Hochschule Rhein Main
Fachbereich Wirtschaftsinformatik
Wintersemester 2019/ 2020

Hochschule RheinMain

Seminararbeit zum Thema

# Data Mining

Vorgelegt von: Nikita Solodovnikov
Matrikelnummer: 1012116
E-Mail: Nikita.Solodovnikov@student.hs-rm.de

Eingereicht bei: Prof. Dr. Dirk Voelz
Abgabetermin: 13.01.2020

# Inhaltsverzeichnis

# 1. Einleitung

„Ohne Big-Data-Analysen sind Unternehmen blind und taub und bewegen sich im Internet wie ein Reh auf der Autobahn“, so schrieb der Autor Geoffrey Moore, sich auf die immer größer werdende Menge an Daten beziehend, welche mit herkömmlichen Methoden nicht mehr analysierbar sei und die Wichtigkeit der Daten-Analyse in der Zukunft verdeutlichend.

Auch eine groß angelegte Studie von IDC bestätigt diese Aussage.

In dem im Jahr 2018 herausgebrachten White Paper wird prognostiziert, dass die Menge an weltweiten Daten sich in den Jahren 2018 bis 2025 von 33 Zettabyte auf 175 Zettabyte vervielfältigen wird [IDC]. Bereits das im Jahr 1965 formulierte Mooresche Gesetz, welches oft in der technischen Fachliteratur zitiert wird, besagt, dass die Gesamtheit an Information in der Welt sich etwa alle 20 Monate verdoppelt [Runkler 2010 Vorwort]. Diese Entwicklung benötigt passende Methoden und Informationssysteme, um mit der wachsenden Datenflut zurecht zu kommen und nützliche Informationen aus ihr ziehen zu können.

In dieser Seminararbeit wird eine dieser Methoden, genannt Data Mining, behandelt.

Im zweiten Kapitel wird der Begriff des Data Mining aufgeführt, in der fachlichen Begriffswelt eingeordnet, der Ablauf des Data Mining Prozesses dargestellt und zwischen den unterschiedlichen Arten des Data Mining unterschieden.

Im dritten Kapitel werden vier populäre Data Mining Methoden vorgestellt.

Das vierte Kapitel beschreibt den in der Wirtschaft etablierten Data Mining Standard CRISP-DM und seinen Ablauf. Fachliche Beispiele aus der Praxis, die sich mit Data Mining auseinandersetzen, werden im fünften Kapitel behandelt.

Im sechsten Kapitel wird Data Mining im Bezug auf die heutige Zeit kritisch hinterfragt und mögliche Entwicklungen werden prognostiziert.

Zum Schluss werden die in dieser Seminararbeit gewonnenen Erkenntnisse und Aufschlüsse zusammengefasst.

Trotz der großen Anzahl an fachlicher Literatur und vorliegender Uneinigkeit in der Wissenschaft zum Thema Data Mining, wird im Folgenden versucht die einzelnen Themen möglichst schlüssig wiederzugeben und die fachbezogene Meinung in der Wissenschaft verständlich zusammen zu fassen.

# 2. Der Begriff des Data Mining

## 2.1 Definition des Data Mining

Data Mining wird häufig als eine Menge bestimmter Datenanalysemethoden definiert, mit dem Ziel gültige, neue, potenziell nützliche und verständliche Muster und entscheidungsrelevante Zusammenhänge autonom aus Daten zu extrahieren und sie dem Anwender zu präsentieren, ohne vom Anwender Aussagen über den gesuchten Inhalt zu fordern [Knobloch&Weidner 2000 S.4; Petersohn 2005 S.8-9].

## 2.2 Einordnung des Data Mining

Data Mining wird als Teilschritt des Gesamtprozesses des Knowledge Discovery in Databases (KDD) eingeordnet [Alpar 2000 S.4-5]. An dessen Anfang steht eine unverarbeitete Datenmenge, in der durch Methoden des Data-Mining Muster und Zusammenhänge erarbeitet werden, und am Ende extrahiertes bzw. entdecktes Wissen steht [Knobloch&Weidner 2000 S.4, Petersohn 2005 S.9].
Es etablierte sich der Begriff Data Mining für den gesamten KDD-Prozess [Alpar 2000 S.4].

Das Datenanalyseproblem kann in 2 grundlegende Typen unterschieden werden:

Die aus der Statistik kommende Methode der hypothesengetriebenen Fragestellung, oder auch Top-Down-Ansatz, versucht von einer bestehenden Annahme oder Theorie aus, die vorliegende Hypothese, mit Hilfe von Daten und Algorithmen entweder zu bestätigen (verifizieren), oder zu verwerfen (falsifizieren). Die in den achtziger Jahren aus dem Bereich der Künstlichen Intelligenz entwickelte datengetriebene Analyse, oder auch Bottom-Up-Vorgehensweise versucht aus vorliegenden Daten neue, zunächst hypothetische Erkenntnisse zu erzeugen, welchen eine Hypothesenprüfung und Interpretation folgt, bevor Handlungsalternativen ausgearbeitet werden können [Knobloch&Weidner 2000 S.4, Alpar 2000 S.3].

## 2.3 Der Data Mining Prozess

Bevor der Data Mining Prozess abläuft, muss der Anwender sich der Ziele, bewusst sein. Des Weiteren muss das Verständnis für den Prozess an sich, sowie das Wissen über die relevanten Daten vorhanden sein und ein konkreter Anlass für das Data Mining vorliegen [Alpar 2000 S.6].

Der erste Schritt des Prozesses ist die Aufgabendefinition. Aus der Problemstellung und dem Wissen über die relevanten und vorhandenen Daten muss der Anwender eine geeignete Data Mining Methode formulieren können. Im nächsten Schritt der Datenselektion wird die zielgerichtete Auswahl von für die Analyse sinnvollen Daten aus dem Rohdatenbestand beschrieben.
Nachdem im Schritt der Vorverarbeitung die Daten mit fehlenden oder falschen Eintragungen korrigiert oder entfernt wurden, wird im Schritt der Transformation die Anzahl der berücksichtigten Variablen minimiert, sowie eine Skalentransformation und Normierung durchgeführt.
Nun werden die Daten in der Data Mining Untersuchung analysiert und im letzten Prozessschritt die gewonnenen Erkenntnisse und Muster interpretiert [Knobloch&Weidner 2000 S.5, Alpar 2000 S.7, Petersohn 2005 S.11-12].

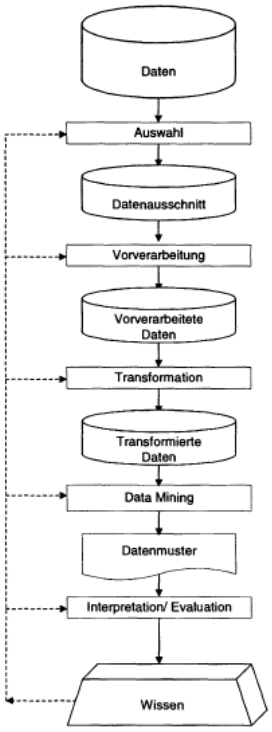


Abb. 1: Schritte im Data Mining Prozess (nach Fayyad et.al. 1996)

### 2.4 Arten des Data Mining

Ursprünglich wurde Data Mining hauptsächlich auf formatierte und strukturierte Daten angewendet. Allerdings liegt heutzutage ein Großteil der Daten in unstrukturierter und unformatierter Form vor. Text Mining beschreibt die Extraktion von Mustern aus unformatierten Datenbeständen [Alpar 2000 S. 5].

Allerdings ist aufgrund der unstrukturierten Datengrundlage die Erkennung von Mustern oft ohne Weiteres nicht möglich. Drees beschränkt den Prozess des Text Mining deshalb darauf, aus unstrukturierten Daten strukturierte Daten zu generieren [Drees 2016 S.53]. Im Falle der Anwendung von Data Mining Methoden im Web spricht man von Web-Mining.

### 2.5 Aufgaben des Data Mining

Die Aufgabe des Data Mining besteht darin, Daten derart zu analysieren, dass Muster und deren Strukturmodelle erkannt werden. Diese Strukturmodelle stellen die schematische Aufbereitung und Typisierung der Daten unter dem Aspekt eines konkreten Analyse- und Anwendungsziels dar [Petersohn 2000 S.11].

Das Auffinden oder Wiederfinden bekannter Information oder Entdecken neuer, unbekannter Information ist die Kernaufgabe des Data Mining [Drees 2016 S.54].

Die Resultate des Data Mining beeinflussen unternehmensstrategische Entscheidungen und tragen zur Verbesserung der Chancen im Wettbewerb bei [Petersohn 2000 S.14].

## 3. Methoden des Data Mining

### 3.1 Klassifikation

Unter dem Oberbegriff der Klassifikation werden die beiden Methoden, Klassenbildung und Klassifizierung, zusammengefasst [Petersohn 2000 S.11].

Für das überwachte Verfahren der Klassenbildung werden eine vorgegebene Problemstellung sowie eine vorher bestimmte Anzahl an Klassen benötigt, in die eine Menge von Objekten zugeteilt wird. [Petersohn 2000 S.25-26].
Die Methode der Klassifizierung ordnet anhand des Wissens vergangener Zuordnungen neue Objekte mit einer unbekannten Klassenzugehörigkeit einer bereits existierenden Klasse zu [Petersohn S.31].

Ein typisches Anwendungsfeld der Klassifikation findet in der Marketingforschung statt. In einem persönlichen Interview, oder durch das Internet, werden potenzielle Kunden befragt und in Kundengruppen mit gemeinsamen Wünschen gegliedert. Anschließend können auf die jeweilige Gruppe individuelle Marketinginstrumente angewendet werden.
Um ihren Kunden eine auf sie zugeschnittene Erfahrung zu ermöglichen, analysieren große Online-Marktplätze das Kundenverhalten mithilfe von Data-Mining-Methoden und erstellen Kundenprofile anhand von Faktoren, wie der Dauer des Aufenthalts eines Kunden auf einer Webseite, des Weges und des Kaufbetrags [Petersohn 2000 S.16-17].

## 3.2 Clustering

Ist die Klassenzugehörigkeit der Objekte nicht gegeben, wird das unüberwachte Lernverfahren des Clustering angewendet.
Die unmarkierten Daten werden nicht wie bei der Klassifizierung vorgegebenen Klassen zugeteilt, stattdessen werden die Klassen bzw. Cluster anhand der Muster in den Daten gebildet.
Die Objekte eines Clusters müssen eine hohe Ähnlichkeit vorweisen, gleichzeitig sollen die einzelnen Cluster zueinander möglichst unterschiedlich sein [Larose 2014 S.12].

Zum Beispiel können Politikberater ganze Wahlbezirke mit Hilfe von Clustering-Methoden analysieren, um die Standorte von Wählergruppen aufzudecken, die auf einen bestimmten Kandidaten besonders positiv reagieren. In diesem Fall werden alle entsprechenden Variablen

(z.B. Einkommen, Herkunft, Geschlecht) in den Clustering-Algorithmus eingegeben, ohne dass Zielvariablen festgelegt werden.
Somit können präzise und effektive Wählerprofile für das Sammeln von Wahlspenden und Wahlwerbung angelegt werden [Larose 2014 S.138-139].

## 3.3 Assoziationsanalyse

Eine weit verbreitete Methode der Musterextraktion ist die Assoziationsanalyse, deren bekannteste Anwendung die Warenkorbanalyse ist.
Bei der Assoziationsanalyse werden im Datenbestand Attribute identifiziert, die häufig gemeinsam auftreten [Drees 2016 S.55].

Die Assoziationsanalyse kann überwacht und unüberwacht ausgeführt werden.
In der Warenkorbanalyse zum Beispiel kann der Anwender die unüberwachte Analyse des Kaufverhaltens durchführen, um mögliche auffällige Beziehungen von zusammen gekauften Waren festzustellen. Somit können solche Waren gezielt beim Kauf eines der Produkte beworben werden.
Eine überwachte Assoziationsanalyse kann hingegen bei einer Untersuchung mit bekanntem Endergebnis angewendet werden.
Beispielsweise wenn politische Meinungsforscher demographische Daten zu einer bereits abgeschlossenen Wahl zusammen mit der Wahlpräferenz des Probanden besitzen, können aus diesem Datensatz Assoziationsregeln gewonnen werden. Diese Regeln können dabei helfen, die Wahlpräferenzen von Wählern mit bestimmten demographischen Charakteristiken in einer überwachten Assoziationsanalyse zu klassifizieren [Larose 2014 S.260-261].

## 3.4 Zeitreihenanalyse

Neben der Analyse von Finanzzeitreihen ist die Optimierung von wirtschaftlichen Geschäftsprozessen durch Analyse und Modellierung von Prozessmustern ein wichtiger Bestand der Data-Mining-Methode der Zeitreihenanalyse. Man betrachtet nicht nur vergangene Daten, sondern versucht auch zukünftige Daten möglichst genau zu prognostizieren [Runkler 2010 S.83].

Durch ausschlaggebende Merkmale, wie zum Beispiel den Ressourcenverbrauchs, können Prozessinstanzen in optimale, durchschnittliche und schlechte Prozessinstanzgruppen zusammengefasst werden. Aus der Analyse der optimalen Prozessinstanzgruppen können Rückschlüsse zur Verbesserung der als schlecht eingestuften Prozesse gezogen werden und Prognosen zur zukünftigen Entwicklung getroffen werden [Petersohn 2000 S.18].

# 4. CRISP - DM

Das CRISP-DM (Cross-Industry Standard Process for Data Mining) Modell ist ein branchenübergreifender Standard-Prozess-Modell für Data Mining.

Der Prozess bestehend aus sechs Phasen, ist adaptiv und iterativ. Jede Phase ist abhängig vom Ergebnis der vorherigen Phase und muss teilweise mehrmals durchlaufen, bevor die nächste Phase beginnen kann.

1. **Business Understanding (Geschäftsverständnis)**

   Im ersten Schritt müssen die Projektziele und Anforderungen klar definiert werden. Anschließend müssen diese Ziele und Einschränkungen in die Formulierung einer Data Mining Problemdefinition übersetzt werden.

   Schließlich ist eine vorläufige Strategie zur Erreichung dieser Ziele auszuarbeiten.

2. **Data Understanding (Datenverständnis)**

   Im folgenden Schritt werden zunächst die zur Verfügung stehenden Daten gesammelt und es wird sich mit ihnen vertraut gemacht. Die Datenqualität wird bewertet und erste Muster analysiert.

3. **Data Preparation (Datenvorbereitung)**

   Diese arbeitsintensive Phase umfasst alle Vorbereitungen des endgültigen Datensatzes aus den ursprünglichen Rohdaten. Es werden alle geeigneten Objekte zur Modellierung ausgewählt und bei Bedarf transformiert und bereinigt.

4. **Modeling (Modellierung)**

   Zu Beginn der Modellierungsphase werden eine oder mehrere geeignete Modellierungstechniken ausgewählt, individuell an die Daten und das Ziel angepasst und angewendet. Möglicherweise muss mehrmals zurück in die Datenvorbereitungsphase gewechselt werden, um die Daten an die Anforderungen einer Data Mining Methode anzupassen.

5. **Evaluation (Evaluation)**

   Nachdem die Modellierungsphase ein oder mehrere Modelle geliefert hat, werden diese auf ihre Qualität und Effektivität überprüft, und es wird bestimmt, welches Modell die Aufgabenstellung am besten erfüllt.

6. **Deployment (Bereitstellung)**

   Zum Schluss werden die Ergebnisse aufbereitet und präsentiert [Larose 2014 S.4-6].

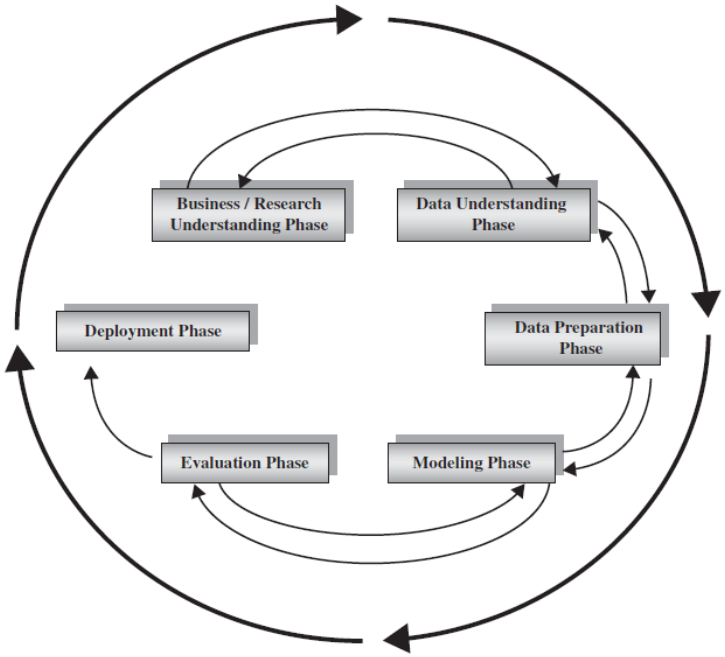


Abb. 2: Die iterativen Phasen des CRISP-DM
(nach Larose 2014)

# 5. Anwendungsbeispiele des Data Mining in der Praxis

## 5.1 Einführung und Anwendung von Data Mining in der Versicherungswirtschaft

Das folgende Praxisbeispiel basiert auf der Untersuchung eines großen deutschen Versicherungsunternehmens, dessen Aufgabe die Auswahl eines Data Mining Tools, sowie die Anwendung auf einen vorbereiteten Datenbestand war.

Der Umfang der vorhandenen Daten, die Komplexität der bestehenden Beziehungen, sowie der Wunsch bisher unbekannte Zusammenhänge und Abhängigkeiten zu ermitteln, führte zur Suche nach einer Alternative zu konventionellen Datenbank-Werkzeugen, wie zB. SQL.

Für das eingesetzte Data Mining Tool wurden folgende Kriterien vorgegeben:

**Mustererkennung**

Die Fähigkeit auf bestimmten Mustern basierende Zusammenhänge in den Daten zu erkennen und diese korrekt zu gewichten ist ein Schlüsselmerkmal eines geforderten Data Mining Tools.

**Unterstützte Methoden**

Da die Data Mining Methoden für diverse Aufgabenstellungen unterschiedlich gut geeignet sind, muss die vom Data Mining Tool benutzte Methode auf die vorhandene Aufgabenstellung abgestimmt sein. Der Einsatz von hybriden Systemen, welche mehrere Methoden unterstützen, ist bei komplexen Aufgabenstellungen zu empfehlen.

**Datenverarbeitung**

Da Data Mining Tools oft sehr große Datenmengen, die teilweise unvollständig, fehlerhaft oder unterschiedlich formatiert sind, verarbeiten müssen, ist es wichtig, dass das System damit umzugehen weiß und diese zuverlässlich verarbeiten kann.

**Anpassung an die Systemlandschaft**

Um auftretende Kosten beim Migrieren der Daten auf neue oder externe Systeme zu vermeiden und aufkommende Speicher- und Ressourcenkosten zu minimieren, ist die Unterstützung von vorhandenen Informationssystemen ein wichtiges Kriterium bei der Auswahl des Data Mining Tools.

**Weitere Kriterien**

Weitere Kriterien, wie das Wiederfinden und Vergleichen verschiedener und vergangener Analyseläufe in Form einer Versionsführung, die Abwägung der Verwendung ausländischer Tools ohne deutschen Support, die für die Fachabteilung wichtige Benutzerfreundlichkeit, sowie die oft außerordentlich hohen Kosten, müssen vor der Implementierung eines Data Mining Tools überprüft und diskutiert werden.

Nach der Auswahl und der Implementierung des Data Mining Tools müssen die zur Verarbeitung benötigten Daten beschafft und bereinigt werden.
Es ist wichtig, dass kompetente, sich mit dem Fachbereich und den Daten auskennende Mitarbeiter für die Datenauswahl sowie der korrekten Interpretation der relevanten Ergebnisse eingesetzt werden.
Für die Auswahl der Daten wird ein zusammen mit dem Fachbereich vorbereiteter Fragekatalog durchgearbeitet, um die Validität der Daten festzustellen.

Es muss zum Beispiel geklärt werden, ob die Datenbasis die entsprechende Qualität und Quantität aufweist, um aussagekräftige Ergebnisse erreichen zu können. Des Weiteren ist die Frage nach der erforderlichen Klassifizierung der Daten, der Stärke der Relation verschiedener Attribute und nach dem Einfluss, z.B. der Nationalität des Versicherungsnehmers, oder der Lage des versicherten Grundstücks, zu klären.

Nun folgt der in Kapitel 2.3 aufgeführte Data Mining Prozess, nach dessen Abschluss folgende Ergebnisse festgehalten wurden.
Die ersten beiden Phasen (Datenselektion und Datenvorverarbeitung) machen ca. 80% des Gesamtaufwandes aus, wobei die Analyseergebnisse umso

schneller und exakter ausgewertet werden können, je besser die Unternehmensdaten vorher bereinigt wurden.
Fachliche Fragestellungen, die nicht vor der Analyse geklärt wurden, verzögern diese stark.
Statt den Data Mining Prozess für alle Bereiche des Unternehmens gleichzeitig aufzubauen und durchzuführen, empfiehlt es sich diesen auf einige Fachbereiche zu beschränken [Alpar 2000 S.2011-224].

## 5.2 Data Mining im Marketing und Controlling

Mithilfe von neuronalen Netzen möchte ein Versandhaus das ertragsoptimale Verhältnis zwischen dem Umsatz und der Werbekosten beim Verschicken von Werbeträgern ermitteln. Aus der Vielzahl an vorliegenden Kundenbeobachtungen sollen mithilfe von selbstständig lernenden Prognosesystemen Schlüsse über die optimale Auflage gezogen werden, um langfristig zu einem möglichst hohen Return on Investment zu führen und die bisherigen Planungs- und Steuerungsprozesse der Werbeträgerplanung zu optimieren.
Als Prognoseziel wird das Kaufwahrscheinlichkeit eines Kunden anhand seines Kaufverhaltens in der Vorsaison festgelegt.

Mit dem Ziel zu erfahren wie die Anzahl der durch dem Warenkatalog angesprochenen Kunden mit der Kennzahl der z.B. Kosten-Umsatz-Relation oder des Deckungsbeitrags zusammenhängt, werden im ersten Schritt alle Kunden mithilfe des neuronalen Netzes anhand ihres Verhaltens in der Vorsaison beurteilt.
Im zweiten Schritt wird für jeden Kunden der Bruttobestellwert, sowie der Nettoumsatz prognostiziert.
Anhand dieser Prognosen können die Kennzahlen in Abhängigkeit von der Anzahl der angesprochenen Kunden aufgestellt werden.
Mit Hilfe dieses Vorgehens können quantitative und qualitative Veränderungen des Kundenstamms bereits in der Planungsphase detailliert berücksichtigt werden [Alpar 2000 S.53-56].

## 5.3 Kundensegmentierung aufgrund von Kassenbons mit Data Mining

Anhand der Verkaufsdaten eines Supermarkts wird eine Segmentierung der Kunden vorgenommen.

Die Abnehmer des Absatzmarktes, aufgeteilt in potenzielle und tatsächliche Kunden, unterscheiden sich hinsichtlich ihrer Bedürfnisse, Präferenzen und finanziellen Mitteln. Deshalb ist die Einteilung des Absatzmarktes in mehrere Segmente, welche anschließend mit individuellen Marketingprogrammen bearbeitet werden können, sinnvoll.

Die nachfolgende Betrachtung beschränkt sich auf die Marktsegmentierung von Konsumgütermärkten, sowie auf die tatsächlichen Kunden.

Aus diesem Grund eignen sich die gewonnen Erkenntnisse für Kundenbindungs- und Umsatzsteigungsprogramme.

Die Marktsegmentierung bringt eine Reihe von Vorteilen mit sich.

Durch die Segmentierung des Marktes wird nicht nur die Kenntnis des Marktes verbessert, welche die Abgrenzung des relevanten Marktes erlaubt, es wird in Folge die zielgerichtete Ansprache der Kunden entsprechend ihrer Bedürfnisse ermöglicht. Der gezielte Einsatz von Marketinginstrumenten sowie eine effizientere Aufteilung des Marketingbudgets und die verbesserte Positionierung von Neuprodukten und Marken wird durch die Segmentierung ebenfalls ermöglicht. Ein weiterer Vorteil ist die Möglichkeit vor der Konkurrenz in etablierte Märkte einzudringen, was durch die Erkennung von Veränderungen in der Segmentstruktur ermöglicht wird.

Der Ablauf einer Segmentierungsstudie kann in drei Schritten beschrieben werden.

1. Im ersten Schritt wird die Datenerhebung durchgeführt. Ein auf der Basis von Interviews erstellter Fragebogen wird bei einer Stichprobe von Verbrauchern angewendet, um die Datenbasis zu ermitteln.
2. Mit Hilfe von z.B. der Clusteranalyse werden die verschiedenen Segmente ermittelt.
3. Im Hinblick auf die in der ersten Phase ermittelten Daten werden Kundenprofile erstellt.

Das Verfahren muss in regelmäßigen Abständen wiederholt werden, da das Verhalten, die Einstellung und die demographischen Daten sich im Laufe der Zeit verändern.

Die Segmentierung kann nach vier Gruppen von Merkmalen verwendet werden.

1. Die erste Gruppe wird in biologische, geographische und sozialdemographische Kriterien aufgeteilt. Wünsche und Präferenzen korrelieren oft mit diesen Kriterien, weiterhin sind diese leicht zu erheben.
2. In der Gruppe Merkmale des beobachtbaren Kaufverhaltens wird das Preisverhalten, die Verbrauchsintensität, die Einkaufshäufigkeit und das Produktwahlverhalten analysiert. Wie auch die erste Gruppe lässt sich aus diesen Kriterien nicht die Ursache für das jeweilige Verhalten ermitteln.
3. Die Gruppe psychographischen Kriterien beschreibt das Kaufverhalten durch subjektive Wahrnehmungen. Diese Gruppe kann in Persönlichkeitsmerkmale und produktspezifische Merkmale unterschieden werden.
4. Für die Ansprachemöglichkeiten der Zielgruppe werden Faktoren wie die Mediennutzung, das Qualitäts- und Preisbewusstsein sowie das Vertrauen zu Betriebsformen des Handels behandelt.

Vor der Ausführung der Marktsegmentierung muss überprüft werden, ob die Unterscheidbarkeit der Segmente im Hinblick auf das Nachfrageverhalten gegeben ist. Ebenso wie die Wirtschaftlichkeit der Marktsegmentierung, muss die Größe der Segmente sich für eine Bearbeitung mit eigenen Marketingprogrammen lohnen.

[Alpar 2000 S.57-66]

# 6. Herausforderungen, Probleme und Kritik am Data Mining

## 6.1 Sicherheit, Privatsphäre und Datenintegrität

Die Gewährleistung der Geheimhaltung von sensiblen Nutzer-Daten beim Data Mining ist ein wichtiges Thema. Eine Methode zur Wahrung der Privatsphäre von Nutzern ist die absichtliche Modifizierung der verarbeiteten Daten. Die

Herausforderung, die sich Data Mining Anwender stellen müssen, ist nicht nur die bisherige Muster- und Wissensextraktion aus großen Datenbeständen, sondern auch die mit der Modifizierung der Daten kommende Datenintegrität [YangWu 2006 S.603].

Mit 57% der Weltbevölkerung (über 4 Milliarden Menschen) im Jahr 2019 [WorldStat] im Internet agierend, ist der Fokus auf der Gewährleistung der Sicherheit im Internet so hoch wie noch nie und muss bei der Weiterentwicklung des Data Mining eine hohe Priorität besitzen.

### 6.2 Entwicklung einer einheitlichen Data Mining Theorie in der Wissenschaft

Viele Data Mining Methoden, wie die Klassifikation oder die Clusteranalyse werden für spezifische Problemfälle angewendet, ohne dass eine universelle Methode existiert [YangWu S.597]. Diese Methoden sind sehr unflexibel und legitimieren die Forderung nach einem neuen innovativen und universellen Lösungsansatz.

### 6.3 Umgang mit großen Datenmengen

Durch den stetig steigenden Zuwachs an Daten, entwickelt sich das Problem der Verarbeitung dieser. Satelliten oder Computer-Netzwerk-Daten zum Beispiel haben im Jahr 2005 mit Leichtigkeit Datenbank-Kapazitäten von über 100 Terabyte erreicht. [YangWu S.598]. Ein Artikel vom August 2019 zeigt uns die heutigen Ausmaße. Das NCEI (National Centers for Environmental Information) speichert demnach monatlich 26 Terabyte Umwelt-Daten, mit einem Gesamtarchiv von 25 Petabyte [NCEI].

Die Gefahr, dass die zukünftige Datenmenge schneller wächst als die für die Verarbeitung dieser Daten zur Verfügung stehende Technik, fordert neue intelligente Lösungen zur Verarbeitung und Speicherung von großen Datenbeständen.

### 6.4 Wissen aus komplexen Daten

Ein weiteres Problem stellt sich bei der Aufgabe nicht relationale Daten wie Texte, Bilder, Multimedia, social Media, Blogs und andere Web-Daten zu verarbeiten. Einfache Data Mining Methoden wie Klassifikation oder Clustering reichen lange nicht mehr aus, um Wissen aus solchen komplexen

Daten zu generieren. Eine wichtige Quelle von Informationen ist zum Beispiel die Erkennung von Bewegungen und vom Verhalten von Objekten und Personen im Internet, um räumliches und zeitliches Wissen und Abhängigkeiten zu entdecken. [YangWu S. 599-600]. Mit steigender Aktivität im Internet und 3,5 Milliarden Social-Media nutzenden Menschen weltweit im Jahre 2019 [DigRep], ist mit Sicherheit zu sagen, dass das Data Mining sich in Richtung der Verarbeitung von komplexen Daten entwickeln wird.

## 7. Zusammenfassung und Fazit

Data Mining ist heutzutage eine viel dokumentierte, in der Fachliteratur beschriebene und in der Wirtschaft und Praxis angewendete Methode der Datenanalyse.

Auch wenn man sich in Zukunft noch dringend mit Themen wie Privatsphäre und eine universelle intelligente Data Mining Methode auseinandersetzen muss und möglicherweise ein Durchbruch in der Art der Speicherung der Daten benötigt wird, um mit ihnen klarkommen zu können, ist es eindeutig, dass mit dem weltweiten Wachstum an Daten, Data Mining für Wirtschaft und Forschung ein sehr großer Bestandteil werden wird und sich vollständig etablieren wird.

## Literaturverzeichnis

**[Petersohn 2005]** H. Petersohn: Data Mining - Verfahren, Prozesse, Anwendungsarchitektur, Oldenbourg Wissenschaftsverlag GmbH, 2005

**[Knobloch&Weidner 2000]** B. Knobloch, J. Weidner, Eine kritische Betrachtung von Data-Mining-Prozessen - Ablauf, Effizienz und Unterstützungspotenziale. In: Jung, R.; Winter, R. (Hrsg.):
Data Warehousing 2000. Methoden, Anwendungen, Strategien, Physica, S. 345-365, 2000

**[Alpar 2000]** P. Alpar, Joachim Niedereichholz (Hrsg.), Data Mining im praktischen Einsatz - Verfahren und Anwendungsfälle für Marketing, Vertrieb, Controlling und Kundenunterstützung, Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, 2000

**[Drees 2016]** B. Drees, Text und Data Mining: Herausforderungen und Möglichkeiten für Bibliotheken, Perspektive Bibliothek 5.1, S. 49-73, 2016

**[Larose 2014]** Daniel T. Larose and Chantal D. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, John Wiley & Sons, Inc., 2014

**[Runkler 2010]** T. Runkler, Data Mining - Modelle und Algorithmen intelligenter Datenanalyse, 2. Auflage, Springer Verlag, 2010

**[YangWu 2006]** Q. Yang, X. Wu, 10 Challenging Problems in Data Mining Research**,** (Hrsg.): International Journal of Information Technology & Decision Making Vol. 5, No. 4, World Scientific Publishing Company, S. 597–604, 2006

**[Fayyad et. Al 1996]** U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996

## Webverzeichnis

**[IDC]** D. Reinsel, J. Gantz, J. Rydning, The Digitization of the World -From Edge to Core https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf , 2018, - abgerufen am 02.01.2020

**[WorldStat]** https://www.internetworldstats.com/stats.htm - abgerufen am 02.01.2020

**[NCEI]** https://www.ncei.noaa.gov/ , https://datamakespossible.westerndigital.com/satellite-data-from-outer-space/ - abgerufen am 02.01.2020

**[DigRep]** - https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates - abgerufen am 02.01.2020