JUST DO IT
HÖHERE ABTEILUNG FÜR
INFORMATIONSTECHNOLOGIE

tgm
Die Schule der Technik

Mündliche Matura

Datamining & Datawarehousing

INSY
5AHITM 2015/16

Daniel Melichar

Version: 0.1
Betreuer: Michael Borko

Inhaltsverzeichnis

1 Einführung

Der Term *data mining* beschreibt das Extrahieren von Information aus relevanten Ansammlungen von Daten. Hierbei wird vor allem auf Muster und Trends geachtet. Das Data Mining ist ein Teil des so gennanten *Knowledge Discovery From Data* oder *KDD* Prozess. Im heutigen Zeitalter sind wir mit enormen Mengen an Informationen, welche als Dateien abgespeichert werden, umgeben. Durch die unterschliedlichsten Branchen der Wirtschaft werden weltweit eine gigantische Anzahl an Information bezüglich Finanzgeschichte, Produktbeschreibungen, Performance Logs, Kundenfeedback und vielem mehr, produziert. Der KDD soll eine Hilfestellung geben, um mehr von den Daten zu erhalten, hierbei werden folgende Schritte durchgeführt:

1. **Daten Reinigung** wo inkonsistente Daten verworfen werden
2. **Daten Integration** wo mehrere Datenquellen zu einer vereinigt werden
3. **Daten Selektion** wo Daten, welche für die analyse relevant sind, von einer Datenbank geholt werden
4. **Data Mining** wo Algorithmen angewandt werden um Datenmuster zu erkennen
5. **Musterevaluierung** wo die *interessanten* Muster endgültig identifiziert werden.
6. **Präsentation** wo die Daten für User dargestellt werden.

Diese Schritte (siehe Abbildung 1) können allgemein für jede Art an Daten angewandt werden, so lang wie diese auch für die Applikation relevant sind. In den meisten Fällen werden die Informationen für die Data Mining Applikationen von Datenbanken, Data Warehouses, und Daten die bei Transaktionen entstehen, bezogen. Es besteht aber auch die Möglichkeit Data Mining bei anderen Arten von Daten durchzuführen, z.B. Datenstreams, sequenzielle Daten, Graph oder Netzwerk Daten, Text Daten, Multimedia Daten und das WWW.

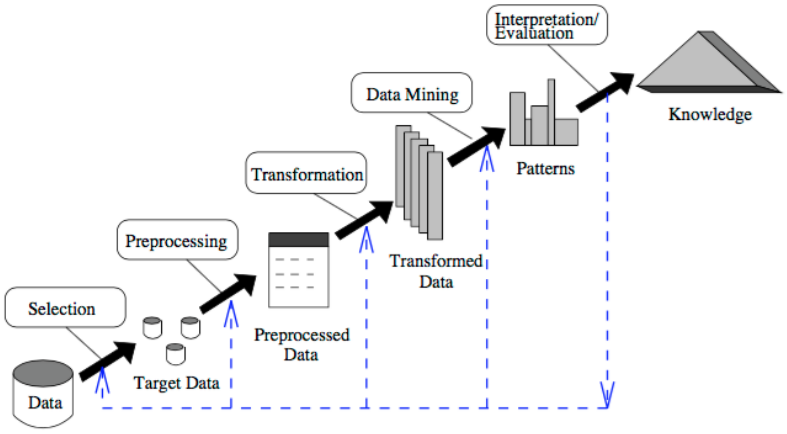


Abbildung 1: Typischer KDD Prozess [6]

Eine Suchmaschine (z.B. Google) erhält hunderte von Anfragen jeden Tag. Jede einzelne dieser Anfragen kann als Transaktion gesehen werden. Bei dieser Transaktion beschreibt der Verwender des Service (auch User gennant), welche Information er erhalten will. Bei Data Mining geht es unter anderem auch darum Informationen aus Daten zu erhalten die nicht offensichtlich sind und bei welchen spezifische Algorithmen mehr rausholen können. So könnten Algorithmen mit den Anfragen herausfinden wo der User wohnt, welchen Beziehungsstatus er momentan hat, sein Alter, und vieles mehr.

In Abbildung 2 kann gesehen werden wie Datamining klassifiziert ist und in welcher Beziehung es zu anderen Konzepten des Datenmanagement steht.

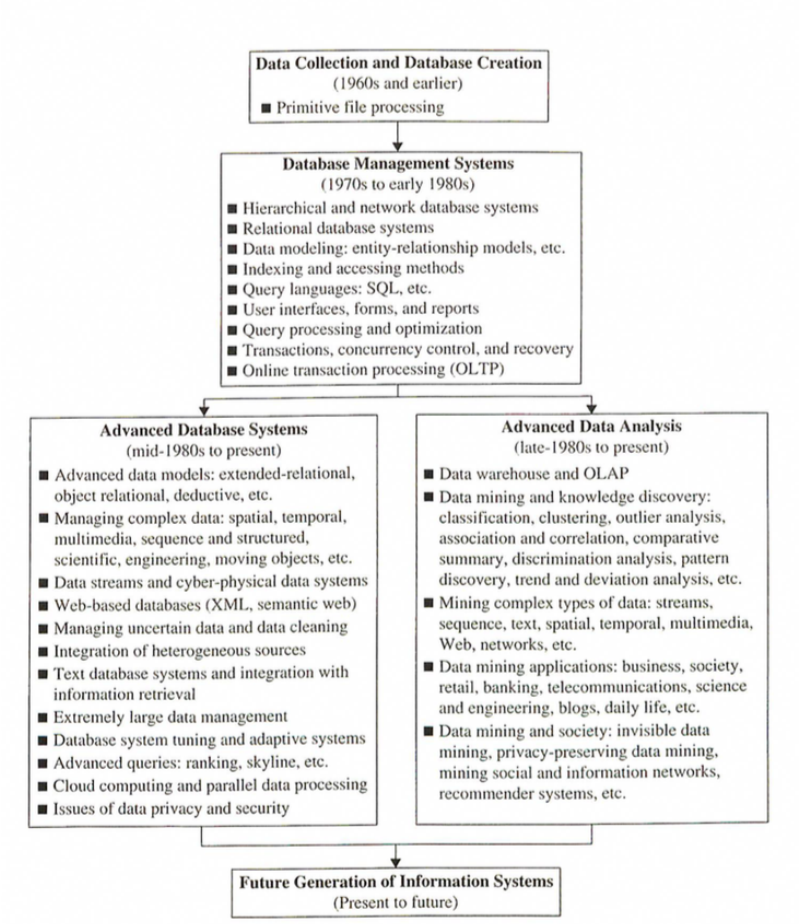


Abbildung 2: Entwicklung von Datenbank Technologien [8]

1.1 Data Mining bei RDBMS

Ein Datenbanksystem, auch als Datenbankmanagementsystem (DBMS) bekannt, besteht aus einer Ansammlung an zusammenhängenden Daten und Software, welche den Zugriff und die Verwaltung dieser ermöglichen. Diese Software stellt Mechanismen zur Verfügung, welche für die Definition einer Struktur der Daten, für die Einhaltung von Konsistenz, für die Sicherheit der Daten selbst bei unerwarteten Abstürzen des Systems, verwendet werden

Grundsätzlich werden bei so genannten relationalen Datenbanken die Daten in Tabellen eingeordnet. Die Tabelle besteht aus Attributen *(columns oder fields)* und speichert normalerweise eine große Anzahl an Tupeln *(records and rows)*. Jede dieser Einträge in der Tabelle ist eindeutig identifizierbar durch einen so genannten *unique key*. Ein semantisches Model, wie ein *entity-relationship (ER)* Model, werden oft erstellt um die Daten und deren Beziehungen in Tabellenform darzustellen.

Der Zugriff auf die relationalen Daten erfolgt mittels Datenbank Queries, welche in einer relationalen Abfragsprache (z.B. SQL) oder mittels einem graphischem Interface geschrieben werden. Diese Abfragen ermöglichen es spezifische Daten aus einer enormen Anzahl zu erhalten. Eine beispielhafte Anfrage an das DBMS ist *„Zeige mir eine Liste von allen Dingen, die im letzten Quartal verkauft wurden“* oder mittels den integrierten Aggregatfunktionen *„Welcher Verkäufer hatte die größte Anzahl an Empfelungen“*.

Mittels Data Mining und den Daten der relationalen Datenbank kann mehr aus den Daten herausgeholt werden, wie aktuelle Trends oder gewisse Muster. Zum Beispiel ist es möglich das Kaufmuster einer Person zu erkennen um die Kreditwürdigkeit basierend auf Alter, Gehalt und bisherige Kredite zu prognostizieren.

1.2 Data Warehouses

W.H. Inmon beschreibt ein Data-Warehouse als Ansammlung verschiedener Information aus vielen unterschiedlichen Quellen, diese sind dann in einer dafür spezifizierten Datenbank gespeichert [5]. Auch hier werden die Daten zuerst gesäubert, dann in die Datenbank integriert, transformiert, geladen und periodisch erneuert. Die Daten werden auch in verschiedene große Kategorien geschlichtet. Es werden aber nicht alle Daten gespeichert, nur jene, die das Objekt am besten beschreiben [8].

Ein Data Warehouse ist normalerweise in einer multidimensionalen Daten Struktur der *Data Cube* gennant. Bei diesem Data Cube ist jede Dimension ein oder mehrere Attribute, jede Zelle speichert den Wert einer Aggregat Funktion der Daten (d.h. vereinfachte Daten).

In Abbildung 3 kann der typischer Aufbau eines Data Warehouses gesehen werden. In dieser werden auch die benötigten Schritte veranschaulicht.

Dank der zur Verfügung gestellten multidimensionalen Datenansicht und der Verwendung von vereinfachten Daten ist ein Support für OLAP (*Online Analytical Processing Operations*) gegeben [8]. Diese erlauben es zu spezifizieren wie sehr die Daten vereinfacht sind, so könnte beispielsweise aus der selben Menge an Information Prognosen auf Bundesebene, bzw. über ganze Staaten gegeben werden. Allgemein haben Data Warehouses zwar einen recht guten Support bezüglich der Analyse von Daten, es müssen aber oft zusätzliche Tools für detaillierte Analysen hinzugefügt werden. Das multidimensionale Data Mining (auch als *Exploratory Multidimensional Data Mining* bekannt [8]) bezieht sich auf Kombinationen von Information mit unterschiedlichen Detailgrad. Daraus entsteht eine potentiell größere Chance interessante Muster zu finden.

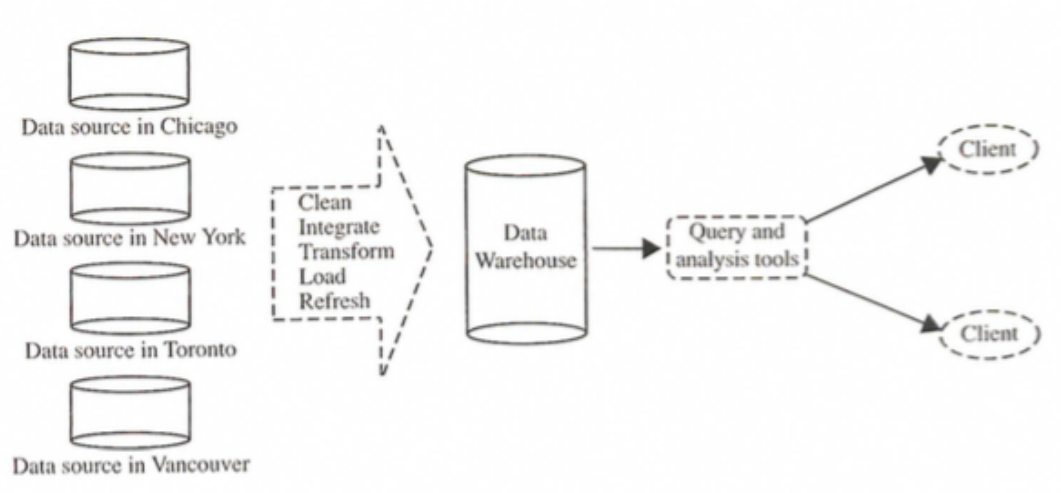


Abbildung 3: Beispielhafter Aufbau eines Data Warehouses [8]

1.3 Data Mining bei Transaktionen

Grundsätzlich ist jeder Eintrag in einer Datenbank für Transaktionen entweder ein Einkauf, ein Klicken auf der Webpage, oder ein Kauf eines Flugtickets. Diese Einträge sind auch eindeutig identifizierbar. Auch die Verwendung von zusätzlichen Tabellen ist gestattet welche Information über die Transaktionen beinhalten, wie z.B. eine Beschreibung des Artikels. Es können auch Analysen durchgeführt werden bezüglich Artikel welche oft zusammen gekauft wurden. Hierdurch könnten dann Rabatte gesetzt werden und somit letztendlich erhält man dann mehr Umsatz.

1.4 Verwendete Technologien

Bei der Entwicklung von Data Mining Applikationen werden oftmals die unterschiedlichsten Konzepte und Technologien verwendet [8]. Unter anderem zählen hierzu: Statistik, Machine Learning, Mustererkennung, Datenbank und Data Warehouse Systeme, Informationsbeschaffung, Visualisierung, und komplexe Algorithmen (siehe Abbildung 4).

1.4.1 Statistiken

Die Statistik beschreibt die Sammlung, Analyse, Interpretation bzw. Erklärung, und die Präsentation von Daten. Ein statistisches Model besteht aus mathematischen Funktionen welche das Verhalten von Objekten in einem gewissen Bereich von zufälligen Variablen beschreiben. Diese Modelle werden oft verwendet um die zu klassifizieren. Diese statistischen Modelle können das Ergebnis von Data Mining sein, oder basieren auf dieser Klassifizierung sein. Mittels der Statistik können verschiedene Muster der selben Daten fest gelegt werden.

Eines dieser statistischen Tools ist das Bayesian Netzwerk, welches die Beziehungen zwischen Variablen darstellt. Ein beispielhaftes Bayesian Netzwerk für ein medizinisches Problem kann in Abbildung 5 gesehen werden. Die einzelnen Knoten repräsentieren Variablen oder Zustände und

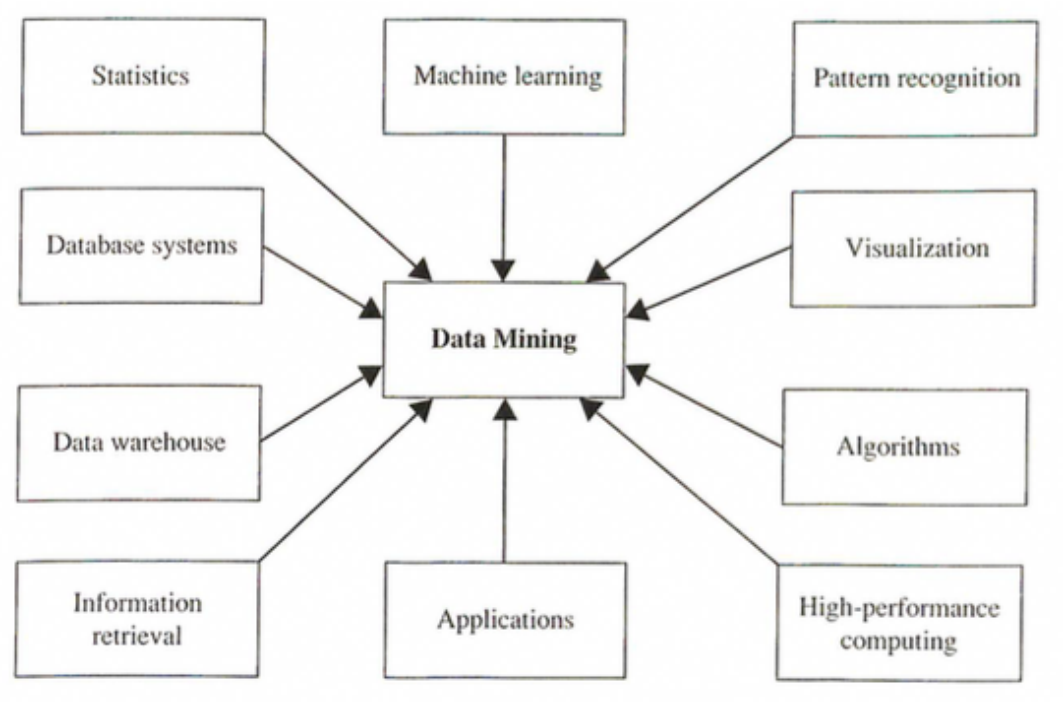


Abbildung 4: Data Mining Technologien [8]

die Abhängigkeiten werden durch die Pfeile dargestellt. Es ist erkennbar, dass das Alter, der Beruf, und die Ernährung einen Einfluss auf die Krankheit nehmen, welches wiederum einige Symptome hervorruft.

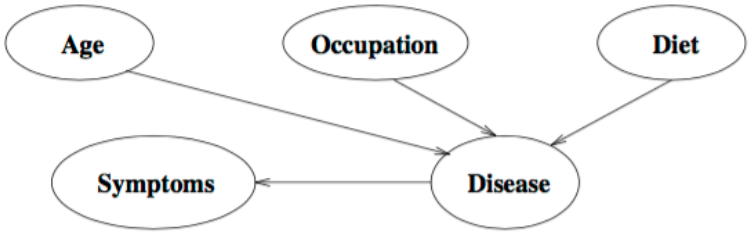


Abbildung 5: Einfaches Bayesian Netzwerk [6]

1.4.2 Machine Learning

Das Machine Learning (oder *maschinelles Lernen*) beschreibt wie Systeme sich etwas beibringen können, oder eine bessere Performance erzielen können, basierend auf Daten. Hierbei ist eine gewisse Selbstständigkeit der Systeme und eine intelligente Entscheidungsfindung von großen Wert. Eine beispielhafte Anwendung von maschinellem Lernen ist die Erkennung der Handschrift einer Person anhand von einigen Stichproben. Die folgenden Punkte beschreiben klassische Probleme von maschinellen Lernen, die auch in Relation zu Data Mining stehen.

Überwachtes Lernen beschreibt die Klassifikation von Daten (*engl.: classification*), wobei diese

bereits allgemein gekennzeichnet sind.

Unüberwachtes Lernen beschreibt die Einteiling von Daten (*engl.: clustering*), wobei diese hier nicht gekennzeichnet sind.

Fast-Überwachtes Lernen beschreibt eine Technik, bei welcher gekennzeichnete und nicht gekennzeichnete Daten zum Einsatz kommen.

Aktives Lernen beschreibt eine Technik, bei welcher User auf den Lernprozess während der Laufzeit Einfluss nehmen können.

Die häufigste Methode des maschinellen Lernens basiert auf Entscheidungsbäumen (*engl.: decision tree*), bei welchem die Klasse eines Objekts anhand von Negierung herausgefunden werden kann. Diese Entscheidungsbäume entstehen durch das durchgeführte Training beim Data Mining. In Abbildung 6 ist beispielhafter Entscheidungsbäum erkennbar. Dieser findet die Kilometeranzahl eines Autos anhand von der Größe, Getriebe, und Gewicht. Die Klassen der Kilometeranzahl wird durch Rechtecke repräsentiert. Anhand des Baums kann z.B. fest gelegt werden, dass ein Mittelgroßes, automatisches Auto eine Mittlere Kilometeranzahl haben wird.

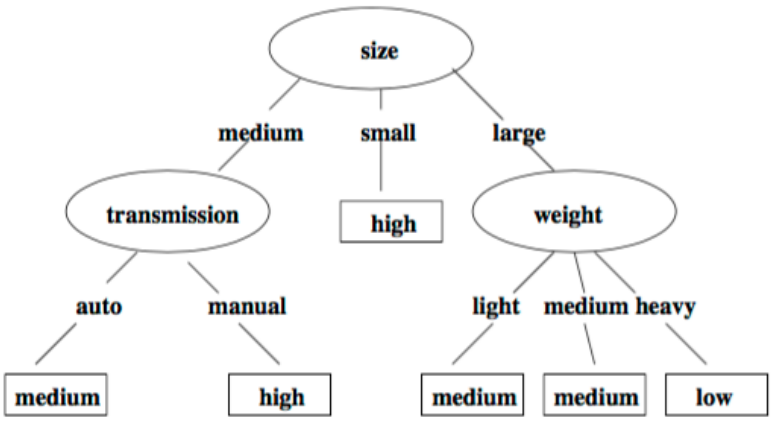


Abbildung 6: Einfacher Entscheidungsbaum [6]

1.4.3 Informationsbeschaffung

Die Informationsbeschaffung (oder *information retrieval*) befasst sich mit der Suche nach einfachen Text- oder Multimediale-Dokumente, bzw. der Information in Dokumenten. Im Gegensatz zu Datenbanksystemen wird hier von Dokumenten ausgegangen, die unstrukturiert und einfach zu verwenden (d.h. ohne Verwendung von komplexen Abfragestrukturen wie SQL) sind. Eine beispielhafte Anwendung ist die Erstellung eines so genannten Wahrscheinlichkeitsmodels, welches die Anzahl an Wörtern in einem Dokument misst, und somit eine Dichtefunktion erzeugt werden kann.

2 Entwurf und Implementierung von Datenbank Applikationen

Der Entwurf von Applikationen, die mittels dem *Knowledge Discovery Process oder KDD* mehr Information aus den Daten erhalten als auf den ersten Blick erkenntlich ist, folgt grundsätzlich dem in Abbildung 1 gezeigten Schema. Eine andere Art der Darstellung zum allgemeinem KDD Prozess, in welchem auch Sichtbar ist wie sich die Daten verändern, kann in Abbildung **??** betrachtet werden.

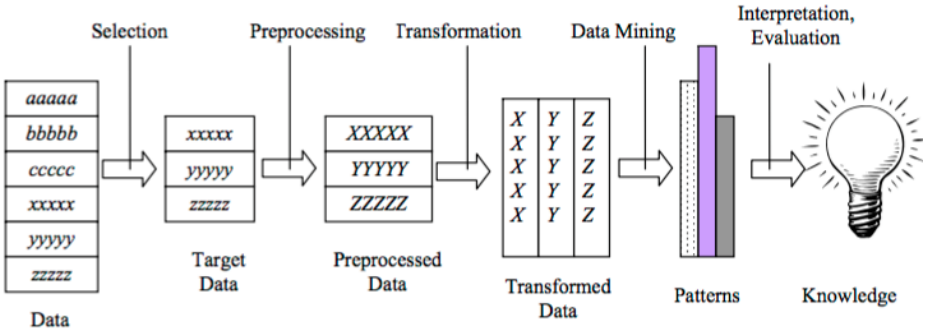


Abbildung 7: KDD Prozess [7]

2.1 Selektion

Grundsätzlich können die Daten für die Entwicklung von Wissen aus internen oder externen Quellen kommen (siehe Tabelle 1). Ein funktionsfähiges Informationssystem verwendet diese Menge an Daten und führt die notwendigen Prozesse durch. Interne Daten werden hauptsächlich für tagtägliche Operationen verwendet, da in den meisten Fällen keine vorherigen Zustände gespeichert werden [1]. Aufgrund dessen werden bei einer schlechten Auswahl der Daten, oder bei einer nicht vorhandenen Überprüfung der Daten, einige Inkonsistenzen auftreten. Die Qualität und Zuverlässigkeit der Daten ist ausschlaggebend für die Analyse. Die internen Daten sollten bei gegebener Dokumentation durch die Firma einen qualitativ höheren Wert haben als externe Daten. Daher muss meist bei Verwendung von externen Quellen ein weiterer Schritt hinzugefügt werden, welcher die Daten auf Richtigkeit und Relevanz testet[1]. Der große Vorteil von externen Quellen ist definitiv die enorme Anzahl von Daten (auch *public data sets* gennant[2]).

2.2 Pre-Processing und Transformation

Nach der Beschaffung der Daten müssen diese für die Analyse, bzw. für das eigentliche Mining, vorbereitet werden. Oft werden die internen und externen Daten in für eine Analyse in einem dafür spezifisch bereitgestellten Data Warehouse gespeichert [1]. Im Warehouse werden die Daten noch

[1]Es sollte nicht davon ausgegangen werden, dass interne Quellen vollständig Richtig sind. Auch bei internen Quellen muss ggf. eine Überprüfung durchgeführt werden.

[2]Sammlung von öffentlichen Datensätzen: `https://github.com/caesar0301/awesome-public-datasets`

Daten Quelle	**Beispiel**	**Charakteristik**
Intern	Datum an welchen ein Produkt erstellt wurde	Kontrolliert durch eine Firma, meist qualitativer.
Extern	Ort an welchem das Produkt gekauft wurde	Könnte Ungenauigkeiten und Inkonsistenzen beinhalten

Tabelle 1: Interne und Externe Datenquellen [1]

analysiert, transformiert und ein Aggregat gebildet. Die Art der Aggregation hat einen großen Einfluss auf das finale Resultat, deswegen muss diese sorgfältig Ausgewählt werden.

In manchen Fällen ist es Sinnvoll die Identifikatoren der Daten in einem Data Warehouse System als Gruppen zusammenzufassen. Bei Produkten aus dem Einzelhandel können Gruppen Anhand von der Art des Produkts (Lebensmittel, Reinigungsmittel, Baustoffe), der Farbe, oder andere allgemeine Eigenschaften, gebildet werden.

Bei diesem reinigenden Prozess werden einzelne Fälle bearbeitet, welche die Qualität des Resultats beeinflussen können. Unter anderen zählen hierzu: fehlende Werte, wobei hier zwischen *echten fehlenden Werten* oder *Werten die nicht gespeichert sind* zu unterscheiden ist; Werte, die stark von anderen abweichen; oder falsche Werte.

2.3 Data Mining Methoden

Die Resultate der Methoden des Data Minings sind Voraussagungen basierend auf den gegebenen Daten. Es gibt unzählige Methoden um auf ein Resultat zu kommen [1] alle mit verschiedenen Anwendungsfällen und Durchführungszeiten. Mittels diesen Voraussagungen sollen Modelle für zukünftige Applikationen entwickelt werden.

2.3.1 Supervised- und Unsupervised Learning

Die Supervised Learning Methode erstellt eine Funktion von bereits gegebenen Daten. Diese bereits gegebenen Daten bestehen grundsätzlich aus der Information und dem Erwartungswert. Das Ergebnis der Funktion ist eine Regression der Daten (also ein Model für einen späteren Zeitpunkt) oder eine Gruppierung der Eingabedaten. Dies ist das generelle Ziel der Methode: die Voraussage zu einem gewissen Zeitpunkt anhand der Trainingsdaten. Die Unsupervised Learning Methode hingegen soll prinzipiell die Organisation von Daten feststellen, also wie diese zueinander in Beziehung stehen. Daher sollte die Unsupervised Learning Methode verwendet werden, wenn Daten *verstanden* werden sollen und die Supervised Learning Methode, wenn Voraussagungen getroffen werden sollen. Der wichtigste Punkt, der bei der Methode des Supervised Learning erfüllt werden muss, ist eine gegebene Zielvariable (z.B. *kauft* oder *kauft nicht*). Im Gegensatz dazu ist beim Unsupervised Learning keine Zielvariable gegeben [10].

Ein beispielhaftes Anwendungsgebiet einer Supervised Learning Methode in einem Data Mining System ist die Politikwissenschaft. Mittels der Methode könnte somit die Frage geklärt werden ob ein Wähler für einen Kandidaten stimmt basierend auf Onlinemedien.

2.3.2 Decision Trees

Mit der Aufstellung eines geordneten und gerichteten Entscheidungsbäum werden Regeln für gewisse Fälle aufgestellt [4]. Sie veranschaulichen hierarchisch, aufeinanderfolgende Entscheidungen. Diese Methode des Data Mining hat viele Anwendungsgebiete, für einige gibt es bereits erstellte Bäume, für andere gibt es bereits formale Regeln. Die Einsatzgebiete reichen von der Diagnosemedizin über die Finanzanalyse bis hin zur Astronomie. Hierbei gilt es zu unterscheiden zwischen Klassifikationsbäumen, welche eine Auswahl von Kategorien und deren Beziehungen darstellt, und Regressionsbäumen, welche Prognosen und die Zuteilung von einem Objekt zu einem Wert darstellt. Jeder Knoten des Baums kann mehrere Unterknoten besitzen.

Bei bereits klassifizierten Daten können Entscheidungsbäume automatisch Erzeugt werden, hierfür wurden verschiedene Algorithmen entwickelt, welche sich durch das Kriterium der Unterteilung, unterscheide [4].und un Einige der implementierten Baumtypen sind CARTs (Classification and Regression Trees), CHAIDs (Chi-Sqare Automatic Interaction Detectors), sowie der ID3-Algorithmus (Iterative Dichotomiser 3). Alle diese Verfahren funktionieren nach demselben Schema, bis auf das Attributauswahlverfahren, wodurch die Unterteilung des Baumes gesteuert wird.

Anhand von Entscheidungsbäumen kann explizites Wissen zur Klassifikation und Vorhersagen getroffen werden [4].

2.3.3 Cluster Analysis

Bei der Cluster Analysis geht es darum die Information der Daten in verschiedene Gruppen einzuteilen. Hierbei ist die Einteilung in den einzelnen Gruppen von essentieller Bedeutung. Die Methode hat kein internes Selektionsverfahren, das bedeutet, dass alle übergebenen Daten in der Analyse inkludiert werden [1].

Eine konkrete Implementierung ist die hierarchische Cluster Analyse. Bei dieser werden die einzelnen Objekte zunächst in eine einzelne Kategorie eingeteilt. Im weiterem Verlauf sucht dann immer wieder der Algorithmus nach ähnlichen Paaren und somit entstehen exponentiell größere Gruppen. Die Ähnlichkeit von zwei Objekten kann anhand von mehreren Methoden gemessen werden, wie z.B. dem aus der Statistik bekannten Median [1].

Eine weitere konkrete Implementierung ist die K-Means Algorithmus, welcher von einer Gruppe mit K Objekten ausgeht. Der Wert von K kann durch die Software bestimmt werden oder durch einen Analysten. Der Algorithmus sucht nach dem nächsten Cluster und verbindet die beiden. Die Methode ist schneller als die hierarchische und kann mit großen Eingabewerten umgehen, da die Anzahl der Vergleiche kleiner ist, und somit die Entscheidung, welche Gruppen kombiniert werden können, rascher getroffen werden kann [1].

2.4 Interpretation und Evaluation

Die Interpretation und Evaluierung beschreibt grundsätzlich die Darstellung in einer simpleren (meist grafischen) Form. Anhand dieser sollen die Ergebnisse leicht Erkennbar sein. Die Visualisierung wird in Sektion 3 beschrieben.

3 Präsentation von Inhalten mittels Web-Applikation

Visuelles Data Mining basiert auf den Techniken der Informationsvisualisierung, einem Gebiet der Computergraphik, welches sich, wie der Name schon sagt, mit der Darstellung von Information beschäftigt. Im Gegensatz zur wissenschaftlichen Visualisierung dient die Informationsvisualisierung nicht zur Darstellung von chemischen oder physikalischen Daten, Messwerten oder Simulationen, sondern vielmehr zur Visualisierung von Beziehungen, Mustern und vor allem Information [8]. Die Methoden der Informationsvisualisierung können in drei Kategorien [8] unterteilt werden: präsentative Techniken, Techniken zur sogenannten *bestätigenden* Analyse (engl.: *confirmative analysis*) und Techniken zur erkundenden Analyse (engl.: *explorative analysis*). Aufgabe der Visualisierung ist es, die Daten in einer geeigneten Form darzustellen, die eine Bestätigung oder Entkräftung der aufgestellten Prognose erlaubt.

3.1 Datentypen und Dimensionalität

Große Datensätze wie beispielsweise Screeningdaten oder Resultate kombinatorischer Experimente bestehen aus einer großen Anzahl an Einzeleinträgen den sogenannten Datenrecords, die sich ihrerseits aus einer definierten Anzahl an Variablen, den Dimensionen, zusammensetzen [12]. In der Informationsvisualisierung wird die Zahl der Variablen auch als Dimensionalität des Datensatzes bezeichnet. Nach Shneiderman [12] können Datensätze ein- zwei- oder auch multi-dimensional sein oder auch aus komplexeren Datentypen wie Texten, Hypertexten, Hierarchien, Graphen oder Algorithmen bestehen.

Eindimensionale Datensätze Typische Vertreter von eindimensionalen Datensätzen sind zeitabhängige Daten. Dabei können jedem Punkt auf der Zeitskala ein oder mehrere Messwerte zugeordnet werden.

Zwei- und dreidimensionale Datensätze Ein typisches Beispiel für solche Datensätze stellen geographische Karten dar. Zwei- und dreidimensionale Datensätze werden in der Regel durch einfache x-y- bzw. x-y-z-Plots visualisiert.

Multidimensionale Datensätze vHäufig bestehen Datensätze aus mehr als drei Dimensionen und können daher nicht mit Hilfe von zwei- oder dreidimensionalen Plots dargestellt werden. Multidimensionale Datensätze können in der Regel mehrere Hundert bis Tausend Dateneinträge enthalten. Sie werden unter anderen durch automatisierte und in relationalen Datenbanken gespeichert. Diese Daten können nur mit Hilfe weiterentwickelter Visualisierungstechniken dargestellt werden, da das effektive Mapping der zahlreichen Dimensionen auf einen zweidimensionalen Bildschirm ein schwieriger Vorgang ist.

Spezielle Datentypen Nicht alle Datentypen können durch Angabe der Dimensionalität beschrieben werden. Dies trifft beispielsweise auf digitale Texte und Hypertexte zu, deren Analyse vor allem im Bereich des World Wide Web hohe Bedeutung beizumessen ist. Diese Datentypen können nicht sinnvoll in Form von Zahlen dargestellt werden, weshalb viele Visualisierungstechniken zur Darstellung dieser Daten nicht eingesetzt werden können. Eine

weitere Klasse von Datensätzen stellen Hierarchien und Graphen dar, die auf Beziehungen zwischen einzelnen Datenpunkten basieren.

3.2 Die Visualisierungstechniken

Im Laufe der letzten dreißig Jahre wurden, wie bereits erwähnt, zahlreiche Techniken zur Visualisierung von Informationen und Daten entwickelt und die Zahl der Visualisierungstechniken wird auch in Zukunft weiter ansteigen. In der folgenden Sektion werden solche Ansätze erwähnt, die der Darstellung von multivariaten und multidimensionalen Datensätzen dienen. Multivariate Visualisierungstechniken können dabei in Abhängigkeit ihrer zugrundeliegenden Visualisierungsprinzipien in fünf Kategorien unterteilt werden [8] - geometrische Techniken, Icon- und Glyph-basierte Techniken, Pixel- und Voxel-orientierte Systeme, hierarchische Techniken und Techniken, die auf sogenannten Graphen basieren. Darüber hinaus existieren auch zahlreiche hybride Ansätze, die sich durch Kombination verschiedener Visualisierungstechniken aus den genannten Bereichen ergeben.

3.2.1 Geometrie-basierte Ansätze

Scatterplots und Scatterplot-Matrizen Scatterplots zählen wahrscheinlich zu den bekanntesten Data Mining-Visualisierungstechniken und werden standardmäßig von vielen Statistik- und Tabellenkalkulationsprogrammen unterstützt. Dabei kommen sowohl zwei- als auch dreidimensionale Darstellungen zum Einsatz. Im Regelfall können mittels Scatterplots zwei bzw. drei Datendimensionen dargestellt werden, wobei jede Datendimension auf eine der zwei- bzw. drei orthogonalen Achsen abgebildet wird. Um Datensätze mit höherer Dimensionalität zu visualisieren, werden unter anderen sogenannte Scatterplot-Matrizen verwendet. Dabei kann beispielsweise ein vierdimensionaler Datensatz durch eine 4x4-Matrix von Scatterplots dargestellt werden (Abbildung 8). Häufig werden Scatterplot-Darstellungen durch interaktive Techniken wie Zoom erweitert, um eine komfortablere Analyse des Datensatzes zu gewährleisten.

Barcharts und Histogramme Balkendiagramme (engl. Barcharts) werden in erster Linie zur Präsentation von Daten eingesetzt. Darüber hinaus existieren jedoch auch weiterentwikkelte Techniken, wie beispielsweise 3D-Barcharts mit variierenden Formen, Farben und Größen, Survey Plots sowie Histogramm-Matrizen, die im Bereich des Data Mining zum Einsatz kommen. eingesetzt.

Parallel Coordinates Der wohl prominenteste Vertreter geometrischer Visualisierungstechniken ist die Parallel Coordinates-Darstellung. Bei dieser Technik werden die einzelnen Dimensionen durch vertikale Achsen repräsentiert, wobei der entsprechende Wertebereich der Variablen entlang der einzelnen Achsen aufgetragen ist. Jeder Datenpunkt bzw. jedes Datenobjekt wird dabei durch eine polygonale Linie dargestellt, welche die Achsen an den entsprechenden Stellen schneidet.

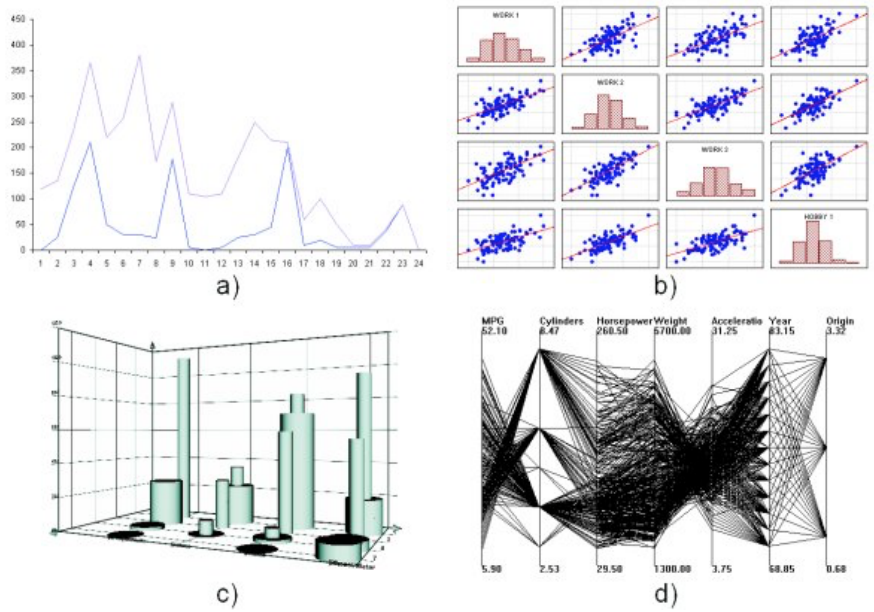


Abbildung 8: Geometrische Visualisierungstechniken: a) Multiple Liniengraphen [12], b) Scatterplot-Matrix [12], c) 3D-Balkendiagramm [12] d) Parallel Coordinates [8].

3.2.2 Pixel- und Voxel-orientierte Techniken

In Pixel-orientierten Ansätzen wird jeder einzelne Dimensionswert einem farbigen Pixel zugeordnet. Die Pixel werden darüber hinaus in Abhängigkeit von den jeweiligen Dimensionen gruppiert und in separaten Regionen dargestellt [9]. Die Werte der einzelnen Dimensionen werden durch die Farbe des Pixels repräsentiert. Da lediglich ein Pixel pro Datenobjekt benötigt wird, können mit Hilfe dieser Technik die derzeit höchste Anzahl an Datenpunkten gleichzeitig dargestellt werden. Die bekanntesten Vertreter dieser Visualisierungsform sind die sogenannte Recursive Pattern-Technik und die Circle Segment-Technik [2]. Der zweidimensionale, Pixel-orientierte Ansatz kann darüber hinaus auch auf drei Raumdimensionen erweitert werden. Das dreidimensionale Analogon zum Pixel ist dabei das sogenannte Voxel. Voxel- bzw. texturbasierte Ansätze erlauben aufgrund der zusätzlichen dritten Dimension die Darstellung noch größerer Datensätze.

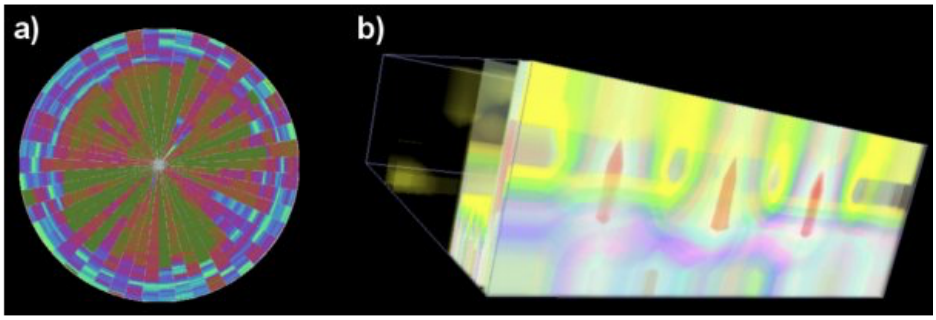


Abbildung 9: Pixel- und Voxel-basierte Visualisierungstechniken [2]: a) Circle Segment-Technik, b) Voxel-basierte Visualisierung.

3.2.3 Icon- und Glyph-basierten Techniken

Eine andere Klasse von explorativen Visualisierungstechniken stellen die sogenannten Icon- bzw. Glyph-basierten Ansätze dar. Diese Techniken werden vor allem zur Darstellung von diskreten, multivariaten Daten eingesetzt. Unter einem Glyphen versteht man dabei ein graphisches Objekt, welches ein einzelnes multivariates Datenobjekt repräsentiert. Bei der Generierung der Icons bzw. Glyphen werden die diversen Datendimensionen eines Datensatzes in systematischer Weise den verschiedenen graphischen Attributen wie Form, Farbe, Größe, Orientierung, Textur, etc. des graphischen Objekts zugeordnet. Dieses Abbilden (engl. Mapping) der Dimensionen auf die sogenannten retinalen Eigenschaften [3] wird auch als visuelles Mapping bezeichnet.

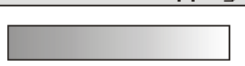
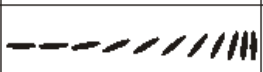
Graphisches Attribut	Dimensionalität	Kontinuierliche Daten Quantitatives Mapping	Diskrete Daten Nominales Mapping
Farbe	*max. 3 Dimensionen (3 bei color opponent)*		
Form	*max. 3 Dimensionen (x, y, z)*		
Orientierung	*3 Dimensionen (x, y, z)*		
Textur	*3 Dimensionen (Kontrast, Größe, Orientierung)*	*Textur-Morphing (ungünstig)*	
Bewegung	*mind. 2-3 Dimensionen*		
Blinken	*1 Dimension*	*Stufenlose Blinkgeschwindigkeit*	*Blinken, Nicht-Blinken, klar definierte Stufen*

Abbildung 10: Eigenschaften in der Glyph-basierten Visualisierung.

3.2.4 Hierarchische und Graph-basierte Techniken

Hierarchische Techniken, auch Stacked Displays genannt, stellen Daten in Form von hierarchisch aufgeteilten Untereinheiten dar. Im Fall von multidimensionalen Datensätzen dienen dabei selektierte Dimensionen zur Aufteilung des Datensatzes und zum Aufbau der Hierarchie. Bekannte Vertreter sind das Dimensional Stacking sowie die Cone Tree-Technik [11].

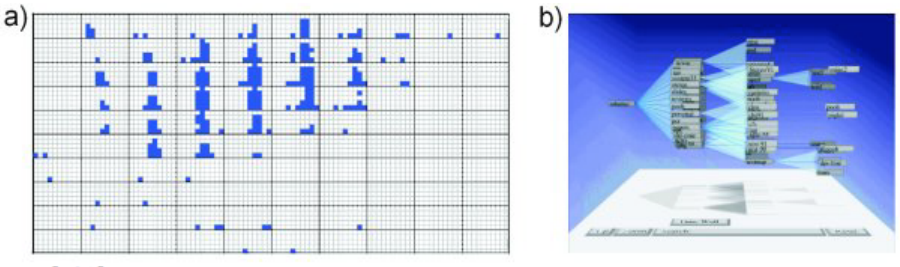


Abbildung 11: Hierarchische Visualisierungstechniken [11]: a) Dimensional Stacking, b) Cone Tree

3.3 Software Implementierung

Bei der Entwicklung von Data Mining Web-Applikationen wird die visualisierung in den meisten Fällen durch eine JavaScript Libary durchgeführt. Hierfür gibt es unzählige Frameworks, die alle für unterschiedliche Anwendungsfälle sinnvoll sind. Die meisten dieser JavaScript Frameworks sind unter GitHub verfügbar. Eine Liste von Libaries und Frameworks, die oft für Data Mining Visualisierungen verwendet werden, kann hier[3] gefunden werden.

4 Datenbankanbindung und Konsistenzerhaltung bei CMS

5 Verwaltung und Optimierung von Informationssystemen

6 Modellierung und Datenintegration bei mobilen Endgeräten

7 Datenhaltung und -weitergabe im zwischenbetrieblichen Umfeld

[3] JavaScript Visualisierungs Libaries und Frameworks: `https://gist.github.com/entaroadun/1515418`

Literatur

[1] S. C. Andrea Ahlemeyer-Stubbe. *A Practical Guide to Data Mining for Business and Industry.* John Wiley and Sons, Ltd, 2014.

[2] D. A. K. H.-P. Ankerst, M.; Keim. *A Technique for Visually Exploring Large Multidimensional Data Sets.* Proceedings Visualization 96, 1996.

[3] J. Bertin. *Semiology of Graphics.* The University of Wisconsin Press, 1983.

[4] P. D. W. Dilger. Data mining - kompaktkurs an der berufsakademie mannheim. Online, 2005. [Accessed on 27 May, 2016] `https://www.tu-chemnitz.de/informatik/KI/scripts/ss05/DM1-stud-05.pdf`.

[5] K. Farkisch. *Data-Warehouse-Systeme kompakt: Aufbau, Architektur, Grundfunktionen.* Springer DE, 2011.

[6] Y. Fu. Data mining: Tasks, techniques and applications. Online. [Accessed on 24 May, 2016] `http://academic.csuohio.edu/fuy/Pub/pot97.pdf`.

[7] B. D. D. Gary M. Weiss. Data mining. Online. [Accessed on 27 May, 2016] `http://storm.cis.fordham.edu/~gweiss/papers/data-mining-chapter-2010.pdf`.

[8] J. P. Jiawei Han, Micheline Kamber. *Data Mining - Concepts and Techniques.* MK - Morgan Kaufmann, third edition edition, 2012.

[9] D. A. Keim. *Pixel-orientated Database Visualizations.* Proceedings Tutorial ACM SIGMOD Int. Conf. on Management of Data, 1996.

[10] S. Purpura. Introduction to applied supervised learning w/nlp. Online, 2010. [Accessed on 27 May, 2016] `https://faculty.washington.edu/jwilker/tft/UWTextToolsConferencePurpura.pdf`.

[11] J. D. Robertson, G. G.; Mackinlay. *Cone Trees: Animated 3D Visualizations of Hierarchical Information.* Proceedings Human Factors in Computing Systems CHI 91 Conf., 1991.

[12] B. Shneiderman. *The Eyes Have It: A Task by Data-type Taxonomy for Information Visualization, In: Proceedings of Visual Languages.* IEEE Computer Science Press, 1996. 336 - 343.

[13] M. T. S.Z. Erdogan. A data mining application in a student database. Online, 2005. [Accessed on 27 May, 2016] `http://www.hho.edu.tr/HutenDergi/2005TEMMUZ/09_ERDOGAN_TIMOR.pdf`.

8 Appendix

8.1 Figuren

8.2 Tabellen

8.3 Listings