Durchschnittswarenkorb interpretiert werden kann, stammt ein *K*-Medoid-Prototyp m_k als realer Warenkorb aus der Menge der von ihm repräsentierten Cluster C_k. Durch diese Eigenschaft des Prototypensystems haben sich *K*-Medoid-Verfahren unter anderem auch in früheren Marketing-Anwendungen (vgl. *Larson/Bradlow/Fader* 2005) als vergleichsweise robust gegenüber Ausreißern erwiesen. Vor diesem Hintergrund erscheint die Verwendung eines *K*-Medoid-basierten Verfahrens für die Warenkorbanalyse gerechtfertigt.

Im nachfolgenden Analysebeispiel verwenden wir eine bekannte iterative, relokationsbasierte Heuristik zur Lösung des *K*-Medoid-Problems (3), die von *Kaufman/Rousseeuw* (2005) unter der Bezeichnung *Partitioning Around Medoids* (PAM) vorgeschlagen wurde. Der Umstand, dass PAM die Konstruktion einer $N \times N$ Distanzmatrix erfordert, lässt das Verfahren für eine Anwendung auf größere Datenumfänge zunächst nur bedingt tauglich erscheinen. Dieser Einschränkung kann allerdings durch eine Zufallsauswahl aus dem zu analysierenden Transaktionsdatenbestand oder diverse Resampling-Ansätze begegnet werden. Beispiele für effiziente Suchstrategien nach fehlerminimalen Medoiden in großen Datensätzen sind etwa die Verfahren CLARA (Clustering LARge Applications; vgl. *Kaufman/Rousseeuw* 2005) oder CLARANS (Clustering Large Applications based upon RANdomized Search; vgl. *Ng/Han* 2002).

Da wir an einer Abbildung von Kaufverbundstrukturen interessiert sind, werden für die *K*-Medoid-Partitionierung des Groceries-Datensatzes nur Transaktionen verwendet, die mindestens zwei verschiedene Warengruppen beinhalten. Weiters werden die als eigene Warengruppe ausgewiesenen Einkaufstüten („shopping bags“) aus der Analyse ausgeschlossen, da diese bei Beendigung des Einkaufsvorganges am Kassenterminal als Transportbehelf sehr häufig mitgekauft werden[7]. Nach dieser Vorauswahl verbleiben 7676 Transaktionen und 168 Warengruppen für die Analyse. Um die Größe der in PAM verwendeten Unähnlichkeitsmatrix zu reduzieren, werden zufällig 2000 Transaktionen ausgewählt. Die Unähnlichkeitsmatrix wird wiederum unter Anwendung des bewährten *Jaccard*-Koeffizienten bestimmt und mittels PAM Clusterlösungen für eine Sequenz von $K = 1,...,8$ Cluster generiert (siehe *Abb. 7* für eine Darstellung dieser Vorgehensweise als R Anweisungen).

Für die Auswahl einer „geeigneten“ Clusteranzahl steht eine reichhaltige Palette an internen Validitätsmaßen zur

```
R> groc <- Groceries[size(Groceries)>1,
      which(itemLabels(Groceries) != "shopping bags")]
R> samp <- sample(groc, 2000)
R> diss <- dissimilarity(samp, method = "Jaccard")

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

Abb. 7: Beispiel zur Verdichtung von Transaktionsdaten zu K-Medoid-Prototypen

Auswahl (einen Überblick dazu findet man bei *Milligan/Cooper* 1985), die großteils auch in R verfügbar sind. Ein solches Maß für die Beurteilung der Qualität von Partitionen, welches für ein breites Spektrum an unterschiedlichen Distanzmaßen (u. a. auch für die hier verwendeten *Jaccard*-Distanzen) angewendet werden kann, stellt der von *Rousseeuw* (1987) vorgeschlagene *Silhouettenkoeffizient* dar. Vereinfacht ausgedrückt quantifiziert dieser die Diskrepanz zwischen den durchschnittlichen Unähnlichkeiten der Datenpunkte innerhalb eines Clusters und den nächstgelegenen Datenpunkten des jeweils benachbarten Clusters. Der Silhouettenkoeffizient kann daher für die Beurteilung der Trennschärfe eine Clusterlösung herangezogen werden (vgl. ausführlicher dazu bei *Kaufman/Rousseeuw* 2005, S. 83 ff.).

Unter Zugrundelegung dieser Heuristik wird für das vorliegende Set von *K*-Medoid-Clusterlösungen eine Clusteranzahl von $K = 5$ empfohlen. Ausgewählte Charakteristika dieser Lösung werden in *Tab. 3* dargestellt. Neben der den einzelnen Clustern zugewiesenen Anzahl an Transaktionen („size“) und der relativen Clustergröße („relative size“) erkennt man, dass sowohl die maximale als auch die durchschnittliche Unähnlichkeit der Datenpunkte innerhalb der Cluster durchwegs sehr hoch sind. Starke klasseninterne Streuungen sind allerdings für hochdimensionale Daten ($J = 168$) nicht ungewöhnlich und deuten nicht unbedingt auf eine schlechte Qualität der Clusterlösung hin. Die Trennschärfe („separation“) einer Clusterlösung ist eine Maßzahl für den Grad der Separiertheit einer Clusterlösung und als geringste Unähnlichkeit zwischen zwei Objekten aus verschiedenen, aber räumlich benachbarten Cluster definiert. Wie aus *Tab. 3* ersichtlich, sind die einzelnen Cluster mit Ausnahme von Cluster Nr. 1 gleichermaßen voneinander separiert.

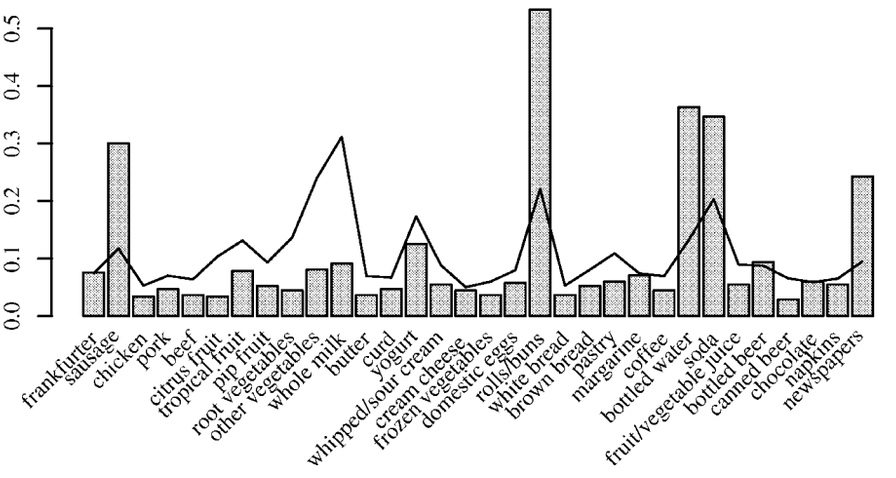
Für eine nähere inhaltliche Beschreibung der einzelnen Warenkorbcluster können nun die *Prototypenprofile* herangezogen werden. Mit Hilfe der Anweisung `itemFrequencyPlot` sind in *Abb. 8* exemplarisch die Prototypenprofile für zwei ausgewählte Cluster mit den Nummerierungen 2 und 5 wiedergegeben. Aus Darstellungsgründen werden dabei nur Warengruppen herangezogen, die in mindestens 5 % der Transaktionen vorkommen. Die Linie kennzeichnet jeweils die relative Kauffrequenzverteilung im gesamten Datensatz und die Balken die Verteilung der warengruppenspezifischen Kaufrequenzen innerhalb der Cluster. Da die Warenkörbe als Binärdaten codiert sind, entsprechen die Prototypenprofile den bedingten Kaufwahrscheinlichkeiten der Warengruppen innerhalb eines Clusters.

Für die cluster- bzw. segmentspezifische Verbundanalyse sind nun insbesondere Warengruppenkombinationen mit starken Abweichungen der bedingten (Balkendiagramm) vom Profil der unbedingten Kaufwahrscheinlichkeiten (Linienplot) interessant. Starke positive (negative) Abweichungen signalisieren komplementäre (substitutive) Verbundbeziehungen zwischen den betreffenden Warengruppen. So stellt Cluster 2 in *Abb. 8* (a) beispielsweise ein prototypisches Einkaufsmuster mit besonders hohen Anteilen in den Warengruppen Wurst, Brötchen, Getränke („bottled water“, „rolls/buns“, „soda“) sowie Zeitungen („newspapers“) dar. Cluster 5 repräsentiert hingegen ein typisches Getränkecluster mit hohen Kaufanteilen in den Warengruppen Limonade („soda“) und Dosenbier („canned beer“). Wenn auch in den beiden betrachteten

Cluster	Size	Relative size	Maximium Dissimilarity	Average Dissimilarity	Separation
1	513	0.257	0.9474	0.8113	0.2857
2	383	0.192	0.9286	0.7700	0.2500
3	578	0.289	1.0000	0.8427	0.2500
4	285	0.143	0.9091	0.7294	0.2500
5	241	0.121	0.9091	0.7146	0.2500

Tab. 3: Ausgewählte Charakteristika einer Clusterlösung mit K = 5

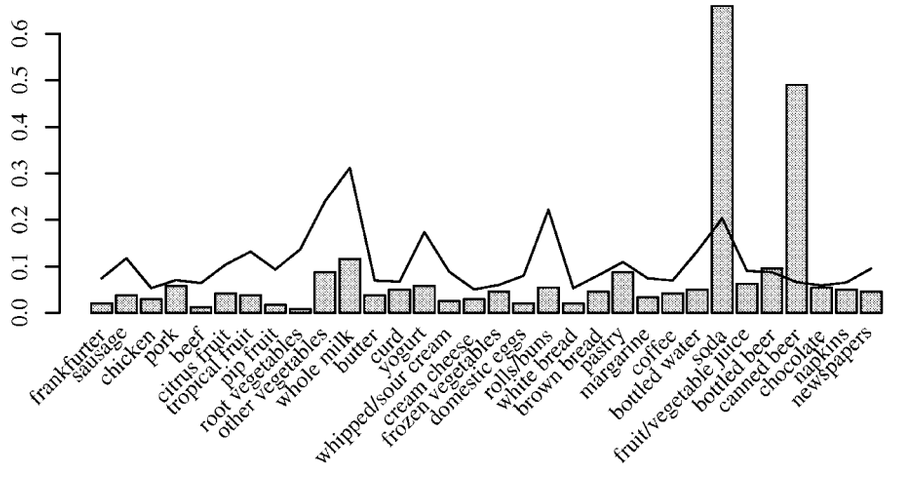
Medoid: {sausage, rolls/buns, bottled water, soda, newspapers}




```
R> itemFrequencyPlot(samp[clust[[5]]$clustering == 2],
         population = groc, support = 0.05)
R> inspect(samp[clust[[5]]$medoids[2]]
```

(a) Cluster 2

Medoid: {soda, canned beer}




```
R> itemFrequencyPlot(samp[clust[[5]]$clustering == 5],
         population = groc, support = 0.05)
R> inspect(samp[clust[[5]]$medoids[5]]
```

(b) Cluster 5

Abb. 8: Prototypenprofile und Medoide segmentspezifischer Transaktionen

Cluster unterschiedlich stark ausgeprägt, können im Vergleich zur aggregierten Sichtweise substitutive Verbundrelationen insbesondere für die diversen Früchte- und Gemüsewarengruppen („fruits“, „vegetables“) sowie Milch („whole milk“) erwartet werden.

Neben den Prototypenprofilen können mit Hilfe der Anweisung `inspect` auch die in den Medoid-Warenkörben der einzelnen Cluster enthaltenen Warengruppen näher inspiziert werden. Wie aus der Illustration in *Abb. 8* klar ersichtlich, beinhalten diese deutlich überdurchschnittlich häufig nachgefragten Warengruppen. Als prototypische Repräsentanten der von ihnen abgebildeten Warenkorbcluster signalisieren die Medoide also symptomatisch wiederkehrende Muster stark komplementärer Verbundrelationen. Diese Eigenschaft ist allerdings weniger dem *K*-Medoid-Verfahren selbst, sondern vielmehr dem auf der Verwendung von *Jaccard*-Koeffizienten basierenden Distanzkonzept der Clusterbildung zuzuschreiben.

Die prototypenbasierte Verbundanalyse liefert also Warenkorbcluster mit intern prägnanter Kaufverbundstruktur und erlaubt daher eine Verbundanalyse auf beliebig disaggregiertem (segmentspezifischem) Niveau. Eine Erweiterung dieses Konzepts in Richtung dynamische Segmentierung von Kundendatenbanken sowie ein Anwendungsbeispiel in der Direktmarketing-Praxis findet man bei *Reutterer et al.* (2006). Häufig erscheint eine derartige auf realen Transaktionsdaten basierende Kundensegmentierung in Kombination mit (meist vorgeschalteten) konventionelleren Segmentierungsanalysen, beispielsweise unter Verwendung von sozio-demographischen oder sog. RFM (Recency, Frequency, Monetary value) Kriterien, sinnvoll. *Malthouse* (2003) bezeichnet einen solchen mehrstufigen Segmentierungsansatz als *‚database subsegmentation‘*, wobei in diesem Zusammenhang als wohl prominentestes Beispiel aus der Handelspraxis auf das „*Tesco*-Clubcard-Programm“ hingewiesen werden kann (vgl. *Humby/Hunt* 2003, S. 143 ff.). Wie *Boztug/Reutterer* (2007) zeigen und diskutieren, 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.

7. Generierung von bedeutsamen Assoziationsregeln

Neuere aus der Data-Mining-Literatur stammende Ansätze zur Warenkorbanalyse zielen ebenfalls auf die Analyse der gemeinsamen Kaufhäufigkeiten für eine (typischerweise sehr große) Auswahl von Warengruppen oder einzelne Artikel ab. Es handelt sich hierbei um Methoden zur Konstruktion und Beurteilung sog. Assoziationsregeln (*Association Rules*; vgl. *Agrawal/Imielinski/Swami* 1993; *Agrawal/Srikant* 1994), die über die kritisierte Einschränkung der Affinitätsanalyse auf paarweise Verbundbeziehungen hinaus gehen und in den beobachteten Transaktionsdaten verborgene Interdependenzstrukturen über einen probabilistischen Messansatz in Form von Regeln zwischen beliebigen Mengen von Artikeln bzw. Warengruppen abbilden. Unter einer solcher Regel wird eine Implikation der Art $\{Brot, Milch\} \Rightarrow \{Butter\}$ verstanden, deren linke Seite (LHS) als Rumpf, Prämisse oder Antezedent und deren rechte Seite (RHS) auch als Kopf oder Konklusion bezeichnet wird.

Für die Beurteilung einer Assoziationsregel $A \Rightarrow B$ von elementarer Bedeutung sind das Signifikanzmaß *Support* und das Qualitätsmaß *Konfidenz*, die wie folgt definiert sind (*Agrawal et al.* 1993):

$$\text{supp}(A \Rightarrow B) = \frac{n_{A \cup B}}{N}, \tag{5}$$

$$\text{conf}(A \Rightarrow B) = \frac{n_{A \cup B}}{n_A} = \frac{\text{supp}(A \Rightarrow B)}{\text{supp}(A)}, \tag{6}$$

wobei N die Anzahl aller Transaktionen, n_A die Anzahl jener Transaktionen, die alle Warengruppen in A beinhalten, und $n_{A \cup B}$ die Anzahl der Transaktionen, welche alle Warengruppen in A und B beinhalten, bezeichnet (es gilt also $n_{A \cup B} \subset n_A \subset N$). Als relative Häufigkeit, mit der ein sog. *Itemset* $\{A,B\}$ beobachtet werden kann, handelt es sich beim Support um ein symmetrisches Verbundmaß. Dem gegenüber entspricht die Konfidenz den bedingten Kaufwahrscheinlichkeiten und ist folglich ein asymmetrisches Verbundmaß, d.h. $\text{conf}(A \Rightarrow B)$ muss nicht gleich $\text{conf}(B \Rightarrow A)$ sein.

Aus einer Vielzahl weiterer Qualitätsmaße ist ein in der einschlägigen Literatur als *Lift* bezeichnetes Maß (vgl. *Brin et al.* 1997) für die Verbundanalyse von besonderem Interesse:

$$\text{lift}(A \Rightarrow B) = \frac{n_{A \cup B}}{n_A n_B} = \frac{\text{conf}(A \Rightarrow B)}{\text{supp}(B)} \tag{7}$$

Das Lift-Maß gibt die Abweichung der gemeinsamen Kaufhäufigkeit der linken und der rechten Seite einer Regel von der unter Annahme stochastischer Unabhängigkeit erwarteten Vorkommenshäufigkeit (gegeben durch das Produkt der Häufigkeiten $n_A n_B$ im Nenner von (7)) an[8]. 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 bedeutsamer Assoziationsregeln erweist sich die mit zunehmender Sortimentsgröße explodierende Menge aller möglichen Itemsets (einschließlich der daraus ableitbaren Regeln). Zur Bewältigung des damit einhergehenden Komplexitätsproblems wurden eine Reihe effizienter Suchstrategien vorgeschlagen. In den meisten Data-Mining-Systemen (wie auch in **arules**) ist eine Variante des populären *APRIORI*-Algorithms (*Agrawal/Srikant* 1994) implementiert, der für ein vorgegebenes minimales Support- und Konfidenzkriterium alle zulässigen Regeln findet.

```
R> rules <- apriori(Groceries, parameter = list(support = 0.001,
       confidence = 0.2), control = list(verbose = FALSE))
R> inspect(SORT(rules, by = "lift")[1:3])
R> rulesBeef <- subset(rules, rhs %in% "beef")
R> inspect(SORT(rulesBeef, by = "conf")[1:3])
```

Abb. 9: Beispiel zur Suche und anschließenden Selektion von Assoziationsregeln

#	LHS		RHS	Support	Confidence	Lift
1	{ bottled beer, red/blush wine }	⇒	{ spirits }	0.0019	0.40	35.72
2	{ hamburger meat, soda }	⇒	{ instant food products }	0.0012	0.21	26.21
3	{ ham, white bread }	⇒	{ processed cheese }	0.0019	0.38	22.93
…	…		…	…	…	…

Tab. 4: Sortierung der Regeln nach absteigendem Lift-Maß

#	LHS		RHS	Support	Confidence	Lift
1	{ root vegetables, whole milk, butter, rolls/buns }	⇒	{ beef }	0.0011	0.48	9.12
2	{ sausage, root vegetables, butter }	⇒	{ beef }	0.0010	0.45	8.66
3	{ root vegetables, butter, yogurt }	⇒	{ beef }	0.0015	0.34	7.52
…	…		…	…	…	…

Tab. 5: Sortierung der Regeln nach absteigendem Konfidenz-Maß

Für den Groceries-Datensatz werden mit der ersten Anweisung in *Abb. 9* alle möglichen Assoziationsregeln mit minimalem Support von 0,1 % und einer Konfidenz von mindestens 20% generiert. Mit diesen Minimalvorgaben werden 21.574 (!) Regeln gefunden. Im Durchschnitt enthalten die den gefundenen Regeln zugrunde liegenden Itemsets 3,593 unterschiedliche Warengruppen.

Für eine nähere Inspektion der Regelliste werden eine Reihe von Selektions- und Auswahloperationen benötigt, welche die Aufdeckung besonders „interessanter" Regeln erlauben (siehe dazu exemplarisch in *Abb. 9*). Beispielsweise können die zuvor gefundenen Regeln nach absteigenden Werten des Lift-Maßes sortiert werden. In *Tab. 4* werden die drei Regeln mit den höchsten Werten dargestellt. Das Ergebnis liefert erwartungsgemäß sehr plausible Beziehungen zwischen typisch verwendungsverbundenen Warengruppen wie beispielsweise Schinken, Weißbrot und Käse (Regel 3 in *Tab. 4*). Die Interpretation des Lift-Maßes impliziert, dass falls Schinken („ham") und Weißbrot („white bread") im Warenkorb enthalten sind, Käse („processed cheese") um einen Faktor von 22,93 Mal häufiger nachgefragt wird, als im gesamten Datensatz (vgl. *Decker/Schimmelpfennig* 2002). Analoges gilt für die anderen Regeln.

Eine weitere Selektionsmöglichkeit besteht etwa in der Einschränkung auf Regeln, die ganz bestimmte Warengruppen beinhalten. Interessieren beispielsweise nur jene Regeln, die auf den Kauf von Rindfleisch hinweisen, können Regeln mit Rindfleisch („beef") im Regelkopf selektiert werden. Die drei Regeln mit der höchsten Konfidenz werden in *Tab. 5* dargestellt. Zwecks Ableitung von Marketing-Maßnahmen kann auf diese Art und Weise bequem die Menge jener Warengruppen ermittelt werden, welche die bedeutsamsten Verbundbeziehungen zur im jeweiligen Planungsfokus stehenden Warengruppe aufweisen.

Mit Hilfe effizienter Suchalgorithmen aus dem Bereich des Data Mining wird das Finden von Assoziationsregeln auch für sehr große Datenmengen und umfangreiche Sortimente selbst auf Artikelebene problemlos möglich und kann zur Entscheidungsunterstützung in den Berei-

chen Sortiments-, Platzierungs- und Werbeplanung hilfreich sein (vgl. *Brijs et al.* 2004, *Van den Poel/De Schamphelaere/Wets* 2004). Allerdings muss auch auf Probleme mit den Maßen Support, Konfidenz und Lift hingewiesen werden. Wie die Experimente von *Hahsler/Hornik/Reutterer* (2006) zeigen, werden Support, Konfidenz und auch Lift systematisch durch die Kaufhäufigkeiten der Produkte beeinflusst, was zu Problemen beim Vergleich von Regeln anhand der Maße führen kann.

8. Schlussfolgerungen

Im vorliegenden Beitrag wurden Einsatzfelder und -potenziale des Einsatzes von Data-Mining-Techniken im Marketing-Bereich diskutiert und anhand der explorativen Analyse von Warenkorbdaten näher vorgeführt. Ein Hauptvorzug von datengetriebenen Analysezutritten der „algorithmischen Bauart" (verstanden als Methoden des Data Mining i.e.S.) besteht in deren Fähigkeit, für große Datenmengen weitestgehend automatisiert robuste Ergebnisse mit vergleichsweise moderatem Analyseaufwand liefern zu können. Diese Eigenschaft erweist sich insbesondere dann als hilfreich, wenn wenig über die dem Datengenerierungsprozess zugrunde liegenden Verteilungsfunktionen bekannt ist oder komplexe nichtlineare Zusammenhangsstrukturen in unübersichtlichen Massendaten untersucht werden sollen. Umgekehrt ist der Einsatz von Data-Mining-Methoden (i.e.S.) weniger geeignet, wenn es darum geht, die in den Daten beobachteten Muster und Phänomene auch strukturell zu erklären.

Wie am Beispiel der explorativen Warenkorbanalyse gezeigt werden konnte, sind bestimmte Data-Mining-Techniken durchaus in der Lage, die von traditionelleren Zutritten auferlegten Restriktionen und Einsatzbeschränkungen zu mildern bzw. komplett zu beseitigen, womit sich auch bis dato meist ungenutzte Verwertungsmöglichkeiten im Rahmen moderner CRM- und Marketing-Managementkonzepte eröffnen. Zusammenfassend dürften daher insbesondere in Kombination mit statistischen Modellierungszutritten genutzte Data-Mining-Methoden nützliche Erweiterungen des konventionellen Methodenvorrats darstellen.

Anmerkungen

[1] Siehe http://www.kdnuggets.com/polls/.

[2] R ist eine frei verfügbare Umgebung für Datenanalyse und Grafik. 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.

[3] An dieser Stelle sei auf das auf dem GNOME-Standard basierende Data-Mining-Werkzeug RATTLE („R Analytical Tool To Learn Easily") verwiesen: http://rattle.togaware.com/

[4] Wir schließen uns damit der Konvention an, die numerische Repräsentation der Transaktionen als Realisationen sog. ‚Pick-Any-Daten' aufzufassen (vgl. *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 bzw. Artikeln.

[5] Für das Einlesen eigener Datensätze steht in **arules** die Funktion `read.transactions()` zur Verfügung. Diese Funktion kann Daten in verschiedenen Formaten bequem 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.

[6] Um eine überproportionale Gewichtung ‚längerer' Transaktionen zu vermeiden, werden die in die Frequenzmatrix eingezählten Warenkörbe gelegentlich mit der Inversen ihrer Warenkorbgrößen gewichtet (vgl. ausführlicher dazu *Merkle* 1981, S. 47 ff.; *Hruschka* 1991).

[7] Stellvertretend für eine Reihe weiterer sehr häufig mitverkaufter, aber nicht das Kerngeschäft des Händlers ausmachende Produkte weisen solche Emballagen auch starke „Mitnahmeeffekte" auf das eigentliche Warensortiment aus, was sich bei der Clusterlösung auch dementsprechend bemerkbar machen würde. Derartige Effekte sind für das Handelsmanagement in der Regel von untergeordneter Bedeutung und liefern daher auch keine besonders substantiellen Erkenntnisse mit Hinblick auf die Identifizierung bedeutsamer Verbundbeziehungen im angebotenen Warensortiment.

[8] Vor dem Hintergrund der Beuteilung von komplementären bzw. substitutionalen Verbundeffekten gelangt diese Annahme auch bereits in der „klassischen" Affinitätsanalyse zur Anwendung (vgl. dazu bereits bei *Böcker* 1981, S. 20 f. und S. 80 f. oder *Hruschka* 1991).

Literaturverzeichnis

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., S. 207–216.

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

Baesens B./Viaene, S./Van den Poel, D./Vanthienen, J./Dedene, G. (2002): Bayesian Neural Network Learning for Repeat Purchase Modelling in Direct Marketing, in: European Journal of Operational Research, Vol. 138, No. 1, S. 191–211.

Balakrishnan, P./Sundar, V./Cooper, M.C./Varghese S.J./Lewis, P.A. (1996): Comparative Performance of the FSCL Neural Net and K-Means Algorithm for Market Segmentation, in: European Journal of Operational Research, No. 93, S. 346–357.

Berry, M./Linoff, G. (1997): Data Mining Techniques for Marketing, Sales and Customer Support, New York.

Blattberg, R.C./Glazer, R./Little, J.D.C. (1994): The Marketing Information Revolution. Boston.

Bock, H.-H. (1999): Clustering and Neural Network Approaches, in: Gaul, W./Locarek-Junge, H. (Eds.): Studies in Classification, Data Analysis, and Knowledge Organization: Classification in the Information Age, Berlin, S. 42–57.

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

Bordemannn, H.-G. (1985): Analyse von Verbundbeziehungen zwischen Sortimentsteilen im Einzelhandel, Duisburg.

Borgelt, Ch. (2003): Efficient Implementations of Apriori and Eclat, in: Goethals, B./Zaki, M.J. (Eds.): Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Melbourne, Florida.

Boztug, Y./Hildebrandt, L. (2006): A Market Basket Analysis Conducted with a Multivariate Logit Model, in: Gaul, W./Kruse, R. (Eds.): Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, S. 558–565.

Boztug, Y./Reutterer, T. (2007): A Combined Approach for Segment-Specific Analysis of Market Basket Data, in: European Journal of Operational Research (forthcoming).

Boztug, Y./Silberhorn, N. (2006): Modellierungsansätze in der Warenkorbanalyse im Überblick, in: Journal für Betriebswirtschaft, 56. Jg., Nr. 2, S. 105–128.

Brachman, R.J./Anand, T. (1996): The Process of Knowledge Discovery in Databases, in: Fayyad, U.M./Piatetsky-Shapiro, G./Smyth, P./Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining, Menlo Park, S. 37–57.

Breiman, L. (2001): Statistical Modeling: The Two Cultures, in: Statistical Science, Vol. 16, No. 3, S. 199–231.

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, No. 1, S. 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, Arizona, S. 255–264.

Buckinx, W./Van den Poel, D. (2005): Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Contractual FMCG Retail Setting, in: European Journal of Operational Research, Vol. 164, No. 1, S. 252–268.

Buckinx, W./Verstraeten, G./Van den Poel, D. (2006): Predicting Customer Loyalty Using the Internal Transactional Database, in: Expert Systems with Applications, Vol. 32, No. 1, S. 125–134.

Bucklin, R.E./Lehmann, D.R./Little, J.D.C. (1998): From Decision Support to Decision Automation: A 2020 Vision, in: Marketing Letters, Vol. 9, No. 3, S. 235–246.

Cooper, L.G./Giuffrida, G. (2000): Turning Datamining into a Management Science Tool: New Algorithms and Empirical Results, in: Management Science, Vol. 46, No. 2, S. 249–264.

Cui, D./Curry, D. (2005): Prediction in Marketing Using the Support Vector Machine, in: Marketing Science, Vol. 24, No. 4, S. 595–615.

Cui, G./Wong, M.L./Lui, H.-K. (2006): Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming, in: Management Science, Vol. 52, No. 4, S. 597–612.

Decker, R. (2005): Market Basket Analysis by Means of a Growing Neural Network, in: The International Review of Retail, Distribution and Consumer Research, Vol. 15, No. 2, S. 151–169.

Decker, R./Monien, K. (2003a): Market Basket Analysis With Neural Gas Networks and Self-Organising Maps, in: Journal of Targeting, Measurement and Analysis for Marketing, Vol. 11, S. 373–386.

Decker, R./Monien, K. (2003b): Support-Vektor-Maschinen als Analyseinstrument im Marketing am Beispiel der Neukundenklassifikation, in: der markt, 42. Jg. Nr. 164, S. 3–13.

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.

DeSarbo, W.S./Manrai, A.K./Manrai, L.A. (1993): Non-Spatial Tree Models for the Assessment of Competitive Market Structure: An Integrated Review of the Marketing and Psychometric Literature, in: Eliashberg, J./Lilien, G.L. (eds.): Handbooks in Operations Research and Management Science, Volume 5: Marketing, Amsterdam, S. 193–257.

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, No. 4, S. 351–379.

Elrod, T. (1991): Internal Market Structure Analysis: Recent Developments and Future Prospects, in: Marketing Letters, Vol. 2, August, S. 253–266.

Elsner, R./Krafft, M./Huchzermeier, A. (2004): Optimizing Rhenania's Direct Marketing Business Through Dynamic Multilevel Modeling (DMLM) in a Multicatalog-Brand Environment, in: Marketing Science, Vol. 23, No. 2, S. 192–206.

Fayyad, U.M./Piatetsky-Shapiro, G./Smyth P. (1996): From Data Mining To Knowledge Discovery: An Overview, in: Fayyad, U.M./Piatetsky-Shapiro, G./Smyth, P./Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining, Menlo Park, S. 1–34.

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, No. 15, S. 1–25.

Hahsler, M./Grün, B./Hornik, K. (2006): Arules: Mining Association Rules and Frequent Itemsets, R package version 0.5–0.

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, Berlin, S. 598–605.

Hand, D./Mannila, H./Smyth, P. (2001): Principles of Data Mining, Cambridge.

Hastie, T./Tibshirani, R./Friedman J. (2001): The Elements of Statistical Learning. Data Mining, Inference, and Prediction. New York.

Haughton, D./Deichmann, J./Eshghi, A./Sayek, S./Teebagy, N./Topi, H. (2003): A Review of Software Packages for Data Mining, in: The American Statistician, Vol. 57, No. 4, S. 290–309.

Herschel, G. (2006): Magic Quadrant for Customer Data Mining, 1Q06, Gardner RAS Core Research Note G00132466, Gardner, Inc.

Hildebrandt, L. (2000): 30 Jahre Forschung im deutschen Sprachraum zum quantitative orientierten Marketing – Koreferat zum Beitrag von Sönke Albers, in: Backhaus, K. (Hrsg.): Deutschsprachige Marketingforschung. Bestandsaufnahme und Perspektiven. Stuttgart, S. 239–248.

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, S. 189–220.

Hippner, U./Küsters/Meyer, M./Wilde, K.D. (2001): Handbuch Data Mining im Marketing. Knowledge Discovery in Marketing Databases, Wiesbaden.

Hornik, K./Stinchcombe, M./White, H. (1989): Multilayer Feedforward Networks are Universal Approximators, in: Neural Networks, Vol. 2, No. 5, S. 359–366.

Hruschka H. (1985): Der Zusammenhang zwischen Verbundbeziehungen und Kaufakt- bzw. Käuferstrukturmerkmalen, in: Zeitschrift für betriebswirtschaftliche Forschung, 37. Jg., Nr. 3, S. 218–231.

Hruschka, H. (1991): Bestimmung der Kaufverbundenheit mit Hilfe eines probabilistischen Meßmodells, in: Zeitschrift für betriebswirtschaftliche Forschung, 43. Jg., Nr. 5, S. 418–434.

Hruschka, H. (1993): Determining Market Response Functions by Neural Network Modeling: A Comparison to Econometric Techniques, in: European Journal of Operational Research, Vol. 66, S. 27–35.

Hruschka, H. (2001): An Artificial Neural Net Attraction Model (ANNAM) to Analyze Market Share Effects of Marketing Instruments, in: Schmalenbach Business Review, Vol. 53, No. 1, S. 27–40.

Hruschka, H./Fettes, W./Probst, M. (2004): An empirical Comparison of the Validity of a Neural Net Based Multinomial Logit Choice Model to Alternative Model Specifications, in: European Journal of Operational Research, Vol. 159, S. 166–180.

Hruschka, H./Fettes, W./Probst, M./Mies, C. (2002): A Flexible Brand Choice Model Based on Neural Net Methodology. A Comparison to the Linear Utility Multinomial Logit Model and its Latent Class Extension, in: OR Spectrum, Vol. 24, S. 127–143.

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

Hruschka, H./Natter, M. (1993): Analyse von Marktsegmenten mit Hilfe konnexionistischer Modelle, in: Zeitschrift für Betriebswirtschaft, 63. Jg., Nr. 5, S. 425–442.

Hruschka, H./Natter, M. (1995): Clusterorientierte Marktsegmentierung mit Hilfe Künstlicher Neuraler Netzwerke, in: Marketing ZFP, 17. Jg., Nr. 4, S. 249–254.

Humby, C./Hunt, T. (2003): Scoring Points. How Tesco is winning customer loyalty, London & Sterling.

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

Kaufman, L./Rousseeuw, P.J. (2005): Finding Groups in Data: An Introduction to Cluster Analysis, New York.

Kumar, A./Rao, V.R./Soni, H. (1995): An Empirical Comparison of Neural Network and Logistic Regression Models, in: Marketing Letters, Vol. 6, No. 4, S. 251–264.

Larson, Jeffrey S./Bradlow, Eric T./Fader, Peter S. (2005): An Exploratory Look at Supermarket Shopping Paths, in: International Journal of Research in Marketing, Vol. 22, No. 4, S. 395–414.

Leeflang, P.S.H./Wittink, D.R./Wedel, M./Naert, P.A. (2000): Building Models for Marketing Decisions, Boston.

Levin, N./Zahavi, J. (1998): Continuous Predictive Modeling: A Comparative Analysis, in: Journal of Interactive Marketing, Vol. 12, No. 2, S. 5–22.

Levin, N./Zahavi, J. (2001): Predictive Modeling Using Segmentation, in: Journal of Interactive Marketing, Vol. 15, No. 2, S. 2–22.

MacQueen, J. (1967): Some Methods for Classification and Analysis of Multivariate Observations, in: Le Cam, L.M., & Neyman, J. (eds.): Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, S. 281–297.

Malthouse, E.C. (2003): Database Sub-Segmentation, in: Iacobucci, D./Calder, B. (eds.): Kellogg on Integrated Marketing, New York, S. 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, No. 2, S. 95–114.

Matsatsinis, N. F./Siskos, Y. (2003): Intelligent Support Systems for Marketing Decisions, Boston.

Mazanec, J.A. (1999): Simultaneous Positioning and Segmentation Analysis with Topologically Ordered Feature Maps: A Tour Operator Example, in: Journal of Retailing and Consumer Services, Vol. 6, No. 4, S. 219–235.

Mazanec, J.A. (2000): Perceptual Market Structure and Strategy Formation, in: Mazanec, J./Strasser, H. (eds.): A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations. Berlin, S. 9–96.

McCann, J.M./Gallagher, J.P. (1990): Expert Systems for Scanner Data Environments, Boston.

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

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, S. 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, S. 159–179.

Müller-Hagedorn, L. (1978): Das Problem des Nachfrageverbundes in erweiterter Sicht, in: Zeitschrift für betriebswirtschaftliche Forschung, 30. Jg., S. 181–193.

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

Natter, M./Reutterer, T./Mild, A./Taudes, A. (2006): An Assortmentwide Decision-Support System for Dynamic Pricing & Promotion Planning, in: Marketing Science (forthcoming).

Neslin, S.A./Gupta, S./Kamakura, W./Lu, J./Mason, C.H. (2006): Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models, in: Journal of Marketing Research, Vol. 43, May, S. 204–211.

Ng, R.T./Han, J. (2002): Clarans: A Method for Clustering Objects for Spatial Data Mining, in: IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 5, S. 1003–1016.

R Development Core Team (2007): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Rangaswamy, A. (1993): Marketing Decision Models: From Linear Programs to Knowledge-based Systems, in: Eliashberg, J./Lilien, G.L. (eds.): Handbooks in Operations Research and Management Science, Volume 5: Marketing, Amsterdam, S. 733–771.

Ravi, V./Raman, K./Mantrala, M.K. (2006): Applications of Intelligent Technologies in Retail Marketing, in: Krafft, M./Matrala, M. K. (eds.): Retailing in the 21st Century. Current and Future Trends, Berlin, S. 127–141.

Reutterer, T./Mild, A./Natter, M./Taudes, A. (2006): A Dynamic Segmentation Approach for Targeting and Customizing Direct Marketing Campaigns, in: Journal of Interactive Marketing, Vol. 20, No. 3/4, S. 43–57.

Reutterer, T./Natter, M. (2000): Segmentation-Based Competitive Analysis with MULTICLUS and Topology Representing Networks, in: Computers & Operations Research, Vol. 27, No. 11/12, S. 1227–1247.

Rossi, P./Allenby, G./McCulloch, R. (2005): Bayesian Statistics and Marketing, Chichester.

Rousseeuw, P.J. (1987): Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, in: Journal of Computational and Applied Mathematics, Vol. 20, S. 53–65.

Russell, G.J./Petersen, A. (2000): Analysis of Cross Category Dependence in Market Basket Selection, in: Journal of Retailing, Vol. 76, S. 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, S. 319–332.

Rygielski, C./Wang, J.-C./Yen, D.C. (2002): Data mining techniques for customer relationship management, in: Technology in Society, Vol. 24, S. 483–502.

Schnedlitz, P./Kleinberg M. (1994): Einsatzmöglichkeiten der Verbundanalyse im Lebensmittelhandel, in: Der Markt, 33. Jg., Nr. 1, S. 31–39.

Schnedlitz, P./Reutterer, T./Joos, W. (2001): Data-Mining und Sortimentsverbundanalyse im Einzelhandel, in: Hippner, U./Küsters/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, No. 3/4, S. 239–254.

Shaw, M.J./Subramaniam, C./Tan, G.W./Welge, M.E. (2001): Knowledge Management and Data Mining for Marketing, in: Decision Support Systems, Vol. 31, S. 127–137.

Sneath, P.H. (1957): Some Thoughts on Bacterial Classification, in: Journal of General Microbiology, Vol. 17, S. 184–200.

Tan P.-N./Steinbach, M./Kumar, V. (2006): Introduction to Data Mining, Boston.

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, No. 1, S. 53–62.

Verhoef, P.C./Spring, P.N./Hoekstra, J.C./Leeflang, P.S.H. (2002): The Commercial Use of Segmentation and Predictive Modeling Techniques for Database Marketing in the Netherlands, in: Decision Support Systems, Vol. 34, S. 471–481.

Wedel, M./Kamakura, W.A. (1999): Market Segmentation. Conceptual and Methodological Foundations, Boston.

West, P.M./Brockett, P.L./Golden, L.L. (1997): A Comparative Analysis of Neural Networks and Statistical Consumer Choice, in: Marketing Science, Vol. 16, No. 4, S. 370–391.

Wierenga, B./Van Bruggen, G.H. (2000): Marketing Management Support Systems. Principles, Tools and Implementation, Boston.
Winer, R./Tuzhilin, A. (2005): Proceedings of the 2005 International Workshop on Customer Relationship Management: Data Mining Meets Marketing. Stern School of Business, New York University.
Witten, I.H./Frank, E. (2005): Data Mining: Practical machine learning tools and techniques, San Francisco.
Zaki, M.J. (2000): Scalable Algorithms for Association Mining, in: IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 3, S. 372–390.
Zielke, S. (2002): Kundenorientierte Warenplatzierung. Modelle und Methoden für das Category Management, Stuttgart.

Summary

Data mining techniques are becoming increasingly popular and widely used add-ons to the more conventional methodological toolbox in modern marketing research and practice. The objective of such data driven analytical approaches is to extract and to represent relevant marketing information hidden in large data warehouses in a managerially meaningful fashion. This contribution discusses interfaces between data mining and marketing and demonstrates the adoption of selected data mining techniques for analyzing market baskets using supermarket transaction data. The employed methods comprise conventional affinity analysis, an algorithm for *k*-medoid clustering, as well as tools for mining and evaluating association rules among categories included in a typical supermarket assortment. The analytical procedures are illustrated using the **arules** package available under R, a freely available language and environment for statistical computing and graphics.

Schlüsselbegriffe

Data Mining, Warenkorbanalyse, Sortimentsverbund

Keywords

Data mining, market basket analysis, cross-category effects