Institut für Parallele und Verteilte Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Diplomarbeit Nr. 3731

Visual Analytics im Kontext der Daten- und Analysequalität am Beispiel von Data Mashups

Michael Behringer

Studiengang:	Informatik
Prüfer/in:	Prof. Dr.-Ing. habil. Bernhard Mitschang
Betreuer/in:	Dipl.-Inf. Pascal Hirmer, M. Sc. Cornelia Kiefer, Dr.-Ing. Christoph Gröger
Beginn am:	15. September 2015
Beendet am:	16. März 2016
CR-Nummer:	H.1.2, H.3.3, H.5.2, I.3.8

Kurzfassung

Viele Prozesse und Geschäftsmodelle der Gegenwart basieren auf der Auswertung von Daten. Durch Fortschritte in der Speichertechnologie und Vernetzung ist die Akquisition von Daten heute sehr einfach und wird umfassend genutzt. Das weltweit vorhandene Datenvolumen steigt exponentiell und sorgt für eine zunehmende Komplexität der Analyse. In den letzten Jahren fällt in diesem Zusammenhang öfter der Begriff *Visual Analytics*. Dieses Forschungsgebiet kombiniert visuelle und automatische Verfahren zur Datenanalyse. Im Rahmen dieser Arbeit werden die Verwendung und die Ziele von *Visual Analytics* evaluiert und eine neue umfassendere Definition entwickelt. Aus dieser wird eine Erweiterung des *Knowledge Discovery*-Prozesses abgeleitet und verschiedene Ansätze bewertet. Um die Unterschiede zwischen Data Mining, der Visualisierung und *Visual Analytics* zu verdeutlichen, werden diese Themengebiete gegenübergestellt und in einem Ordnungsrahmen hinsichtlich verschiedener Dimensionen klassifiziert. Zusätzlich wird untersucht, inwiefern dieser neue Ansatz im Hinblick auf Daten- und Analysequalität eingesetzt werden kann. Abschließend wird auf Basis der gewonnenen Erkenntnisse eine prototypische Implementierung auf Basis von *FlexMash*, einem an der Universität Stuttgart entwickelten Data Mashup-Werkzeug, beschrieben. Data Mashups vereinfachen die Einbindung von Anwendern ohne technischen Hintergrund und harmonieren daher ausgezeichnet mit *Visual Analytics*.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

"Hältst Du mich für einen gelehrten, belesenen Mann?"
"Gewiss", antwortete Zi-gong. "So ist es doch?"
"Keineswegs", sagte Konfuzius. "Ich habe einfach einen Faden aufgegriffen, der mit dem Rest zusammenhängt."

- Sima Qian, *Konfuzius*

1.1 Ausgangslage und Motivation

Nach diesem alten Zitat ist der Weg zur Weisheit sehr einfach. Es legt nahe, dass mit dem richtigen Einstiegspunkt der nachfolgende Pfad lediglich bis zur gewünschten Information nachverfolgt werden muss. In der Tat ist das Konzept eines zu verfolgenden Pfades naheliegend für die Beschreibung von Analysen, doch der angesprochene initiale Faden ist heute schwerer als jemals zuvor zu finden oder zu verfolgen.

Viele Prozesse und Geschäftsmodelle der Gegenwart basieren auf der Auswertung von Daten. Durch Fortschritte in der Speichertechnologie ist die Akquisition der Daten heute problemlos möglich und wird umfassend genutzt. Um möglichst viele Daten vorrätig zu haben werden diese in den meisten Fällen jedoch nur abgespeichert, wobei die spätere Auswertung selbiger nicht berücksichtigt wird [KAF+08]. Seit vielen Jahren steigt das Volumen der gespeicherten und zu verarbeitenden Daten folglich weltweit an und die aktuelle Epoche wird historisch als Informationszeitalter eingeordnet, in welchem Daten als der essentielle Rohstoff betrachtet werden.

Im Jahr 2012 generierte *Facebook*[1] 500 Terabyte an neuen Daten, während *Amazon*[2] bis zu 26 Millionen Artikel verkaufte und drei Milliarden Suchanfragen von *Google*[3] beantwortet wurden – täglich [Con, Ama12, Goo12]. *YouTube*[4] verarbeitete 2013 pro Minute über 100 Stunden neues Videomaterial [You13]. Das 2014 vorhandene digitale Datenvolumen wird auf 4,4 Zettabyte – 4,4 Billionen Gigabyte – taxiert und verdoppelt sich alle 20 – 24 Monate [EMC, MR10].

[1] http://www.facebook.com
[2] http://www.amazon.com
[3] http://www.google.com
[4] http://www.youtube.com

Dieses exponentielle Wachstum – Van Wijk [Wij05] spricht gar von einer Datenexplosion – führt zunehmend zu Schwierigkeiten durch eine immer komplexer werdende Analyse. Die menschliche Analysefähigkeit bleibt zeitgleich nahezu konstant, weshalb die Lücke zwischen vorhandenen Daten und deren Analyse weiter anschwillt [MR10].

Dieses Phänomen ist als *Information Overload*-Problem bekannt und referenziert die Gefahr sich in diesen Datenmengen zu verlieren, da diese entweder irrelevant für die derzeitige Aufgabe sind oder auf die falsche Art verarbeitet bzw. präsentiert werden [KKM+10a]. Erschwerend kommt hinzu, dass nach Schätzungen zwischen 80 % [Gri08] und 90 % [DGS99] aller Daten unstrukturiert sind.

Um Informationen aus diesen großen Datenbeständen (vgl. Abschnitt 2.1 – Big Data) zu gewinnen existieren zwei verschiedene Ansätze. Data Mining-Algorithmen (vgl. Abschnitt 2.2 – Data Mining) suchen nach bisher unbekannten Mustern in großen Datenmengen, während die Visualisierung die hochentwickelte menschliche Wahrnehmung zur Mustererkennung verwendet und entsprechend eine Repräsentation des Datensatzes generiert. Beide Ansätze haben verschiedene Vor- und Nachteile (vgl. Abschnitt 4.5 – Bewertung der verschiedenen Ansätze) und können die angesprochenen Probleme nicht eigenständig lösen, derzeit werden infolgedessen nur etwa fünf Prozent der Daten analysiert [EMC].

In den letzten Jahren wird versucht, durch eine Kombination automatisierter und visueller Methoden, individuellere, schnellere und genauere Einsichten in vorhandene Daten zu erlangen. Dieser Ansatz wird als *Visual Analytics* bezeichnet.

Das übergeordnete Ziel dieser Arbeit ist es den unterschiedlich definierten und verwendeten Begriff *Visual Analytics* einzuordnen und gegenüber verwandten Themengebieten abzugrenzen. Hierzu werden zunächst die verwendeten Begriffe und Forschungsdisziplinen definiert und die grundsätzliche Vorgehensweise beschrieben. Auf Basis einer umfangreichen Literaturrecherche wird der Begriff *Visual Analytics* hinsichtlich seiner Verwendung klassifiziert und sowohl eine neue Definition, als auch ein erweiterter Prozessablauf entwickelt.

Weiterhin werden auf exemplarisch verschiedene Ansätze sowohl für *Visual Analytics*, als auch für eine interaktive und visuelle Erweiterung des *Knowledge Discovery*-Prozesses vorgestellt und unter Berücksichtigung der neu entwickelten Definition evaluiert. Der Fokus liegt auf den drei verschiedenen Verfahren zur Datenanalyse – Data Mining, Visualisierung und *Visual Analytics*. Diese werden aufgrund identifizierter Vor- und Nachteile nach unterschiedlichen Dimensionen in einem Ordnungsrahmen klassifiziert und gegeneinander abgegrenzt. In der Folge wird evaluiert, inwiefern *Visual Analytics* die Auswertung bezüglich der Daten- und Analysequalität unterstützen kann.

Abschließend werden die gewonnenen Erkenntnisse prototypisch implementiert. In diesem Zusammenhang werden Data Mashups genutzt, welche es ermöglichen auch Anwender ohne tieferen technischen Hintergrund einbeziehen. Dieses Konzept bietet diverse Vorteile (vgl. Abschnitt 2.13 – Data Mashups) und harmoniert daher ausgezeichnet mit *Visual Analytics*.

1.2 Aufbau dieser Arbeit

Diese Arbeit ist wie folgt gegliedert:

Kapitel 2 – Grundlagen definiert die verwendeten Fachbegriffe und gibt einen Überblick über Themengebiete die im Bereich *Visual Analytics* relevant sind.

Kapitel 3 – Verwandte Arbeiten erläutert verschiedene Umsetzungen von *Visual Analytics*, sowie interaktiver, visueller Unterstützung des *Knowledge Discovery*-Prozesses.

Kapitel 4 – Visual Analytics – Definition, Abgrenzung und Ordnungsrahmen entwickelt auf Basis der unterschiedlichen Verwendung des Begriffes *Visual Analytics* eine umfassende Definition und erweiterten Prozess. Vor diesem Hintergrund erfolgt eine Bewertung der, in Kapitel 3 dargelegten, Ansätze. Zudem wird in diesem Kapitel eine Abgrenzung zwischen *Visual Analytics*, Visualisierung und Data Mining vorgenommen und diese in einem Ordnungsrahmen klassifiziert.

Kapitel 5 – Visual Analytics im Kontext der Daten- und Analysequalität zeigt verschiedene Varianten zur Visualisierung der Datenqualität und beurteilt den Beitrag von *Visual Analytics* hinsichtlich der Daten- und Analysequalität.

Kapitel 6 – Visual Analytics und Data Mashups beschreibt eine prototypische Implementierung von *Visual Analytics* auf Basis von Data Mashups unter Berücksichtigung der gewonnenen Erkenntnisse.

Kapitel 7 – Zusammenfassung, Fazit und Ausblick fasst die Erkenntnisse dieser Arbeit zusammen und gibt einen Ausblick auf zukünftige Entwicklungen.

2 Grundlagen

In diesem Kapitel werden die einzelnen Fachbegriffe vorgestellt und definiert, sowie die für das Verständnis der nachfolgenden Kapitel wichtigen Begriffe und Themengebiete anhand verwandter Arbeiten genauer beleuchtet.

2.1 Big Data

Big Data ist in den letzten Jahren in unterschiedlichen Zusammenhängen ein allgegenwärtiger Begriff, beispielsweise in Bezug auf das generierte Datenvolumen von sozialen Netzwerken. Der Begriff *Big Data* ist jedoch differenzierter zu betrachten als die wörtliche Bedeutung.

Die häufigste und bekannteste Definition geht zurück auf die von Laney [Lan01] propagierten Herausforderungen für eCommerce und ist in der Literatur meist als "3Vs" bekannt.

Diese beinhalten die vorhandene Datenmenge (*Volume*), die Geschwindigkeit mit der neue Daten generiert werden (*Velocity*), sowie die Art der Daten (*Variety*). Der letzte Punkt wird in Abschnitt 2.12 ausgeführt.

De Mauro et al. [DGG15] untersuchten die Verwendung des Begriffs *Big Data* hinsichtlich unterschiedlichem Kontext und unterteilen in drei Gruppen von Definitionen.

Die erste Gruppe beschäftigt sich hierbei mit der Charakteristik die Daten im Bereich *Big Data* definieren, wie die obige "3V"-Definition. Die zweite Gruppe umfasst die technischen Voraussetzungen (Speicherplatz, Rechenkraft), die für *Big Data* erforderlich sind. Die letzte Gruppe bezieht sich auf den aus *Big Data* generierten Mehrwert und dem Einfluss auf Industrie und Gesellschaft.

Aus diesen drei Gruppen wird abschließend eine möglichst allgemeingültige Definition gebildet, welche alle obigen Bereiche vereint:

> *"Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."*
>
> De Mauro et al. [DGG15]

Der Begriff *Big Data* wird in dieser Arbeit im Zusammenhang mit Datenbeständen verwendet, welche aufgrund des Umfangs, der Varianz oder der Änderungsgeschwindigkeit spezielle Technologien und Methoden benötigen, um einen Mehrwert zu generieren.

2.2 Data Mining

Der Begriff Data Mining beschreibt ein interdisziplinäres Forschungsgebiet, welches u. a. auf den Gebieten der Datenbanktechnologien, des maschinellen Lernens, der Statistik, der Mustererkennung und der künstlicher Intelligenz aufsetzt [HK06].

Es existiert für Data Mining keine eindeutige, allgemeingültige Definition, weshalb im Folgenden eine kleine Auswahl an Definitionen vorgestellt wird:

> *"Data Mining is a term coined to describe the process of sifting through large databases for interesting patterns and relationships."*
>
> Oded Maimon und Lior Rokach [MR10, S. 2]

> *"[Data Mining] is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories."*
>
> Jiawei Han und Micheline Kamber [HK06, S. 7]

> *"[Data Mining] is the application of specific algorithms for extracting patterns from data."*
>
> Fayyad et al. [FPSS96c, S. 39]

Diese Definitionen haben gemein, dass sich Data Mining grundsätzlich mit großen Datenmengen beschäftigt, um die darin verborgenen Muster und Regelmäßigkeiten aufzudecken. Weiterhin wird der Begriff Data Mining häufig als automatisierter oder semi-automatisierter Prozess bezeichnet [OLW08].

In der Literatur wird Data Mining häufig als Synonym für den *Knowledge Discovery*-Prozess (vgl. Abschnitt 2.4) verwendet [HK06], andererseits nur als ein einzelner Schritt innerhalb des Prozesses gesehen [FPSS96c]. Dies ist insofern relevant, da bei Verzicht auf die vorhergehenden Schritte des *Knowledge Discovery*-Prozesses, wie *Data Cleansing* (vgl. Abschnitt 2.9), unbedeutende und falsche Muster erkannt werden [FPSS96c].

Data Mining splittet sich in deskriptive, prognostische und präskriptive Verfahren [HK06, Grö15], wobei die Grenze nicht absolut gezogen werden kann [FPSS96c]. Nachfolgend werden die gebräuchlichsten klassischen Verfahren kurz vorgestellt [FPSS96c, KBM10, HW01, Grö15]:

Deskriptive Analyseverfahren
Deskriptive Data Mining-Verfahren treffen Aussagen über den aktuellen Zustand des Datensatzes. Diese können in die folgenden Klassen aufgeteilt werden:

Gruppenbildung
Die Gruppenbildung hat zum Ziel ähnliche Daten/Objekte zu identifizieren und in Clustern zusammenzufassen. Die Eigenschaften, nach denen die Gruppen gebildet werden, stehen zu Beginn nicht fest und werden dem Verfahren überlassen. Hierbei sind Elemente innerhalb einer Klasse möglichst ähnlich, zwischen Klassen möglichst unterschiedlich [FPSS96b, KBM10, HW01].

Assoziation
Das Verfahren der Assoziation dient dem Finden von Abhängigkeiten zwischen Daten. Ein häufig verwendetes Beispiel für dieses Verfahren ist die Warenkorb-Analyse, welche häufig zusammen erworbene Produkte identifiziert [KBM10]. Für die Berechnung der Assoziationsregel sind zwei Metriken besonders relevant:

Support
Der Support einer Assoziationsregel beschreibt die Häufigkeit in der alle Elemente der Assoziationsregel in einer Transaktion (einem Tupel) auftreten. Ein Support von 10 Prozent bzw. 0,1 bei einer Assoziationsregel (A -> B) bedeutet, dass in 10 Prozent aller Transaktionen sowohl A als auch B enthalten sind. Weiterhin kann Support entweder absolut (Anzahl der Vorkommen) oder relativ (im Verhältnis zur Gesamtzahl der Transaktionen) berechnet werden[HK06].

Confidence
Die Confidence einer Assoziationsregel beschreibt die prozentuale Häufigkeit mit der eine Regel korrekt ist. Eine Assoziationsregel (A -> B) mit Confidence 60 Prozent bzw. 0,6 bedeutet, dass in 60 Prozent der Fälle, in denen A eintritt auch B eintritt [HK06].

Abweichungsanalyse
Mit Hilfe dieses Verfahrens sollen atypische Werte identifiziert werden, welche nicht der sonstigen Charakteristik entsprechen, beispielsweise eine Verwendung der Kreditkarte an einem unüblichen Ort [KBM10] oder Abweichungen von Planwerten [HW01].

Deskription
Deskription versucht durch deskriptive, statistische Verfahren eine Beschreibung für eine Teilmenge an Daten zu finden [FPSS96b, KBM10], beispielsweise Mittelwert und Standardabweichung [FPSS96b]. Dies wird vor allem für explorative Datenanalyse [KBM10] und automatische Reports verwendet [FPSS96b].

Prädiktive Analyseverfahren
Prognostische Data Mining-Verfahren versuchen noch nicht vorhandene oder zukünftige Zustände vorherzusagen.

Klassifikation
Das Klassifikationsverfahren generiert anhand eines Trainingsdatensatzes verschiedene Klassen mit bestimmten Eigenschaften. Neue Elemente können anschließend in diese Klassen eingeordnet werden. Ein Beispiel hierfür ist die Kreditwürdigkeit von Personen basierend auf Kriterien wie Alter und Einkommen [KBM10, FPSS96c, HW01].

Wirkungsprognose
Eine Wirkungsprognose versucht basierend auf bereits bekannten Werten eine Funktion zu definieren, welche neuen Werten einen Erwartungswert zuweist. Dieser muss in den bisherigen Daten noch nicht vorkommen [KBM10, FPSS96c].

Präskriptive Analyseverfahren
Präskriptive Analyseverfahren werden eingesetzt, um Handlungsempfehlungen aus den vorhandenen Daten zu entwickeln und den Anwender mit konkreten Verbesserungsmaßnahmen zu unterstützen [Grö15].

In der weiteren Arbeit wird Data Mining nach Fayyad et al. [FPSS96c] als ein Schritt innerhalb des *Knowledge Discovery*-Prozesses aufgefasst, der möglichst automatisiert abläuft und in großen Datenmengen enthaltene Muster aufspüren kann.

2.3 Text Mining

Data Mining beschäftigt sich mit strukturierten, numerischen Daten, jedoch liegen die meisten Daten unstrukturiert in Textform vor (vgl. Abschnitt 2.12) oder können in diese überführt werden [DGS01] und sind somit nicht ohne Weiteres zu analysieren [Küs01]. Hier kommt das eng verwandte Text Mining zum Zug, welches durch Interpretation natürlicher Sprache [DGS99] in Texten enthaltene Information analysieren [DGS01] kann.

Dörre et al. [DGS99, DGS01] identifizieren zwei grundsätzliche Anwendungen:

Feature-Extraktion
Die erste Anwendung von Text Mining befasst sich mit der Extraktion von Informationselementen aus einem Text. Hierbei wird beispielsweise eine Klassifizierung einzelner Worte hinsichtlich der Bedeutung (Person, Ort) vorgenommen oder Beziehungen zwischen verschiedenen Begriffen erkannt. Ein weiteres Einsatzgebiet ist die Schlüsselwortextraktion. In dieser werden charakterisierende Wörter des Textes extrahiert und so ein Fingerabdruck generiert. Letztlich kann eine Art Zusammenfassung erstellt werden, indem die aussagekräftigsten Sätze kombiniert werden [DGS01].

Analyse von Textkollektionen
Die zweite Anwendung von Text Mining befasst sich mit einer großen Menge an Dokumenten, die hinsichtlich ihres Inhaltes charakterisiert werden. Dies umfasst einerseits die Einordnung auf Basis des Inhalts in vorgegebene Kategorien, anderseits die aus dem Data Mining bekannte Gruppenbildung. Diese identifiziert verschiedene Klassen und ordnet die Dokumente in diese ein [DGS01].

Text Mining kann wie folgt definiert werden:

> *"The extraction of codified information (features) from single documents as well as the analysis of the feature distribution over whole collections to detect interesting phenomena, patterns, or trends. Any non-trivial application of 'text mining' necessarily involves both of those mining phases."*
>
> Dörre et al. [DGS99]

Die Vorteile des Text Mining liegen darin auch große Bestände textueller Daten schnell zu verarbeiten, was für den Menschen nicht möglich ist. Zudem können Routineaufgaben automatisiert werden oder Dokumente direkt an den zuständigen Sachbearbeiter weitergeleitet werden [DGS01].

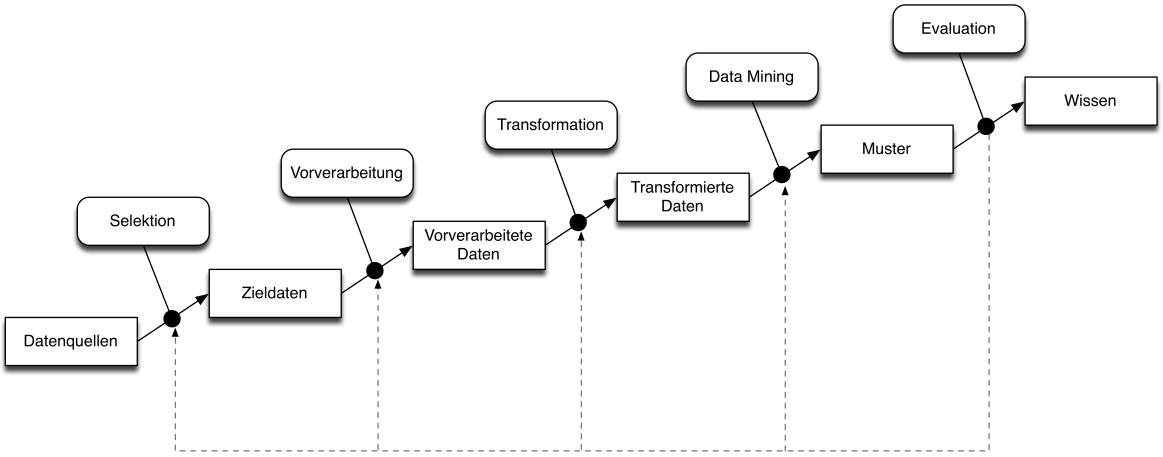


Abbildung 2.1: Stufen des *Knowledge Discovery*-Prozesses (angelehnt an [FPSS96b, S. 10])

2.4 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) – manchmal auch als *Knowledge Discovery from Data* [HK06] oder *Knowledge Discovery and Data Mining* [PBT+10] bezeichnet – beschreibt die Wertschöpfungskette von Rohdaten zu verwertbarem und wertvollem Wissen. Eine in der Literatur häufig verwendete Definition stammt von Fayyad et al.:

> *"Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."*
>
> Fayyad et al. [FPSS96b, S. 6]

Es handelt sich demnach um einen nicht-trivialen Prozess, der in einem Datensatz gültige, bisher nicht bekannte, nützliche und letztlich verständliche Muster extrahiert [FPSS96a].

An dieser Stelle sei nochmals darauf hingewiesen, dass Data Mining im Kontext dieser Arbeit lediglich als ein einzelner Schritt innerhalb dieses Prozesses definiert ist, jedoch in der Literatur häufig synonym verwendet wird (vgl. Abschnitt 2.2).

Der Prozess der Wissensgenerierung ist in Abbildung 2.1 schematisch dargestellt, die einzelnen Stufen werden im Folgenden kurz erklärt [FPSS96b]:

1. Selektion
   Aus einer vorhandenen Datenmenge (verschiedene Datenquellen oder Teilmenge einer Datenquelle) werden alle Daten ausgewählt mit denen der Prozess durchgeführt werden soll. Wenn nicht alle benötigten Daten integriert werden kann der Prozess fehlschlagen [MR10]. Die Untermenge der Daten wird Zieldaten genannt.
2. Vorverarbeitung
   In diesem Schritt werden Daten für die weitere Analyse vorbereitet, um eine möglichst gute Datenqualität sicherzustellen (vgl. Abschnitt 2.9). Die Daten werden anschließend als vorverarbeitete Daten bezeichnet [FPSS96b, MR10].

3. Transformation
   Im Schritt der Transformation werden die Daten für den Data Mining-Schritt vorbereitet. Hierzu gehören unter anderem die folgenden Verfahren:

   Attribut-Selektion
   Ein Datensatz mit hoher Anzahl an Attributen kann das Ergebnis und die Effizienz von Mining-Algorithmen beeinträchtigen, weshalb diese reduziert werden sollten [ES00].

   Diskretisierung und Generalisierung
   Je nach Data Mining-Algorithmus kann es sinnvoll oder notwendig sein, dass Transformationen vorgenommen werden, beispielsweise von numerischen auf kategorische Werte. Die Generalisierung verändert die Granularität der Daten, indem beispielsweise die genaue Adresse auf die Stadt oder das Land reduziert wird. Auf diese Weise kann die Anzahl der Attributwerte reduziert werden [ES00, Küs01, HK06].

   Normalisierung
   Die Normalisierung beschreibt die Abbildung von Werten auf eine begrenzte Skala, beispielsweise von 0,0 bis 1,0. Diese Vorgehensweise ist bei entfernungsbasierten Algorithmen sinnvoll, da ansonsten die höheren Skalenwerte überwiegen und das Ergebnis verfälscht wird [HK06].

   Attribut-Konstruktion
   In manchen Fällen kann es nötig sein auf Basis vorhandener Werte neue Attribute zu konstruieren, die den Data Mining-Prozess unterstützen [HK06].

   Dieser Schritt ist nach Maimon und Rokach [MR10] häufig entscheidend für den Erfolg, gleichzeitig jedoch stark von der jeweiligen Aufgabe abhängig. Im Falle einer fehlerhaften oder ungenügenden Transformation entstehen Ergebnisse, die Hinweise auf die nötigen Transformationen geben. Diese können anschließend in der nächsten Iteration berücksichtigt werden [MR10].

4. Data Mining
   In diesem Schritt des *Knowledge Discovery*-Prozesses wird ein Data Mining-Algorithmus (vgl. Abschnitt 2.2) in Abhängigkeit des Ziels ausgewählt und nach Mustern im vorbereiteten Datensatz gesucht.

5. Interpretation
   Die gefundenen Muster werden dem Anwender präsentiert und von diesem interpretiert.

Der gesamte *Knowledge Discovery*-Prozess ist iterativ aufgebaut [MR10, HK06], d. h. jeder Schritt kann beliebig oft wiederholt werden bzw. wenn nötig kann auch zu einem vorhergehenden Schritt zurück gesprungen werden und von diesem Punkt – mit der gewonnenen Einsicht – fortgesetzt werden.

2.5 Visualisierung

Das Themengebiet der Visualisierung wird in der Literatur unterteilt in verschiedene Definitionen, zum einen der Überbegriff der Visualisierung, zum anderen im Hinblick auf die zu visualisierenden Daten.

Eine häufig verwendete Definition für den Überbegriff stammt von Card et al.:

> *"The use of computer-supported, interactive, visual representations of data to amplify cognition."*
>
> Card et al. [CMS99, S. 6]

Daten unterscheiden sich nach Card et al. [CMS99] im Hinblick auf ihren Ursprung. Einerseits haben Daten physikalischen Bezug, beispielsweise in der Biologie, der Chemie oder der Physik. Diese werden mit Hilfe der wissenschaftlichen Visualisierung dargestellt:

> *"The use of computer-supported, interactive, visual representations of scientific data, typically physically based, to amplify cognition."*
>
> Card et al. [CMS99, S. 7]

Im Gegensatz dazu wird von Informationsvisualisierung gesprochen, wenn die Daten ihren Ursprung in nicht-physikalischen Bereichen haben, beispielsweise Finanzdaten, Kennzahlen oder abstrakte Daten. Die Definition ändert sich somit geringfügig:

> *"The use of computer-supported, interactive, visual representations of abstract, nonphysically based data to amplify cognition."*
>
> Card et al. [CMS99, S. 7]

Eine andere Definition zur Unterteilung zwischen wissenschaftlicher und Informationsvisualisierung stammt von Munzner [Mun08]:

> *"It's infovis when the spatial representation is chosen, and it's scivis when the spatial representation is given."*
>
> Tamara Munzner [Mun08]

Weiterhin gilt zu unterscheiden, ob das Ziel der Visualisierung die Präsentation der Daten für eine andere Zielgruppe oder die explorative bzw. konfirmative Analyse ist [Wij05, SM00].

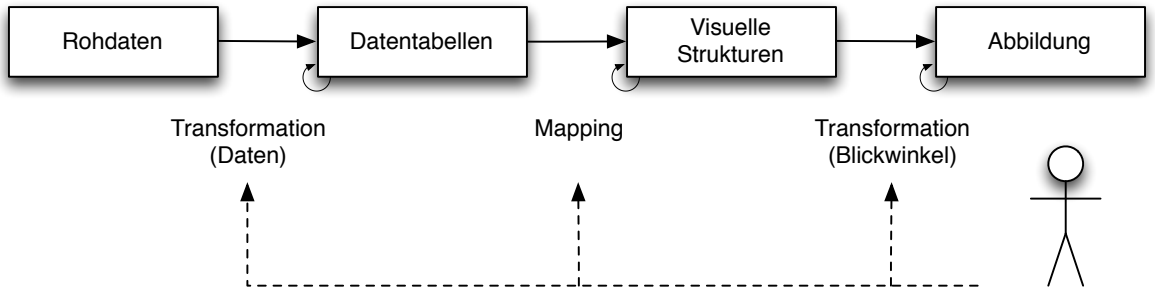


Abbildung 2.2: Referenzmodell für die Visualisierung (angelehnt an [CMS99, S. 17])

Für die Visualisierung existiert ein von Card et al. eingeführtes Referenzmodell (vgl. Abbildung 2.2). Dieses beschreibt die Abfolge von Operationen ausgehend von Rohdaten bis hin zu einer visuellen Repräsentation dieser Daten. Die dazu vorgesehenen Schritte werden im Folgenden kurz skizziert:

Transformation (Daten)
Daten liegen zunächst in einem beliebigen Format vor, als sogenannte Rohdaten. Diese sollten zunächst in relationale Beschreibungen, die Datentabellen, überführt werden. Die Rohdaten sind anschließend strukturiert und lassen sich somit in den nächsten Schritten leichter verarbeiten. Im gleichen Schritt können die in Rohdaten häufig enthaltenen fehlerhaften oder fehlenden Werte korrigiert oder ermittelt werden. Weiterhin können weitere Informationen berechnet und den Daten hinzugefügt werden.

Abhängig von der Art der Daten bieten sich später unterschiedliche Visualisierungen an. Eine ausführliche Erklärung dieser Transformationen ist bei Card et al. [CMS99] nachzulesen. Eine Interaktion durch den Anwender kann beispielsweise über verschiedene Slider erfolgen und auf diese Weise der Umfang der Rohdaten eingeschränkt werden.

Mapping
In diesem Schritt werden die Datentabellen auf visuelle Strukturen abgebildet, wofür in den meisten Fällen mehrere Möglichkeiten existieren. Ein sinnvolles Mapping zu finden ist eine nicht-triviale Aufgabe die zwei Voraussetzungen erfüllen muss. Einerseits muss das Mapping *'expressive'* sein, d. h. alle Daten müssen abgebildet sein, während gleichzeitig keine weiteren Daten hinzukommen dürfen, beispielsweise visuelle Beziehungen, die in den Daten nicht vorhanden sind.

Andererseits sollte eine Visualisierung *'effective'* sein, d. h. möglichst schnell zu interpretieren. Als Beispiel hierfür nennen Card et al. [CMS99] in Farben kodierte Funktionswerte einer Sinuskurve, anstatt die erwartete Kodierung über die Position, was deutlich schwerer interpretierbar ist.

Entsprechend spielt die menschliche Wahrnehmung eine entscheidende Rolle für die Effizienz einer Visualisierung. Die präattentive Wahrnehmung [HBE96] ist hierbei bevorzugt zu berücksichtigen.

Dieses Konzept beschreibt die Fähigkeit des Menschen verschiedene visuelle Eigenschaften ohne explizite Aufmerksamkeit zu erkennen. Der Anwender kann auf diese Weise in einem Sekundenbruchteil evaluieren, ob in einer Menge an Quadraten ein Dreieck enthalten ist. Healey et al. [HBE96] stellen auf Basis einer Literaturrecherche verschiedene visuelle Eigenschaften (u. a. Größe, Länge, Form) vor, welche diesem Prinzip genügen.

Die Interaktion durch den Anwender kann an dieser Stelle entweder über ein separates Interface erfolgen, bei welchem der Anwender das Mapping verändern kann, oder direkt in der Visualisierung, indem beispielsweise durch Klick auf eine Achse der Visualisierung deren Referenz verändert wird.

Transformation (Blickwinkel)
Im letzten Schritt des Visualisierungsprozesses kann der Blick auf die Abbildungen interaktiv verändert werden, um mehr Informationen als aus einem statischen Abbild zu erhalten. Hierbei werden drei grundsätzliche Konzepte identifiziert:

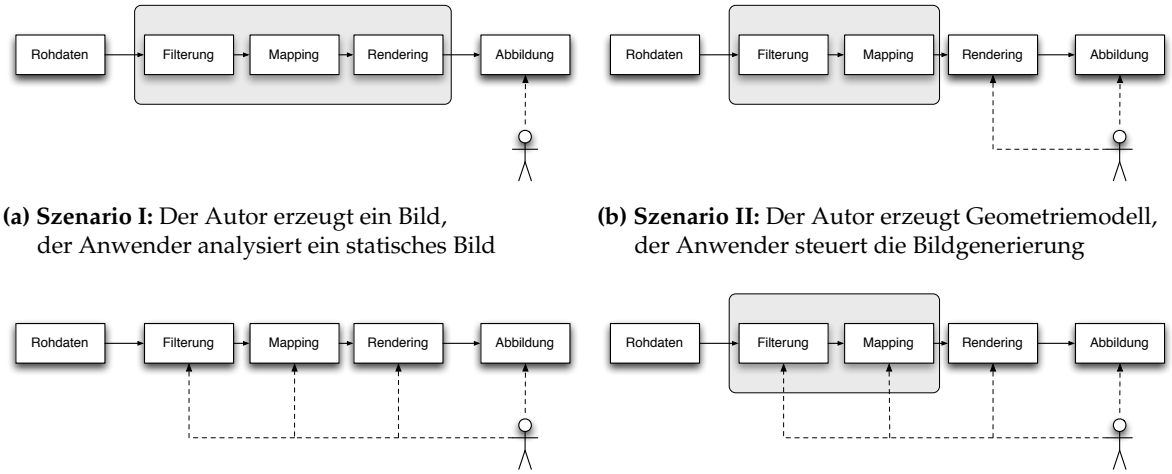
Location Probes
Location Probes ermöglichen zusätzliche Informationen, indem beispielsweise durch das Markieren eines Elements nähere Informationen angezeigt werden. Alternativ kann statt einem Detail-Fenster auch die umgebende Region auf eine andere Art, etwa stark vergrößert, dargestellt werden.

Viewpoint Controls
Viewpoint Controls ermöglichen eine Veränderung des Blickes auf die Daten. Beispiele hierfür sind *Zooming* (Vergrößerung eines Bereiches) und *Panning* (Verschieben des Bildschirmausschnittes).

Verzerrung
Eine Verzerrung ermöglicht die Einbindung von Details innerhalb der vorhandenen visuellen Struktur. Hierbei wird ein Bereich der Daten vergrößert hervorgehoben. Durch Verwendung dieser Technik kann ein Teilbereich evaluiert werden, während der Kontext erhalten bleibt.

Die einzelnen Schritte können auch als Filterung (Datenauswahl und Aufbereitung), Mapping (Abbildung nicht-geometrischer Daten auf geometrische Objekte und Farben) sowie Rendering (Bildgenerierung) bezeichnet werden. Weiterhin kann der Visualisierungsprozess angelehnt an Wood et al. [WBW96]/Schumann und Müller [SM00] auf verschiedene Nutzer aufgeteilt werden (vgl. Abbildung 2.3).

Diese Aufteilung wurde ursprünglich für die verteilte Visualisierung über das Internet entwickelt, lässt sich jedoch auch für eine Aufteilung zwischen Experte/Autor und Anwender verwenden. Im Folgenden werden diese vier Szenarien kurz erläutert:



(a) Szenario I: Der Autor erzeugt ein Bild, der Anwender analysiert ein statisches Bild

(b) Szenario II: Der Autor erzeugt Geometriemodell, der Anwender steuert die Bildgenerierung

(c) Szenario III: Der Anwender erzeugt die Visualisierung

(d) Szenario IV: Autor erzeugt ein Geometriemodell unter der Kontrolle des Anwenders

Abbildung 2.3: Mögliche Ausprägungen des Visualisierungsprozesses (angelehnt an [SM00, WBW96])

Szenario I: Autor erzeugt ein Bild
Im einfachsten Fall (vgl. Abbildung 2.3a) erzeugt ein Experte mit Hintergrundwissen eine Visualisierung, die anschließend von einem Anwender interpretiert werden kann. Für diesen gibt es jedoch keine Möglichkeit mit dieser Visualisierung zu interagieren.

Szenario II: Autor erzeugt ein Geometriemodell
In diesem Szenario (vgl. Abbildung 2.3b) wird dem Anwender die Kontrolle über das Rendering überlassen. Somit kann innerhalb der Visualisierung eine Interaktion stattfinden, da der Experte lediglich das grundsätzliche Geometriemodell erstellt. Der Anwender kann jedoch keinen Einfluss auf das Mapping nehmen, welches der entscheidende Schritt für eine effektive Visualisierung ist.

Szenario III: Anwender erzeugt die Visualisierung
Im dritten Szenario (vgl. Abbildung 2.3c) erhält der Anwender volle Kontrolle über die einzelnen Schritte des Visualisierungsprozesses. Der Experte stellt lediglich die Rohdaten zur Verfügung. Somit hat der Anwender die größtmögliche Freiheit der Analyse, benötigt jedoch auch das entsprechende Wissen und die Rechenkapazität.

Szenario IV: Autor erzeugt ein Geometriemodell unter Kontrolle des Anwenders
Im letzten Szenario (vgl. Abbildung 2.3d) sollen die Nachteile von Szenario II und III adressiert werden. Hierbei wird wie in Szenario II vorgegangen, jedoch erhält der Anwender über Schnittstellen die Möglichkeit Parameter für das Filtering und das Mapping einzustellen.

Szenario IV ähnelt verbreiteten Definitionen des Visualisierungsprozesses (vgl. Ware [War12]) und wird somit für die weitere Arbeit als Grundlage für Erläuterungen und Bewertungen aufgefasst.

Grundsätzlich lässt sich das empfohlene Vorgehen der interaktiven Visualisierung nach Shneiderman, auch bekannt als das *Visual Information Seeking Mantra*, wie folgt beschrieben:

> *"Overview first,*
> *zoom and filter,*
> *then details-on-demand"*
>
> Ben Shneiderman [Shn96]

Der Anwender soll zuerst einen Überblick über die Daten erhalten, anschließend interessante Bereiche genauer betrachten (*zoom*) bzw. uninteressante herausfiltern (*filter*) und nähere Informationen bei Bedarf erhalten (*details-on-demand*) können.

2.6 Visual Analytics

Visual Analytics ist ein relativ junges, interdisziplinäres Forschungsgebiet. Frühe Verwendungen dieses Terms gehen in das Jahr 2004 zurück [KMS+08, PT04]. Entsprechend existieren verschiedene Definitionen:

> *"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces."*
>
> James J. Thomas und Kristin A. Cook [TC05, S. 4]

> *"Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets."*
>
> Keim et al. [KAF+08, S. 175]

Beide Definitionen haben gemein, dass eine interaktive, visuelle Komponente beteiligt ist, der Umfang ist jedoch strittig. Eine Anwendung erfüllt den Begriff *Visual Analytics* einerseits bereits bei Verwendung einer interaktiven Oberfläche, andererseits wird zusätzlich ein Wechsel zwischen automatischen und visuellen Verfahren benötigt. Um den Begriff *Visual Analytics* präziser zu spezifizieren wird im Verlauf dieser Arbeit die Verwendung in der Literatur evaluiert und eine neue Definition entwickelt (vgl. Abschnitt 4.1).

Da das auszuwertende Datenvolumen immer weiter und schneller ansteigt [Wij05] ist das Mantra der Visualisierung *Overview first, zoom and filter, details on demand* in vielen Fällen nicht praktikabel, da ein Überblick nur mit Informationsverlust darstellbar ist [KKM+10b]. Der Anwender kann auf diese Weise nicht erkennen, welche Bereiche der Daten für eine weitere und nähere Erkundung sinnvoll sind [KKM+10b]. Visuelle Verfahren reichen folglich alleine nicht mehr aus [ABM07]. Es ist sinnvoll bereits vor der initialen Visualisierung die interessanten Bereiche zu finden und den Anwender auf diese aufmerksam zu machen [KKM+10b].

Entsprechend erweitern Keim et al. [KMSZ06b] das Mantra der Visualisierung für *Visual Analytics*:

> *"Analyse First –*
> *Show the Important –*
> *Zoom, Filter and Analyse Further –*
> *Details on Demand"*
>
> Keim et al. [KMSZ06b, S. 6]

2.7 Human In The Loop

Das Konzept des *Human In The Loop* wird häufiger in der Literatur erwähnt (u. a. [EHR+14, WIR+10, AS94]), jedoch nicht gesondert definiert. Entsprechend wird der Term in dieser Arbeit gemäß der wörtlichen Bedeutung verwendet, d. h. der Anwender wird – bei einem auf diesem Konzept basierenden Ansatz – innerhalb eines Schleifendurchlaufs beliebig oft zu einer Tätigkeit aufgefordert.

In Verbindung mit einer interaktiven Visualisierung kann ein Anwender zusätzlich zu dem in den Daten enthaltenen, expliziten Wissen sein implizites Wissen einbringen [WJD+09]. Nach Wagner [Wag15] ist das implizite Wissen oft nötig um die Daten zu verstehen.

2.8 Datenqualität

In der Literatur existieren im Bezug auf die Datenqualität die Begriffe *data quality* und *information quality*. Über die exakte Verwendung besteht kein allgemeiner Konsens [MWLZ09, PLW02]. In der deutschsprachigen Literatur ist der Term zur Definition unterteilt in Daten und Qualität [ABEM15]. Nach Garvin [Gar84] lässt sich Qualität anhand von fünf verschiedenen Ansätze unterscheiden, wovon zwei im Hinblick auf Datenqualität anwendbar sind:

Produktorientierter Ansatz
Der produktorientierte Ansatz betrachtet Qualität als präzise und messbare Größe. Die Qualität des Produktes setzt sich aus den einzelnen Produkteigenschaften zusammen, ändert sich die Qualität so muss sich folglich eine Produkteigenschaft verändert haben. Dieser Ansatz ist aus diesem Grund losgelöst von subjektiven Wahrnehmungen und somit objektiv bestimmbar [Gar84, ABEM15].

Anwenderorientierter Ansatz
Der anwenderorientierte Ansatz begründet sich auf der Beziehung zwischen Anwender und dem Produkt. Jeder Anwender hat unterschiedliche Erwartungen oder Anforderungen, entsprechend derer sich die Qualität unterschiedlich darstellt. Als Folge hiervon kann für einen Anwender das Produkt qualitativ hochwertig sein, während ein anderer Anwender die Qualität als unzureichend beurteilt. Dieser Ansatz ist infolgedessen hochgradig subjektiv [Gar84, ABEM15].

Ähnliche Unterteilungen finden sich in der englischsprachigen Literatur. Daten haben u. a. eine innere, sowie eine kontextabhängige Qualität [WS96]. Mögliche Dimensionen für die innere Qualität sind Genauigkeit, Objektivität oder auch Reputation der Quelle [WS96]. Für die kontextabhängige Qualität sind Beispiele für verwendete Dimensionen ob eine Relevanz für die Aufgabe gegeben ist oder die Datenmenge ausreichend ist [WS96].

In der weiteren Arbeit wird die Datenqualität auf Basis dieser beiden Interpretationen aufgefasst. Eine Auswahl an Definitionen für die subjektive Qualität ist:

> *"Qualität ist ein mehrdimensionales Maß für die Eignung von Daten, den an ihre Erfassung/Generierung gebundenen Zweck zu erfüllen. Diese Eignung kann sich über die Zeit ändern, wenn sich die Bedürfnisse ändern."*
>
> Volker Gerhard Würtele [Wür03, S. 21]

> *"Data Quality [is] data that [is] fit for use by data consumers."*
>
> Richard Y. Wang und Diane M. Strong [WS96, S. 6]

> *"[Exactly] the right data and information in exactly the right place at the right time and in the right format to complete an operation, serve a customer, make a decision, or set and execute a strategy."*
>
> Thomas C. Redman [Red13, S. 4]

Alle Definitionen haben gemein, dass die Daten aus Sicht desjenigen der die Daten verarbeitet *"fit for use"* sein müssen, d. h. den Einsatzweck unterstützend, um als qualitativ bezeichnet zu werden.

Um die Datenqualität zu bewerten kommen Metriken zum Einsatz, welche auf Basis unterschiedlicher Ansätze (theoretisch, empirisch oder intuitiv) hergeleitet werden können [BS06, WS96] und sich hinsichtlich der gefundenen Dimensionen entsprechend unterscheiden. Eine Auswahl abgeleiteter Metriken und Dimensionen finden sich beispielsweise bei Wang und Strong [WS96], Redman [Red12] oder Price und Shanks [PS04].

Weiterhin existieren für die einzelnen Metriken weitere Unterscheidungen über die genaue Auffassung je nach Perspektive. Die Vollständigkeit ist entsprechend hinsichtlich des Schemas, der Spalten oder des Bestandes definierbar [PLW02]. Das Schema muss alle vorkommenden Attribute enthalten, um als vollständig zu gelten, während auf Datenebene die Anzahl der fehlenden Werte als Maß für die Vollständigkeit verwendet wird. Letztlich kann über die Repräsentation in den Attributausprägungen ein Wert ermittelt werden [PLW02]. Die Berechnung der Vollständigkeit nach der dritten Perspektive wird nachfolgend anhand der US-Bundesstaaten beispielhaft beschrieben [PLW02]:

Ein Attribut repräsentiert den jeweiligen US-amerikanischen Bundesstaat und kann somit theoretisch 50 verschiedene Werte beinhalten. Für den Fall, dass lediglich 43 verschiedene Bundesstaaten mindestens einmalig repräsentiert sind ist der Datensatz in dieser Perspektive unvollständig [PLW02]. Die Vollständigkeit ist für diesen Fall mit 0.86 definiert.

Auch Kombinationen verschiedener Perspektiven sind in diesem Zusammenhang vorstellbar. Batini und Scannapieco [BS06] beschreiben mögliche Perspektiven ausführlich für verschiedene Metriken.

2.9 Data Cleansing

Der Term *Data Cleansing* (auch *Data Cleaning* [RD00, GMV96]) wird im Kontext der Datenbereinigung verwendet. Üblicherweise liegen Daten nicht in bestmöglicher Qualität vor, sondern unterliegen verschiedenen Mängeln (fehlende Werte, unterschiedliche Datentypen, doppelte Einträge, Inkonsistenzen, etc. [RD00, HK06, GMV96]), welche den weiteren Verlauf des *Knowledge Discovery*-Prozesses (vgl. Abschnitt 2.4) beeinflussen können. Entsprechend empfiehlt es sich zu analysierende Daten zuerst zu bereinigen, um belastbare Analysen zu ermöglichen [HK06]. In der Literatur wird *Data Cleansing* typischerweise – jedoch nicht ausschließlich – verwendet, um eine integrierte und konsistente Datenhaltung in einem Data Warehouse vorzubereiten [HK06, HW01], insbesondere wenn mehrere, heterogene Datenquellen zusammengefasst werden [RD00, MM10]. Weitere Einsatzgebiete sind der *Knowledge Discovery*-Prozess [FPSS96c, MM10] oder das Qualitätsmanagement bei Daten und Informationen [MM10].

Aufgrund der unterschiedlichen Einsatzbereiche gibt es keine eindeutige, allgemeingültige Definition [MM10]. Maletic und Marcus [MM10] untersuchten verschiedene Einsatzgebiete von *Data Cleansing* und spezifizierten drei generelle Phasen des *Data Cleansing*-Prozesses. Zunächst die werden die Fehlertypen definiert, anschließend die Fehler identifiziert und abschließend korrigiert. Die ersten beiden Schritte können mit spezialisierten Methoden und Technologien (Statistik, Clustering, Mustererkennung) automatisiert werden, während die automatisierte Korrektur abseits klar definierter Aufgabenbereiche sehr kompliziert ist [MM10].

Ausführlich beschrieben wird *Data Cleansing* beispielsweise von Han und Kamber [HK06].

2.10 Data Wrangling

In vielen Fällen muss für die Analyse sichergestellt werden, dass die Daten in dem hierfür geforderten Format und entsprechender Qualität vorliegen. Hierzu existieren verschiedene automatische Verfahren, diese sind in den meisten Fällen jedoch nicht interaktiv oder verzichten auf eine visuelle Unterstützung [KHP+11]. An dieser Stelle kommt das Prinzip des *Data Wrangling* zum Einsatz, welches wie folgt definiert werden kann:

> *"Data Wrangling is the process of iterative data exploration and transformation that enables analysis."*
>
> Kandel et al. [KHP+11]

Data Wrangling ist nach Kandel et al. [KHP+11] der Prozess aus Rohdaten für die Analyse verwertbare und somit wertvolle Daten zu gewinnen. Dieser Prozess ist iterativ, d. h. eine stetige Wiederholung verschiedener Schritte, um sich mit Hilfe der gewonnenen Erkenntnisse der bestmöglichen Lösung zunehmend anzunähern.

Eine ausführlichere Beschreibung von *Data Wrangling* und den Herausforderungen in diesem Bereich findet sich bei Kandel et al. [KHP+11].

2.11 Analysequalität

Analysequalität beschreibt die Qualität der Ergebnisse, die durch die Analyse erreicht werden können. Aus den in Abschnitt 2.8 beschriebenen Definitionen für Qualität folgt, dass Datenqualität nicht nur objektiv, sondern auch subjektiv hinsichtlich des Zweckes bewertet werden muss. Dennoch ist die "Datenqualität essentiell für die Entscheidungsfindung" [PM08, S. 1]. Analysequalität kann definiert werden als Qualität der erzeugten Ergebnisse und mit Hilfe der folgenden Metriken angegeben werden [MRS08, OD08]:

Relevanz
Die Relevanz eines Ergebnisses ist gegeben, wenn der Anwender dieses als informativ erachtet. Da dies hochgradig subjektiv ist kann die Übereinstimmung zwischen verschiedenen Menschen durch den Kappa-Koeffizienten gemessen werden:

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$P(A)$ ist hierbei die tatsächliche Übereinstimmung und $P(E)$ die erwartete Übereinstimmung. Letztere kann beispielsweise über die Klassenverteilung abgeschätzt werden.

Effektivität
Die Effektivität beschreibt die Qualität der Ergebnisse und kann durch nachfolgende Metriken angegeben werden, wobei gilt:

tp = korrekt positiv, tn = korrekt negativ, fp = falsch positiv, fn = falsch negativ

Precision
Precision beschreibt, welcher Anteil der Ergebnismenge die Anfrage richtig beantwortet:

$$P = \frac{tp}{tp + fp}$$

Recall
Recall beschreibt, welcher Anteil aller korrekten Ergebnisse im Resultat enthalten ist:

$$R = \frac{tp}{tp + fn}$$

Accuracy
Accuracy beschreibt den Anteil der korrekt erkannten Elemente gegenüber allen Elementen:

$$A = \frac{tp + tn}{tp + tn + fp + fn}$$

F-measure
F-measure ist eine Metrik, welche die obigen Werte Recall und Precision durch Verwendung des harmonischen Mittelwerts (üblicherweise mit $\beta = 1$) ins Verhältnis setzt:

$$F_{\beta=1} = \frac{2 * P * R}{P + R}$$

Für die Bestimmung der Qualität eines Analyseverfahrens wird somit ein Datensatz benötigt, bei welchem die einzelnen Elemente bereits hinsichtlich der gesuchten Information klassifiziert sind und obige Metriken auf das Ergebnis angewendet werden können. Dieser Datensatz wird als *gold standard* bezeichnet und wird idealerweise durch Menschen mit hoher Übereinstimmung nach dem Kappa-Koeffizienten festgelegt [MRS08].

2.12 Strukturierte und unstrukturierte Daten

Daten können grundsätzlich in zwei verschiedenen Formen, strukturiert und unstrukturiert, vorliegen.

Strukturierte Daten sind nach Weglarz [Weg04] alle Daten, welche sich in aus atomaren Datentypen zusammensetzen lassen, während unstrukturierte Daten entweder geschriebene Sprache (Dokumente, E-Mails, etc.) oder nicht-sprachbasierte Objekte (Bild, Video, Audio) umfassen. Letztere lassen sich beispielsweise durch Spracherkennung in textuelle Dokumente überführen [DGS01] und werden in der folgenden Arbeit aus diesem Grund nicht gesondert betrachtet.

Inmon und Nesavich [IN07] unterscheiden auf Grundlage der erforderlichen Disziplin. Demnach werden strukturierte Daten gemäß einem Schema in einer Datenbank gespeichert, die für Organisation und Indizierung zuständig ist. Entsprechend können die Daten jederzeit unter verschiedenen Gesichtspunkten (z. B. durch Einteilung in verschiedene Zeiträume wie täglich, monatlich, etc.) betrachtet werden. Gegenteilig können unstrukturierte Daten ohne Berücksichtigung einer erzwungenen Form erzeugt werden.

Apel et al. [ABEM15] klassifizieren unstrukturierte Daten als diejenigen, *"die sich nicht adäquat in herkömmlichen Datenbanken mit Zeilen und Spalten abbilden lassen"* [ABEM15, S. 99].

Nach Schätzungen sind zwischen 80 % [Gri08] und 90 % [DGS99] aller Daten unstrukturiert.

Nach Blumberg und Atre [BA03] ist der Begriff semi-strukturierte Daten für viele Bereiche zutreffender, da mit Ausnahme reiner Textdokumente strukturierte Metadaten (Autor, Datum, etc.) enthalten sind.

2.13 Data Mashups

Mashup is ein häufig vorkommender Begriff in unterschiedlicher Verwendung, der sich nach Daniel und Matera [DM14a] wie folgt definieren lässt:

> *"A mashup is a composite application developed starting from reusable data, application logic, and/or user interfaces typically, but not mandatorily, sourced from the web."*
> Florian Daniel und Maristella Matera [DM14a, S. 3]

Weiterhin erläutern Daniel und Matera [DM14b] den spezialisierten Fall der Data Mashups:

> *"Data Mashups [...] fetch data from different data services or resources, process them, and return an integrated result set (the output of the data mashup)."*
> Florian Daniel und Maristella Matera [DM14b, S. 143]

Dieser Prozess beinhaltet u. a. die Bereinigung, Reformatierung, Trennung oder Kombination von Daten zum Zwecke der Zusammenlegung verschiedener, heterogenener Datenquellen [DM14b].

Aus der Verwendung von Mashups ergeben sich verschiedene Vorteile [DM14a, HM16]:

Individuelle Lösungen
Mashups erlauben durch geringere Entwicklungskosten auch weniger nachgefragte Lösungen zu entwerfen, welche für den Massenmarkt nicht relevant sind und aus diesem Grund nicht entwickelt würden.

Flexibilität
Mashups vereinfachen die Formulierung einer Analyse unter anderem durch grafische Modellierung. Hiermit kann bei einfachen Problemstellungen die Entwicklungsabteilung umgangen werden und der Anwender kann seine Fragestellung selbst beantworten. Mashups erhöhen demzufolge die Flexibilität gegenüber spezifischen Verfahren.

Wissenstransfer
Mashups erlauben es den Anwendern einfache Lösungen für Problemstellungen selbst zu entwickeln, welche in der Entwicklungsabteilung bisher nicht bekannt sind. Auf diese Weise kann das Innovationspotential der Anwender genutzt werden.

Schnelles Prototyping
Mashups bieten die Möglichkeit durch Verwendung bereits existierender Funktionalitäten schnell einen Prototyp zu entwickeln und verhindert, dass für jede Applikation erneut von vorne begonnen werden muss.

Reduzierte Kosten
Mashups ermöglichen es dem Endanwender die benötigten Applikationen selbst zu entwickeln, wodurch für die Entwicklungsabteilung aufwändige Untersuchungen über nachgefragte Produkte und die Anwenderzufriedenheit entfallen.

Visuelle Modellierung
Mashups werden üblicherweise mit visueller Unterstützung definiert, indem Datenflüsse und Abläufe gezeichnet werden. Der Anwender benötigt so keine tieferen Kenntnisse über die Implementierung und Ausführung.

Förderung der Neugierde
Mashups bieten dem Anwender die Möglichkeit ohne Detailwissen über die genaue Implementierung eigene Applikationen zu entwickeln und fördern auf diese Weise dessen Neugierde.

Höhere Zufriedenheit
Mashups ermöglichen eine engere Zusammenarbeit zwischen Entwickler und Endanwender. Diese führt zu effektiveren und nützlicheren Applikation sowie höherer Zufriedenheit des Anwenders.

Data Mashups integrieren demnach verschiedene heterogene Datenquellen und bieten durch grafische Modellierung eine Lösung von Entwicklungsabteilungen bei offenen Problemstellungen. Durch dieses Konzept steigt die Flexibilität und Unabhängigkeit der Anwender und bietet so ein großes Potential für bessere Analysen.

3 Verwandte Arbeiten

In diesem Kapitel werden drei Anwendungen vorgestellt, welche *Visual Analytics* zugeordnet werden. Im Anschluss wird evaluiert in welchen Schritten des *Knowledge Discovery-Prozesses* visuelle Verfahren zum Einsatz kommen können.

3.1 Visual Analytics – Anwendungsszenarien

3.1.1 Szenario I – Meinungsanalyse

Ein Ansatz für die visuelle Textanalyse auf unstrukturierten Daten stammt von Keim et al. [KMOZ08]. Das Ziel dieses Ansatzes ist es Nachrichten in positive und negative Aussagen zu kategorisieren und anhand verschiedener Titelseiten einer Tageszeitung zu visualisieren.

Abbildung 3.1: Visuelle Analyse von Zeitungsartikeln [KMOZ08]

Für die Kategorisierung kommt ein einfacher Algorithmus zum Einsatz, welcher vorgegebene, als meinungsbildend identifizierte, Wörter zählt. Von Bedeutung sind in diesem Zusammenhang Signalwörter wie "wundervoll", "Problem" oder "schlecht".

In Abbildung 3.1 ist ein Ausschnitt der Visualisierung über zwei Wochen dargestellt. Die Zeilen stehen hierbei für Wochen, die Spalten für verschiedene Tage. Einzelne Abschnitte sind in verschiedenen Abstufungen rot und grün eingefärbt, um dem Anwender die Tendenz des jeweiligen Satzes zu offenbaren. Gut sichtbar ist auf diese Weise, dass einzelne Absätze zwischen positiv und negativ schwanken.

3.1.2 Szenario II – Jigsaw

Ein weiteres Beispiel für die visuelle Analyse unstrukturierter Daten ist *Jigsaw*[1] von Stasko et al. [SGL08]. Im Kontrast zu Beispiel I zielt *Jigsaw* auf die Analyse von Verbindungen innerhalb einer Menge an Dokumenten, um dem Anwender ein besseres Verständnis über die beinhalteten Themen und Fakten zu ermöglichen. Für die in diesem Abschnitt dargestellten Abbildungen wurde hierzu der mitgelieferte Datensatz verwendet. Dieser beinhaltet die IEEE InfoVis- und VAST-Konferenzbeiträge[2] von 1994 bis 2015. Das *Jigsaw*-System ermöglicht alternativ auch das Einlesen von Textdokumenten oder CSV-Dateien.

Als Vorbereitung müssen die Entitätstypen und Entitäten jedes Dokuments identifiziert und extrahiert werden. *Jigsaw* bietet hierzu sowohl statistische Verfahren, als auch die Möglichkeit eigene Entitätstypen über eine Wortliste zu spezifizieren.

In der Nachbearbeitung werden verschiedene Optionen zur Verfügung gestellt, um beispielsweise nur einmalig auftretende Entitäten zu entfernen, Tippfehler zu korrigieren oder Aliase anzulegen. Weiterhin können einzelne Entitäten, die nicht erkannt wurden, manuell hinzugefügt oder bei einer falschen Erkennung entfernt werden.

Jigsaw bietet auf Basis der extrahierten Entitäten verschiedene Visualisierungen, ein Ausschnitt derselben wird auf den nachfolgenden Seiten vorgestellt.

[1] http://www.cc.gatech.edu/gvu/ii/jigsaw
[2] http://ieeevis.org

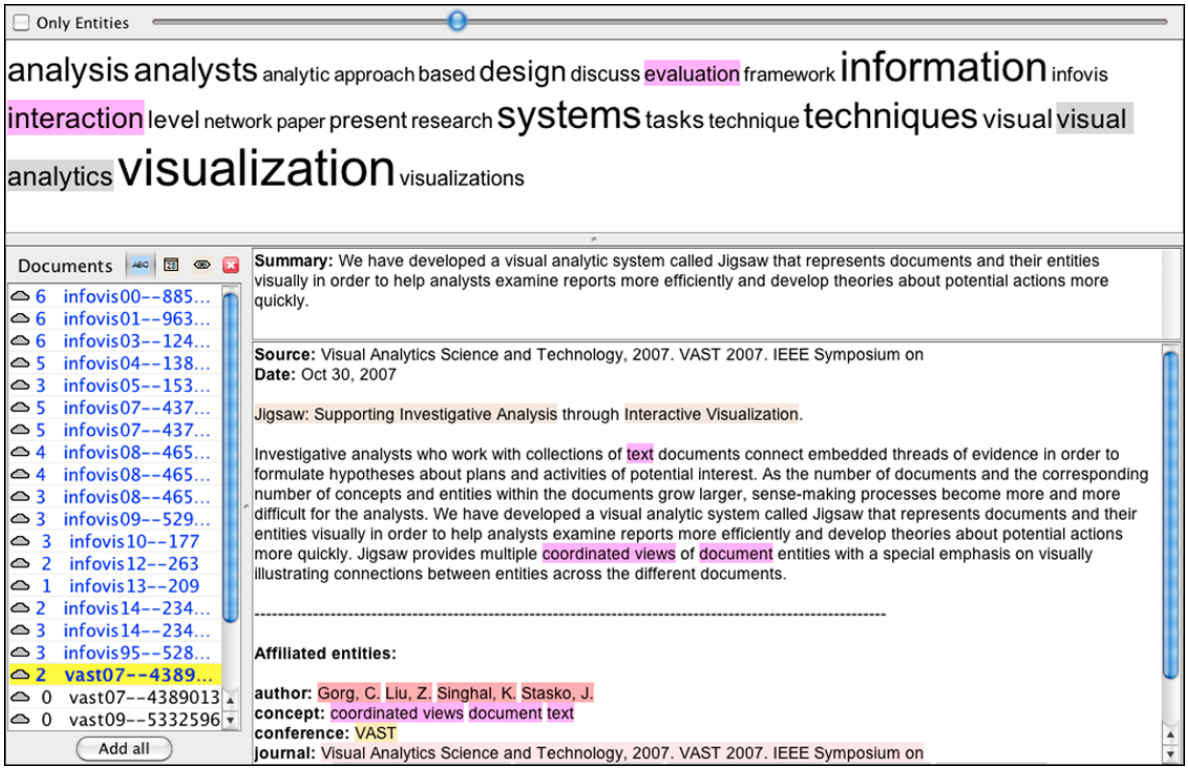


Abbildung 3.2: Document View

Document View
Diese Darstellung ermöglicht dem Anwender einen Überblick über ein einzelnes Dokument. In Abbildung 3.2 sind die vier Bereiche und die jeweiligen Inhalte zu sehen. Auf der linken Seite sind die derzeit betrachteten Dokumente aufgelistet und können nach verschiedenen Gesichtspunkten sortiert werden. Der obere Bereich nutzt eine sogenannte *WordCloud*, welche vorkommende Wörter anhand der Auftrittshäufigkeit unterschiedlich groß darstellt. Dies versetzt den Anwender in die Lage die wichtigsten Schlagwörter der Dokumentensammlung auf einen Blick zu erkennen und die betrachteten Dokumente entsprechend einzugrenzen.

Der rechte Bereich bezieht sich lediglich auf das aktuell selektierte Dokument und stellt den originalen Text sowie die gefundenen Entitäten dar. Diese werden zudem im originalen Text je nach Entitätstyp farblich unterlegt und können bearbeitet werden. Zudem identifiziert *Jigsaw* für jedes Dokument einen Satz, welcher den Inhalt möglichst umfassend beschreibt und stellt diesen gesondert heraus.

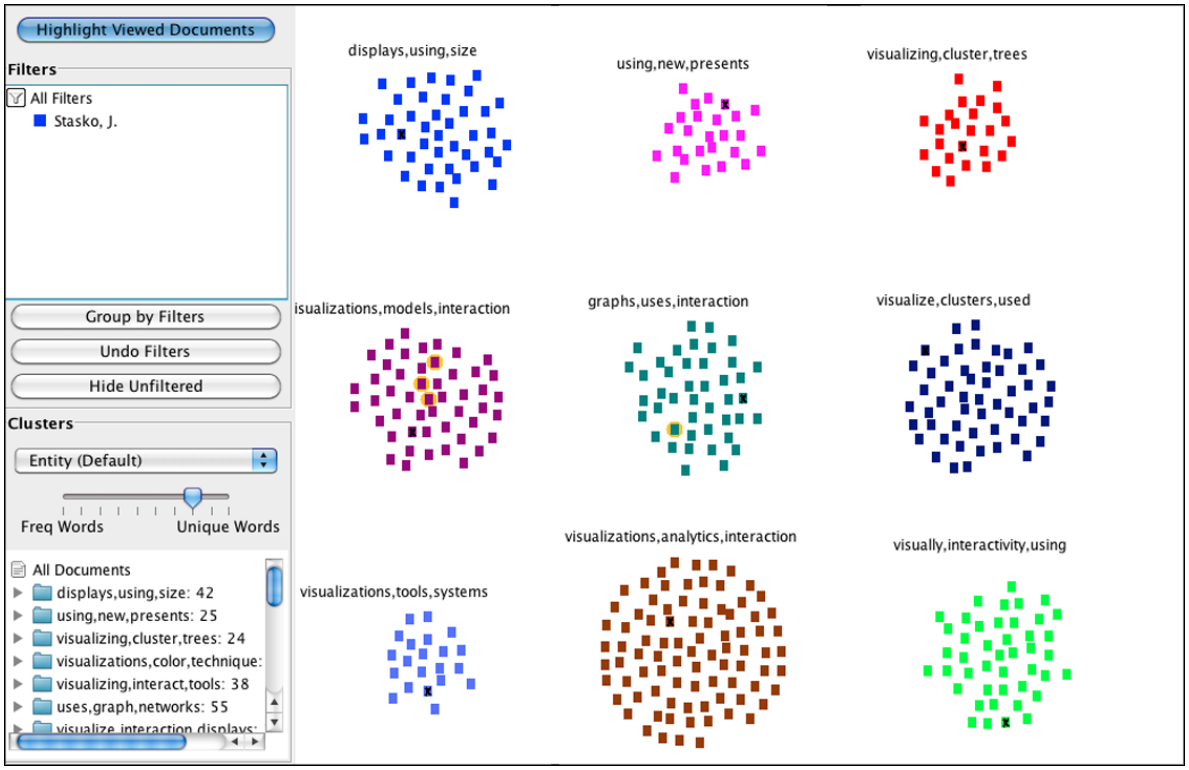


Abbildung 3.3: Document Cluster View

Document Cluster View
Einen vollständigen Überblick über alle ausgewählten Dokumente erhält der Anwender durch den *Document Cluster View*. In Abbildung 3.3) ist ein automatisches Clustering der Dokumente, anhand der Ähnlichkeit der identifizierten Entitäten oder des beinhalteten Textes, dargestellt. Jedes Dokument wird hierbei als kleines Rechteck repräsentiert und zeigt als Tooltip den Satz innerhalb des Dokumentes an, welcher dieses bestmöglich beschreibt.

Im linken Bereich hat der Anwender die Möglichkeit nach frei definierbaren Filtern, beispielsweise bestimmte Autoren oder Schlagworte, das Clustering zu verändern und die betreffenden Dokumente farblich hervorzuheben.

In der unteren Hälfte dieser Spalte werden die einzelnen Gruppen und die Anzahl, der darin zusammengefassten Dokumente, aufgelistet. Der Anwender kann die verwendete Gruppenbeschreibung anpassen, wobei die häufigsten bzw. möglichst seltenen Wörter in mehreren Abstufungen zur Verfügung stehen.

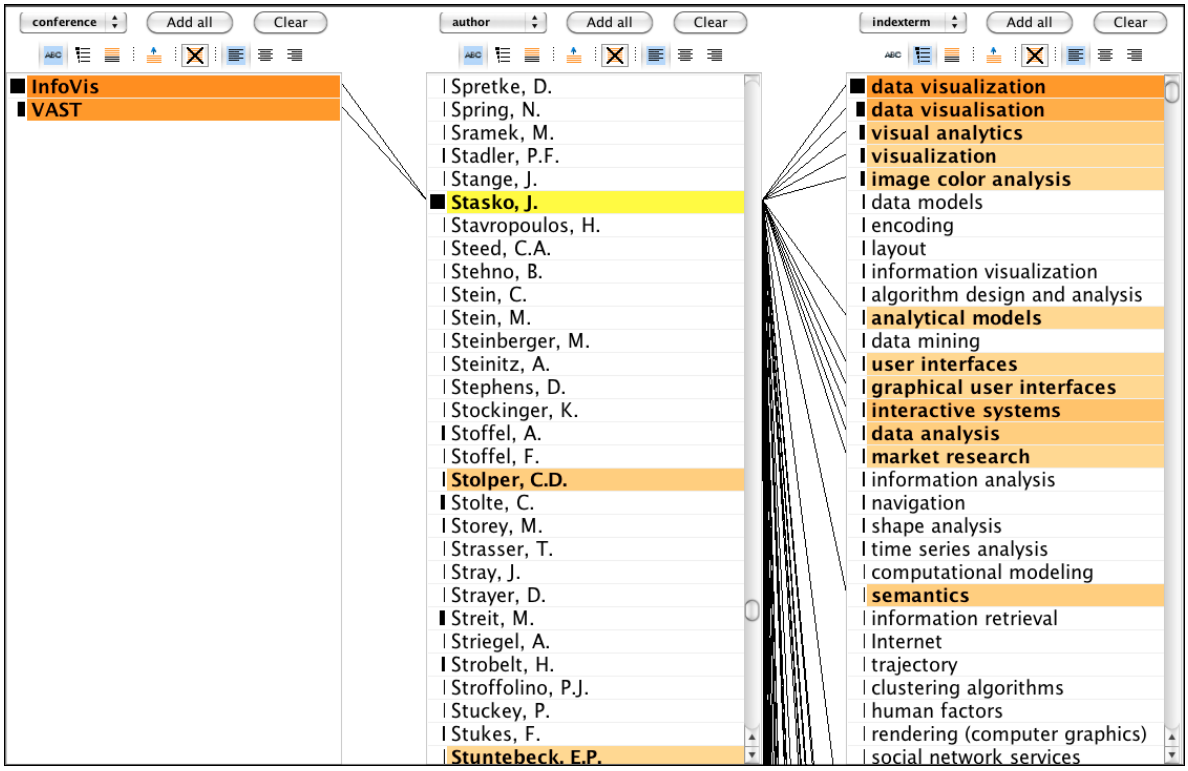


Abbildung 3.4: List View

List View
Der *List View* erlaubt Zusammenhänge und Relationen zwischen beliebigen Entitäten herauszuarbeiten. Hierzu kann der Anwender eine beliebige Anzahl an Tabellen erstellen und diese einem Entitätstyp zuweisen, woraufhin diese mit den verschiedenen Entitäten gefüllt werden.

In Abbildung 3.4 wurden drei Tabellen hinzugefügt, jeweils eine für Konferenzen, Autoren und Schlagworte. Jede Tabelle kann unabhängig von den anderen sortiert werden, so sind in der Abbildung die Autoren alphabetisch, die Schlagworte dagegen nach Häufigkeit sortiert. Die relative Häufigkeit ist hierbei durch einen schwarzen Balken vor dem jeweiligen Eintrag dargestellt.

Bei Selektion einer Tabellenzelle werden in allen Tabellen die Relationen farblich hervorgehoben. So ist beispielsweise in der Abbildung der Autor von *Jigsaw* (J. Stasko) mit beiden Konferenzen verbunden, arbeitet zudem in verschiedenen Themengebieten und mit mehreren anderen Autoren zusammen. Je gesättigter der Farbton umso häufiger tritt diese Relation auf.

Weiterhin wird durch Linien zwischen benachbarten Tabellen auf Zusammenhänge hingewiesen, auch wenn diese im derzeitigen Bildausschnitt nicht sichtbar sind.

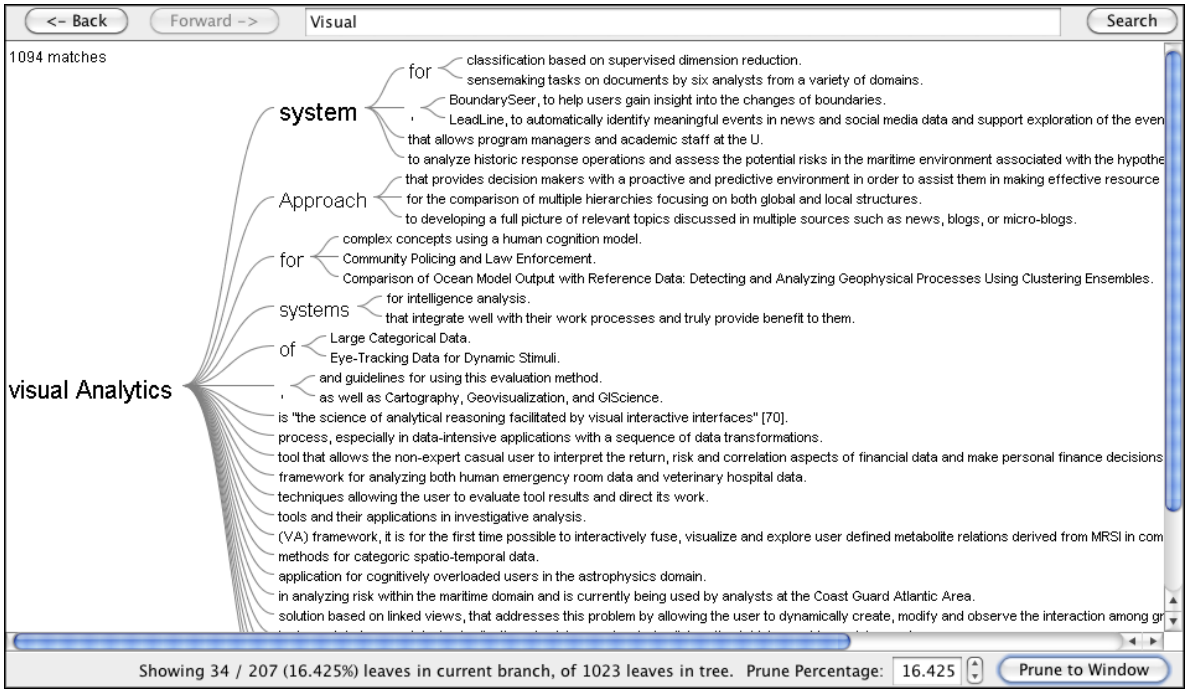


Abbildung 3.5: Word Tree View

Word Tree View
Der *Word Tree View* visualisiert dem Anwender den Kontext eines Schlagwortes. Zunächst wird nach einem beliebigen Term gesucht und der Baum baut sich ausgehend von diesem Term auf. In diesem werden ausgehend des spezifizierten Terms die nachfolgenden Worte angezeigt.

Je nach Häufigkeit des Vorkommens werden diese in ansteigender Schriftgröße und Schriftstärke abgebildet. Die Anzahl der betrachteten Pfade kann vom Anwender spezifiziert und beispielsweise auf den vorhandenen Bildschirmplatz beschränkt werden. Weiterhin kann manuell durch Selektion eines Knoten durch den Baum navigiert werden.

In Abbildung 3.5 wurde initial nach dem Schlagwort "Visual" gesucht, woraufhin *Jigsaw* als häufigsten Kontext die Terme "Analytics", "Analysis" und "Exploration" vorschlägt. Nach weiterer Selektion nach "Analytics" entsteht der abgebildete Wortbaum.

Die einzelnen Ansichten sind hierbei miteinander verknüpft. Der Anwender kann beispielsweise anhand des *List View* die Veröffentlichungen eines Autors zu einem bestimmten Thema identifizieren und anschließend direkt aus dieser Ansicht alle Veröffentlichungen dieses Autors innerhalb des *Document View* evaluieren.

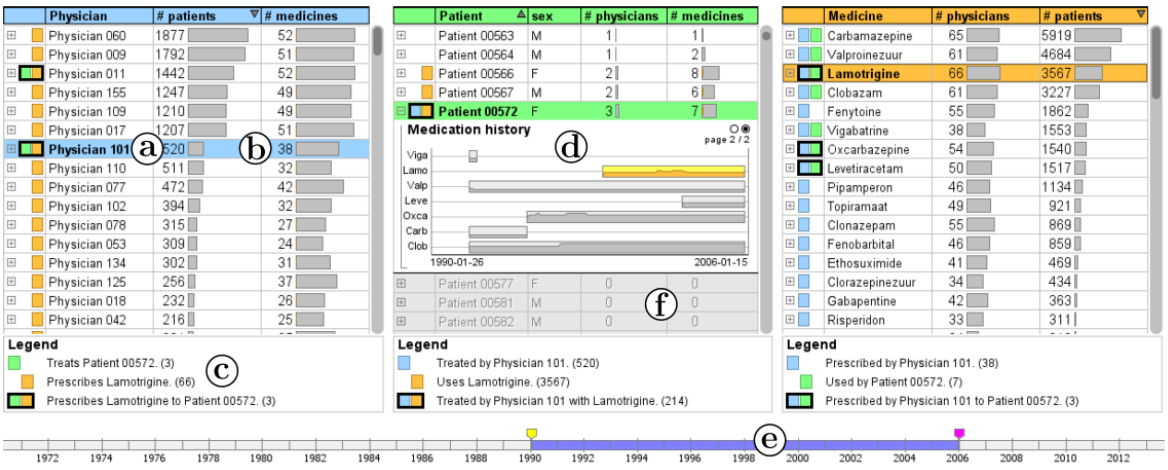


Abbildung 3.6: Visualisierung und Analyse ärztlicher Medikamentenverordnungen [CAW14]

3.1.3 Szenario III – Gesundheitswesen

Van der Corput et al. [CAW14] untersuchten, wie sich mit Hilfe von visueller, interaktiver Analyse die Zusammenhänge zwischen Ärzten, an Epilepsie erkrankten Patienten und verschriebenen Medikamenten analysieren lassen. Hierfür wurden vier Anwendungsfälle formuliert:

- Wie ändert sich die Verwendung eines bestimmten Medikamentes über einen gewissen Zeitraum?
- Hat sich die Menge der verwendeten Medikamente pro Patient in den letzten 20 Jahren verändert? Ist hierbei ein Trend erkennbar?
- Wie unterscheiden sich die Ärzte untereinander hinsichtlich der Verschreibung von Medikamenten?
- Welche Medikamente werden als erste Behandlung bevorzugt?

Entsprechend den Empfehlungen von van Wijk [Wij05] wird eine sinnvolle, initiale Visualisierung angeboten, in diesem Fall eine dreiteilige, verlinkte Tabelle, da ein Graph bei steigender Anzahl an Datensätzen zu unübersichtlich wäre.

Die initiale Oberfläche ist in Abbildung 3.6 dargestellt und bietet verschiedene Bereiche und Möglichkeiten:

Ärzte
Die linke Spalte der Tabelle listet alle Ärzte innerhalb des Datensatzes auf *(a)* und nennt zusätzlich noch die Anzahl sowohl der behandelten Patienten als auch der verschriebenen Medikamente *(b)*. Bei Selektion eines Arztes wird diese Zeile blau markiert und eine Neuberechnung ausgelöst.

Patienten
Die mittlere Spalte ist den Patienten gewidmet. Hier werden genauere Angaben wie das Geschlecht, die Anzahl der behandelnden Ärzte, die Zahl der verschiedenen Medikamente oder die Diagnose angezeigt. Die Farbe des markierten Patienten ist in diesem Abschnitt grün.

Medikamente
Die verschiedenen verschriebenen Medikamente des gesamten Datensatzes werden in der dritten Tabelle dargestellt. Hier erhält der Anwender die Information über die Anzahl der Ärzte, die dieses spezifische Medikament verordnet haben und die Anzahl der Patienten die selbiges erhielten. Wird ein Medikament ausgewählt wird die Zeile orange markiert.

Legende
Alle drei obig genannten Bereiche besitzen eine Spalte, um die Beziehungen zwischen Ärzten, Patienten sowie Medikamenten darzustellen. Hierbei werden bei Selektion eines Medikamentes die anderen beiden Bereiche aktualisiert und alle Ärzte bzw. Patienten mit einem kleinen orangefarbenen Viereck gekennzeichnet. Diese Markierung symbolisiert bei einem Arzt die Verordnung bzw. die Einnahme des Medikamentes auf Patientenseite. Sollte ein Arzt sowohl das markierte Medikament, als auch den ausgewählten Patienten behandeln, so werden die Vierecke zusätzlich schwarz umrahmt *(c)*. Weiterhin wird in der Legende die Häufigkeit angezeigt.

Details/Gruppierung/neue Fenster
Sollte der Anwender Interesse an einem bestimmten Datensatz zeigen kann die jeweilige Zeile erweitert werden, um beispielsweise die Medikamenten-Historie eines Patienten einzusehen *(d)*. Alternativ können die einzelnen Bereiche auch gruppiert werden, um so etwa alle Patienten mit einer bestimmten Diagnose zusammenzufassen.

Zeitraum
Letztlich kann der Zeitraum eingegrenzt oder erweitert werden *(e)*. Diejenigen Einträge, welche nicht mehr berücksichtigt werden können somit in der Übersicht ausgegraut werden *(f)*.

3.2 Visuelle Verfahren im Knowledge Discovery-Prozess

Um trotz der größtmöglichen Analysefreiheit ein strukturiertes Vorgehen zu ermöglichen bietet sich ein schrittweiser, aufeinander aufbauender Prozess an. Eine umfassende Beschreibung der notwendigen Schritte für die gesamte Analyse inklusive der Vorbereitung der zu untersuchenden Daten ist der *Knowledge Discovery*-Prozess (vgl. Abschnitt 2.4).

Auf den folgenden Seiten wird untersucht, inwiefern Teilschritte desselben visuell unterstützt werden können und hierzu ein Auszug verschiedener Ansätze vorgestellt.

3.2.1 Verfahren für unstrukturierte Daten

Unstrukturierte Daten können mit Hilfe des Text Mining strukturiert werden. In Abschnitt 3.1.2 wurde mit *Jigsaw* eine Anwendung vorgestellt, welche mit unstrukturierten Daten arbeitet und mit Hilfe automatischer Verfahren Text strukturieren kann.

Weiterhin bietet *Jigsaw* verschiedene Möglichkeiten die Datenqualität zu erhöhen, beispielsweise durch die Zusammenfassung der Entitäten oder der Beseitigung von vermuteten Tippfehlern. *Jigsaw* zeigt exemplarisch, wie eine visuelle Exploration und Vorverarbeitung von unstrukturierten Daten aussehen könnte.

3.2.2 Verfahren für die Vorverarbeitung

Für die interaktive und visuelle Aufbereitung von Daten entwickelten Kandel et al. [KPHH11] *Wrangler*. Dieses ermöglicht dem Anwender umfassende Transformationen an einem Datensatz vorzunehmen.

Die Benutzeroberfläche von *Wrangler* ist in Abbildung 3.7 abgebildet. Diese ist zweigeteilt und bietet sowohl eine Tabelle mit den Daten, als auch die bisher ausgeführten Aktionen. Hierbei soll der Anwender seine Ziele möglichst schnell und intuitiv erreichen können anstatt mit komplizierten regulären Ausdrücken hantieren zu müssen. Weiterhin können anhand der durchgeführten Schritte Regeln abgeleitet werden, welche exportiert und später automatisiert auf neue Daten angewendet werden können.

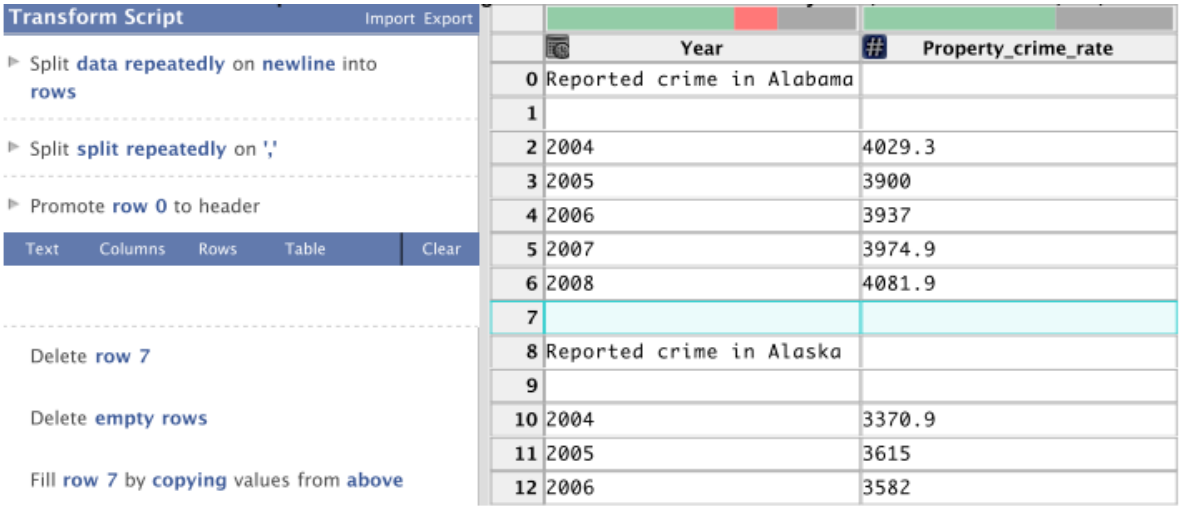


Abbildung 3.7: Wrangler-Interface zur Aufbereitung eines Datensatzes [KPHH11]

Im Folgenden werden die Möglichkeiten kurz skizziert.

Einfügen eines Datensatzes
Der Anwender wählt eine Datei aus oder kopiert den Inhalt selbiger in *Wrangler*. Der darin enthaltende Datensatz wird anschließend in einer Tabelle dargestellt.

Vorschläge
Bei der Auswahl eines Elementes (Zelle, Spalte, Zeile) werden dem Anwender verschiedene Vorschläge unterbreitet, beispielsweise Entfernen der Zeile, Entfernen gleicher Zeilen, Kopieren, Aufteilen, etc.

Qualitätsanzeige
Oberhalb jeder Spalte wird die gegenwärtige Qualität der Daten innerhalb dieser Spalte angezeigt. Diese wird anhand der Datentypen, Vollständigkeit oder Plausibilität berechnet.

Text-Extraktion
Der Anwender kann einen Text innerhalb einer Zeile markieren und in eine neue Spalte extrahieren. Basierend auf der Charakteristik der Markierung, beispielsweise der Text nach einem Leerzeichen, wird diese Extraktion auf allen Zellen dieser Spalte durchgeführt.

Leere Zellen
Zellen ohne Werte können durch den nächsten Wert von oben/unten gefüllt oder gelöscht werden.

Reshaping
Der Anwender kann die Anordnung der Tabelle reorganisieren, indem mehrere Spalten auf Schlüssel-Werte-Paare reduziert (*fold*) oder neue Spalten auf Basis vorhandener Datenwerte erstellt werden (*unfold*).

Lookup-Tables
Um einerseits die Datenqualität zu prüfen und andererseits Zuordnungen auf andere Aggregationsebenen vornehmen zu können, unterstützt *Wrangler Lookup-Tables*. So können beispielsweise Postleitzahlen auf Korrektheit geprüft werden oder auf Landkreise/Bundesländer zugeordnet werden.

Regeln in natürlicher Sprache
Die Regeln werden in natürlicher Sprache dargestellt und sind infolgedessen leicht verständlich und editierbar. Weiterhin können diese Regeln im fortgeschrittenen Verarbeitungsstatus verändert werden, beispielsweise anstatt einer Kopie eine Interpolation durchgeführt werden. In diesem Fall werden alle nachfolgenden Aktionen auf Basis dieser Änderung erneut berechnet.

Weitere Funktionen
Wrangler unterstützt verschiedene weitere Aktionsmöglichkeiten wie das Sortieren des Datensatzes, Schlüsselgenerierung (Skolemisierung) oder mathematische Aggregierungsfunktionen (Minimum, Maximum, Durchschnitt, Summe, etc.).

3.2.3 Verfahren für Data Mining

Visual Classification

Ein durchgehend interaktiver Ansatz für die visuelle Konstruktion von Entscheidungsbäumen stammt von Ankerst et al. [AEEK99]. Hierbei wird zunächst der komplette Datensatz visualisiert und ein leerer Entscheidungsbaum angelegt. Anschließend wählt der Anwender ein Attribut und eine beliebige Anzahl an Trennintervallen. Diese können durch einen Slider genauer spezifiziert und so das Trennverhalten des Entscheidungsbaumes festgelegt werden. Jedes Intervall kann entweder zu einem Blatt oder zu einem neuen Knoten führen. Durch die Selektion eines Knotens wird die Visualisierung des Datensatzes aktualisiert und beinhaltet nur Attribute, welche auf diesem Pfad noch nicht verarbeitet wurden. Einem Blatt wird eine Bezeichnung zugewiesen und der Pfad endet an dieser Stelle. Wenn jeder Pfad an einem Blatt endet ist der Entscheidungsbaum vollständig und das Verfahren abgeschlossen.

Diese Vorgehensweise ermöglicht eine freie Konstruktion des Entscheidungsbaumes, jedoch ohne die Kombination mit einem automatischen Verfahren. Somit genügt dieser Ansatz alleine nicht für die in dieser Arbeit verwendete Definition von *Visual Analytics*, jedoch könnte durch automatische Verfahren ein Trennattribut und Intervall vorgeschlagen werden und der Anwender anschließend seine Entscheidung treffen. Dieser Limitierung wirkt eine Weiterentwicklung [AEK00] dieses Konzeptes entgegen, in der die prinzipielle Vorgehensweise erhalten bleibt und zusätzlich automatische Verfahren implementiert sind. Dies bietet drei Vorteile:

Vorschlag
Das System schlägt auf Basis verschiedener vom Anwender selektierter Attribute jenes Attribut vor, welches die optimale Teilung verspricht und weiterhin den genauen Wert an welchem getrennt werden sollte.

Vorschau
Bevor ein Attribut getrennt wird kann das System vorausberechnen wie ein Teilbaum auf Basis dieser Trennung aussehen könnte. Der Anwender kann hierbei optional zu berücksichtigende Parameter (z. B. maximale Tiefe) angeben. Hierdurch kann die Auswahl des zu trennenden Attributes unterstützt und spätere Korrekturen vermieden werden.

Vervollständigung
Der Anwender kann dem System die Konstruktion des aktuellen Teilbaumes überlassen. In diesem Fall kommen herkömmliche Algorithmen zum Einsatz, welche die benötigte Zeit gegenüber der manuellen Konstruktion drastisch reduzieren können. Hierfür können äquivalent zur Vorschau-Funktion Parameter angegeben werden um den generierten Baum zu beeinflussen.

Das Fazit der Autoren legt nahe, dass der Einsatz automatischer Verfahren für die optimale Genauigkeit nötig sind, jedoch das semi-automatisierte Verfahren dem vollständig automatisierten vorzuziehen ist und bessere Ergebnisse auf verschiedenen Testdaten liefert.

Visual Clustering

Clustering-Verfahren zielen darauf ähnliche Elemente zu gruppieren bzw. Unterschiede aufzudecken. Dies ist ebenfalls eine Stärke der menschlichen Wahrnehmung – insbesondere Position und relative Nähe, Form und Abgleich mit Mustern [OW11]. Es existieren somit unzählige verschiedene Verfahren. Hinneburg [Hin14] unterteilt Clustering-Verfahren in vier Ansätze, welche in unterschiedlichem Ausmaß visuelle und automatische Verfahren verbinden. Davon unterstützen zwei das *Visual Analytics*-Prinzip und werden folgend ausgeführt:

Steuerung automatisierter Algorithmen
Im Gegensatz zu vollautomatisierten Algorithmen wird die Black Box geöffnet und der aktuelle Zustand visualisiert. Hierdurch soll der Anwender ein besseres Verständnis über die gebildeten Cluster erhalten und kann zudem interagieren um die Clusterbildung zu steuern. Generell wird dies erreicht, indem ein Teil des automatischen Verfahrens durch eine interaktive, visuelle Prozedur ersetzt wird.

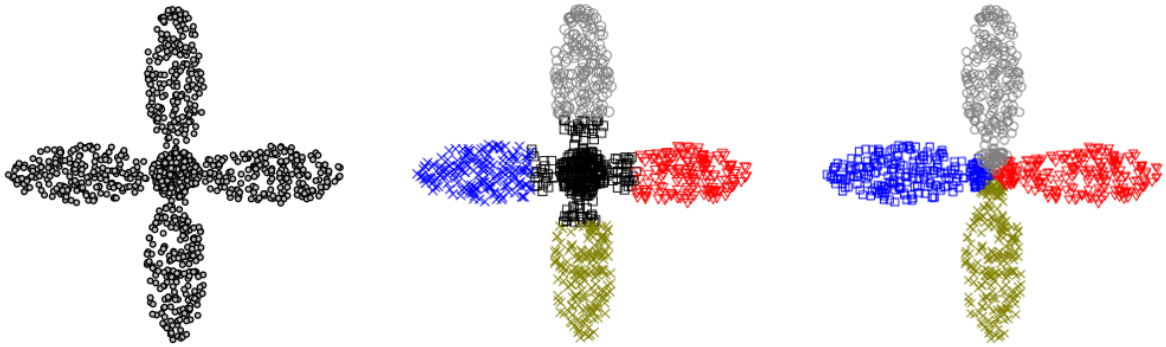
Modellauswahl
Dieser Ansatz berechnet eine Vielzahl von Clustern durch unterschiedliche Algorithmen mit unterschiedlichen Parametern und präsentiert diese dem Anwender. Hierzu wird eine Distanzmetrik berechnet und anschließend als Scatterplot visualisiert. Mit Hilfe dessen können die unterschiedlichen Interpretationen evaluiert und das verwendete Modell ausgewählt werden

Ein Beispiel für die Steuerung von Clustering-Verfahren basiert auf dem Scatter/Gather-Prinzip, welches von Cutting et al. [CKPT92] beschrieben wurde. Dieses sieht vor, dass zunächst dem Anwender ein Überblick über einzelne, initiale Cluster präsentiert wird. Anschließend kann eine beliebige Anzahl Cluster selektiert werden (*gather*) und auf dieser Teilmenge erneut Cluster gebildet werden (*scatter*). Dieser Wechsel zwischen Selektion und Neuberechnung erlaubt es dem Anwender die entstehenden Cluster an die eigenen Vorstellungen anzupassen.

Endert et al. [EHR+14] bzw. Hossain et al. [HOG+12] veranschaulichen diese interaktive Variante und belegen die entstehenden Vorteile, welche sich durch Interaktion mit dem Anwender ergeben.

In Abbildung 3.8a ist der initiale Datensatz visualisiert. Das menschliche Gehirn ist darauf ausgelegt in dieser Punktmenge in wenigen Augenblicken die vorhandenen Formen und zusammenhängende Blöcke zu erkennen. Offensichtlich ähnelt diese Punktmenge einem Windrad oder einer Blume und setzt sich aus fünf Bereichen (4 Blätter, Stiel) zusammen. Der Einsatz eines automatischen Clustering-Verfahrens (k-means [Mir11]) liefert jedoch keine zufriedenstellenden Ergebnisse:

k-means (k=5)
Die für den Menschen leicht zu erkennenden Cluster werden nicht erkannt (vgl. Abbildung 3.8b). Stattdessen werden Elemente der Blätter ebenfalls dem Stiel zugerechnet.



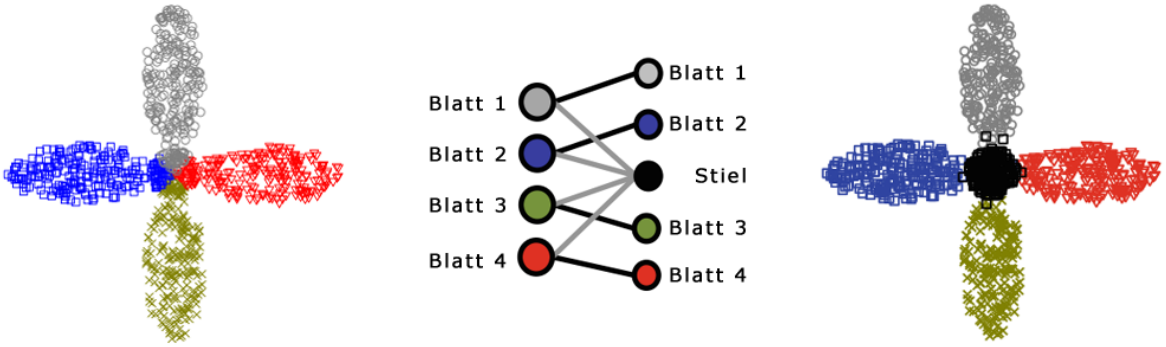
(a) Ausgangsdaten **(b)** k-means (k=5) **(c)** k-means (k=4)

Abbildung 3.8: Überblick über Clustering-Ergebnisse mit k-means (angelehnt an [EHR+14])

k-means (k=4)
Wenn k-means mit verringertem k brechnet wird, so beinhalten die Cluster der Blätter zusätzlich Teile des Stiels (vgl. Abbildung 3.8c).

Beide in Frage kommenden Parameter generieren demnach zwar korrekte Unterteilungen der Punktmenge nach objektiven Maßstäben, dennoch entsprichen die Ergebnisse nicht dem subjektiven Empfinden und spiegeln somit nicht die Wahrnehmung wider.

An dieser Stelle setzt das Scatter/Gather-Verfahren an. Zunächst wird mit einem automatischen Verfahren ein initiales Clustering erstellt (vgl. Abbildung 3.9a). Anschließend kann der Anwender beliebig häufig eine Untermenge an Clustern wählen (gather) und diese durch Neuberechnung von Clustern auf dieser Teilmenge (scatter) restrukturieren. Eine sinnvolle Menge an Operationen ist in Abbildung 3.9b dargestellt.



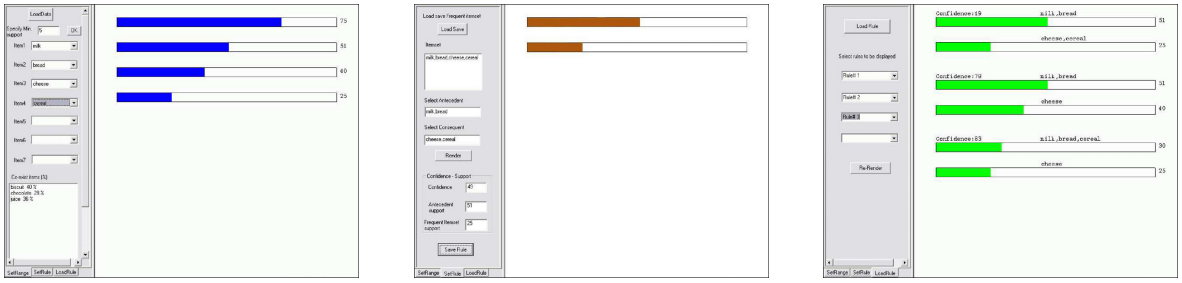
(a) Automatisches Verfahren **(b)** Operationen **(c)** Ergebnis

Abbildung 3.9: Scatter/Gather-Clustering (angelehnt an [EHR+14])

Die Punkte von Blatt 1 werden demnach aufgeteilt in Blatt 1 und Stiel, indem das blau eingefärbte Cluster selektiert und neu berechnet werden. Die Aufteilung der anderen Blätter erfolgt analog. Abschließend werden die entstandenen mittleren Cluster verbunden. Der genaue Ablauf inklusive Nutzeroberfläche und mathematischen Grundlagen werden von Hossain et al. [HOG+12] ausgeführt. Am Ende des interaktiven Ansatzes steht eine Menge an Clustern, welche der menschlichen Wahrnehmung entsprechen (vgl. Abbildung 3.9c) und von vollautomatisierten Verfahren nicht abgedeckt werden können.

Visual Association Rules

Ein visueller Ansatz zur Generierung von Assoziationsregeln stammt von Techapichetvanich und Datta [TD04] und ist in Abbildung 3.10 dargestellt.



(a) Schritt 1 **(b)** Schritt 2 **(c)** Schritt 3

Abbildung 3.10: Visuelle Generierung von Assoziationsregeln [TD04]

Im ersten Schritt spezifiziert der Anwender den minimalen Support, auf dessen Basis die Attribute hinsichtlich ihrer Häufigkeit sortiert und durch – je nach Support unterschiedlich gefüllten Balken – visualisiert werden (vgl. Abbildung 3.10a). Mit diesen Informationen kann der Anwender anschließend im linken Bereich der Nutzerschnittstelle Attribute hinzufügen und wird zusätzlich auf weitere Attribute hingewiesen, welche häufig mit den zu diesem Zeitpunkt selektierten Attributen zusammen auftreten. Der Anwender bestimmt somit selbst das Frequent Itemset, auf welchem weitere Analysen ausgeführt werden.

In einem weiteren Schritt kann der Anwender beliebige Kombinationen innerhalb des Frequent Itemset angeben und anschließend Support und Confidence dieser Kombination berechnet werden. Die Visualisierung beschränkt sich in diesem Zwischenschritt auf einen Balken für den Support der linken Seite. Ein weiterer Balken stellt den Support der gesamten Regel dar (vgl. Abbildung 3.10b). Auf diesem Wege kann der Anwender beliebige Assoziationsregeln generieren und bei Bedarf speichern.

Abschließend werden die in Schritt 2 erstellten Regeln zusammen visualisiert (vgl. Abbildung 3.10c). Hierbei können einzelne Regeln ausgeblendet oder die Anordnung verändert werden, um dem Anwender einen Überblick und den Vergleich zwischen verschiedenen Assoziationsregeln zu ermöglichen. Die Visualisierung besteht erneut aus zwei Balken je Assoziationsregel für den Support der linken Seite und den Support der vollständigen Regel, zusätzlich ist die Confidence textuell angegeben.

Dieser Ansatz erlaubt vollständige Kontrolle des Prozesses und die Freiheit die aus Anwendersicht interessanten Assoziationsregeln zu erstellen. Zudem werden im Gegensatz zu automatisierten Verfahren nur die Assoziationsregeln berechnet, nach denen der Anwender explizit verlangt. Jedoch wird hierbei außer Acht gelassen, dass nach der Definition von Data Mining (vgl. Abschnitt 2.2) bisher verborgene Muster und Regelmäßigkeiten gefunden werden sollen.

3.2.4 Verfahren für die Präsentation

Für die Präsentation der Ergebnisse sind keine gesonderten visuellen Verfahren nötig, da es sich hierbei um die Kernfunktion der Visualisierung handelt. Entsprechend existieren viele unterschiedliche Möglichkeiten um dem Anwender gefundene Muster oder Ergebnisse zu präsentieren. Im folgenden Abschnitt werden am Beispiel von Assoziationsregeln verschiedene Visualisierungsmöglichkeiten [BD08] aufgezeigt.

Die einfachste Art der Visualisierung ist eine Tabelle (vgl. Abbildung 3.11a). Diese unterteilt sich in einen Bereich für die linke und rechte Seite der Assoziationsregel, sowie je einer Spalte für Support und Confidence (vgl. Abschnitt 2.2). Diese Darstellung ermöglicht dem Anwender die Assoziationsregeln nach verschiedenen Gesichtspunkten/Spalten zu sortieren, jedoch können lediglich einzelne Ausschnitte evaluiert werden [BD08].

In Abbildung 3.11b werden Assoziationsregeln als Matrix dargestellt. Die z-Achse stellt die Elemente der linken, die x-Achse die der rechten Seite der Assoziationsregeln dar. Die y-Achse gibt hierbei die berechnete Confidence der einzelnen Regeln an, während die Einfärbung der Säulen den Support widerspiegelt. Hierbei lassen sich jedoch lediglich 1-zu-1-Beziehungen sinnvoll visualisieren [BD08]. Weitere Probleme dieser Visualisierung bestehen darin, dass sich einzelne Säulen überdecken und geringe Unterschiede nur schwer erkennbar sind [WWT99].

Um die zuvor geschilderten Schwierigkeiten zu vermeiden wurde die in Abbildung 3.11c dargestellte Visualisierung entwickelt. Hierbei wird eine andere Perspektive gewählt und versucht zusätzlich zusammengesetzte Assoziationsregeln zu unterstützen. Die Zeilen stehen für die einzelnen Elemente, während jede Spalte eine Assoziationsregel repräsentiert. Die Höhe der Markierung differenziert auf welcher Seite der Regel ein Element auftritt. In den hinteren beiden Zeilen werden Support und Confidence durch entsprechende Höhe der Markierung angegeben.

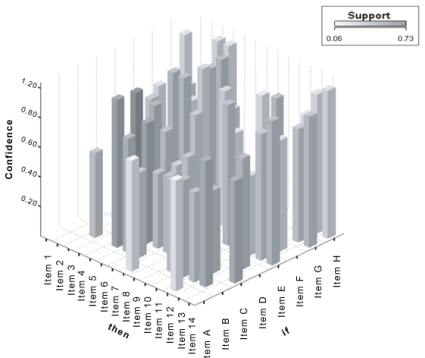
Diese Visualisierung erlaubt dem Anwender auch sehr viele Assoziationsregeln zu überblicken und so einen Gesamteindruck zu gewinnen. Weiterhin ist eine Analyse sowohl hinsichtlich einer bestimmten Regel als auch eines festgelegten Elements möglich. Dennoch könnten einzelne Markierungen verdeckt werden, auch wenn die Wahrscheinlichkeit hierfür, durch die Platzierung der tendenziell hohen Support- und Confidence-Säulen, deutlich reduziert wird. [BD08, WWT99].

Eine weitere Möglichkeit Assoziationsregeln darzustellen ist das bewährte Konzept der Parallelen Koordinaten [Ins85]. Hierbei sind die Elemente auf einer vertikalen Achse angeordnet (vgl. Abbildung 3.11d). Die Breite einer Linie symbolisiert den Support, die Farbe die Confidence (nach Yang [Yan03]). Zunächst werden die Elemente der linken Regelseite abgetragen, dann getrennt durch eine Pfeilspitze die Elemente der rechten Regelseite. Es ist offensichtlich, dass dieser Ansatz für eine große Anzahl Assoziationsregeln nicht geeignet ist [BD08, Yan03].

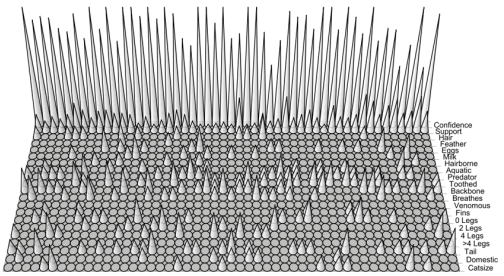
Die obig beschriebenen Ansätze sind lediglich ein Ausschnitt der Literatur zu dieser Problemstellung. Bruzzese und Davino [BD08] kommen zu dem Fazit, dass jede Visualisierung Vor- und Nachteile besitzt, es jedoch keine perfekte Visualisierung für jeden Anwendungszweck gibt. Die Stärke der Visualisierungen liegt in der kombinierten Anwendung verschiedener Techniken. Weiterhin wird vorgeschlagen, dass die Verbindung zwischen automatischen und visuellen Verfahren gesteigert werden sollte.

	A	B	C	D	E	F	G
1	**Antecedent items**				**Consequence**	**Confidence**	**Support**
2	Breathes	Toothed			Backbone	1.00	0.47
3	Backbone	Milk	Toothed		Breathes	1.00	0.40
4	Breathes	Milk	Toothed		Backbone	1.00	0.40
5	0 Legs	Backbone			Tail	0.95	0.18
6	Backbone	Hair	Milk		Breathes	1.00	0.39
7	Breathes	Hair	Milk		Backbone	1.00	0.39
8	Backbone	Breathes	Hair	Toothed	Milk	1.00	0.38
9	0 Legs	Catsize			Tail	0.86	0.06
10	0 Legs	Predator			Eggs	0.76	0.13
11	Eggs	Fins	Predator	Toothed	Tail	1.00	0.09
12	Predator	Tail	Toothed	Venomous	Eggs	0.67	0.02
13	Tail				Toothed	0.69	0.51
14	>4 Legs	Eggs			Breathes	0.67	0.08
15	>4 Legs	Hairborne			Hair	0.67	0.04
16	0 Legs	Aquatic			Backbone	0.94	0.17
17	2 Legs	Aquatic	Eggs		Hairborne	0.83	0.05
18	2 Legs	Aquatic	Tail		Eggs	0.86	0.06

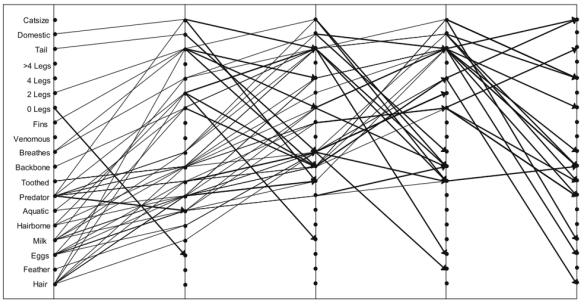
(a) Tabelle [BD08]



(b) 3-dimensionales Säulendiagramm (angelehnt an [BD08])



(c) 3-dimensionale Matrix [BD08]



(d) Parallele Koordinaten [BD08]

Abbildung 3.11: Visualisierungen von Assoziationsregeln

4 Visual Analytics – Definition, Abgrenzung und Ordnungsrahmen

In diesem Kapitel wird zunächst eine neue Definition für *Visual Analytics* entwickelt und die vorgestellten Szenarien darauf basierend bewertet. In einem weiteren Schritt wird *Visual Analytics* gegenüber der Visualisierung und dem Data Mining abgegrenzt sowie die jeweiligen Vor- und Nachteile zusammengefasst. Abschließend wird *Visual Analytics* auf Basis verschiedener Dimensionen in einem Ordnungsrahmen strukturiert.

4.1 Definition

In Abschnitt 2.6 wurden verschiedene Definitionen für *Visual Analytics* vorgestellt. Um diesen Begriff zu spezifizieren, wird nachfolgend *Visual Analytics* hinsichtlich der Ziele, des Prozesses und der verwandten Gebiete näher evaluiert.

Ziele
Visual Analytics hat in der Literatur übereinstimmend das Ziel, den Prozess der Sinnstiftung/Entscheidungsfindung zu verbessern [KMS+08, EFN12, TC05]. Um dieses zu erreichen sollen die außergewöhnlichen menschlichen Fähigkeiten der Informationsverarbeitung [ABM07], Wahrnehmung [TC05], Flexibilität [KMOZ08], Intuition [PT04, EFN12] und dem vorhandenen Hintergrundwissen [KMOZ08] mit der enormen maschinellen Rechenkraft [ABM07] und Speicherkapazitäten [KMOZ08] durch mathematische/statistische Verfahren [PT04, KBC+07, TJKMW98] kombiniert werden um Einsicht in große Datenmengen [KMS+08, EHR+14] zu erhalten.

Das übergeordnete Ziel ist es die "qualitativ hochwertige menschliche Urteilsfähigkeit zu nutzen, jedoch bei möglichst geringer aufzuwendender Zeit des Analysten-[TC05] und dabei sowohl erwartete Ergebnisse, als auch unerwartete, neue Einsichten zu gewinnen (*detect the expected, discover the unexpected*) [PT04, TC05, KKM+10b]. Endert et al. [EHR+14] sehen dabei den Anwender in der entscheidenden Rolle.

Prozess
Den *Visual Analytics*-Prozess definieren Keim et al. wie folgt:

> *"Visual Analytics is an iterative process that involves information gathering, data preprocessing, knowledge representation, interaction and decision making."*
> Keim et al. [KMS+08]

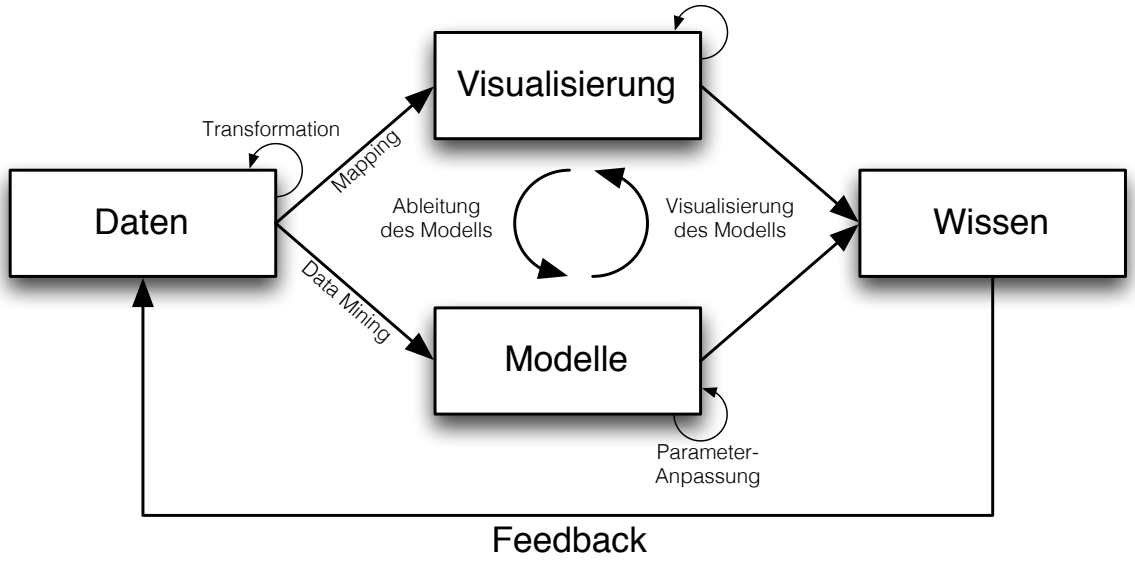


Abbildung 4.1: Visual Analytics-Prozess (angelehnt an [KMOZ08, S. 10])

Diese Definition beschreibt *Visual Analytics* als umfassenden Prozess der Datenanalyse über mehrere Stufen. In Abbildung 4.1 ist der *Visual Analytics*-Prozess nach Keim et al. [KMOZ08] dargestellt und wird im Folgenden erläutert.

Der Visual Analytics-Prozess nach Keim et al. ist in Abbildung 4.1 abstrakt dargestellt und erweitert den Prozess der Visualisierung durch die Möglichkeit einer automatischen Analyse. Die einzelnen Möglichkeiten und die Abfolge werden im Folgenden erläutert [KKM+10b, KMS+08]:

Daten
Daten liegen üblicherweise nicht in der benötigten Qualität oder in einer singulären Quelle vor, daher kann optional durch die *Transformation* eine Vorverarbeitung (*Data Cleansing*, Normalisierung, Gruppierung, Integration, etc.) vorgenommen werden. Dieser Schritt kann so lange ausgeführt werden bis die gewünschte, erforderliche Datenqualität erreicht ist. Sofern die Daten anschließend in der benötigten Form vorliegen kann der Anwender zwischen automatischer und visueller Analyse wählen.

Visualisierung
Entscheidet sich der Anwender für die visuelle Analyse, so werden die Daten auf visuelle Repräsentationen abgebildet und dargestellt (vgl. Abschnitt 2.5). Auf Basis dieser Visualisierung können neue Hypothesen aufgestellt werden.

Interaktion
Der Anwender steht im Mittelpunkt des *Visual Analytics*-Prozesses und hat daher Einfluss auf die Analyse. Einerseits können die Parameter des Data Mining-Algorithmus durch Interaktion mit der Visualisierung angepasst werden, oder eine andere Analysemethode gewählt werden. Ebenfalls können neue Hypothesen durch die Interaktion aufgestellt werden, indem die Visualisierung nach neuen Einsichten durchsucht wird (vgl. Abschnitt 2.5 – Visualisierung).

Data Mining
Entscheidet sich der Anwender für die automatische Analyse werden die Daten durch Data Mining-Methoden analysiert und das Resultat zur Evaluation visualisiert.

Ableitung & Visualisierung des Modells
Eine Kernfunktionalität des Visual Analytics-Prozess ist der Wechsel zwischen visueller und automatischer Analyse. Entsprechend können die Ergebnisse der visuellen Analyse für eine Verbesserung der Modelle genutzt werden. In die andere Richtung kann die visuelle Darstellung genutzt werden, um die Resultate der automatischen Analyse zu veranschaulichen, zu evaluieren, durch Verfeinerung der Parameter zu verbessern oder einen anderen Algorithmus zu wählen. Durch den Wechsel und den iterativen Ablauf können vorhergehende Ergebnisse immer weiter verfeinert werden.

Wissen
Neue Einsichten werden entsprechend im *Visual Analytics*-Prozess durch eine Kombination visueller und analytischer Verfahren gewonnen. Diese können durch den Nutzer beliebig oft iteriert werden.

Verwandte Gebiete
Der interdisziplinäre Charakter zeigt sich an der Vielzahl verschiedener Forschungsgebiete, auf denen *Visual Analytics* basiert und zurückgreift. Diese umfassen beispielsweise die wissenschaftliche Visualisierung und Informationsvisualisierung [PT04], Wissensmanagement [PT04], Statistik [PT04], Mensch-Computer-Interaktion [KAF+08], Kognitionswissenschaft [PT04, ABM07], Datenanalyse/Data Mining [KAF+08], Datenanalyse/Data Mining [KAF+08], menschliche Wahrnehmung und menschliches Denken [KKM+10b] oder Entscheidungstheorie [PT04].

Basierend auf der erfolgten Literaturrecherche umfasst keine Definition vollständig die obigen Aspekte. Vor diesem Hintergrund wird *Visual Analytics* wie folgt definiert:

> *Visual Analytics ist ein interdisziplinäres Forschungsgebiet, welches einerseits neue Einsichten in große Datenmengen, andererseits aber auch die Überprüfung von Hypothesen ermöglichen soll. Es werden iterativ Erkenntnisse und Verfahren verschiedenster Fachgebiete, z. B. Visualisierung, Data Mining oder Kognitionswissenschaft, (kombiniert) angewendet. Innerhalb des Prozesses ist der Anwender – und dessen vorhandenes Wissen – immer die zentrale und wichtigste Komponente für die Analyse.*

4.2 Evaluation der vorgestellten Anwendungen

In Kapitel 3 – Abschnitt 3.1 wurden exemplarisch Lösungen vorgestellt, welche unter dem Schlagwort *Visual Analytics* verortet werden.

Im folgenden Abschnitt wird zunächst die Umsetzung von *Visual Analytics* in den einzelnen Anwendungen bewertet. Anschließend wird evaluiert, inwiefern essentielle Dimensionen im Bezug auf *Visual Analytics* genutzt werden.

Szenario I – Meinungsanalyse
Dieser Lösungsansatz ähnelt Szenario I der identifizierten Ausprägungen des Visualisierungsprozesses (vgl. Abbildung 2.3a). Es findet jedoch weder ein wiederholter Wechsel zwischen automatischen und visuellen Verfahren statt, noch besitzt der Anwender einen Einfluss auf den Prozess. Mit obiger Definition oder dem *Visual Analytics*-Prozess kann diese Lösung demzufolge nur schwer in Verbindung gebracht werden.

Szenario II – Jigsaw
Die Applikation *Jigsaw* überlässt dem Anwender initial die Kontrolle über die verwendeten Daten und stellt umfangreiche Analysemöglichkeiten zur Verfügung. Auf diese Weise ist der Anwender der zentrale Punkt der Analyse und kann diese nach eigenen Gesichtspunkten beeinflussen. Weiterhin ist die Verknüpfung zwischen automatischer Analyse und visueller Kontrolle gut umgesetzt. *Jigsaw* ist jedoch stark auf Dokumentsammlungen und somit unstrukturierte Daten ausgelegt.

Szenario III – Gesundheitswesen
Dieses Beispiel nutzt das *Visual Analytics*-Mantra (vgl. Abschnitt 2.6). Zunächst werden die Daten analysiert und visualisiert. Der Anwender erhält anschließend die Möglichkeit verschiedene Analysen durchzuführen, jedoch ist die automatische Analyse und die Nutzereinbindung über den gesamten Prozess nicht so stark ausgeprägt, wie es die Definition oder Prozess von *Visual Analytics* vermuten lässt (vgl. Abschnitt 4.1).

Für die Umsetzung von *Visual Analytics* sind vier Dimensionen essentiell. Die vorgestellten Beispiele werden nachfolgend unter diesen Gesichtspunkten verglichen:

Einbindung des Anwenders
Visual Analytics stellt den Anwender in den Mittelpunkt innerhalb des Analyseprozesses. In Szenario I – Meinungsanalyse ist diese Voraussetzung nicht gegeben, da erst die finale Visualisierung präsentiert wird und somit kein Einfluss auf diese besteht. In Szenario II – Jigsaw und Szenario III – Gesundheitswesen hat der Anwender im Gegensatz dazu deutlich mehr Möglichkeiten den Analyseprozess zu beeinflussen, insbesondere Szenario II – Jigsaw ermöglicht eine freie Analyse im Rahmen des abgedeckten Umfangs.

Wechsel zwischen automatischen und visuellen Verfahren
Visual Analytics zeichnet sich durch einen stetigen Wechsel zwischen automatischen und visuellen Verfahren aus, solange bis der Anwender das gewünschte Ergebnis erreicht hat. In Szenario I – Meinungsanalyse ist dies nicht gegeben, da lediglich eine Visualisierung durch ein automatisches Verfahren erzeugt wird. Auch in Szenario III – Gesundheitswesen wird dieser Wechsel nur geringfügig umgesetzt, da dies auf Neuberechnung verschiedener Werte beschränkt ist und das Modell nicht verändert werden kann. Szenario II – Jigsaw dagegen zeigt eindrucksvoll, welche Möglichkeiten sich durch konstanten Wechsel ergeben und erlaubt so dem Anwender den Datensatz nach unterschiedlichsten Gesichtspunkten zu evaluieren.

Umfang der Analyse
Hinsichtlich der Möglichkeiten die sich für den Anwender ergeben schneidet Szenario I – Meinungsanalyse erneut am schlechtesten ab. Hier existiert genau ein Analyseziel. In Szenario III – Gesundheitswesen hingegen kann der Anwender viele verschiedene Analysen ausführen und verschiedene Kombinationen der einzelnen Entitäten auswerten. Szenario II – Jigsaw geht noch weiter und erlaubt eine Zusammenfassung von Entitäten, Aufspaltung in neue Entitätstypen, sowie die Anwendung verschiedener automatischer Verfahren, ist jedoch auf unstrukturierte Daten ausgelegt.

Adaptivität für andere Szenarien
Ein weiterer Punkt der für *Visual Analytics* spricht ist die Anpassungsfähigkeit auf verschiedene Szenarien. Szenario I – Meinungsanalyse und Szenario III – Gesundheitswesen sind durch die nicht vorhandene Einbindung des Anwenders auf einen festgelegten Anwendungsfall beschränkt. Für eine Anpassung müsste der Autor die Auswahl des Datensatzes verändern und die Oberfläche anpassen. Dies ist sowohl zeitlich aufwändig als auch kostenintensiv. Szenario II – Jigsaw hingegen nutzt die durch *Visual Analytics* entstehenden Möglichkeiten, lässt den Anwender über die Analyse entscheiden und kann eine Vielzahl unterschiedlicher Dateiformate einlesen.

Auch wenn sich die vorgestellten Lösungen das Themengebiet teilen ist der Grad der Umsetzung sehr verschieden und verdeutlicht, dass *Visual Analytics* meist im Zusammenhang mit einer spezifischen Problemstellung verwendet wird. Puloamäki et al. [PBT+10] sehen eine Notwendigkeit für generische Werkzeuge sowie *Visual Analytics*-Methoden u. a. für *Data Cleansing* und Integration.

Visual Analytics muss entsprechend derzeitig eher als eine Erweiterung der Visualisierung aufgefasst werden, was aufgrund der historischen Entwicklung des Terms nicht weiter verwundert. Das Hauptaugenmerk liegt in vielen Fällen auf der Visualisierung und der Erweiterung der Visualisierungspipeline und nicht, wie der *Visual Analytics*-Prozess vorgibt, auf einem stetigen Wechsel dieser Möglichkeiten. Das Konzept des *Human In The Loop* (vgl. Abschnitt 2.7) wird somit nicht konsequent genutzt.

Neben diesen drei exemplarischen Lösungen wurden in Kapitel 3 (vgl. Abschnitt 3.2) visuelle Verfahren für einzelne Stufen des *Knowledge Discovery*-Prozesses vorgestellt. Dabei zeigt sich, dass viele Schritte bereits durch visuelle Ansätze unterstützt werden können. Insbesondere das *Wrangler*-Verfahren (vgl. Abschnitt 3.2.2) unterstützt die entwickelte Definition von

Visual Analytics umfangreich. Einerseits wird der Anwender über die visuelle Oberfläche und umfangreiche Interaktionsmöglichkeiten in den Prozess eingebunden, gleichzeitig im Hintergrund Regeln generiert, welche zu einem späteren Zeitpunkt automatisch ausgeführt werden können.

Die evaluierten Verfahren zeigen, dass die Kombination von automatischen und visuellen Verfahren sehr unterschiedlich umgesetzt werden kann. Bertini und Lalanne [BL10] untersuchten diese Diversität an Ansätzen genauer und unterteilen drei Klassen:

Erweitertes Mining
Hierzu gehören Ansätze bei denen automatische Algorithmen die grundsätzliche Datenanalyse durchführen, jedoch visuelle Verfahren zur Validierung und dem Verständnis verwendet werden.

Erweiterte Visualisierung
Hierzu gehören Ansätze, welche grundsätzlich visuellen Charakter haben, jedoch automatische Verfahren zur Unterstützung der Visualisierung verwenden.

Integrierte Visualisierung und Mining
Hierzu gehören Ansätze, welche visuelle und automatische Verfahren auf eine Art kombinieren, so dass kein Teilbereich eine dominante Rolle einnimmt.

Nach der in dieser Arbeit entwickelten Definition von *Visual Analytics* sollte in jedem Schritt der Analyse der Anwender die entscheidende Rolle spielen und somit die obigen integrierten Ansätze verwendet werden. Ein bewährtes Verfahren um eine schrittweise Analyse durchzuführen ist der *Knowledge Discovery*-Prozess. In der Literatur existiert die Idee einer Verbindung von *Visual Analytics* mit dem *Knowledge Discovery*-Prozess vereinzelt bereits seit längerem. Über die Art und Weise wie dieses Ziel erreicht werden soll besteht jedoch Uneinigkeit. Brunk et al. [BKK97] sowie Bertini und Lalanne [BL10] nennen als Ziel einen interaktiven *Knowledge Discovery*-Prozess. Bei Puolamäki et al. [PBT+10] ist der *Knowledge Discovery*-Prozess dagegen als ein Hilfsmittel für *Visual Analytics* zu verstehen. Einen ganz anderen Weg gehen Berthold et al. [BCD+08] indem visuell ein Arbeitsablauf erstellt und die passende Visualisierung aus einem Katalog ausgewählt werden kann.

Ein Konsens besteht dagegen hinsichtlich der Effizienz einer Kombination automatischer und visueller Verfahren [PBT+10, PB10, TD04]. Als großes Hindernis hierbei wird die Geschwindigkeit der automatisierten Verfahren genannt, welche mit den Anforderungen an Interaktivität kollidieren [PB10]. Nach Heer und Shneiderman [HS12] ist es nötig, dass *Visual Analytics* mit der Geschwindigkeit der menschlichen Gedanken mithalten kann. Puolamäki et al. [PBT+10] spezifizieren die hierfür benötigte Reaktionszeit auf unter eine Sekunde. Techapichetvanich und Datta [TD04] sehen dagegen auch bei einem langsamen kombinierten Prozess die Vorteile gegenüber rein automatischen Verfahren deutlich überwiegen. Um die Reaktionszeit zu erhöhen schlagen Brunk et al. [BKK97] eine verteile Server/Client-Lösung vor, nach der die Interaktivität auf dem Client, die komplizierten und aufwändigen Berechnungen dagegen auf einem leistungsstarken Server ausführt.

4.3 Abgrenzung zu Data Mining und Visualisierung

In diesem Abschnitt wird *Visual Analytics* unter Berücksichtigung der entwickelten Definition gegenüber der Visualisierung und dem Data Mining abgegrenzt.

4.3.1 Abgrenzung zwischen Visualisierung und Visual Analytics

Die Grenzen zwischen diesen Termen sind nicht absolut, sondern die Anwendungsgebiete überschneiden sich. Nach Soukup und Davidson [SD02] übernimmt der Anwender bei der Visualisierung die Rolle der Data Mining-Engine und verliert so deren Vorteile, während bei *Visual Analytics* die jeweiligen Stärken zwischen Mensch und Maschine kombiniert werden. Die herkömmliche Visualisierung nach dem Mantra "*Overview first, zoom and filter, details on demand*" stößt bei großen Datenmengen an ihre Grenzen. Der Platz auf einem Display ist begrenzt, entsprechend kann bei einem zu großen Datenvolumen nicht länger ein sinnvoller Überblick gewährleistet sein, sodass für den Anwender nicht erkennbar ist welcher Bereich genauer erkundet werden sollte [KKM+10b]. An dieser Stelle kann *Visual Analytics* durch den ersten Schritt (*Analyze first*) interessante Muster finden und die Aufmerksamkeit des Anwenders auf diese lenken. Demzufolge ist *Visual Analytics* "*mehr als Visualisierung*" [KMSZ06a].

Ein weiteres Unterscheidungsmerkmal ist die Einbindung des Anwenders in den Analyseprozess. Während bei der Visualisierung unterschiedliche Umfänge denkbar sind, u. a. die reine Präsentation eines Ergebnisses (vgl. Abschnitt 2.5), so ist bei *Visual Analytics* der Anwender immer der zentrale Punkt des Prozesses. Ohne den Anwender ist *Visual Analytic*s infolgedessen unmöglich, während die Visualisierung gleichwohl zu einem abschließend zu interpretierenden Ergebnis gelangt.

4.3.2 Abgrenzung zwischen Data Mining und Visual Analytics

Die Unterschiede zwischen Data Mining und *Visual Analytics* lassen sich indessen klarer abgrenzen. Data Mining liefert die Verfahren für die automatische Analyse innerhalb des *Visual Analytics*-Prozesses (vgl. Abschnitt 4.1) und ist somit ein Teilschritt, ähnlich wie Data Mining ein Teilschritt des *Knowledge Discovery*-Prozesses ist. Data Mining ist demzufolge ein entscheidendes Hilfsmittel für *Visual Analytics*, um große Datenvolumen verarbeiten und analysieren zu können.

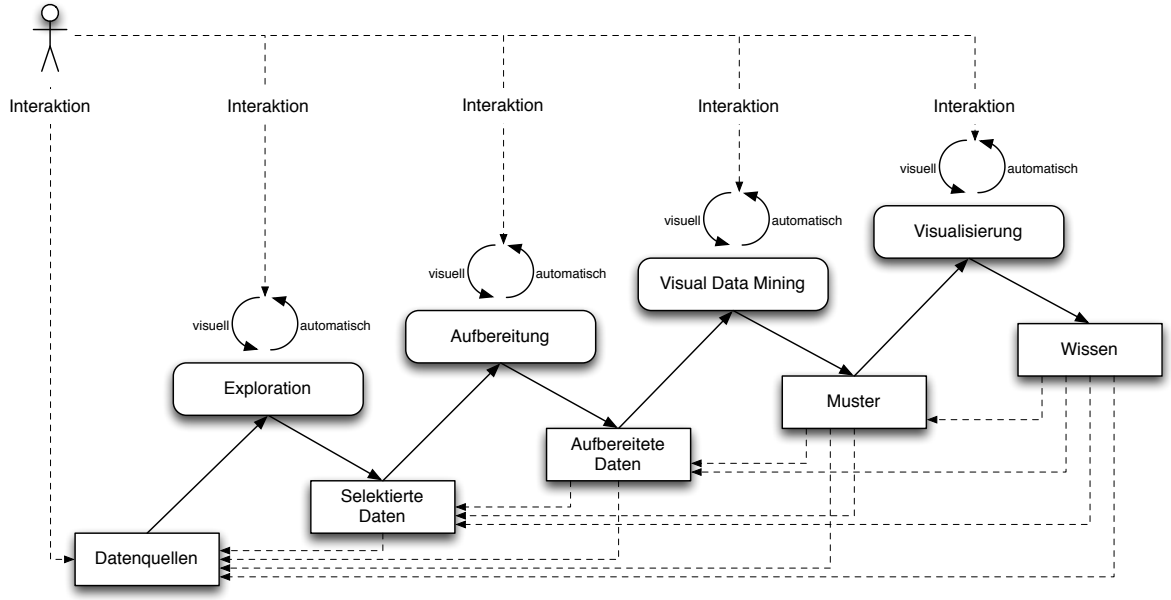


Abbildung 4.2: Erweiterter Visual Analytics-Prozess für die Datenanalyse

4.4 Visual Analytics-Prozess auf Basis des Knowledge Discovery-Prozesses

Neben *Visual Analytics*, welches sich aus der Visualisierung begründet, existiert der Begriff *Visual Data Mining*. Hierbei steht die Idee im Vordergrund Data Mining durch Visualisierung zu erweitern. Nach Ankerst [Ank01] betrifft *Visual Data Mining* die letzten beiden Stufen des *Knowledge Discovery*-Prozesses und identifiziert hierbei drei verschiedene Ausprägungen, in denen eine visuelle Unterstützung verwendet werden kann. Nach Soukup und Davidson [SD02] erlaubt *Visual Data Mining* die Interaktion mit der Visualisierung eines Data Mining-Modells, um die Resultate zu verstehen und zu überprüfen. Han und Kamber [HK06] verstehen *Visual Data Mining* als Verknüpfung von Visualisierung und Data Mining.

Die Idee hinter *Visual Analytics* und *Visual Data Mining* ist folglich sehr ähnlich und zielt auf die Einbindung des Anwenders, um den Analyseprozess zu verbessern.

Der konventionelle *Visual Analytics*-Prozess (vgl. Abschnitt 4.1) wurde nicht im Hinblick auf den *Knowledge Discovery*-Prozess entwickelt und bildet diesen entsprechend unzureichend ab. Bei diesem startet der Anwender die Analyse durch Transformation der Datenquelle, führt anschließend visuelle und automatische Verfahren im Wechsel aus und beginnt mit der gewonnenen Erkenntnis bei Bedarf von vorne. In diesem Prozess erhält der Anwender keine Möglichkeit nur einzelne Schritte zu wiederholen, sondern muss bei falscher Wahl der Datenquelle zuerst die Analyse ausführen, um mit der gewonnenen Erkenntnis erneut zu beginnen.

Nachfolgend wird ein Prozess entwickelt, welcher *Visual Analytics* und den *Knowledge Discovery*-Prozess kombiniert und hierbei nach Möglichkeit die von Bertini und Lalanne [BL10] integrierten Verfahren (vgl. Abschnitt 4.2 verwenden soll.

Der resultierende Prozess ist in Abbildung 4.2 dargestellt und basiert grundsätzlich auf dem *Knowledge Discovery*-Prozess (vgl. Abbildung 2.1).

Die Weiterentwicklung besteht hauptsächlich darin, dass in jedem Schritt ähnlich des konventionellen *Visual Analytics*-Prozesses eine Kombination von automatischen und visuellen Verfahren zum Einsatz kommt, welche durch den Anwender gesteuert werden können. Dies trägt einerseits dem Konzept des *Human In The Loop* (vgl. Abschnitt 2.7) Rechnung, anderseits wird die Kombinationsmöglichkeit verschiedener Verfahren innerhalb des Analyseprozesses hervorgehoben. Entgegen dem konventionellen *Visual Analytics*-Prozess existiert eine deutlich gesteigerte Zahl an Rückkopplungen, um die Idee des *Data Wrangling* (vgl. Abschnitt 2.10) umzusetzen. Die einzelnen Schritte werden im Folgenden kurz vorgestellt:

Exploration
Im ersten Schritt soll der Anwender die Auswahl der Daten vornehmen. Hierzu bietet es sich an, diesem bereits ein Gefühl über Inhalt und Qualität zu vermitteln. Da eine visuelle Exploration von willkürlichen Datenquellen nur schwer umsetzbar wäre ist eine Möglichkeit hierfür eine Anzahl an Datenquellen in einem Katalog vorzuhalten. Bei der Definition selbiger sollte auch eine geeignete Visualisierung für diesen Zweck spezifiziert werden. Über eine visuelle Schnittstelle können die geeigneten Daten anschließend selektiert oder durch Veränderung der Parameter die Visualisierung für weitere Exploration verändert werden. Weiterhin könnte eine visuelle Schnittstelle für die manuelle Spezifikation einer Datenquelle und passender Visualisierung eingebunden werden, um den Anwender nicht auf vorgegebene Datenquellen einzuschränken.

Aufbereitung
Im zweiten Schritt müssen die selektierten Daten aufbereitet und je nach Anwendungsfall aus verschiedenen Datenquellen integriert werden. Ein Ansatz in diese Richtung ist das, in Abschnitt 3.2.2 vorgestellte, *Wrangler*-Verfahren. Für einen einzelnen Datensatz beschreibt dieses eine Aufbereitung mit Hilfe des Anwenders und Generierung wiederverwertbarer Regeln.

Visuelles Data Mining
Der Data Mining-Schritt des *Knowledge Discovery*-Prozesses wird durch einen visuellen Data Mining-Schritt ersetzt. Verschiedene, bereits existierende Verfahren dieses Bereichs wurden in Abschnitt 3.2.3 vorgestellt und ermöglichen dem Anwender die automatischen Verfahren entsprechend des eigenen Ziels zu beeinflussen und so bessere Ergebnisse zu erhalten.

Visualisierung
Die gefundenen Muster werden wie gewohnt visualisiert. Um den Anwender auch in diesen Schritt einzubinden und die Interpretation der Ergebnisse zu unterstützen sollte diese Visualisierung interaktiv gestaltet sein. Hierdurch kann der Anwender die Ergebnisse nach eigenem Ermessen filtern und so spezifischere Erkenntnisse erlangen.

Rückkopplung
Der vorgestellte Prozess verlangt in jeder Stufe eine hohe Anzahl an Interaktionen. Der Anwender soll hierbei die Charakteristik der Daten kennenlernen und sich auf diese Weise der optimalen Lösung annähern. Es ist entsprechend unwahrscheinlich, dass bereits in der ersten Iteration dieses Resultat erreicht wird, vielmehr ist es nötig die gewonnenen Erkenntnisse zu verwenden. Hierbei ist es jedoch nicht sinnvoll die Analyse von Neuem zu beginnen, sondern vielmehr lediglich die von der Änderung beeinflussten Stufen zu berücksichtigen. Dies ist eine weitere Verbesserung hinsichtlich des konventionellen *Visual Analytics*-Prozesses. Im Optimalfall können ähnlich der wiederverwertbaren Regeln des *Wrangler*-Ansatzes die folgenden Schritte ebenfalls direkt berechnet werden und so den Zeitaufwand und mehrfache Wiederholungen der gleichen Tätigkeit vermeiden. Lediglich im Falle eines nicht automatisch lösbaren Konflikts oder einer Unzufriedenheit mit dem entstehenden Ergebnis ist der Anwender erneut gefordert.

Ablaufplan
Um ein strukturiertes Vorgehen zu ermöglichen ist eine Verknüpfung mit einem Modellierungswerkzeug empfehlenswert. In diesem können Datenquellen, Analyseschritte und verschiedene Visualisierungen durch eine grafische Oberfläche erstellt werden und somit Änderungen mit geringem Aufwand vorgenommen werden.

Diese Änderungen kombinieren die Freiheiten von *Visual Analytics* und die dadurch entstehenden Vorteile mit einem bewährten, etablierten und strukturiertem Prozess. Der Anwender wird somit schrittweise durch die Analyse geführt und dies bei Anwendung geeigneter visueller Unterstützung ohne tiefere IT-Kenntnisse. Durch die Visualisierungen und wiederholte Einbindung ist davon auszugehen, dass implizites Wissen eingebracht und so ein besseres Resultat erzielt werden kann.

4.5 Bewertung der verschiedenen Ansätze

Wie zuvor aufgezeigt umfasst *Visual Analytics* einen größeren Bereich als Data Mining und Visualisierung und verbindet diese beiden etablierten Verfahren. Im folgenden Abschnitt werden jeweils die Vor- und Nachteile kurz dargestellt.

4.5.1 Data Mining

Vorteile:

Vielfalt an Datenquellen
Data Mining kann mit einer Vielzahl von verschiedenen Datenquellen verwendet werden, da diese in der Vorverarbeitung auf eine einzelne Tabelle vereinigt werden und erst anschließend ein Data Mining-Algorithmus zum Einsatz kommt [HW01, Rei99].

Datenvolumen
Data Mining kann automatisiert ausgeführt werden und eignet sich dadurch für extrem große Datenmengen, die für den Menschen alleine nicht mehr zu überblicken sind. Die gefundenen Muster können anschließend leicht vom Anwender ausgewertet werden.

Objektive und bewährte Verfahren
Data Mining-Verfahren sind klassifiziert hinsichtlich einer klaren Zielsetzung (vgl. Abschnitt 2.2). Weiterhin basieren diese auf mathematischen Grundlagen.

Schnell und zuverlässig bei bekannten Problemstellungen
Data Mining funktioniert zuverlässig und vollautomatisch bei bekannten, gut verstandenen Problemen [KAF+08]. Ein Beispiel für eine solche Problemstellung sind Entscheidungsbäume, bei welchen nach initialer Erstellung des Baumes weitere Entscheidungen zuverlässig und automatisiert getroffen werden können.

Nachteile:

Prozess nicht transparent
Data Mining-Verfahren kommunizieren die internen Vorgänge nicht an den Anwender und werden für diesen somit häufig zu einer Art *Black Box* [PBT+10].

Starke Abhängigkeit von der Datenqualität
Data Mining sucht nach Mustern in einem Datensatz, entsprechend ist die Qualität der zu analysierenden Daten entscheidend für den Erfolg. Diese Problematik ist bei jeder Datenverarbeitung gegeben, das Data Mining hierbei jedoch besonders stark betroffen. Während bei der Visualisierung der Anwender über die Berücksichtigung einzelner Werte entscheiden kann, ist Data Mining auf die automatische Evaluation angewiesen. Einzelne falsche Werte können das erkannte Muster stark beeinflussen und somit zu falschen Schlussfolgerungen führen.

4.5.2 Visualisierung

Vorteile:

Menschliche Wahrnehmung wird berücksichtigt
Die Visualisierung versucht die menschliche Wahrnehmung bestmöglich auszunutzen um die Analyse der Daten zu unterstützen. Hierzu zählen beispielsweise die Gestaltgesetze oder die präattentive Verarbeitung (vgl. Abschnitt 2.5) [War12, HBE96].

Überblick über die Daten
Die Visualisierung erlaubt eine große Menge Daten in sehr kurzer Zeit zu interpretieren wenn die Darstellung sinnvoll gewählt wurde [War12].

Viele Visualisierungen
In der Literatur existiert eine Vielzahl unterschiedlicher Visualisierungen für unterschiedliche Zwecke und unterschiedliche Quelldaten [BBHK10, KK11]. Jede Visualisierung bietet Vor- und Nachteile, beispielsweise ist ein Kreisdiagramm gut geeignet zur Abschätzung des Verhältnisses eines Elementes gegenüber dem gesamten Datensatz, jedoch schlecht geeignet für den Vergleich zwischen zwei Elementen [KK11]. Die vorhandene Vielfalt bietet infolgedessen für viele Problemstellungen eine geeignete Visualisierung.

Nachteile:

Lie-Factor
Visualisierungen können den Anwender zu falschen Rückschlüssen verleiten. Diese Problematik ist unter dem Begriff *Lie-Factor* bekannt und beschreibt die Differenz zwischen visueller Darstellung und den zugrunde liegenden Daten [Tuf01].

Change Blindness
Das Phänomen der *Change Blindness* beschreibt die Schwierigkeiten des Menschen auftretende Veränderungen wahrzunehmen [VLF04].

Große Datenvolumen
Die Visualisierung ist für die im Bereich *Big Data* anfallenden Datenmengen nicht besonders geeignet [KKM+10b, ABM07].

Intransparenter Prozess
In Abschnitt 2.5 wurde die Visualisierungspipeline vorgestellt. Diese ist üblicherweise nicht direkt durch den Anwender kontrolliert und somit werden die Abläufe innerhalb des Analyseprozesses nicht kommuniziert. Entsprechend ist nicht offensichtlich auf welche Weise die Visualisierung zustande kommt.

Weiteres
Eine umfassende, systematische Literaturrecherche in diesem Kontext wurde von Bresciani und Eppler [BE15] durchgeführt und eventuell auftretende Probleme, sowohl bei der Gestaltung als auch bei der Evaluation einer Visualisierung, klassifiziert.

4.5.3 Visual Analytics

Vorteile:

Kombination der Stärken von Mensch und Maschine
Visual Analytics nutzt visuelle und automatische Verfahren zur Kombination der jeweiligen Stärken. Auf diese Weise kann die Interpretation durch den Anwender übernommen werden, während die großen Datenmengen durch automatische Verfahren verarbeitet werden.

Frühe Erkennung von Fehlern
Durch die starke Einbindung des Anwenders können fehlerhafte Resultate bereits im frühen Status der Analyse erkannt und diese entweder verhindert oder passend reagiert werden.

Transparenter Prozess
Der Anwender ist in jeden Schritt der Analyse eingebunden und gewinnt hierdurch ein tieferes Verständnis über die Daten und die Abläufe.

Stark interdisziplinär
Visual Analytics zeichnet sich durch seinen interdisziplinären Charakter aus und ermöglicht somit über die Grenzen des eigenen Forschungsgebietes heraus das für den jeweiligen Zweck sinnvollste Verfahren auszuwählen und somit den Analyseprozess bestmöglich zu unterstützen.

Implizites Wissen
Mit Hilfe von *Visual Analytics* kann das implizite Wissen des Anwenders genutzt werden und so auf eine nicht erfasste, zusätzliche Datenquelle zurückgegriffen werden.

Nachteile:

Erfolg abhängig vom Anwender
Visual Analytics bindet den Anwender in den vollständigen Analyseprozess mit ein und wird von diesem gesteuert. Dies ist insofern ein Nachteil, da der Anwender mit der Analysefreiheit überfordert werden könnte und in diesem Fall kein belastbares Ergebnis der Analyse zustande kommt. Diese Abhängigkeit ist gleichzeitig jedoch auch die große Stärke von *Visual Analytics*.

Neutralität abhängig vom Anwender
Ein weiterer Nachteil ist, dass durch die große Kontrolle des Analyseprozesses durch den Anwender dieser auch sinnvolle, erkannte Muster als nicht hilfreich verwerfen kann, da diese nicht die Erwartungshaltung hinsichtlich der gewünschten Ergebnisse widerspiegeln. Dieser Bias ist in menschlicher Wahrnehmung immer enthalten [PC05].

Hohe Kosten für Neuberechnung
Visual Analytics bietet während des gesamten Analyseprozesses stetig Visualisierungen des derzeitigen Zustandes an und bedingt dadurch dauerhafte Neuberechnungen. Bereits bei einer interaktiven Visualisierung ist sind mehrere Sekunden nötig, um die Informationen zu berechnen, zu erfassen und kognitiv zu verarbeiten [Wij05].

	Data Mining	Visualisierung	Visual Analytics
Interaktionsvolumen	keine	niedrig	hoch
Automatisierungsgrad	hoch	mittel	variabel
Geschwindigkeit	hoch	mittel	variabel
Laufzeitkosten	gering	mittel	hoch
Verwendung (unstrukturierte Daten)	nein	ja	ja
Verwendung (strukturierte Daten)	ja	ja	ja
Umfang im KDD-Prozess	gering	gering	vollständig
Datenvolumen	hoch	mittel	hoch
Objektivität	hoch	mittel	gering
Reproduzierbarkeit	hoch	mittel	gering
Flexibilität der Analyse	gering	mittel	hoch
Komplexität der Analyse	gering	mittel	hoch
Verbesserung der Datenqualität	gering	gering	hoch
Einfluss schlechter Datenqualität	hoch	mittel	gering

Tabelle 4.1: Vergleich von Data Mining, Visualisierung und Visual Analytics

4.6 Ordnungsrahmen

Auf den vorigen Seiten wurden die Vor- und Nachteile der drei Themengebiete Data Mining, Visualisierung und *Visual Analytics* erläutert. Zur Verdeutlichung werden diese hinsichtlich verschiedener Dimensionen in einem Ordnungsrahmen gegenübergestellt. Der entstandene Ordnungsrahmen ist in Tabelle 4.1 zusammengefasst und nachfolgend ausgeführt.

Interaktionsvolumen
Bezüglich der Interaktivität, also der Einbindung des Anwenders in die Analyse, unterscheiden sich die Verfahren deutlich. Data Mining arbeitet in den meisten Fällen autonom und liefert direkt zu interpretierende Ergebnisse. Die Visualisierung bezieht den Anwender in die Analyse mit ein, jedoch im Allgemeinen zur Aufstellung und Überprüfung von Hypothesen. Der Ansatz von *Visual Analytics* ist unter diesem Gesichtspunkt der umfassendste. Der Anwender ist in jeden Schritt der Analyse eingebunden und kann diese entscheidend beeinflussen.

Automatisierungsgrad
Der Automatisierungsgrad ist bei Data Mining am stärksten ausgeprägt. Hier wird von einem Experten das grundsätzliche Verfahren festgelegt und vom Anwender anschließend lediglich ausgeführt. Die Visualisierung kann ebenso in weiten Teilen automatisiert werden, auch wenn der Nutzer letztlich für die Analyse benötigt wird. Im Bereich *Visual Analytics* ist eine Einordnung bezüglich des Automatisierungsgrades nicht eindeutig. Der Anwender entscheidet, ob ein Schritt im Analyseprozess automatisch oder visuell ausgeführt wird. Entsprechend ist der Automatisierungsgrad variabel.

Geschwindigkeit
Die Geschwindigkeit bzw. Dauer einer Analyse korreliert mit dem Automatisierungsgrad und den damit einhergehenden Laufzeitkosten. Data Mining wird automatisch ausgeführt und die Geschwindigkeit ist somit lediglich abhängig von dem ausführenden System. Mit Visualisierung können ebenfalls schnelle Analysen ermöglicht werden, wenn das Mapping gut gewählt wurde. Durch die Interaktion mit dem Anwender und dem größeren Umfang jedoch tendenziell langsamer. *Visual Analytics* umfasst den größten Umfang und die größte Interaktion, entsprechend zeitintensiv ist die Analyse.

Laufzeitkosten für Prozess
Algorithmen und Verfahren werden häufig anhand der Laufzeit oder des Speicherverbrauchs angegeben. Dies ist explizit nur für einzelne Algorithmen, nicht jedoch für den Vergleich von ganzen Verfahren möglich. Dennoch können prinzipielle Aussagen getroffen werden. Data Mining nimmt eine Eingabe und berechnet daraus mit einem beliebigen Verfahren in einem Durchgang ein bestimmtes Ergebnis. Für die statische Visualisierung gilt dasselbe, jedoch ist die Interaktion hinsichtlich der wiederholten Neuberechnung teuer [Wij05]. In diesem Zusammenhang wird nach jeder Interaktion eine neue Visualisierung unter Berücksichtigung der Änderungen benötigt, entsprechend steigt der Aufwand gegenüber dem Data Mining. Insbesondere für *Visual Analytics* bedeutet dies, dass durch das Konzept des *Human In The Loop* (vgl. Abschnitt 2.7) in jeder Iteration die Visualisierung erneut berechnet werden muss und der Analyseprozess in dieser Hinsicht entsprechend aufwändig ist.

Verwendung mit unstrukturierten Daten
Unstrukturierte Daten können mit Data Mining nicht verarbeitet werden, hierfür existiert stattdessen das eng verwandte Text Mining (vgl. Abschnitt 2.3). Der Visualisierungsprozess sieht hingegen im ersten Schritt eine Umwandlung von unstrukturierten Daten in strukturierte Datentabellen vor, weshalb die Visualisierung mit unstrukturierten Daten verwendet werden kann. Da *Visual Analytics* durch den interdisziplinären Ansatz auf beliebige andere – visuelle oder automatische – Verfahren zurückgreifen kann können unstrukturierte Datenquellen ebenfalls für die Analyse herangezogen werden.

Verwendung mit strukturierten Daten
Strukturierte Daten können mit allen drei Verfahren analysiert werden.

Umfang innerhalb des Knowledge Discovery-Prozesses
Der *Knowledge Discovery*-Prozess ist ein bewährtes Analyseverfahren. Während Data Mining in diesem Prozess lediglich einer von mehreren durchzuführenden Schritten ist und die Visualisierung für die Präsentation verwendet wird kann *Visual Analytics* grundsätzlich in jeder Phase eingesetzt werden (vgl. Abschnitt 4.4).

Datenvolumen
Eine weitere Differenzierung kann auf Basis des analysierbaren Datenvolumens vorgenommen werden. Data Mining wurde explizit für große Datenmengen entwickelt und der Bezug ist bereits in der Definition des Data Mining enthalten (vgl. Abschnitt 2.2). Bei der Visualisierung ist dagegen die Menge der Daten, die analysiert werden können, begrenzt, da bei entsprechend großer Datenmenge kein singuläres Bild auf einem üblichen Bildschirm sinnvoll dargestellt werden kann. *Visual Analytics* verbindet beides und kann somit ebenfalls mit großen Datenvolumen verwendet werden.

Objektivität
Die Objektivität ist für alle Verfahren nicht abschließend zu beurteilen. Im Bereich des Data Mining wird das Verfahren üblicherweise von einem Fachexperten definiert und die entsprechenden Parameter festgesetzt. Somit bleibt das Verfahren zwar einerseits unabhängig vom Anwender objektiv, jedoch hat der Fachexperte subjektiven Einfluss auf das Verfahren. Bei der Visualisierung trifft der Anwender Annahmen über die Daten und wertet diese aus, entsprechend können Muster in den Daten auch ignoriert bzw. bewusst übersehen werden. Noch geringer ist die Objektivität bei *Visual Analytics*, da der Nutzer hierbei bereits von Anfang an Einfluss auf die Analyse nehmen kann, beispielsweise bei der Auswahl der Daten. Im Vergleich zur Visualisierung und *Visual Analytics* ist das Data Mining demnach das objektivste Verfahren.

Reproduzierbarkeit
Die Reproduzierbarkeit, also die gleichen Ergebnisse bei den gleichen Quelldaten, variiert ebenfalls je nach angewendetem Verfahren. Data Mining liefert aufgrund der Anwendung bestehender Modelle die gleichen Resultate bei jeder Ausführung. Visualisierung liefert grundsätzlich bei gleichen Daten und Verfahren die gleiche visuelle Repräsentation, die Analyse muss jedoch nicht zwingend die gleichen Ergebnisse liefern [Wij05]. Ein anderer Anwender oder eine Änderung des Vorwissens oder der Absicht könnte zu anderen Ergebnissen führen. Kaum reproduzierbar ist eine Analyse mit Hilfe von *Visual Analytics*, da jede Aktion des Anwenders sich auf das Ergebnis auswirken kann.

Flexibilität der Analyse
Die Flexibilität des Anwenders eine Analyse zu steuern ist bei Data Mining durch das automatische Verfahren nicht gegeben. Die Visualisierung ermöglicht eine Interpretation hinsichtlich verschiedener Gesichtspunkte im Rahmen der abgedeckten Funktionen. *Visual Analytics* ermöglicht eine freie Wahl des Analyseziels und den damit verbundenen Schritten. Der Anwender erhält auf diese Weise die größtmögliche Flexibilität.

Komplexität der Analyse
Die Komplexität der Analyse die aus Anwendersicht durch die Verfahren ermöglicht wird ist durch die Freiheiten bei *Visual Analytics* am größten. Bei Data Mining findet die Analyse dagegen automatisch statt und ist somit für den Nutzer sehr einfach durchzuführen. Die Visualisierung platziert sich in dieser Kategorie in der Mitte.

Verbesserung der Datenqualität
Kemper et al. [KBM10] unterteilen Datenmängel in drei Klassen – automatisierte Erkennung und Korrektur, automatisierte Erkennung und manuelle Korrektur, manuelle Erkennung. Eine automatisierte Qualitätssteigerung ist hierbei lediglich in der ersten Klasse möglich und somit im Bereich von Data Mining und Visualisierung umsetzbar. *Visual Analytics* erlaubt dagegen durch Einbindung des Anwenders auch die restlichen Fehlerklassen zu erkennen und zu beseitigen.

Einfluss schlechter Datenqualität
Data Mining ist von einer schlechten Datenqualität besonders betroffen, da Daten nicht in allen Fällen automatisiert bereinigt werden können [KBM10]. Fehlerhafte Daten werden hierdurch in das jeweilige Modell einbezogen. Data Mining ist für den Anwender ein intransparenter Prozess und dieser ist folglich über diese Problematik nicht informiert. Die Visualisierung erlaubt dem Anwender je nach Art der Darstellung beispielsweise abweichende Werte zu erkennen und diese nicht in die Schlussfolgerung einzubeziehen. *Visual Analytics* kann durch die Einbindung des Anwenders mit allen drei Mängelklassen umgehen und die Datenqualität infolgedessen bereits im Vorfeld der Analyse steigern um diesen Einfluss möglichst stark zu begrenzen. Letztlich sind jedoch alle Verfahren von der Datenqualität abhängig, lediglich die Auswirkungen können unterschiedlich stark beeinflusst werden.

Aus dem obig erstellten Ordnungsrahmen ergibt sich, dass es *Visual Analytics*, nach der in dieser Arbeit verwendeten Definition, gelingt die Vorteile von Data Mining (hohes Datenvolumen) und Visualisierung (Interaktivität, menschliche Wahrnehmung) zu verbinden und zu verbessern (Flexibilität, Komplexität). Die Steigerung der Datenqualität ist in diesem Zusammenhang besonders hervorzuheben, da diese für die Analysequalität entscheidend ist.

Nachteile entstehen durch die Anwenderintegration und damit einhergehende Reduzierung der Objektivität, Reproduzierbarkeit und Geschwindigkeit.

Als Fazit dieser Gegenüberstellung bleibt, dass *Visual Analytics* nur dann die Stärken gegenüber dem Data Mining und der klassischen Informationsvisualisierung bietet, wenn der Anwender unabhängiger von einem Domänen-Experten wird. In diesem Fall müssen die Ziele erst während der Analyse festgelegt werden und nicht bereits bei der Entwicklung der Software spezifiziert sein.

5 Visual Analytics im Kontext der Daten- und Analysequalität

In diesem Kapitel wird untersucht inwiefern *Visual Analytics* den Anwender hinsichtlich der Daten- und Analysequalität unterstützen kann.

5.1 Datenqualität

Ein Vorteil der Visualisierung ist, dass die menschliche Wahrnehmung für die Interpretation genutzt wird (vgl. Abschnitt 4.5.2). *Visual Analytics* setzt für die Kommunikation mit dem Anwender ebenfalls auf dieses Konzept, was in der Literatur größtenteils als sinnvoll und unterstützend angesehen wird.

Ware [War12] nennt verschiedene Vorteile von einer gut gewählten Visualisierung, beispielsweise die Möglichkeit große Datenmengen schnell zu überblicken und dabei unerwartete Muster zu erkennen. Auf diesem Wege fallen Unregelmäßigkeiten in einem Datensatz sofort ins Auge und bieten demnach einen unschätzbaren Wert hinsichtlich der Qualitätskontrolle [War12, KHP+11].

Obige Aussagen gelten jedoch nur, wenn eine der Problemstellung angemessene Visualisierung verwendet wird. Einen Überblick der Schwierigkeiten hinsichtlich der Erstellung und Evaluation von Visualisierungen geben Bresciani und Eppler [BE15].

Die durch Visualisierung identifizierbaren Unregelmäßigkeiten in den Daten haben ihren Ursprung in unzureichender Datenqualität. Diese liegen in den meisten Fällen nicht in perfekter Qualität vor und beinhalten zudem eine gewisse Unsicherheit über die Zuverlässigkeit der Quelle, mangelnde Präzision oder auch fehlender Werte [GS05]. Zudem kann dieser Effekt an jeder Stelle innerhalb der Visualisierungspipeline auftreten [PWL97]. Diese Ungewissheit über die tatsächlich vorhandene Qualität wird unter dem Term *Uncertainty* zusammengefasst [GS06].

Nach Zuk [Zuk08] kann *Uncertainty* in den Daten auch bewusst integriert werden, beispielsweise wenn lediglich eine Kategorisierung auf Ja oder Nein benötigt wird und somit Informationen verloren gehen. Ware [War12] hält die Visualisierung von *Uncertainty* für wichtig, wenngleich schwer erreichbar. Dies liegt an der Neigung des Anwenders Visualisierungen üblicherweise als korrekt einzuschätzen. Watkins [Wat00] und Zuk [Zuk08] behandeln dieses Thema ausführlich.

Die folgenden Abschnitte beschränken auf die Möglichkeiten *Uncertainty* innerhalb einer Visualisierung darzustellen und so den Anwender über diese in Kenntnis zu setzen. Die Grundidee ist es, sowohl die Daten, als auch die *Uncertainty* selbiger, in einer Visualisierung zeitgleich abzubilden [PWL97].

Griethe und Schumann [GS06] identifizieren basierend auf existierenden Ansätzen zwei grundsätzliche Wege:

Verwendung nicht verwendeter Attribute
Die erste Darstellungsform verändert die Eigenschaften der grafischen Primitive, indem bisher nicht verwendete Attribute anhand der berechneten *Uncertainty* gesetzt werden, beispielsweise Farbe, Form, Transparenz, Schärfe oder Textur. In Abbildung 5.1d ist ein Beispiel für diesen Ansatz dargestellt. Die einzelnen Konturen symbolisieren verschiedene räumliche Messwerte. Die *Uncertainty* ist durch eine Unterbrechung der Kontur visualisiert, je größer die *Uncertainty*, desto größer die vorhanden Lücken.

Verwendung neuer Elemente
Die zweite Darstellungsform ergänzt die Visualisierung um neue Elemente. Dies können Fehlerbalken, Glyphen oder auch ein überlagerndes Gitternetz sein. Abbildung 5.1c zeigt eine kartografierte Oberfläche, die *Uncertainty* wird unterdessen durch Quader symbolisiert.

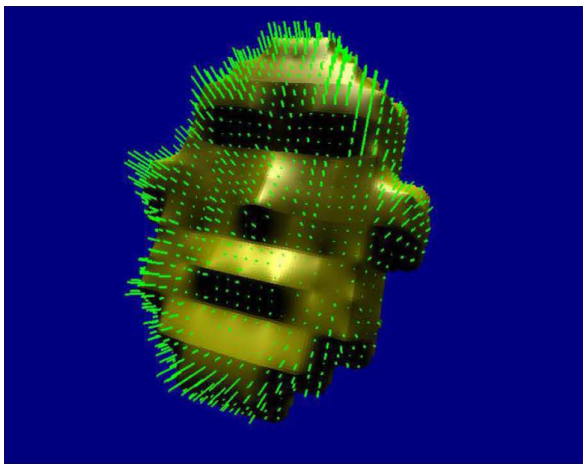
Einen direkten Vergleich zwischen diesen beiden Ansätzen liefern die Abbildungen 5.1a und 5.1b, anhand der durch unterschiedliche Interpolationsverfahren entstehenden *Uncertainty*. Abbildung 5.1a verwendet zusätzliche Elemente um diese Differenz darzustellen, während Abbildung 5.1b die Position der Fläche anpasst. Weitere Wege zur Darstellung von *Uncertainty* basieren auf Animationen, Interaktivität durch Mausbewegung oder sprechen weitere menschliche Sinne an [GS06].

In Kapitel 3 wurden verschiedene Applikationen vorgestellt, die teilweise die Datenqualität visualisieren. Das Tool *Wrangler* (vgl. Abschnitt 3.2.2) fügt beispielsweise für jedes Attribut des Datensatzes einen Fehlerbalken für die Qualität hinzu, sowohl bezüglich fehlender Werte, als auch für inkonsistente Datentypen.

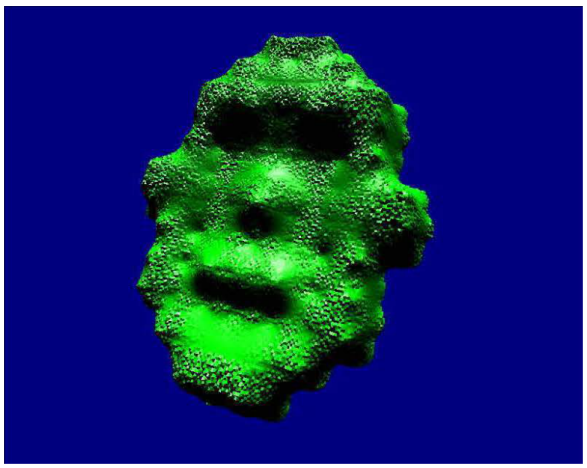
An dieser Stelle bietet die Kombination von automatischen und visuellen Verfahren unter Einbindung des Anwenders einen entscheidenden Vorteil für die Datenqualität. Nach Kemper [KBM10] können drei Mängelklassen unterschieden werden, von denen lediglich eine automatisch zu korrigieren ist.

Visual Analytics bietet hier weitergehende Möglichkeiten als durch automatische Verfahren alleine möglich sind. Zunächst können Fehler wie gehabt durch bereits bekannte Regeln korrigiert werden, andererseits Fehler erkannt und der Anwender auf diese hingewiesen werden, z. B. durch Plausibilitätskontrollen oder Abweichungsanalysen [KBM10].

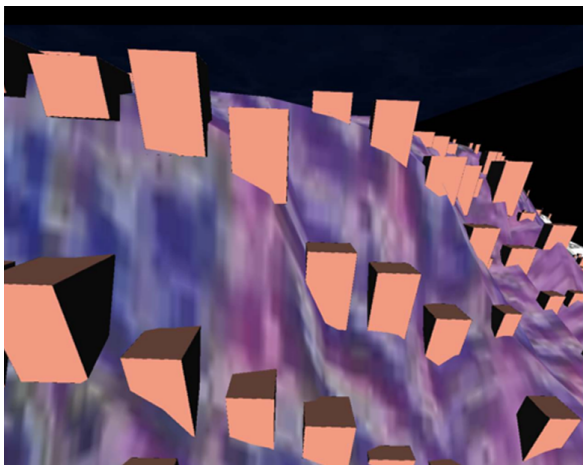
Die dritte Fehlerklasse (inkorrekte Datenwerte) nach Kemper [KBM10] kann nicht automatisch erkannt werden, jedoch möglicherweise durch den Anwender identifiziert und korrigiert werden.



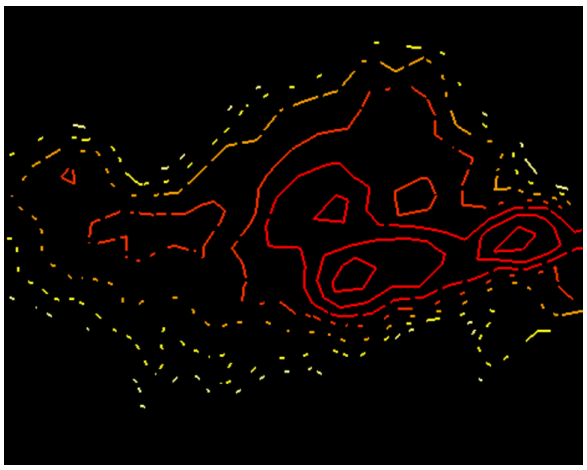
(a) Unterschiede zwischen verschiedenen Interpolationsalgorithmen durch neue Elemente [PWL97]



(b) Unterschiede zwischen verschiedenen Interpolationsalgorithmen durch Veränderung der Position [PWL97]



(c) Glyph-basierte Visualisierung [SSB+04]



(d) Visualisierung über Veränderung der Form [Pan01]

Abbildung 5.1: Überblick über verschiedene Visualisierungen von *Uncertainty*

In diesem Kontext ist die Visualisierung der Datenquelle essentiell und ermöglicht dem Anwender einerseits verschiedene Interpretationsmöglichkeiten [ZC07], sowie andererseits diese positiv zu beeinflussen. *Visual Analytics* bietet demnach mehr Möglichkeiten als die reine Visualisierung oder Data Mining alleine, erst durch die Kombination entstehen Vorteile, die für eine Steigerung der Datenqualität sorgen können.

5.2 Analysequalität

Die Analysequalität hängt unmittelbar mit der Datenqualität zusammen (vgl. Abschnitt 2.11) und entsprechend relevant ist es, dass der Anwender sich über eventuell in den Daten vorhandene Qualitätsdefizite bewusst ist. Wird die *Uncertainty* nicht berücksichtigt führt dies im Verlauf des Analyseprozesses zu zunehmend größeren Defiziten [TC05, ZC07], jedoch ohne dass sich der Anwender hierüber im Klaren ist [ZC07]. Im *Knowledge Discovery*-Prozess wird der *Uncertainty* häufig keine gewichtige Rolle zugewiesen, sondern angenommen, dass alle relevanten Daten vorhanden sind [BL09].

Nach Thomas und Cook [TC05] müssen Qualität und *Uncertainty* während der Analyse jederzeit verfügbar und durch Visualisierung dem Anwender präsentiert werden. Nur auf diese Weise sei gewährleistet, dass der Anwender sich über die möglichen Fehler und Ungenauigkeiten bewusst ist, was zwingend erforderlich ist um die korrekten Rückschlüsse zu ziehen [TC05].

Durch die Einbindung des Anwenders kann *Visual Analytics* die Analysequalität steigern. In Abschnitt 3.2.3 wurde ein visuelles Clustering-Verfahren vorgestellt und gezeigt, dass durch die Interaktion bessere Ergebnisse erreichbar sind. Weiterhin bietet *Visual Analytics* eine umfangreiche Einbeziehung des Anwenders kann dieser sein implizit vorhandenes Hintergrundwissen in den Analyseprozess einbringen (vgl. Abschnitt 4.5.3) und so die Analyse mit diesem zusätzlichen Wissen in die gewünschte Richtung steuern. Auch die Autoren des vorgestellten visuellen Klassifikationsverfahren erwarten durch die Interaktivität Vorteile bessere Resultate.

Weiterhin zeigt Szenario II – Jigsaw (vgl. Abschnitt 3.1.2), dass es möglich ist dem Anwender die weitgehende Kontrolle zu überlassen und so eine freie Analyse zu ermöglichen.

Die Möglichkeit für den Anwender durch Interaktion mit den Daten zu "spielen" ermöglicht nach Bertini und Lalanne [BL10] ein besseres Verständnis über die Daten. Hierdurch könnten möglicherweise Qualitätsmängel in diesen erkannt und bei Schlussfolgerungen aus der Analyse einbezogen werden. Auch Savikhin et al. [SME08] bestätigen bei einer Nutzerstudie, dass interaktive Elemente zu verbesserten Resultaten führen.

Es darf im Hinblick auf die Analysequalität jedoch nicht vernachlässigt werden, dass durch die starke Einbindung des Anwenders auch Nachteile entstehen. Das obig erwähnte und in Abschnitt 3.2.3 vorgestellte Clustering-Verfahren kann durch den Anwender nicht nur bessere, sondern auch schlechtere Resultate ergeben. Shneiderman [Shn02] sieht dies Problematik, wenn eine Hypothese belegt werden soll. In diesem Fall könnten Muster ignoriert werden, wenn diese nicht zu dieser Zielsetzung passen.
Ein weiterer Nachteil der wiederholten Einbindung des Anwenders ist die ständig zu aktualisierende Visualisierung, welche hohen Zeitaufwand durch Neuberechnung und Interpretation bedingt [Wij05].

Um die Analysequalität zu messen, existieren für automatische Verfahren die in Abschnitt 2.11 vorgestellten Metriken auf festgelegten und klassifizierten Testdaten. Für die Evaluation einer Visualisierung werden üblicherweise Nutzerstudien durchgeführt [Pla04, And06]. Hierbei kommen verschiedene Techniken zum Einsatz. Der *Thinking-Aloud*-Test [BR00] fordert die Teilnehmer auf ihre Gedanken zu formulieren und versucht auf diese Weise den Gedankengang nachzuvollziehen [And06]. Eine weitere Möglichkeit ist der Vergleich verschiedener Visualisierungen anhand Metriken wie Anzahl der Interaktionen oder der benötigten Zeit [And06]. *Visual Analytics* nach der in dieser Arbeit entwickelten Definition kombiniert diese beiden Gebiete und benötigt neue Methoden. Blascheck et al. [BJK+16] erörtern für *Visual Analytics* einen Evaluationsansatz, der mehrere Dimensionen (Eye-Tracking, Interaktionen und *Thinking-Aloud*-Test) zeitgleich berücksichtigt.

Nach Keim et al. [KMT10] erlaubt *Visual Analytics* eine höhere Skalierbarkeit auf größere und komplizierte Problemstellungen.

5.3 Anwendungsszenarien

Im Folgenden werden mögliche Szenarien, in welchen *Visual Analytics* einen Mehrwert gegenüber herkömmlichen Verfahren bietet, beschrieben.

Als Ausgangslage wird angenommen, dass eine vertrauenswürdige, qualitativ hochwertige Datenquelle nicht umfangreich genug ist und zudem eine zweite, umfassendere Datenquelle unbekannter Qualität vorhanden ist. Der Anwender wird durch *Visual Analytics* in die Lage versetzt einerseits beide Datenquellen zu berücksichtigen, durch eine Visualisierung zu vergleichen und – anhand der Übereinstimmung der doppelt vorhandenen Daten – die Vertrauenswürdigkeit und Qualität der zweiten Datenquelle abzuschätzen. Auf diese Weise stehen möglicherweise anschließend deutlich mehr Daten zur Verfügung und erlauben eine detailliertere Analyse.

Eine weitere Möglichkeit *Visual Analytics* im Zusammenhang mit Datenqualität zu verwenden ist, dass über eine visuelle Schnittstelle die berechnete Mindestqualität eingestellt werden kann, woraufhin die anschließend qualifizierten Daten neu visualisiert werden. Dies ermöglicht dem Anwender die Qualität selbst zu bestimmen und das Resultat der folgenden Analyse diesbezüglich zu bewerten.

5.4 Bewertung

In diesem Kapitel wurden verschiedene Möglichkeiten aufgezeigt wie *Visual Analytics* die Qualität der Daten und der Analyse verbessern kann. Insbesondere für die Erhöhung der Datenqualität ist der Ansatz visuelle und automatische Verfahren zu kombinieren hilfreich und erlaubt auftretende Datenmängel zu korrigieren.

Weiterhin kann dem Anwender im Analyseverfahren ein Gefühl für die Datencharakteristik vermittelt werden und über deren Zuverlässigkeit informieren. Es ist zu erwarten, dass hierdurch die Qualität der durchgeführten Analysen steigt und spezifischere Resultate erzielt werden können. Bereits im Jahr 2005 zogen Thomas und Cook [TC05] die Schlussfolgerung, dass *Visual Analytics*-Ansätze in den meisten Fällen *Uncertainty* nicht berücksichtigen. Diese Schlussfolgerung konnte im Verlauf der Literaturrecherche für diese Arbeit nicht begründet widerlegt werden.

Für die Evaluation der Analysequalität wurden verschiedene Verfahren für automatische und visuelle Analysen aufgezeigt. Die in dieser Arbeit entwickelte Auffassung von *Visual Analytics* ist umfassender als in der Literatur üblich. Nach dieser Definition ist kein in sich abgeschlossenes Programm erforderlich, sondern vielmehr eine Kopplung verschiedener in sich abgeschlossener Einheiten. Durch die wiederholten und unterschiedlichen Schritte – mit jeweils freier Kombination von automatischem und visuellem Verfahren – ist es nicht zielführend lediglich das schlussendliche Resultat zu evaluieren. Dieses kann einerseits je nach Analysepfad stark variieren, anderseits auf unterschiedlich aufwendigen Wegen erreicht werden.

Empfehlenswert ist demnach entweder eine kontinuierliche oder zumindest regelmäßige Messung der Qualität. Während der Literaturrecherche konnte diesbezüglich kein Verfahren in Zusammenhang mit *Visual Analytics* identifiziert werden.

Visual Analytics bietet nach den vorgestellten theoretischen Grundlagen das Potential die Daten- und Analysequalität gegenüber herkömmlichen Verfahren zu erhöhen.

6 Visual Analytics und Data Mashups

In diesem Kapitel wird die Anwendung von *Visual Analytics* im Kontext von Data Mashups evaluiert. Hierzu wird eine prototypische Implementierung für die Datenaufbereitung und die Assoziationsanalyse entwickelt, welche in Verbindung mit Data Mashups dem Anwender eine möglichst große Analysefreiheit bieten.

6.1 Szenario

In Abschnitt 3.1.3 wurde mit Szenario III – Gesundheitswesen ein heutiges Tool im Bereich *Visual Analytics* vorgestellt. Dieses löst die eigene Problemstellung elegant und umfassend, jedoch wird der Anwender nicht in die Auswahl der Datenquellen eingebunden und ist somit nicht die zentrale Komponente der Analyse. Aus diesem Umstand ergibt sich für die Analyse das Defizit, dass bereits eine kleine Veränderung des Anwendungsfalles eine Aktivität seitens des ursprünglichen Autors verlangt.

Nach der in dieser Arbeit entwickelten Definition von *Visual Analytics* ist es für die Ausschöpfung des vollen Potentials nötig, dass diese Adaptivität direkt durch den Anwender vorgenommen werden kann.

Data Mashups wurden zur freien Kombination mehrerer, heterogener Datenquellen entwickelt und ermöglichen hohe Flexibilität und individuelle Lösungen (vgl. Abschnitt 2.13).

Der in Szenario III – Gesundheitswesen beschriebene Anwendungsfall wird zur Demonstration der Möglichkeiten, welche durch die Kombination von *Visual Analytics* und Data Mashups entstehen:

Es existieren mehrere medizinische Einrichtungen mit jeweils eigenständiger Datenhaltung, jedoch gleicher Ausrichtung. Für eine umfangreiche Analyse der übergeordneten Zusammenhänge (z. B. zwischen Arzt und Medikament) auf einer möglichst breiten Datenbasis müssen diese heterogenen Datenquellen zunächst integriert werden. In einem anschließenden Schritt soll eine Assoziationsanalyse diese Zusammenhänge entdecken und visualisieren. Jeder Schritt soll dabei den Anwender und dessen implizites Wissen in den Prozess einbinden.

Die Datenquellen enthalten hierbei generierte Werte, die jedoch auf realen Medikamenten und Krankheiten basieren.

6.2 FlexMash

Der grundsätzliche Gedanke hinter Data Mashups wurde bereits in Abschnitt 2.13 erläutert. An der Universität Stuttgart wird das Tool *FlexMash* [HM16, HRWM15] entwickelt, welches als technische Basis für die prototypische Implementierung dieser Arbeit dient.

FlexMash besteht aus zwei Schritten. Zunächst wird der Ablaufplan mit Hilfe einer visuellen Schnittstelle definiert, wofür keine technischen Kenntnisse erforderlich sind. Anschließend wird dieser Ablaufplan in ein ausführbares Modell überführt und gestartet.

Für die Modellierung stellt *FlexMash* dem Anwender zwei verschiedene Komponenten zur Verfügung. Zunächst die *Data Source Descriptions* (DSD), welche eine Datenquelle abstrakt beschreiben. Diese Beschreibung umfasst beispielsweise die URL, den Port oder den API-Key und ermöglicht somit die Verwendung erlauben ohne genauere Kenntnis der technischen Implementierung. Weiterhin existieren verschiedene *Data Processing Descriptions* (DPD), welche die einzelnen Operationen (z. B. *Merge* oder *Filter*) beschreiben.

Diese Beschreibungen werden in einem Katalog vorgehalten und können somit problemlos wiederverwendet und nahezu beliebig kombiniert werden. Die genaue technische Implementierung ist für die Modellierung des Ablaufplans dabei noch irrelevant und wird erst bei der Ausführung des Mashups benötigt.

Die Aufteilung zwischen Modellierung und Spezifikation bietet im Hinblick auf *Visual Analytics* dem Anwender die Möglichkeit einen Analyseablauf zu definieren ohne dabei Kenntnisse über die explizite Umsetzung zu erfordern. Dieses Konzept steigert die Unabhängigkeit des Anwenders von einem Domänenexperten sowie ferner die Freiheit bei der Analyse.

Als Datenaustauschformat zwischen den einzelnen Knoten des Ablaufplans kommt die *JavaScript Object Notation* (JSON) zum Einsatz.

6.3 Integration von Visual Analytics in FlexMash

Vor der Modellierung eines Ablaufplans müssen die verwendeten *Data Source Descriptions* und *Data Processing Descriptions* definiert sein. Für das obig beschriebene Szenario werden zwei neue *Data Source Descriptions* definiert, welche die Daten der jeweiligen medizinischen Einrichtungen beschreiben. In diesem konkreten Fall wird jeweils ein mit Zufallswerten gefülltes JSON-Objekt als Datenquelle definiert, was hinsichtlich der Generizität nicht relevant ist.

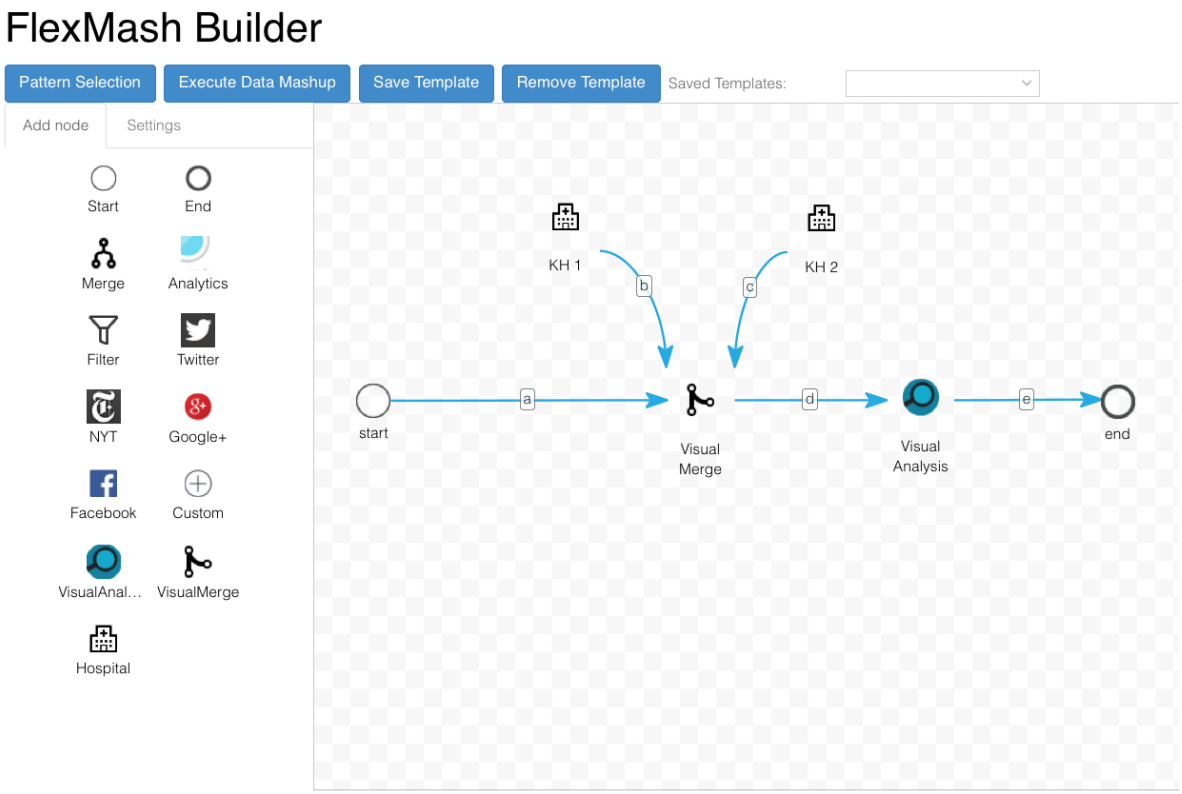


Abbildung 6.1: *FlexMash*-Ablaufplan für das gewählte Szenario

Weiter werden zwei *Data Processing Descriptions* benötigt, welche die unterschiedlichen Operationen definieren. Die in diesem Prototyp verwendeten *Data Processing Descriptions* wurden mit Hilfe von JavaScript, jQuery[1] und Bootstrap[2] implementiert. Als Eingabe und Ausgabe wird entsprechend zu den Anforderungen von *FlexMash* jeweils ein JSON-Objekt übergeben.

In Abbildung 6.1 ist die Modellierung des Ablaufplans für das beschriebene Szenario dargestellt. Zunächst wird der Einstiegspunkt in den Ablaufplan eingefügt. Dieser ist mit dem sogenannten *Visual Merge Node* verbunden (*a*). Dieser Knoten erhält in diesem Fall zwei weitere Eingaben (*b*, *c*) mit den jeweiligen Datenquellen für die medizinischen Einrichtungen. Nach dem Abschluss der Integration werden die Daten an den *Visual Analysis Node* weitergereicht (*d*). Das Ende des Ablaufplans wird durch den Endknoten (*e*) repräsentiert.

[1]http://www.jquery.com
[2]http://www.getbootstrap.com

6.4 Visual Merge Node

Wie ausführlich beschrieben liegen selten alle sinnvollen Daten auch an einem singulären Speicherort vor. Weiterhin benötigen Data Mining-Algorithmen üblicherweise als Eingabemenge eine einzige Tabelle. Die verwendeten Daten müssen somit unter Berücksichtigung des Analyseziels ausgewählt, aufbereitet und zusammengeführt werden.

In Abschnitt 5.1 wurde dargestellt, dass *Visual Analytics* große Vorteile bieten kann um eine gute Datenqualität zu gewährleisten. Die an dieser Stelle dargelegte Visualisierung durch einen Fehlerbalken auf Basis fehlender Werte oder Inkompatibilität wird auch für die prototypische Implementierung verwendet.

Am Ende der Datenaufbereitung soll entsprechend eine singuläre Datentabelle mit möglichst hoher Datenqualität zur Verfügung stehen, welche anschließend entsprechend des modellierten Ablaufs weiterverwendet werden kann.

Um diese Ziele zu verwirklichen sollen die folgenden grundlegenden Funktionen der Datenbereinigung, Datentransformation und Schema-Integration integriert werden [HK06]:

Fehlende Werte
Die folgenden drei Methoden zeigen Möglichkeiten des Umgangs mit fehlenden Werten innerhalb eines Tupels und betreffen die jeweiligen Attribute:

Ignorieren
Die einfachste Variante ist es die betreffenden Attribute zu ignorieren. Diese Vorgehensweise ist nicht besonders effektiv, sofern ein Tupel nicht mehrere leere Attribute beinhaltet. Dieser Punkt wird in der Implementierung durch das Löschen des entsprechenden Attributes umgesetzt.

Manuelle Korrektur
Fehlende Werte können von Hand eingetragen werden. Diese Vorgehensweise ist für große Datensätze nicht praktikabel, wird für einzelne Felder durch den *Visual Merge Node* unterstützt.

Globale Konstante
Mit Hilfe einer globalen Konstante können alle leeren Felder auf den gleichen Wert festgelegt werden. Für die Implementierung existiert eine Ersetzen-Funktion für gleiche Werte innerhalb eines Attributes.

Entität-Identifikation
Diese Funktionalität bezieht sich auf die Bestimmung der Semantik unterschiedlicher Entitäten. Für eine Integration müssen in allen Datenquellen die gleichen Attribute vorhanden sein. Es muss in diesem Zusammenhang sicher gestellt sein, dass der gleiche Bezeichner auch dieselbe Bedeutung hat.

Generalisierung
Die Generalisierung beschreibt die Veränderung der Granularität, beispielsweise indem einzelne Staaten dem entsprechenden Kontinent zugeordnet werden.

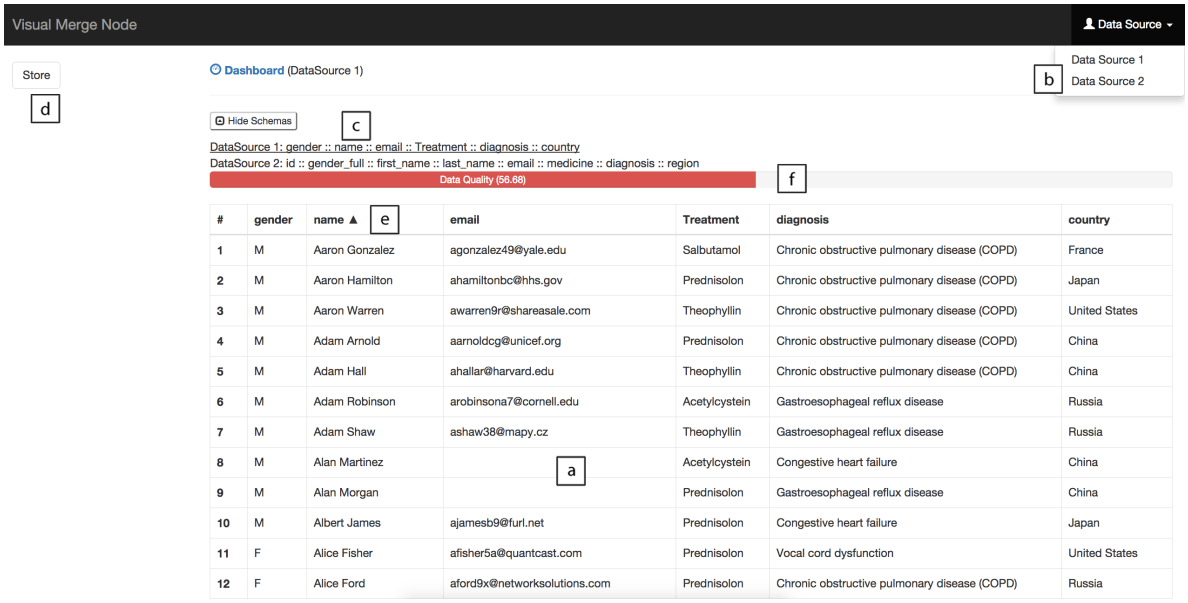


Abbildung 6.2: Visual Merge Node: Grafische Oberfläche

Attribut-Konstruktion
Neue Attribute werden aus den bereits vorhandenen konstruiert, beispielsweise durch Verknüpfung oder Trennung.

Die Oberfläche des *Visual Merge Node* ist in Abbildung 6.2 dargestellt und an das in Abschnitt 3.2.2 vorgestellte Tool *Wrangler* angelehnt. Im Gegensatz zu diesem werden von dieser Implementierung auch mehrere Datenquellen unterstützt.

Die grafische Oberfläche besteht aus mehreren Komponenten und wird nachfolgend erläutert:

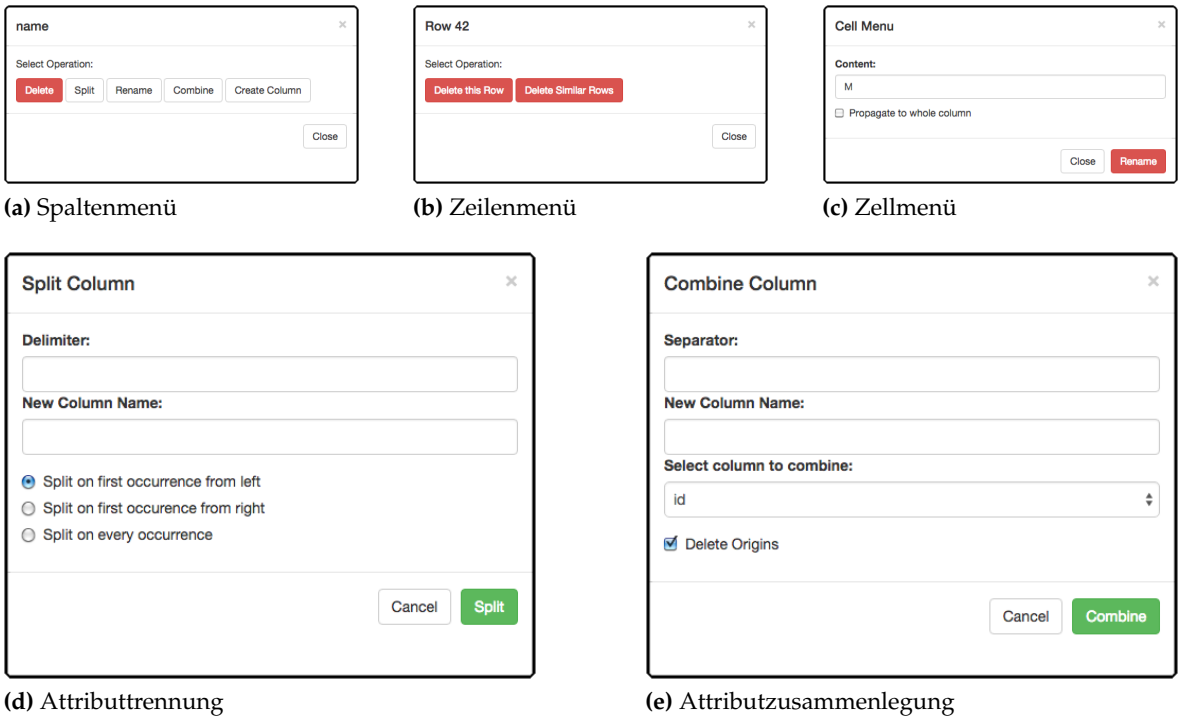
Visualisierung des Datensatzes
Den größten Teil der Oberfläche vereinnahmt die Visualisierung des gegenwärtigen Zustandes der Datenquelle (*a*). Dieser wird in Tabellenform dargestellt und automatisch aus dem eingehenden JSON-Objekt erstellt.

Auswahl der Datenquelle
Der Anwender kann zwischen beliebig vielen Datenquellen wechseln (*b*), woraufhin sich die Oberfläche aktualisiert und somit die ausgewählte Datenquelle bearbeitet werden kann.

Darstellung des Schemas
Die derzeitigen Schemata der einzelnen Datenquellen können oberhalb des ausgewählten Datensatzes bei Bedarf eingeblendet werden (*c*). Dies ermöglicht dem Anwender jederzeit die verbleibenden Unterschiede zwischen den Datenquellen im Blick zu behalten und zu beseitigen. Die aktive Datenquelle ist dabei unterstrichen.

Qualitätsanzeige
Weiterhin existiert eine zentrale Qualitätsanzeige (*f*), welche dem Anwender den aktuellen Fortschritt bei der Angleichung der Datenquellen signalisiert. Dieser Wert berechnet sich anhand der Übereinstimmung der Schemata zwischen den Datenquellen (40 %) , sowie der Vollständigkeit (vgl. Abschnitt 2.8) der Datensätze (60 %). Je nach Qualität wechselt die Farbe stufenweise von rot über orange nach grün.

Bestätigung
Ist die Qualität ausreichend und eine Übereinstimmung gegeben, so kann der Anwender die Aufbereitung/Verknüpfung der Daten beenden (*d*). In diesem Fall werden die Datenquellen in einen singulären Datensatz zusammengeführt und ein neues JSON-Objekt erzeugt. Mit diesem wird der Ablaufplan entsprechend der Modellierung in *FlexMash* fortgesetzt.



(a) Spaltenmenü **(b)** Zeilenmenü **(c)** Zellmenü

(d) Attributtrennung **(e)** Attributzusammenlegung

Abbildung 6.3: Visual Merge Node: Überblick über die Interaktionsmöglichkeiten

Für die Interaktion mit dem Datensatz stehen dem Anwender verschiedene Operationen zur Verfügung:

Sortieren
Bei einem Klick auf die Kopfzeile wird der Datensatz nach dem darunterliegenden Attribut sortiert. Um leere Felder leicht erkennen zu können werden selbige immer zuerst dargestellt. Das der aktuellen Sortierung zugrunde liegende Attribut wird durch ein kleines Symbol neben dem Namen angezeigt (vgl. Abbildung 6.2).

Zellmenü
Bei einem Rechtsklick auf eine beliebige Zelle der Tabelle wird das Zellmenü (vgl. Abbildung 6.3c) geöffnet. Hier eröffnet sich dem Anwender die Option einzelne Werte zu verändern und diese Änderung auf Wunsch auf die komplette Spalte anzuwenden. So kann beispielsweise eine Änderung der Kodierung (m -> 0, w -> 1) mit wenigen Aktionen auf den gesamten Datensatz angewendet werden.

Spaltenmenü
Bei einem Rechtsklick auf eine beliebige Spalte der Tabelle wird das Spaltenmenü (vgl. Abbildung 6.3a) geöffnet. Dieses bietet dem Anwender eine Vielzahl von Möglichkeiten die nun im Folgenden aufgeführt sind:

Löschen
Der Anwender kann eine ganze Spalte ersatzlos löschen. Dies ist immer dann sinnvoll, wenn ein Datensatz ein Attribut enthält, welches nicht in den anderen Datensätzen vorhanden ist und auch nicht berechnet werden kann.

Trennen
Wenn ein Attribut zusammengesetzt ist (z. B. Datum als TT-MM-JJJJ) kann dieses aufgetrennt werden (in Tag, Monat, Jahr). Hierzu muss der Anwender ein Trennzeichen angeben sowie den Namen des neuen Attributs. Weiterhin lässt sich spezifizieren an welchen Stellen das Trennzeichen berücksichtigt wird, entweder das erste Vorkommen von links/rechts oder bei jedem Auftreten (vgl. Abbildung 6.3d). Für den Fall, dass mehr als eine Trennung stattfindet wird der angegebene Name des neuen Attributes automatisch mit ansteigender Nummer versehen und kann in einem späteren Schritt umbenannt werden.

Zusammenlegung
Wenn ein Attribut in seine einzelnen Bestandteile zerlegt ist (z. B. Datum als Tag, Monat, Jahr) kann dieses zusammengelegt werden (in TT-MM-JJJJ). Hierzu kann der Anwender ein optionales Trennzeichen angeben, sowie den Namen des neuen Attributes. Weiter muss die anzuhängende Spalte angegeben werden (vgl. Abbildung 6.3e). Optional können die Spalten aus denen sich das neue Attribut zusammensetzt direkt gelöscht werden. Für den Fall, dass mehr als eine Konkatenation stattfinden soll, so kann die Operation mehrfach ausgeführt werden bis das gewünschte Ergebnis erreicht wurde.

Umbenennen
Attribute müssen über alle Datensätze gleich benannt sein, damit diese verknüpft werden können. Hierzu kann der Anwender durch Eingabe des gewünschten Bezeichners diese umbenennen.

Spalte erstellen
Sollte ein weiteres Attribut benötigt werden, welches nicht mit den obigen Operationen hergeleitet werden kann ist es möglich eine weitere Spalte zu erzeugen.

Zeilenmenü
Bei einem Rechtsklick auf die Zeilennummer öffnet sich das Zeilenmenü (vgl. Abbildung 6.3b). Dieses bietet dem Anwender zwei unterschiedliche Möglichkeiten:

Zeile löschen
In diesem Fall wird die ausgewählte Zeile und der zugrunde liegende Tupel aus dem Datensatz entfernt.

Ähnliche Zeilen löschen
In diesem Fall wird die ausgewählte Zeile, sowie alle ähnlichen Zeilen entfernt. Im aktuellen Entwicklungsstand werden hierbei lediglich die leeren Zellen berücksichtigt, d. h. es werden alle Zeilen entfernt, bei denen exakt die gleichen Attribute unbelegt sind.

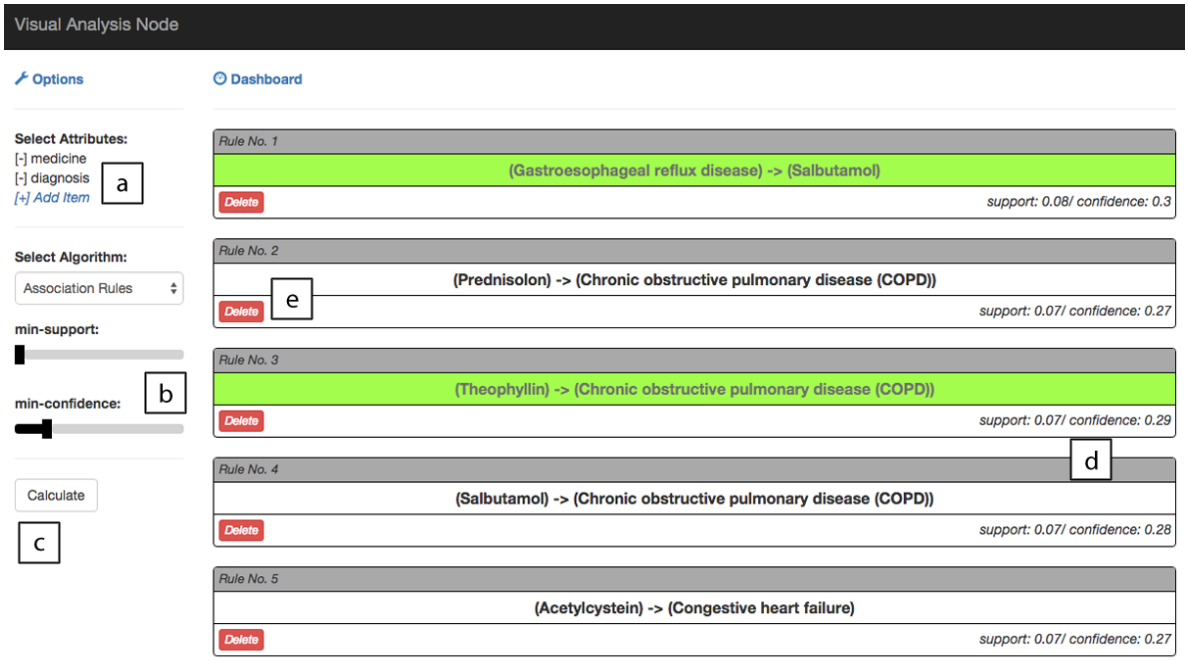


Abbildung 6.4: Visual Analysis Node: Grafische Oberfläche

6.5 Visual Analysis Node

In einem zweiten Schritt sollen die Zusammenhänge in den Daten durch eine Assoziationsanalyse aufgedeckt werden. Diese Funktionalität stellt der *Visual Analysis Node* zur Verfügung und veranschaulicht die Kopplung zwischen einem automatischen und visuellem Verfahren. Die Oberfläche ist in Abbildung 6.4 dargestellt. Nachfolgend werden die einzelnen Funktionen anhand der chronologischen Reihenfolge skizziert:

Auswahl der Attribute
In einem ersten Schritt kann der Anwender die für die Analyse zu berücksichtigenden Attribute hinzufügen oder entfernen (*a*). Dies ist notwendig, da der vorausgehende *Visual Merge Node* eine möglichst umfangreiche Datenbasis für verschiedene Analysen gewährleisten soll – ohne Berücksichtigung der expliziten Analyse. Dies ermöglicht dem Anwender somit ohne erneute Datenaufbereitung die Analyse dynamisch zu verändern. Diese Attribut-Selektion verringert die benötigte Laufzeit der Analysealgorithmen und erlaubt eine Eingrenzung auf für den gewünschten Zweck relevante Attribute.

Spezifikation der Analyse
Anschließend ist vorgesehen, dass der Anwender das gewünschte Analyseverfahren auswählen kann und die Visualisierung entsprechend angepasst wird. In diesem Prototyp beschränkt sich die Funktionalität auf die Assoziationsanalyse und erlaubt die Spezifikation erforderlichen Parameter. Diese wurden in Abschnitt 2.2 aufgeführt und erläutert. Über zwei Schieberegler (*b*) kann der Anwender so den gewünschten minimalen Support und die minimale Confidence spezifizieren. Für die Berechnung der häufig auftretenden Regeln wird die relative Support-Definition zugrunde gelegt.

Visualisierung der Assoziationsregeln
Wenn der Anwender die beiden obigen Schritte abgeschlossen hat kann die Berechnung gestartet (*c*) und die Assoziationsregeln basierend auf den spezifizierten Parametern mit Hilfe des Apriori-Algorithmus [AS94] berechnet werden. Die gefundenen Assoziationsregeln werden nach ihrem Support absteigend sortiert und ausgegeben (*d*). Weiterhin werden die 10 % der Assoziationsregeln mit der höchsten Confidence, d. h. der höchsten Eintrittswahrscheinlichkeit, grün unterlegt und farblich hervorgehoben (vgl. Abbildung 6.4). Bei Bedarf können die Parameter angepasst und eine Neuberechnung der Assoziationsregeln gestartet werden.

Entfernen von Assoziationsregeln
Der Anwender kann einzelne – für irrelevant befundene – Assoziationsregeln entfernen (*e*), worauf die Visualisierung neu berechnet wird. Mit dieser Funktionalität wird der Subjektivität des Anwenders Rechnung getragen.

6.6 Bewertung

Der vorgestellte Prototyp verbindet exemplarisch *Visual Analytics* mit Data Mashups und bietet dem Anwender zwei neue Knoten.

Der *Visual Merge Node* erweitert die bisher vorhandene Verbundoperation durch neue Konzepte. Die bisherige Kombination wurde durch die Spezifikation eines Verbundattributes und ohne Einbeziehung des Anwenders durchgeführt. Die Implementierung auf Basis von *Visual Analytics* bindet den Anwender in diesen Prozess ein und ermöglicht auf diese Weise ein Verständnis für die vorhandenen Daten.

Weiterhin kann durch die interaktive Visualisierung eine höhere Datenqualität erreicht werden als dies mit automatischen Verfahren möglich ist (vgl. Abschnitt 5.1). Als Hilfsmittel wird ein Fortschrittsbalken dargestellt, der zu jedem Zeitpunkt die Datenqualität visualisiert. Die komplizierte Schemaintegration wird vereinfacht, da der Anwender die Semantik hinter Attributbezeichnern identifizieren kann. Neben der Behebung von Datenmängeln kann der Datensatz "*fit for use*" (vgl. Abschnitt 2.11) gemacht werden, beispielsweise durch Konstruktion neuer Attribute.

Der *Visual Analysis Node* kombiniert eine interaktive Oberfläche mit automatischen Verfahren im Hintergrund. Auf diese Weise entsteht ein neuer Knoten der im Zusammenspiel mit dem Anwender sinnvolle Assoziationsregeln berechnet. Durch diese Kombination kann die Analyse durch Anpassung der Parameter schrittweise zu sinnvollen Resultaten geführt werden, die durch einen fixen Standardwert alleine nicht zu erreichen sind.

Durch farbliche Hervorhebung der stärksten Assoziationsregeln springen diese dem Anwender direkt ins Auge und es wird das Konzept der präattentiven Wahrnehmung (vgl. Abschnitt 2.5) umgesetzt.

In vielen Fällen sind die stärksten Assoziationsregeln verhältnismäßig trivial. Der Testdatensatz des Szenarios findet beispielsweise einfache Assoziationsregeln des Schemas *Land -> Geschlecht*, welche zumindest im Kontext der Problemstellung irrelevant sind. Der Anwender wird hierbei durch *Visual Analytics* in die Lage versetzt einzelne Assoziationsregeln aus der Ergebnismenge zu entfernen und diese auf die interessante Untermenge zu reduzieren.

Die Verbindung von *Visual Analytics* mit Data Mashups bietet demnach einen integrierten Ansatz für umfassende und flexible Analysen. Der Anwender erhält weitgehende Freiheiten und kann sein implizites Hintergrundwissen in den verschiedenen Schritten des Analyseprozesses einbringen. Durch die Kombination dieser Forschungsgebiete ist es für den Anwender infolgedessen möglich ohne technische Kenntnisse oder hinzugezogene Experte eine Analyse frei kombinierbarer Daten durchzuführen.

7 Zusammenfassung, Fazit und Ausblick

In diesem Kapitel wird zunächst die Arbeit rekapituliert und ein Fazit über die gewonnenen Erkenntnisse gezogen, sowie schlussendlich ein Ausblick auf zukünftige Arbeiten geworfen.

7.1 Zusammenfassung

In dieser Arbeit wurden die Themengebiete *Visual Analytics*, Visualisierung und Data Mining gegenübergestellt, abgegrenzt und in einem Ordnungsrahmen klassifiziert. Ein besonderes Augenmerk lag dabei auf dem noch sehr jungen Forschungsgebiet *Visual Analytics*, welches automatische und visuelle Verfahren mit dem Ziel einer besseren Analyse verknüpft. Auf Basis einer Literaturrecherche wurden die verschiedenen Ziele und Konzepte identifiziert, sowie anschließend in einer neuen und umfassenderen Definition zusammengefügt. In Verbindung hierzu wurde der *Visual Analytics*-Prozess erweitert und mit dem *Knowledge Discovery*-Prozess der automatischen Datenverarbeitung verknüpft. Weiterhin wurden die Möglichkeiten, die sich durch diesen neuen Ansatz bieten, im Kontext der Daten- und Analysequalität beleuchtet und die entstehenden Chancen dargelegt. Abschließend wurde auf Basis von Data Mashups eine prototypische Implementierung entwickelt, welche die Datenaufbereitung und Datenanalyse durch Anwendung von *Visual Analytics* veranschaulicht.

7.2 Fazit

Visual Analytics war zuvor ein weit gefasster Begriff in sehr unterschiedlicher Auslegung. Durch die neu entwickelte Definition wird *Visual Analytics* und das dahinter liegende Konzept des *Human In The Loop* umfassend abgegrenzt und der für die Nutzung des Potentials nötige Prozess erläutert. Der dargelegte Ordnungsrahmen unterstützt das Verständnis der hierdurch entstehenden Vorteile. Es wird deutlich, dass bisherige Verfahren die im Zusammenhang mit *Big Data* auftretenden Herausforderungen eigenständig nicht lösen können.

Visual Analytics bietet gerade in der Verbindung mit Data Mashups in diesem Bereich Vorteile. Der Anwender löst sich von der Notwendigkeit technischer Expertise und kann sein implizites Hintergrundwissen in die Analyse einbringen. Die prototypische Implementierung veranschaulicht diese Annahme und belegt, dass für eine Analyse keine Programmierkenntnisse erforderlich sind, sondern der ganze Prozess über visuelle Schnittstellen gesteuert werden kann.

Hierdurch entsteht eine einzigartige Möglichkeit maßgeschneiderte, flexible Analysen durchzuführen. Auf diese Weise werden Analysen möglich, die zuvor aufgrund mangelnder Nachfrage oder Unkenntnis über den Bedarf durch IT-Abteilungen nicht umgesetzt wurden. Durch die Integration des Anwenders können auch semantische Datenmängel erkannt und korrigiert werden. Die Resultate der Assoziationsanalyse können durch Anpassung der Parameter eingegrenzt und manuell nachbearbeitet werden. Entsprechend ist eine höhere Daten- und Analysequalität durch Umsetzung von *Visual Analytics* zu erwarten.

Bei allen Vorteilen und Chancen *Visual Analytics* an zwei Nachteilen. Einerseits die Abhängigkeit von den Kenntnissen des Anwenders. Durch automatische Verfahren bestehen weniger Eingriffsmöglichkeiten und folglich weniger Konsequenzen aus fehlerhaften Entscheidungen. Zusätzlich besitzt jeder Mensch einen Bias und könnte dadurch sinnvolle Muster übersehen, welche automatische Verfahren erkennen würden. Anderseits steigt die für eine Analyse benötigte Zeit durch die wiederholten Interaktionen.

Das Fazit aus dieser Arbeit lautet entsprechend, dass *Visual Analytics* ein großes Potential bietet, wobei der größte Vorteil – der Anwender – gleichzeitig das größte Hindernis für eine erfolgreiche Analyse darstellt.

7.3 Ausblick

Auf Basis der in dieser Arbeit gewonnenen Erkenntnisse ergeben sich für die Zukunft diverse Möglichkeiten das Konzept, nach dem neu entwickelten Verständnis von *Visual Analytics*, umzusetzen.

Die Verwendung von Data Mashups erlaubt dem Anwender individuelle Analysen zu planen und interaktiv durchzuführen. In der vorgestellten Implementierung ist es notwendig den vollständigen Datensatz vorzubereiten und dies bei jeder Analyse, was zu gleichförmigen Wiederholungen führt. Die Neugier auf neue Erkenntnisse wird dabei tendenziell stark reduziert. Das in den verwandten Arbeiten vorgestellte Werkzeug *Wrangler* generiert im Hintergrund aus den durchgeführten Aktionen wiederverwertbare Regeln. Es ist zu prüfen, inwiefern ähnliche Konzepte auch in anderen Schritten der Analyse angewendet werden können. Im besten Fall könnte *Visual Analytics* in Kombination mit Data Mashups zur Erstellung eigenständiger (semi-)automatischer Analyseabläufe führen und den Entwicklungsaufwand reduzieren.

Visual Analytics erfordert für den Erfolg, dass der Anwender ein Gefühl für die Charakteristik der Daten gewinnt. Im Zusammenhang mit Big Data ist selbiges mit dem vorgestellten Prototyp nicht zu erwarten. Hierbei ist zu prüfen, ob die Anzahl der durch den Anwender aktiv zu bearbeitenden Datensätze auf eine charakteristische Teilmenge reduziert werden kann, beispielsweise durch Äquivalenzklassen.

Um den Erfolg von *Visual Analytics* zu evaluieren reichen die identifizierten Methoden nicht aus. An dieser Stelle ist es nötig nach neuen Ansätzen zu suchen und diese in den Prozess zu integrieren.

Literaturverzeichnis

[ABEM15] D. Apel, W. Behme, R. Eberlein, C. Merighi. *Datenqualität erfolgreich steuern: Praxislösungen für Business-Intelligence-Projekte*. dpunkt.verlag, Heidelberg, 3. Auflage, 2015. (Zitiert auf den Seiten 25 und 29)

[ABM07] W. Aigner, A. Bertone, S. Miksch. Tutorial: Introduction to Visual Analytics. In A. Holzinger, Herausgeber, *HCI and Usability for Medicine and Health Care SE - 41*, Band 4799 von *Lecture Notes in Computer Science*, S. 453–456. Springer-Verlag, Berlin Heidelberg, 2007. doi:10.1007/978-3-540-76805-0{_}41. URL http://dx.doi.org/10.1007/978-3-540-76805-0{_}41. (Zitiert auf den Seiten 24, 51, 53 und 62)

[AEEK99] M. Ankerst, C. Elsen, M. Ester, H.-P. Kriegel. Visual Classification: An Interactive Approach to Decision Tree Construction. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, S. 392–396. ACM, New York, NY, USA, 1999. doi:10.1145/312129.312298. URL http://doi.acm.org/10.1145/312129.312298. (Zitiert auf Seite 44)

[AEK00] M. Ankerst, M. Ester, H.-P. Kriegel. Towards an Effective Cooperation of the Computer and the User for Classification. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00*, S. 179–188. ACM Press, New York, New York, USA, 2000. doi:10.1145/347090.347124. URL http://portal.acm.org/citation.cfm?doid=347090.347124. (Zitiert auf Seite 44)

[Ama12] Amazon.com Inc. *For The Eighth Consecutive Year, Amazon Ranks #1 In Customer Satisfaction During The Holiday Shopping Season*. Amazon.com Inc., 2012. URL http://phx.corporate-ir.net/phoenix.zhtml?c=176060{&}p=irol-newsArticle{&}ID=1769785. (Zitiert auf Seite 9)

[And06] K. Andrews. Evaluating Information Visualisations. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, S. 1–5. ACM, 2006. (Zitiert auf Seite 73)

[Ank01] M. Ankerst. *Visual Data Mining*. Dissertation, Ludwig-Maximilians-Universität München, 2001. (Zitiert auf Seite 58)

[AS94] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules. In J. B. Bocca, M. Jarke, C. Zaniolo, Herausgeber, *Proceedings of 20th International Conference on Very Large Data Bases*, S. 487 – 499. Morgan Kaufmann, 1994. (Zitiert auf den Seiten 25 und 84)

[BA03] R. Blumberg, S. Atre. The Problem with Unstructured Data. *DM REVIEW*, S. 42 – 46, 2003. (Zitiert auf Seite 29)

[BBHK10] M. R. Berthold, C. Borgelt, F. Höppner, F. Klawonn. Data Understanding. In *Guide to Intelligent Data Analysis - How to Intelligently Make Sense of Real Data*, Kapitel 4, S. 33–79. Springer-Verlag, London, 2010. doi:10.1007/978-1-84882-260-3{_}4. URL http://link.springer.com/10.1007/978-1-84882-260-3{_}4. (Zitiert auf Seite 62)

[BCD+08] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel. *KNIME: The Konstanz Information Miner*, Kapitel 38, S. 319–326. Springer-Verlag, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-78246-9{_}38. URL http://dx.doi.org/10.1007/978-3-540-78246-9{_}38. (Zitiert auf Seite 56)

[BD08] D. Bruzzese, C. Davino. Visual Mining of Association Rules. In S. J. Simoff, M. H. Böhlen, A. Mazeika, Herausgeber, *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*, Kapitel 8, S. 103 – 122. Springer-Verlag, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-71080-6{_}8. (Zitiert auf den Seiten 48 und 49)

[BE15] S. Bresciani, M. J. Eppler. The Pitfalls of Visual Representations: A Review and Classification of Common Errors Made While Designing and Interpreting Visualizations. *SAGE Open*, 5(4), 2015. doi:10.1177/2158244015611451. URL http://sgo.sagepub.com/lookup/doi/10.1177/2158244015611451. (Zitiert auf den Seiten 62 und 69)

[BJK+16] T. Blascheck, M. John, K. Kurzhals, S. Koch, T. Ertl. VA2 : A Visual Analytics Approach for Evaluating Visual Analytics Applications. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):61–70, 2016. doi:10.1109/TVCG.2015.2467871. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7192649. (Zitiert auf Seite 73)

[BKK97] C. Brunk, J. Kelly, R. Kohavi. MineSet: An Integrated System for Data Mining. 1997. (Zitiert auf Seite 56)

[BL09] Y. Benjamini, M. Leshno. Statistical Methods for Data Mining. In *Data Mining and Knowledge Discovery Handbook*, S. 523–540. Springer US, Boston, MA, 2009. doi:10.1007/978-0-387-09823-4{_}25. URL http://link.springer.com/10.1007/978-0-387-09823-4{_}25. (Zitiert auf Seite 72)

[BL10] E. Bertini, D. Lalanne. Investigating and Reflecting on the Integration of Automatic Data Analysis and Visualization in Knowledge Discovery. *ACM SIGKDD Explorations Newsletter*, 11(2):9 – 18, 2010. doi:10.1145/1809400.1809404. URL http://portal.acm.org/citation.cfm?doid=1809400.1809404. (Zitiert auf den Seiten 56, 59 und 72)

[BR00] T. Boren, J. Ramey. Thinking Aloud: Reconciling Theory and Practice. *IEEE Transactions on Professional Communication*, 43(3):261–278, 2000. doi:10.1109/47.867942. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=867942. (Zitiert auf Seite 73)

[BS06] C. Batini, M. Scannapieco. Data Quality Dimensions. In *Data Quality - Concepts, Methodologies and Techniques*, Kapitel 2, S. 19 – 49. Springer-Verlag Berlin Heidelberg, 1. Auflage, 2006. doi:10.1007/3-540-33173-5{_}2. URL http://link.springer.com/10.1007/3-540-33173-5{_}2. (Zitiert auf Seite 26)

[CAW14] P. van der Corput, J. Arends, J. J. van Wijk. Visualization of Medicine Prescription Behavior. *Computer Graphics Forum*, 33(3):161–170, 2014. doi:10.1111/cgf.12372. URL http://doi.wiley.com/10.1111/cgf.12372. (Zitiert auf Seite 39)

[CKPT92] D. R. Cutting, D. R. Karger, J. O. Pedersen, J. W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *15th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, S. 318–329. ACM Press, New York, New York, USA, 1992. doi:10.1145/133160.133214. URL http://portal.acm.org/citation.cfm?doid=133160.133214. (Zitiert auf Seite 45)

[CMS99] S. K. Card, J. D. Mackinlay, B. Shneiderman. Information Visualization. In S. K. Card, J. D. Mackinlay, B. Shneiderman, Herausgeber, *Readings in Information Visualization: using vision to think*, Kapitel 1, S. 1–34. Morgan Kaufmann, 1999. (Zitiert auf den Seiten 19 und 20)

[Con] J. Constine. How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day. *techcrunch.com*, (22.08.2012). URL http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/. (Zitiert auf Seite 9)

[DGG15] A. De Mauro, M. Greco, M. Grimaldi. What is Big Data? A Consensual Definition and a Review of Key Research Topics. In *Proceedings of the 4th International Conference on Integrated Information*, Band 1644, S. 97–104. 2015. doi:10.1063/1.4907823. URL http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.4907823. (Zitiert auf Seite 13)

[DGS99] J. Dörre, P. Gerstl, R. Seiffert. Text mining. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, S. 398–401. ACM Press, New York, New York, USA, 1999. doi:10.1145/312129.312299. URL http://doi.acm.org/10.1145/312129.312299http://portal.acm.org/citation.cfm?doid=312129.312299. (Zitiert auf den Seiten 10, 16 und 29)

[DGS01] J. Dörre, P. Gerstl, R. Seiffert. Text Mining. In H. Hippner, U. Küsters, M. Meyer, K. D. Wilde, Herausgeber, *Handbuch Data Mining im Marketing*, Kapitel 12, S. 465 – 488. Vieweg Verlagsgesellschaft, Braunschweig/Wiesbaden, 1. Auflage, 2001. (Zitiert auf den Seiten 16 und 29)

[DM14a] F. Daniel, M. Matera. Introduction. In *Mashups - Concepts, Models and Architectures*, Kapitel 1, S. 1–12. Springer-Verlag Berlin Heidelberg, 1. Auflage, 2014. doi:10.1007/978-3-642-55049-2{_}1. URL http://link.springer.com/10.1007/978-3-642-55049-2{_}1. (Zitiert auf Seite 30)

[DM14b] F. Daniel, M. Matera. Mashups. In *Mashups - Concepts, Models and Architectures*, Kapitel 6, S. 137–181. Springer-Verlag Berlin Heidelberg, 1. Auflage, 2014. doi:10.1007/978-3-642-55049-2{_}6. URL http://link.springer.com/10.1007/978-3-642-55049-2{_}6. (Zitiert auf Seite 30)

[EFN12] A. Endert, P. Fiaux, C. North. Semantic Interaction for Visual Text Analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, S. 473–482. ACM, New York, NY, USA, 2012. doi:10.1145/2207676.2207741. URL http://doi.acm.org/10.1145/2207676.2207741. (Zitiert auf Seite 51)

[EHR+14] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, C. Andrews. The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*, S. 1–25, 2014. doi:10.1007/s10844-014-0304-9. (Zitiert auf den Seiten 25, 45, 46 und 51)

[EMC] EMC Corporation. Digital Universe Invaded By Sensors. (09.04.2014). URL http://www.emc.com/about/news/press/2014/20140409-01.htm. (Zitiert auf den Seiten 9 und 10)

[ES00] M. Ester, J. Sander. *Knowledge Discovery in Databases*. Springer-Verlag, Berlin, Heidelberg, 2000. doi:10.1007/978-3-642-58331-5. URL http://link.springer.com/10.1007/978-3-642-58331-5. (Zitiert auf Seite 18)

[FPSS96a] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39(11):27–34, 1996. (Zitiert auf Seite 17)

[FPSS96b] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Herausgeber, *Advances in Knowledge Discovery and Data Mining*, Kapitel 1, S. 1–34. AAAI Press [u.a.], 1996. URL http://www.gbv.de/dms/goettingen/190022256.pdf. (Zitiert auf den Seiten 14, 15 und 17)

[FPSS96c] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37, 1996. (Zitiert auf den Seiten 14, 15, 16 und 27)

[Gar84] D. A. Garvin. What Does Product Quality Really Mean. *MIT Sloan management review*, 26(1):25–43, 1984. (Zitiert auf Seite 25)

[GMV96] I. Guyon, N. Matić, V. Vapnik. Discovering Informative Patterns and Data Cleaning. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Herausgeber, *Advances in knowledge discovery and data mining*, Kapitel 7, S. 181–203. 1996. URL http://dl.acm.org/citation.cfm?id=257944. (Zitiert auf Seite 27)

[Goo12] Google Inc. Zeitgeist 2012, 2012. URL http://www.google.com/zeitgeist/2012/{#}the-world. (Zitiert auf Seite 9)

[Gri08] S. Grimes. Unstructured data and the 80 percent rule. *Carabridge Bridgepoints*, 2008. URL http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/. (Zitiert auf den Seiten 10 und 29)

[Grö15] C. Gröger. *Advanced Manufacturing Analytics: Datengetriebene Optimierung von Fertigungsprozessen*. Josef Eul, 2015. (Zitiert auf den Seiten 14 und 16)

[GS05] H. Griethe, H. Schumann. Visualizing Uncertainty for Improved Decision Making. In *Proceedings of the 4th International Conference on Business Informatics Research BIR 2005*. 2005. (Zitiert auf Seite 69)

[GS06] H. Griethe, H. Schumann. The Visualization of Uncertain Data: Methods and Problems. In *Proceedings of Simulation and Visualization*. 2006. (Zitiert auf den Seiten 69 und 70)

[HBE96] C. G. Healey, K. S. Booth, J. T. Enns. High-Speed Visual Estimation Using Preattentive Processing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(2):107–135, 1996. (Zitiert auf den Seiten 21 und 62)

[Hin14] A. Hinneburg. Concepts of Visual and Interactive Clustering. In C. C. Aggarwal, C. K. Reddy, Herausgeber, *Data Clustering: Algorithms and Applications*, Kapitel 19, S. 483–503. Boca Raton, FL, 2014. (Zitiert auf Seite 45)

[HK06] J. Han, M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2. Auflage, 2006. (Zitiert auf den Seiten 14, 15, 17, 18, 27, 58 und 78)

[HM16] P. Hirmer, B. Mitschang. *FlexMash - Flexible Data Mashups Based on Pattern-Based Model Transformations*, Kapitel 2, S. 12–30. Springer International Publishing, Cham, 2016. doi:10.1007/978-3-319-28727-0{_}2. URL http://dx.doi.org/10.1007/978-3-319-28727-0{_}2. (Zitiert auf den Seiten 30 und 76)

[HOG+12] M. S. Hossain, P. K. R. Ojili, C. Grimm, R. Muller, L. T. Watson, N. Ramakrishnan. Scatter/Gather Clustering: Flexibly Incorporating User Feedback to Steer Clustering Results. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2829–2838, 2012. doi:10.1109/TVCG.2012.258. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6327289. (Zitiert auf den Seiten 45 und 47)

[HRWM15] P. Hirmer, P. Reimann, M. Wieland, B. Mitschang. Extended Techniques for Flexible Modeling and Execution of Data Mashups. In *Proceedings of 4th International Conference on Data Management Technologies and Applications*, S. 111–122. SCITEPRESS - Science and and Technology Publications, 2015. doi:10.5220/0005558201110122. URL http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005558201110122. (Zitiert auf Seite 76)

[HS12] J. Heer, B. Shneiderman. Interactive Dynamics for Visual Analysis. *Queue*, 10(2):30, 2012. doi:10.1145/2133416.2146416. URL http://doi.acm.org/10.1145/2133416.2146416http://dl.acm.org/citation.cfm?doid=2133416.2146416. (Zitiert auf Seite 56)

[HW01] H. Hippner, K. D. Wilde. Der Prozess des Data Mining im Marketing. In H. Hippner, U. Küsters, M. Meyer, K. D. Wilde, Herausgeber, *Handbuch Data Mining im Marketing*, Kapitel 2, S. 21–91. Vieweg Verlagsgesellschaft, Braunschweig/ Wiesbaden, 1. Auflage, 2001. (Zitiert auf den Seiten 14, 15, 27 und 61)

[IN07] W. H. Inmon, A. Nesavich. *The Environments of Structured Data and Unstructured Data*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1. Auflage, 2007. (Zitiert auf Seite 29)

[Ins85] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985. doi:10.1007/BF01898350. URL http://dx.doi.org/10.1007/BF01898350. (Zitiert auf Seite 49)

[KAF+08] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melançon. Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. Stasko, J.-D. Fekete, C. North, Herausgeber, *Information Visualization*, Band 4950, Kapitel 7, S. 154–175. Springer-Verlag, Berlin Heidelberg, 2008. doi:10.1007/978-3-540-70956-5{_}7. URL http://dx.doi.org/10.1007/978-3-540-70956-5{_}7http://link.springer.com/chapter/10.1007{%}2F978-3-540-70956-5{_}7. (Zitiert auf den Seiten 9, 24, 53 und 61)

[KBC+07] M. Krishnan, S. Bohn, W. Cowley, V. Crow, J. Nieplocha. Scalable Visual Analytics of Massive Textual Datasets. In *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*, S. 1–10. IEEE, 2007. (Zitiert auf Seite 51)

[KBM10] H.-G. Kemper, H. Baars, W. Mehanna. *Business Intelligence – Grundlagen und praktische Anwendungen*. Vieweg+Teubner, Wiesbaden, 2010. doi:10.1007/978-3-8348-9727-5. URL http://link.springer.com/10.1007/978-3-8348-9727-5. (Zitiert auf den Seiten 14, 15, 67 und 70)

[KHP+11] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011. doi:10.1177/1473871611415994. URL http://ivi.sagepub.com/lookup/doi/10.1177/1473871611415994. (Zitiert auf den Seiten 27 und 69)

[KK11] M. Khan, S. S. Khan. Data and Information Visualization Methods, and Interactive Mechanisms: A Survey. *International Journal of Computer Applications*, 34(1):1–14, 2011. (Zitiert auf Seite 62)

[KKM+10a] D. A. Keim, J. Kohlhammer, F. Mansmann, T. May, F. Wanner. Introduction. In D. A. Keim, J. Kohlhammer, G. Ellis, F. Mansmann, Herausgeber, *Mastering*

the Information Age - Solving Problems with Visual Analytics, Kapitel 1, S. 1–6. Eurographics Association, Goslar, 2010. (Zitiert auf Seite 10)

[KKM+10b] D. A. Keim, J. Kohlhammer, F. Mansmann, T. May, F. Wanner. Visual Analytics. In D. A. Keim, J. Kohlhammer, G. Ellis, F. Mansmann, Herausgeber, *Mastering the Information Age - Solving Problems with Visual Analytics*, Kapitel 2, S. 7–18. Eurographics Association, Goslar, 2010. (Zitiert auf den Seiten 24, 51, 52, 53, 57 und 62)

[KMOZ08] D. A. Keim, F. Mansmann, D. Oelke, H. Ziegler. Visual Analytics: Combining Automated Discovery with Interactive Visualizations. In *Discovery Science*, S. 2–14. Springer-Verlag, 2008. (Zitiert auf den Seiten 33, 51 und 52)

[KMS+08] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, H. Ziegler. Visual Analytics: Scope and Challenges. In S. Simoff, M. Böhlen, A. Mazeika, Herausgeber, *Visual Data Mining*, Band 4404, Kapitel 6, S. 76–90. Springer-Verlag, Berlin Heidelberg, 2008. doi:10.1007/978-3-540-71080-6{_}6. URL http://dx.doi.org/10.1007/978-3-540-71080-6{_}6. (Zitiert auf den Seiten 24, 51 und 52)

[KMSZ06a] D. Keim, F. Mansmann, J. Schneidewind, H. Ziegler. Challenges in Visual Data Analysis. In *Tenth International Conference on Information Visualization (IV'06)*, S. 9–16. IEEE, 2006. doi:10.1109/IV.2006.31. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1648235. (Zitiert auf Seite 57)

[KMSZ06b] D. A. Keim, F. Mansmann, J. Schneidewind, H. Ziegler. Challenges in Visual Data Analysis. In *Tenth International Conference on Information Visualization (IV'06)*, S. 9–16. 2006. doi:10.1109/IV.2006.31. (Zitiert auf Seite 24)

[KMT10] D. A. Keim, F. Mansmann, J. Thomas. Visual Analytics: How Much Visualization and How Much Analytics? *ACM SIGKDD Explorations Newsletter*, 11(2):5, 2010. doi:10.1145/1809400.1809403. URL http://doi.acm.org/10.1145/1809400.1809403http://portal.acm.org/citation.cfm?doid=1809400.1809403. (Zitiert auf Seite 73)

[KPHH11] S. Kandel, A. Paepcke, J. Hellerstein, J. Heer. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, S. 3363–3372. ACM, 2011. (Zitiert auf den Seiten 41 und 42)

[Küs01] U. Küsters. Data Mining Methoden: Einordnung und Überblick. In H. Hippner, U. Küsters, M. Meyer, K. D. Wilde, Herausgeber, *Handbuch Data Mining im Marketing*, Kapitel 3, S. 95 – 130. Vieweg Verlagsgesellschaft, Braunschweig/Wiesbaden, 1. Auflage, 2001. (Zitiert auf den Seiten 16 und 18)

[Lan01] D. Laney. 3D Data Management: Controlling Data Volume, Velocity, Variety. Technischer Bericht, META Group Inc., 2001. (Zitiert auf Seite 13)

[Mir11] B. Mirkin. K-Means and Related Clustering Methods. In *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*, S. 221–281. Springer London, London, 2011. doi:10.1007/978-0-85729-287-2{_}6. URL http://link.springer.com/10.1007/978-0-85729-287-2{_}6. (Zitiert auf Seite 45)

[MM10] J. I. Maletic, A. Marcus. Data Cleansing: A Prelude to Knowledge Discovery. In O. Maimon, L. Rokach, Herausgeber, *Data Mining and Knowledge Discovery Handbook*, Kapitel 2, S. 19–32. Springer US, Boston, MA, 2. Auflage, 2010. doi:10.1007/978-0-387-09823-4. URL http://link.springer.com/10.1007/978-0-387-09823-4. (Zitiert auf Seite 27)

[MR10] O. Maimon, L. Rokach. Introduction to Knowledge Discovery and Data Mining. In O. Maimon, L. Rokach, Herausgeber, *Data Mining and Knowledge Discovery Handbook*, Kapitel 1, S. 1–15. Springer US, Boston, MA, 2. Auflage, 2010. doi:10.1007/978-0-387-09823-4. URL http://link.springer.com/10.1007/978-0-387-09823-4. (Zitiert auf den Seiten 9, 10, 14, 17 und 18)

[MRS08] C. D. Manning, P. Raghavan, H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. (Zitiert auf den Seiten 28 und 29)

[Mun08] T. Munzner. Process and Pitfalls in Writing Information Visualization Research Papers. In *Information Visualization*, S. 134–153. Springer-Verlag, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-70956-5{_}6. URL http://link.springer.com/10.1007/978-3-540-70956-5{_}6. (Zitiert auf Seite 19)

[MWLZ09] S. E. Madnick, R. Y. Wang, Y. W. Lee, H. Zhu. Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality*, 1(1):1–22, 2009. doi:10.1145/1515693.1516680. URL http://portal.acm.org/citation.cfm?doid=1515693.1516680. (Zitiert auf Seite 25)

[OD08] D. L. Olson, D. Delen. *Advanced Data Mining Techniques*. Springer-Verlag, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-76917-0. URL http://link.springer.com/10.1007/978-3-540-76917-0. (Zitiert auf Seite 28)

[OLW08] S. Olafsson, X. Li, S. Wu. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429–1448, 2008. doi:10.1016/j.ejor.2006.09.023. URL http://linkinghub.elsevier.com/retrieve/pii/S037722170600854X. (Zitiert auf Seite 14)

[OW11] R. W. Oldford, A. Waddell. Visual Clustering of High-dimensional Data by Navigating Low-dimensional Spaces. In *58th Congress of the International Statistical Institute, Special Topics Session*, Band 57. 2011. (Zitiert auf Seite 45)

[Pan01] A. Pang. Visualizing Uncertainty in Geo-spatial Data. In *Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology*, S. 1–14. National Research Council Arlington, VA, 2001. (Zitiert auf Seite 71)

[PB10] K. Puolamäki, A. Bertone. Introduction to the Special Issue on Visual Analytics and Knowledge Discovery. *ACM SIGKDD Explorations Newsletter*, 11(2):3–4, 2010. (Zitiert auf Seite 56)

[PBT+10] K. Puolamäki, A. Bertone, R. Therón, O. Huisman, J. Johansson, S. Miksch, P. Papapetrou, S. Rinzivillo. Data Mining. In D. A. Keim, J. Kohlhammer, G. Ellis, F. Mansmann, Herausgeber, *Mastering the Information Age - Solving Problems with Visual Analytics*, Kapitel 2, S. 39–56. Eurographics Association, Goslar, 2010. (Zitiert auf den Seiten 17, 55, 56 und 61)

[PC05] P. Pirolli, S. Card. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. *Proceedings of International Conference on Intelligence Analysis*, 2005:2–4, 2005. (Zitiert auf Seite 63)

[Pla04] C. Plaisant. The Challenge of Information Visualization Evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces - AVI '04*, S. 109. ACM Press, New York, New York, USA, 2004. doi:10.1145/989863.989880. URL http://portal.acm.org/citation.cfm?doid=989863.989880. (Zitiert auf Seite 73)

[PLW02] L. L. Pipino, Y. W. Lee, R. Y. Wang. Data Quality Assessment. *Communications of the ACM*, 45(4):211–218, 2002. (Zitiert auf den Seiten 25 und 26)

[PM08] N. Prat, S. Madnick. Measuring Data Believability: A Provenance Approach. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, S. 393–393. IEEE, 2008. doi:10.1109/HICSS.2008.243. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4439098. (Zitiert auf Seite 28)

[PS04] R. Price, G. Shanks. A Semiotic Information Quality Framework. In *Proceedings of the International Conference on Decision Support Systems*, S. 658–672. 2004. (Zitiert auf Seite 26)

[PT04] Pak Chung Wong, J. Thomas. Visual Analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004. doi:10.1109/MCG.2004.39. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1333623. (Zitiert auf den Seiten 24, 51 und 53)

[PWL97] A. T. Pang, C. M. Wittenbrink, S. K. Lodha. Approaches to Uncertainty Visualization. *The Visual Computer*, 13(8):370–390, 1997. (Zitiert auf den Seiten 69, 70 und 71)

[RD00] E. Rahm, H. H. Do. Data Cleaning: Problems and Current Approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 23(4):3–13, 2000. (Zitiert auf Seite 27)

[Red12] T. C. Redman. Data Quality Management Past, Present, and Future: Towards a Management System for Data. In *Handbook of Data Quality*, S. 15–40. Springer-Verlag, Berlin, Heidelberg, 2012. doi:10.1007/978-3-642-36257-6{_}2. URL http://link.springer.com/10.1007/978-3-642-36257-6{_}2. (Zitiert auf Seite 26)

[Red13] T. C. Redman. Introduction. In *Data Driven: Profiting from Your Most Important Business Asset*, S. 1 – 10. Harvard Business Press, 2013. (Zitiert auf Seite 26)

[Rei99] T. Reinartz. *Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-world Domains*. Springer-Verlag, Berlin, Heidelberg, 1999. (Zitiert auf Seite 61)

[SD02] T. Soukup, I. Davidson. Introduction to Data Visualization and Visual Data Mining. In *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, Kapitel 1, S. 15 – 34. John Wiley & Sons, 2002. (Zitiert auf den Seiten 57 und 58)

[SGL08] J. Stasko, C. Görg, Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008. (Zitiert auf Seite 34)

[Shn96] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, S. 336–343. IEEE Comput. Soc. Press, 1996. doi:10.1109/VL.1996.545307. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=545307. (Zitiert auf Seite 23)

[Shn02] B. Shneiderman. Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization*, 1(1):5–12, 2002. doi:10.1057/palgrave.ivs.9500006. URL http://ivi.sagepub.com/content/1/1/5.abstract. (Zitiert auf Seite 72)

[SM00] H. Schumann, W. Müller. *Visualisierung: Grundlagen und allgemeine Methoden*. Springer-Verlag, Berlin, Heidelberg [u.a.], 2000. doi:10.1007/978-3-642-57193-0. URL http://link.springer.com/10.1007/978-3-642-57193-0. (Zitiert auf den Seiten 19 und 22)

[SME08] A. Savikhin, R. Maciejewski, D. S. Ebert. Applied Visual Analytics for Economic Decision-Making, 2008. doi:10.1109/VAST.2008.4677363. (Zitiert auf Seite 72)

[SSB+04] G. Schmidt, Sue-Ling Chen, A. Bryden, M. Livingston, L. Rosenblum, B. Osborn. Multidimensional Visual Representations for Underwater Environmental Uncertainty. *IEEE Computer Graphics and Applications*, 24(5):56–65, 2004. doi:10.1109/MCG.2004.35. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1333628. (Zitiert auf Seite 71)

[TC05] J. J. Thomas, K. A. Cook. *Illuminating the Path - The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005. URL http://vis.pnnl.gov/pdf/RD{_}Agenda{_}VisualAnalytics.pdf. (Zitiert auf den Seiten 24, 51, 72 und 74)

[TD04] K. Techapichetvanich, A. Datta. Visual Mining of Market Basket Association Rules. In *Computational Science and Its Applications–ICCSA 2004*, S. 479–488. Springer-Verlag, 2004. (Zitiert auf den Seiten 47 und 56)

[TJKMW98] S. T. Teoh, T. J. Jankun-Kelly, K.-L. Ma, S. F. Wu. Visual Data Analysis for Detecting Flaws and Intruders in Computer Network Systems. *IEEE/ACM Transactions on Networking*, 6(5):515–528, 1998. (Zitiert auf Seite 51)

[Tuf01] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2. Auflage, 2001. (Zitiert auf Seite 62)

[VLF04] D. A. Varakin, D. Levin, R. Fidler. Unseen and Unaware: Implications of Recent Research on Failures of Visual Awareness for Human-Computer Interface Design. *Human-Computer Interaction*, 19(4):389–422, 2004. doi:10.1207/s15327051hci1904{_}9. URL http://www.tandfonline.com/doi/abs/10.1207/s15327051hci1904{_}9. (Zitiert auf Seite 62)

[Wag15] M. Wagner. Integrating Explicit Knowledge in the Visual Analytics Process. In *Doctoral Consortium on Computer Vision, Imaging and Computer Graphics Theory and Applications (DCVISIGRAPP 2015)*. SCITEPRESS Digital Library, Berlin, 2015. URL http://mc.fhstp.ac.at/sites/default/files/publications/Wagner{_}IntegratingExplicitKnowledgeInTheVisualAnalyticsProcess.pdf. (Zitiert auf Seite 25)

[War12] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., 3. Auflage, 2012. (Zitiert auf den Seiten 23, 62 und 69)

[Wat00] E. T. Watkins. Improving the Analyst and Decision-Maker's Perspective Through Uncertainty Visualization. Technischer Bericht, 2000. (Zitiert auf Seite 69)

[WBW96] J. Wood, K. Brodlie, H. Wright. Visualization over the World Wide Web and Its Application to Environmental Data. In *Proceedings of the 7th Conference on Visualization '96*, VIS '96, S. 81 – 86. IEEE Computer Society Press, Los Alamitos, CA, USA, 1996. URL http://dl.acm.org/citation.cfm?id=244979.245010. (Zitiert auf Seite 22)

[Weg04] G. Weglarz. Two Worlds Data-Unstructured and Structured. *DM Review*, 14(9):19–22, 2004. (Zitiert auf Seite 29)

[Wij05] J. J. van Wijk. The value of visualization, 2005. doi:10.1109/VISUAL.2005.1532781. (Zitiert auf den Seiten 10, 19, 24, 39, 63, 65, 66 und 72)

[WIR+10] J. van Wijk, T. Isenberg, J. B. Roerdink, A. C. Telea, M. Westenberg. Evaluation. In D. A. Keim, J. Kohlhammer, G. Ellis, F. Mansmann, Herausgeber, *Mastering the Information Age - Solving Problems with Visual Analytics*, Kapitel 8, S. 131 – 144. Eurographics Association, Goslar, 2010. (Zitiert auf Seite 25)

[WJD+09] X. Wang, D. H. Jeong, W. Dou, S.-w. Lee, W. Ribarsky, R. Chang. Defining and Applying Knowledge Conversion Processes to a Visual Analytics System. *Computers & Graphics*, 33(5):616–623, 2009. (Zitiert auf Seite 25)

[WS96] R. Y. Wang, D. M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996. doi: 10.1080/07421222.1996.11518099. URL http://www.tandfonline.com/doi/full/10.1080/07421222.1996.11518099. (Zitiert auf den Seiten 25 und 26)

[Wür03] V. G. Würthele. *Datenqualitätsmetrik für Informationsprozesse*. Dissertation, Eidgenössische Technische Hochschule Zürich, 2003. (Zitiert auf Seite 26)

[WWT99] P. C. Wong, P. Whitney, J. Thomas. Visualizing Association Rules for Text Mining. In *Proceedings of the 1999 IEEE Symposium on Information Visualization*, S. 120–123. IEEE Computer Society, 1999. (Zitiert auf Seite 48)

[Yan03] L. Yang. Visualizing Frequent Itemsets, Association Rules, and Sequential Patterns in Parallel Coordinates. In V. Kumar, M. L. Gavrilova, C. J. K. Tan, P. L'Ecuyer, Herausgeber, *Proceedings of the 2003 international conference on Computational science and its applications: Part I*, Band 2667 von *Lecture Notes in Computer Science*, S. 21–30. Springer-Verlag, Berlin, Heidelberg, 2003. doi:10.1007/3-540-44839-X. URL http://link.springer.com/10.1007/3-540-44839-X{_}3http://link.springer.com/10.1007/3-540-44839-X. (Zitiert auf Seite 49)

[You13] YouTube - Official Blog. Here's to eight great years. *YouTube, LLC*, (19.05.2013), 2013. URL http://youtube-global.blogspot.de/2013/05/heres-to-eight-great-years.html. (Zitiert auf Seite 9)

[ZC07] T. Zuk, S. Carpendale. Visualization of Uncertainty and Reasoning. In A. Butz, B. Fisher, A. Krüger, P. Olivier, S. Owada, Herausgeber, *Smart Graphics SE - 15*, Band 4569 von *Lecture Notes in Computer Science*, S. 164–177. Springer-Verlag, Berlin Heidelberg, 2007. doi:10.1007/978-3-540-73214-3{_}15. URL http://dx.doi.org/10.1007/978-3-540-73214-3{_}15. (Zitiert auf den Seiten 71 und 72)

[Zuk08] T. D. Zuk. *Visualizing uncertainty*. Dissertation, University of Calgary, 2008. (Zitiert auf Seite 69)

Alle URLs wurden zuletzt am 15. 03. 2016 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift