Nicht nur im Einzelhandel

Die vielfältigen Einsatzgebiete der Assoziationsanalyse

von Peter Gerngroß

Was ist Assoziationsanalyse?

Das früheste bekannte Beispiel für die Nutzung von Assoziationsanalysen - auch Warenkorbanalyse genannt - stammt aus den späten 1980er Jahren und wird dem US-Einzelhandelskonzern Walmart zugeschrieben. Analytiker hatten dort herausgefunden, dass an Freitagabenden auffallend häufig Babywindeln und Bier zusammen gekauft werden. Diesen Effekt nutzten die Walmart-Manager, indem sie Bier und Windeln in den Supermärkten nebeneinander platzierten und so den Absatz nochmals deutlich steigern konnten.

Inzwischen ist zwar bekannt, dass diese Anekdote dem Bereich urbaner Legenden[1] zuzuordnen ist und sich so nicht zugetragen hat, dennoch macht dieses einfache Beispiel deutlich, worum es bei der Assoziationsanalyse geht: Es gilt, aus einer Menge von gruppierten Ereignissen (z.B. Ereignis: Kauf eines Produktes, Gruppierung: Einkaufskorb im Supermarkt) häufig vorkommende Ereigniskombinationen zu finden.

Das Verfahren wurde erstmals im Einzelhandel genutzt und wurde auch dadurch bekannt. Heute kennt jeder das wichtigste Feature der Amazon-Website (siehe Abb. 1): "Kunden, die diesen Artikel gekauft haben, kauften auch ...". Amazon wertet für diese Funktion jeweils sehr zeitnah ("near time") große Mengen an Kundentransaktionsdaten aus und ist so in der Lage, zum richtigen Zeitpunkt - wenn der Kunde sowieso im Begriff ist, einen Kauf zu tätigen – zusätzliche Kaufimpulse zu setzen.

Ergebnisse von Assoziationsanalysen bei Amazon.de[2]

Die Assoziationsanalyse ist historisch betrachtet die erste neue Data-Mining-Verfahrensklasse[3]. Sie ist zugleich die transparenteste – und die am meisten unterschätzte. So lassen sich die Verfahren nicht nur – wie man vermuten könnte – im Einzelhandel anwenden. Sie liefern vielmehr, wie in diesem White Paper noch gezeigt werden wird, in vielen Branchen und Geschäftsbereichen wichtige Einsichten bei unterschiedlichsten Fragestellungen. Sehr oft werden die gefundenen Zusammenhänge schließlich operativ eingesetzt.

Bei der Anwendung von Assoziationsverfahren gibt es nur wenige Parameter und Kennzahlen, über die das Ergebnis beeinflusst werden kann. Auch braucht man kein Statistiker zu sein, um die Ergebnisse zu verstehen: Die gelieferten Resultate sind fast intuitiv verständlich und können direkt sowohl von Fachanwendern als auch von Kunden genutzt werden, wie nicht zuletzt das Beispiel Amazon zeigt.

Durch die Entwicklung von einerseits effizienten Algorithmen und andererseits immer leistungsfähigerer Computern sind auch große Datenmengen mit kostengünstiger Standardhardware verarbeitbar. In der Praxis sind Zeilenzahlen im mehrstelligen Millionenbereich nicht unüblich.

Entscheidend war dabei die Entwicklung effizienter Algorithmen: Auch mit den leistungsfähigsten Rechnern könnten nicht alle Kombinationen durchgerechnet werden. Beispielsweise beträgt bei nur 100 Produkten die Anzahl aller möglichen Produktkombinationen (die Menge aller Produkt-Teilmengen) $2^{100} \approx 10^{30}$. Im Jahr 1993 wurde der erste Algorithmus veröffentlicht, der das Problem effizient löst: der A-priori-Algorithmus[4,5]. Dieser wurde in der Folge weiter verbessert, und auch neue Algorithmen wurden gefunden. Heute können so Datenbestände mit 7-stelligen Anzahlen unterschiedlicher Produkte in Minuten oder sogar Sekunden analysiert werden.

1 siehe z.B. [Fi]
2 Quelle: http://www.amazon.de

3 Andere Verfahren, die heute dem Data Mining zugerechnet werden, wie Entscheidungs- oder Clusterverfahren, entwickelten sich aus der klassischen Statistik
4 Original-Artikel zum Apriori-Algorithmus: [AgImSw1993]
5 In diesem Whitepaper wird auf die Algorithmen nicht weiter eingegangen. Gute Darstellungen der wichtigsten Algorithmen finden sich z.B. in [TaStKu2005]

In Abschnitt 2 dieses Whitepapers werden zunächst die wichtigsten Varianten der Assoziationsanalyse kurz vorgestellt. Abschnitt 3 enthält Anwendungsszenarien aus unterschiedlichen Branchen, die die vielseitige Anwendbarkeit der Verfahren illustrieren. Im abschließenden Abschnitt 4 werden die wichtigsten Kennzahlen und Parameter der einfachen Assoziationsanalyse vorgestellt und anhand eines einfachen Beispiels erläutert.

Varianten der Assoziationsanalyse

Hinter dem Begriff "Assoziationsanalyse" verbergen sich eine ganze Reihe unterschiedlicher Problemstellungen, denen gemein ist, dass sie häufiges gemeinsames Auftreten von Elementen oder Ereignissen untersuchen. Die Verfahren lassen sich unterscheiden nach der Art der Beziehungen, in der die Ereignisse stehen. Man unterscheidet einfache Relationen, Assoziationen und Sequenzen.

Relationen

Eine Relation ist ein gemeinsames Auftreten von Elementen, die dieselbe Bezugsgröße aufweisen. Die Bezugsgröße, die die Elemente verbindet, wird in diesem Zusammenhang auch "Transaktion" genannt. Beispielsweise verbindet ein Einkaufskorb ("Transaktion") die sich darin befindenden Produkte Butter und Brot ("Elemente"). Die Relation kommt hier dadurch zustande, dass die beiden Artikel von einem Kunden gemeinsam gekauft werden.
Bei der Analyse von Relationen sucht man nach Kombinationen, die relativ häufig vorkommen. Z.B. könnte ein Ergebnis der Relationsanalyse sein, dass Butter gemeinsam mit Brot in 40% der untersuchten Einkaufskörbe lagen. Solche "Relationsmuster" können u.a. im Marketing dazu verwendet werden, um interessante Produktbündel zu identifizieren.
In der Abbildung „Ergebnisse von Assoziationsanalyse bei Amazon.de[2]“ ist der obere Teil ("Wird oft zusammen gekauft") das Ergebnis einer Relationsanalyse. Der untere Teil (Kunden, die diesen Artikel gekauft haben, kauften auch ...") hingegen ist typisch für eine Assoziationsanalyse.

Assoziationen

Im Unterschied zu den Relationen wird bei Assoziationen nicht nur das gemeinsame Auftreten von Elementen, sondern auch deren Abhängigkeiten betrachtet. Das Ergebnis sind Aussagen wie: "Zwei Drittel der Kunden, die Butter kaufen, kaufen gleichzeitig auch Brot". Es wird also auch die "Richtung" der Beziehung betrachtet.
Aber auch die Relationshäufigkeit ist hierbei interessant: Assoziationen müssen auch mengenmäßig relevant sein, d.h. das gemeinsame Auftreten der betrachteten Elemente sollte eine bestimmte Mindesthäufigkeit aufweisen.
Die Analyse von Assoziationen führt im Ergebnis zu sogenannten Assoziationsregeln. Im obigen Beispiel würde die gefundene Regel lauten: "Wenn Butter, dann Brot", oder kurz: "Butter -> Brot". Das Ziel einer Assoziationsanalyse ist also, möglichst aussagekräftige, "interessante" Regeln zu finden.

Kassenbon Nr.	Produkt
4711	Butter
4711	Käse
4711	Brot
4712	Käse
4712	Brot
4712	Müsli
4712	Milch
4713	Butter
4713	...
...	...

Tabelle 1: Input-Datensatz für Assoziationsanalyse

Assoziationsregeln müssen aber nicht aus Paaren von Elementen bestehen. Sowohl die linke als auch die rechte Seite der Regel kann mehr als ein Element enthalten. Beispielsweise wäre eine weitere zulässige Regel: "Brot, Käse -> Wein, Butter". Um Relationen und Assoziationen analysieren zu können, werden lediglich zwei Datenmerkmale

benötigt[6]: zum einen das Element (z.B. gekauftes Produkt) und zum anderen die Bezugsgröße, dem das Element zugeordnet ist (z.B. Kassenbon-Nummer). In Tabelle 1 ist ein typischer Input-Datensatz für Relations- und Assoziationsanalysen dargestellt.

Sequenzen

Ein dritter Aspekt kommt bei der Analyse von Sequenzen hinzu: die zeitliche Reihenfolge der Elemente (hier treffender "Ereignisse" genannt). Damit wird nicht nur das gemeinsame Auftreten von Ereignissen betrachtet, sondern auch die zeitliche Abfolge dieser Ereignisse.
Die damit generierten Regeln haben beispielsweise die Form "Kauf von Produkt A gefolgt vom Kauf von Produkt B führt zum Kauf von Produkt C" (kurz: "A, B -> C"). Im Gegensatz zur Assoziationsanalyse sind bei der Sequenzanalyse die Regeln "A, B -> C" und "B, A -> C" unterschiedlich.
Neben dem Ereignis und der Bezugsgröße wird bei der Sequenzanalyse ein drittes Datenmerkmal benötigt: der Zeitpunkt des Ereigniseintritts. Bei der Anwendung im Einzelhandel würde man beispielsweise folgende drei Merkmale wählen:

- Ereignis: Kauf eines Produktes
- Bezugsgröße: Kunde
- Zeitpunkt: Kaufdatum

Am Beispiel eines Versandunternehmens für Sportartikel könnte der Input-Datensatz folgendermaßen aussehen:

Kunde Nr.	Produkt	Kaufdatum
1234	Hantel-Set	03.05.2013
1234	Gymnastik-Matte	23.05.2013
1234	Langhantel	25.05.2013
1235	Laufschuhe	01.06.2013
1235	Laufhose	01.06.2013
1235	Laufjacke	23.09.2013
1235	Laufschuhe	24.10.2013
1236	Hallenturn-schuhe	09.07.2013
1236	...	...
...	...	...

Tabelle 2: Input-Datensatz für Sequenzanalyse

Ein zusätzliches Ergebnis der Sequenzanalyse sind Aussagen über Zeiträume zwischen den Ereignissen. Beispielsweise kann im obigen Beispiel für die Regel "Hantelset, Gymnastik-Matte -> Langhantel" der durchschnittliche zeitliche Abstand zwischen dem Kauf des Hantel-Sets und der Gymnastik-Matte sowie zwischen der Gymnastik-Matte und der Langhantel berechnet werden. Diese Information kann beispielsweise im Direktmarketing für die zeitliche Steuerung von Werbeimpulsen genutzt werden.

Anwendungsszenarien

Der klassische Anwendungsfall der Assoziationsanalyse ist - wie schon mehrfach erwähnt - der (Einzel-) Handel. Für diese Branche wurden die Verfahren entwickelt, hier werden sie schon seit langem operativ eingesetzt. Anwendungsbeispiele werden in der Literatur ausführlich diskutiert[7].
Einsatzszenarien gibt es jedoch auch in vielen anderen Branchen. In diesem Kapitel werden exemplarisch einige davon beschrieben.

Beispiel 1: Optimierung der Ersatzteilversorgung medizinischer Großgeräte

Ausgangslage

Ein weltweit führender Hersteller von Medizintechnik hat für die von ihm hergestellten Geräte Wartungsverträge mit Ärzten und Kliniken abgeschlossen. Es handelt sich dabei um hochwertige Großgeräte wie Röntgengeräte, Computer- oder Magnetresonanz-Tomographen, deren Preise ty-

6 Zusätzlich könnte auch eine Taxonomie berücksichtigt werden. Als Produkt-Taxonomie bezeichnet man die hierarchische Zusammenfassung von Produkten in Gruppen. Am Beispiel eines Supermarktes könnte eine Taxonomiekette folgendermaßen aussehen: "Kerrygold Butter 250g" -> "Butter" -> "Milchprodukte" -> "Kühlprodukte" -> "Lebensmittel"

7 siehe z.B. [NeKn2005] S. 222 ff

pischerweise im ein- bis zweistelligen Millionenbereich liegen.
Bei Ausfall eines Gerätes reist ein Techniker zum jeweiligen Standort, um das Gerät instandzusetzen. Meist sind dafür Ersatzkomponenten notwendig, die meist sehr hochwertig sind und deren Transport sehr kostenintensiv ist.
Im Normalbetrieb senden die Geräte permanent technische Daten über den Zustand des jeweiligen Gerätes an den Hersteller. Diese Sensordaten werden in einem Data Warehouse gespeichert und stehen zur Auswertung zur Verfügung.

Anforderungen
Bei Ausfall eines Gerätes nimmt der Techniker wenige, häufig notwendige Standard-Ersatzkomponenten mit. In vielen Fällen stellt er aber vor Ort fest, dass für die Instandsetzung andere Komponenten benötigt werden, was erstens hohe Transportkosten für die Spezialteile nach sich zieht und zweitens einen Zeitverzug mit sich bringt, der zu einer langen "Downtime" des Gerätes führt.
Im Rahmen des Projektes sollte im Vorhinein, also nach Geräteausfall und vor Abreise des Technikers, analytisch vorhergesagt werden, welche Ersatzkomponenten für die Instandsetzung nötig sind.

Lösung
Auf Basis der im Data Warehouse gespeicherten Log-Daten werden mit Hilfe der Assoziationsanalyse Regeln generiert, die bei einem Geräteausfall mit hoher Treffsicherheit den Satz benötigter Ersatzkomponenten bestimmen.
Die Log-Daten sind in Form von Event-Datensätzen gespeichert, die die zwei wesentlichen Informationen Event-ID und Geräte-ID enthalten. Außerdem wird der Zeitstempel des Events gespeichert.
Um die Regeln zu generieren, werden Geräteausfälle aus der Vergangenheit betrachtet. Von diesen Ausfällen weiß man (im Nachhinein), welche Ersatzkomponenten benötigt wurden. Jede Ersatzkomponente hat eine ID. Für ausgefallene Geräte werden aus dieser Information "künstliche" Event-Datensätze erzeugt. Diese enthalten die Komponenten-ID und die Geräte-ID.
Mittels Assoziationsanalyse werden nun häufig vorkommende Kombinationen der Form
Event 1 + Event 2 + … + Event N -> Komponente A + … + Komponente X
gesucht, d.h. auf Basis einer Kombination von Log-Events kann auf die für die Instandsetzung benötigten Ersatz-Komponenten geschlossen werden.

Ergebnis
Durch die Einführung des beschriebenen Verfahrens konnte die Quote richtig mitgenommener Ersatzkomponenten stark erhöht werden. Die Folge waren drastische Einsparungen bei den Instandsetzungskosten und deutlich geringere Downtimes.

Beispiel 2: Cross Selling beim Multikanalvertrieb von Finanzdienstleistungen

Ausgangslage
Ein Finanzdienstleister hat mehrere Teilgesellschaften, die für Neukundenakquisition sowie für den Produktvertrieb an Bestandskunden jeweils unterschiedliche Vertriebswege nutzen. Die Vertriebswege sind organisatorisch teilweise strikt getrennt. Der Produktbesitz der Kunden unterscheidet sich zwischen den Vertriebswegen deutlich, jedoch wurde in den letzten Jahren das Cross Selling zwischen den Teilgesellschaften forciert.

Anforderungen
Das zentrale Marketing in der Holding des Finanzdienstleisters ist daran interessiert, den Einsatz von Cross Selling zu intensivieren. Dafür sind tiefere Erkenntnisse über Einstiegsprodukte, Folgekäufe sowie Produktabschluss-Ketten nötig.
Das Ziel ist die automatische Generierung von kundenindividuellen Produktempfehlungen für den Inbound-Vertrieb, die Internet-Seite sowie für die Filialen.

Lösung
Es wurde ein zentraler Datenbestand aufgebaut, der die Produktabschlüsse der Kunden aller Teilgesellschaften umfasst. Dieser enthält die Kundennummer, den Produktcode einschließlich Taxonomie, sowie das Produktabschlussdatum.

Basierend auf diesem Datenbestand wurde eine Sequenzanalyse durchgeführt. Dadurch konnten zum einen typische Einstiegsprodukte ermittelt werden. Noch wichtiger aber ist die Ermittlung häufig vorkommender Produktkaufketten. Ein Beispiel wäre

1. Festgeld + 2. Girokonto
führt zu
3. Wertpapierdepot

Solche Ketten können sehr gut in CRM-Softwarelösungen oder operative Call-Center-Systeme integriert werden und so den Vertriebsmitarbeitern wichtige Hinweise auf das derzeit bestmögliche Produktangebot für den Kunden geben.

Ergebnis
Die Effektivität des Telefon-Inbound- sowie des Filialvertriebs konnte deutlich gesteigert werden. Ebenso konnten die Banner-Klickraten auf der Website erhöht werden. Letztlich liefert das Verfahren einen wichtigen Beitrag zur Erreichung des Ziels einer kontinuierlichen Steigerung der konzernweiten Cross-Selling-Quote.

Weitere Projektschschritte
In die Analyse sollen weitere „Warenkorbobjekte" einbezogen werden, die streng genommen keine Produkte sind, wie z.B. Produktkündigungen, Wettbewerbsinformationen etc.. Dadurch sollen einerseits die Prognosequalität erhöht werden, andererseits können weitere Ergebnisse wie die Kündigung von Kundenbeziehungen modelliert werden.
Eine andere mögliche Erweiterung ist die sogenannte „umgekehrte Modellierung". Beispielsweise könnte interessant sein, in welchen Fällen bei Girokonto-Kunden nicht der gleichzeitige Produktabschluss eines Kreditkartenvertrages vorliegt.

Beispiel 3: Website-Optimierung eines Versicherungskonzerns

Ausgangslage
Ein Versicherungskonzern betreibt eine Website, die zum einen der Information bestehender und potenzieller Kunden über die Leistungen und Produkte dient, zum anderen aber auch den Online-Abschluss von Versicherungsverträgen ermöglicht. Es wird also auf den verschiedenen Unterseiten eine große Menge an Informationen bereitgestellt. Gleichzeitig soll aber der Direkt-Abschluss von Verträgen auf der Website so einfach wie möglich gemacht werden.
Es werden bereits Bewegungsdaten der Website-Besucher anonymisiert erhoben und im Data Warehouse abgelegt. Diese Daten enthalten eine Session-ID (ein Schlüssel, der einen Website-Besuch vom Betreten bis zum Verlassen der Website eindeutig identifiziert), die innerhalb der Session besuchten Unterseiten sowie den jeweiligen Zeitpunkt.

Anforderungen
Die Website soll optimiert werden. Im Zuge dessen soll der Zugang zu den Produktabschluss-Seiten so einfach wie möglich gemacht werden.
Zu diesem Zweck soll analysiert werden, ob es häufig benutzte und umständliche Klickpfade gibt, die zu einer Produktabschluss-Seite führen und die vereinfacht werden können.

Lösung
Die im Data Warehouse gespeicherten Web-Daten wurden mit einer Sequenzanalyse untersucht. Alle dafür benötigten Daten waren vorhanden. Zusätzlich musste definiert werden, welche Unterseiten als „Produktabschluss-Seiten" gelten.
Die Sequenzanalyse liefert Regeln der Art

Seite 1
gefolgt von Seite2
gefolgt von ...
gefolgt von Seite M
führt zu
Seite N

Interessant waren demnach Regeln, bei denen „Seite N“ eine der Produktabschluss-Seiten ist. Außerdem musste die Regel eine bestimmte Mindestlänge haben (d.h. es wurden „umständliche“ Klickpfade gesucht) und sie sollte relativ häufig vorkommen.
Regeln, die diesen Anforderungen genügen, konnten mittels Sequenzanalyse automatisch gefunden werden. Ganz ohne menschliches Zutun geht es aber doch nicht: die in Frage kommenden Regeln (d.h. Klickpfade) mussten anschließend inhaltlich geprüft werden, um zu beurteilen, ob und ggfs. wie der Weg zum Produktabschluss vereinfacht werden kann.

Ergebnis
Interessenten mit der Absicht, ein Versicherungsprodukt anzuschließen, wurde der Weg zum Online-Abschluss in vielen Fällen erheblich erleichtert. Nach der Website-Optimierung konnte durch eine erneute Analyse der Klickpfade festgestellt werden, dass sich die Quote der „Abbrecher“, die sich vorher aufgrund der unnötigen und komplizierten Pfade nicht bis zur Abschluss-Seite durchklickten, deutlich verringerte. Die Effektivität der Website als Vertriebskanal konnte dadurch signifikant gesteigert werden.

Interessante Assoziationen

Zentrale Kennzahlen: Support und Confidence

Wie beurteilt man nun, ob eine Assoziationsregel interessant ist? Dafür gibt es zunächst einige zentrale Kennzahlen für Regeln, mithilfe derer potenziell interessante Regeln gefunden werden können. Umgekehrt ausgedrückt: aus der riesigen Zahl möglicher Regeln wird ein winziger Bruchteil herausgefiltert; diese können weiter untersucht werden, während die allermeisten Regeln mit gutem Gewissen ignoriert werden können.
Die erste Kennzahl, der Support (deutscher Ausdruck: Reichweite), gibt die relative Häufigkeit der in der Regel enthaltenen Elementkombinationen an.
Angenommen, es soll die Regel „A -> B“ untersucht werden. Um den Support der Regel zu berechnen, wird gezählt, in wie vielen Transaktionen sowohl A als auch B enthalten sind. Diese Zahl wird durch die Gesamtzahl der Warenkörbe N geteilt.

$$\text{Support } (A \rightarrow B) = \frac{\#(\text{A und B})}{N}$$

Mit dem Support lässt sich die mengenmäßige Relevanz einer Regel beurteilen. In den meisten Anwendungsfällen ist eine Regel uninteressant, wenn sie nur in wenigen Transaktionen vorkommt. Die Confidence ist die zweite zentrale Kennzahl bei der Assoziationsanalyse. Sie gibt die Zuverlässigkeit einer Assoziationsregel an.
Betrachten wir wieder die Regel „A -> B“. Wie beim Support wird zunächst wieder die Anzahl der Transaktionen gezählt, in denen beide Elemente A und B vorkommen. Dann wird gezählt, in wie vielen Transaktionen das Element A vorkommt. Die Confidence ist der Quotient aus beiden Zahlen:

$$\text{Confidence}(A \rightarrow B) = \frac{\#(\text{A und B})}{\#(A)}$$

Die Confidence lässt sich als bedingte Wahrscheinlichkeit interpretieren: Wenn bekannt ist, dass sich das Element A in einem Warenkorb befindet, wie groß ist dann die Wahrscheinlichkeit, dass auch B enthalten ist? Sie gibt die Zuverlässigkeit der Regel an: Wie sicher kann ich sein, dass die Regel zutrifft?
In den meisten Data-Mining-Tools sind Support und Confidence die einzigen Parameter, die bei einer Assoziationsanalyse eingestellt werden müssen. Man gibt jeweils einen Mindestwert an und erhält im Ergebnis nur Regeln, bei denen Support bzw. Confidence den jeweiligen Mindestwert übersteigen. Das Setzen dieser Schranken ist essenziell für die effiziente Ermittlung von Regeln; die in den Tools implementierten Algorithmen nutzen die Schranken für den Ausschluss großer Teile der möglichen Regelmenge.
Viele Data-Mining-Tools geben für die gefundenen Regeln weitere Kennzahlen an, die verwendet werden können, um die Menge guter Regeln weiter einzuschränken. Ein Beispiel dafür ist der Lift, der angibt, um welchen Faktor die Con-

fidence einer Regel die einfache (unbedingte) Häufigkeit übersteigt[8].

$$\text{Lift } (A \rightarrow B) = \frac{\text{Confidence } (A \rightarrow B)}{P(B)}$$

Beispiele für interessante und uninteressante Assoziationen

Nicht alle Assoziationen, die algorithmisch ermittelt werden, sind auch interessant. Interessante Regeln müssen zunächst zuverlässig und mengenmäßig relevant sein, was durch Filtern über die entsprechenden Parameter Confidence und Support automatisch gewährleistet wird.
Ein weiteres Kriterium für eine gute Assoziationsregel ist, dass sie nicht offensichtlich ist. Beim Supermarkt-Beispiel wäre z.B. "Bier -> Mineralwasser" eine offensichtliche Regel. Um diese Regel aufzustellen, braucht man keine Data-Mining-Analyse, es reicht gesunder Menschenverstand. In solchen Fällen wäre die Analyse bestenfalls geeignet, die offensichtliche Regel zu bestätigen oder zu widerlegen.
Es sollten auch Regeln ausgeschlossen werden, die systematisch entstehen. Beispiel: Ein Online-Versand legt jeder Bestellung eine Broschüre bei. Diese wird in den Kassenbondaten gespeichert. Werden diese Positionen nicht vor der Analyse aus den Daten eliminiert, so werden einige - je nach Support-Schranke auch viele - Regeln der Form "Produkt A -> Broschüre" entstehen, die alle eine Confidence von 100% haben.
Manchmal sind nur Regeln einer bestimmten Form interessant. Im obigen Sensordaten-Beispiel sind nur Regeln erwünscht, bei denen auf der rechten Seite ausschließlich Ersatz-Komponenten stehen. Regeln der Form "Event 1 -> Event 2" sind in diesem Zusammenhang uninteressant.
Allgemein kann man sagen, dass gute Regeln folgende Eigenschaften aufweisen sollten:

- zuverlässig
- mengenmäßig relevant
- bisher unbekannt
- nicht systematisch bedingt
- inhaltlich relevant

8 Zu Kennzahlen und den bekanntesten Algorithmen siehe z.B. [TaStKu2005] S. 327 ff

Rechenbeispiel Supermarkt

In einem Lebensmittelmarkt werden die Einkäufe von fünf Kunden notiert:

Kunde	Einkaufskorb
Anton	Butter, Käse, Brot
Berta	Käse, Bier, Müsli, Milch
Carl	Butter, Brot, Bier
Dora	Butter, Käse, Milch
Erich	Cola, Müsli, Milch

Tabelle 3: Rechenbeispiel

Insgesamt haben die Kunden sechs verschiedene Produkte gekauft. Welche Regeln lassen sich nun daraus generieren? Drei der fünf Kunden haben Butter in ihrem Einkaufskorb (Anton, Carl und Dora). Anton und Carl haben außerdem auch Brot gekauft.
Es soll nun die folgende Regel untersucht werden: ”Wenn ein Kunde Butter kauft, dann auch Brot“. Die zentralen Kennzahlen werden berechnet:

Support

$$\text{Support } (\text{Butter} \rightarrow \text{Brot})$$

$$= \frac{\text{Anzahl Warenkörbe mit Butter und Brot}}{\text{Anzahl aller Warenkörbe}}$$

$$= 40\%$$

Confidence

$$\text{Confidence } (\text{Butter} \rightarrow \text{Brot})$$

$$= \frac{\text{Anzahl Warenkörbe mit Butter und Brot}}{\text{Anzahl der Warenkörbe mit Butter}}$$

$$= \frac{2}{3} \approx 66{,}7\%$$

Diese Confidence kann interpretiert werden als die (bedingte) Wahrscheinlichkeit für den Kauf von Brot, wenn bekannt ist, dass ein Kunde Butter kauft.

Lift
Die einfache Wahrscheinlichkeit für den Kauf von Brot ist

$$P(\text{Brot}) = \frac{\text{Anzahl Warenkörbe mit Brot}}{\text{Anzahl aller Warenkörbe}} = \frac{2}{5} = 40\%$$

woraus sich der Lift ableitet:

$$\text{Lift}\ (\text{Butter} \rightarrow \text{Brot})$$

$$= \frac{\text{Confidence}\ (\text{Butter} \rightarrow \text{Brot})}{P(\text{Brot})} = \frac{\frac{2}{3}}{\frac{2}{5}} \approx 1{,}67$$

Literatur

[Fi] Donald Fisk: Beer and Nappies – A Data Mining Urban Legend. Quelle: http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html

[DMS2010] mayato Data Mining Studie 2010 - Praxistest und Benchmarking. Quelle: http://www.mayato.com/de_DE/kompetenzen/fokusthemen/data-mining-studie.html

[AgImSw1993] R. Agrawal, T. Imieliński, A. Swami: Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207

[TaStKu2005] P. Tan, M. Steinbach, V. Kumar: Introduction to Data Mining. Addison Wesley 2005

[NeKn2005] Peter Neckel, Bernd Knobloch: Customer Relationship Analytics: Praktische Anwendung des Data Mining im CRM. Dpunkt Verlag 2005

Der Autor

Peter Gerngroß ist Experte für Data Mining, Business Intelligence und Analytisches CRM beim Analysten und Beraterhaus mayato. Er beschäftigt sich mit der qualitativen Optimierung und der Veredelung von Kundendaten sowie deren vertrieblicher Nutzung.
Peter Gerngroß | mayato GmbH
peter.gerngross@mayato.com

Über mayato

mayato ist als Analysten- und Beraterhaus spezialisiert auf Business Intelligence und Business Analytics. Von Niederlassungen in Berlin, Bielefeld und Heidelberg aus arbeitet ein Team von erfahrenen IT- und BI-Architekten, Statistikern, Analysten sowie fachlichen Experten für spezielle Themen wie Betrugserkennung, Data Mining und Analytisches CRM. Zu den Kunden von mayato zählen namhafte Unternehmen aus unterschiedlichen Branchen. Als Partner mehrerer Softwareanbieter ist mayato grundsätzlich der Neutralität und in erster Linie der Qualität seiner eigenen Dienstleistungen verpflichtet. Nähere Infos unter: www.mayato.com.

mayato GmbH

Am Borsigturm 9
D-13507 Berlin
T +49 30.4174.8657
info@mayato.com
www.mayato.com