Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Simon Dreyer

Digitale Erzählstrukturen am Beispiel eines Datamining basierten Prognosesystems zur Parkhausauslastung

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Simon Dreyer

Digitale Erzählstrukturen am Beispiel eines Datamining basierten Prognosesystems zur Parkhausauslastung

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Klaus-Peter Schoeneberg

Eingereicht am: 1. August 2016

Simon Dreyer

Thema der Arbeit

Digitale Erzählstrukturen am Beispiel eines Datamining basierten Prognosesystems zur Parkhausauslastung

Stichworte

Digitale Erzählstrukturen, Datenjournalismus, Open-Data, Open-Government-Data, Data-Mining, Big-Data, Prädiktion

Kurzzusammenfassung

Diese Arbeit skizziert den aktuellen wissenschaftlichen Diskurs über den Wandel des Journalismus innerhalb der digitalen Gesellschaft. Es wird ein Arbeitsablauf für das Erstellen von datengetriebenen Geschichten auf der Basis von Data-Mining Prozessen erarbeitet und das Potential von Open-Data auf den Datenjournalismus beschrieben. Im zweiten Teil dieser Arbeit wird anhand eines Prognosesystems zur Auslastung der Parkhäuser in Hamburg der zuvor entwickelte Workflow an einem realen Beispiel verifiziert und Möglichkeiten und Problematiken der Umsetzung erörtert.

Simon Dreyer

Title of the paper

Digital narrative structures using the example of a data mining based forecasting system for car park usage

Keywords

Digital storytelling, Data journalism, Data-driven journalism, Open data, Open government data, Data mining, Big data, Prediction

Abstract

This thesis outlines the current academic discourse about the changes in journalism in the digital society. The first half of this thesis focuses on a proposed workflow for data-driven journalism through the use of open data and data mining processes. The previously developed workflow is then verified by developing a forecast system which predicts the usage of car parks in Hamburg, followed by an analysis of possibilities and problems.

Danksagung

Diese Arbeit stellt nur das Ende meines Bachelor-Studiums dar. Auf dem Weg dorthin haben mich vielen Menschen unterstützt und begleitet. Ohne sie hätte ich nicht die Möglichkeit gehabt, dort anzukommen, wo ich jetzt bin; oder es wäre zumindest schwieriger und sehr viel langweiliger gewesen.

Der erste Dank geht an meiner Familie, die mich mein ganzes Leben lang in dem wohin ich wollte und was ich Tat bedingungslos unterstützten und halfen.

Als Nächstes möchte ich allen Freundinnen und Freunden danken, die immer solidarisch zu mir standen, mich verpatzte Klausuren vergessen ließen und mich durch den ein oder anderen Anstoß und Diskus zu dieser Arbeit brachten.

Der Firma Subsehll, insbesondere Jan Rieke und Vasilis Ikonomou, möchte ich für die initialen Ideen und Anregungen danken, die mich zu dieser Arbeit geführt haben.

Prof. Dr. Klaus-Peter Schoeneberg und besonders Prof. Dr. Kai von Luck möchte ich für ihre Zeit und Mühe, sowie die vielen inspirierenden Ideen und Kontakte, die ich wären dieser Ausarbeitung von ihnen bekam, danken.

Auch Allen, die mich wären dieser Arbeit, sei es durch Korrekturen, Anmerkungen, Gespräche oder das Verzeihen, der manchmal etwas zu knappen Zeit, unterstützt haben, gebührt mein Dank.

Ein ganz besonders lieben Dank geht an dieser Stell an meine Freundin, für alles was sie mir in dieser Zeit gegeben und ausgehalten hat.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einführung

Daten und Informationen spielen in nahezu allen gesellschaftlichen Bereichen eine wichtige Rolle. Noch nie hatte die Menschheit Zugriff auf soviel gespeichertes Wissen wie in der heutigen Zeit; und diese „Datenberge“ werden vermutlich weiter wachsen. Eine der großen Aufgaben der Zukunft wird es sein, diese zu verarbeiten um so sinnvolle Schlüsse aus ihnen ziehen zu können. Informationen werden aber auch zusehends zu einem der umkämpftesten Rohstoffe unserer Gesellschaften. Daten werden gesammelt und ausgewertet um Macht- und Wirtschaftspositionen zu sichern und auszubauen. Sie bieten aber auch die Möglichkeit Zusammenhänge, Entscheidungen und Einflüsse in Gesellschaft, Wirtschaft und Politik offener und transparenter zu gestalten und können so dazu beitragen, die Welt um uns herum besser zu verstehen. Mit ihnen können Geschichten erzählt werden, die komplexe Themenbereiche verständlich, interessant und nachprüfbar darstellen. Wie Howard (2014) in seinem Aufsatz über die Kunst und Wissenschaft des daten-getriebenen Journalismus *(The Art and Science of Data-driven Journalism)* betonte, besteht nicht länger die Frage, ob Daten, Computer und Algorithmen für journalistische Arbeit im öffentlichen Interesse genutzt werden können, sondern wie, wann, wo, warum und von wem dies geschieht.

Als eine der Pionierinnen in diesem Bereich wird häufig die britischen Tageszeitung *The Guardian* genannt. Im Jahr 2009 konnte der Spendenskandal um das britische Parlament aufgedeckt werden, indem sie Daten aus öffentlichen Regierungsdokumenten auslasen, diese in strukturierter Form publizierten und mit Hilfe ihrer Leser*innen[1] auswerteten.

Spätestens aber mit der Analyse der geheimen Kriegstagebücher des Afghanistan-Krieges *(War Logs)* unter anderem durch den *Guardian*, *The New York Times* und *Spiegel Online* gilt der Datenjournalismus als einer der Hoffnungsträger der neuen Medienlandschaft.

Anders als Interviewpartner*innen oder Pressestellen kommunizieren Daten ihre Geschichten jedoch nicht von selbst. Es braucht Methoden und Fähigkeiten aus Informatik und Datenanalyse, um sie ihnen zu entlocken.

[1] Um alle Geschlechter zu berücksichtigen, wurde in dieser Arbeit das Sternchen (*) als Wildcard für alle Geschlechtsidentitäten verwendet *(Gendergap)*.

Offenheit und Transparenz sind Werte und Maßstäbe, die immer mehr Menschen einfordern. Diese Ideale werden in der digitalen Welt bereits seit Jahrzehnten von der Open Source Communitiy gefordert und gelebt. Ihre Prinzipien umfassen neben der Kostenfreiheit auch den Ausschluss jedweder Hindernisse bei der individuellen Anpassung und Weitergabe. Mit diesem Anspruch an Offenheit sind einige der weltweit verbreitetsten und anerkanntesten Softwarelösungen entstanden. Das Betriebssystem *Linux* und der weltweit meistbenutzte Webserver *Apache* sind nur zwei der unzähligen Beispiele erfolgreicher freier Software.

Auch Wikipedia, die umfangreichste und bekannteste Enzyklopädie der Welt, hat ohne Frage das Potenzial freier Lizenzen bewiesen.

Unter anderem durch diese Erfolge ist das Verlangen nach freiem und uneingeschränktem Zugang auf Informationen in verschiedenen Bereichen gestiegen. *Open-Data*, *Open Government Data*, *Open Medicine*, *Linked Open Data* und *Open Education* sind hier nur einige Schlagwörter. Sie alle vereint die Forderung nach der Öffnung von, meist aus Steuergeldern finanzierten, Datenpools für die Öffentlichkeit. Dadurch sollen unter anderem Innovation, Effizienz und Transparenz gesteigert werden.

Viele bereits erfolgreiche Projekte, wie das hamburgische Transparenzgesetz mit seinem aktiven Veröffentlichungsansatz, das bundesweite Datenportal *GovData* oder die Open-Data Portale der EU-Institutionen zeigen, dass die Notwendigkeit von Offenheit und Transparenz auch im breiten politischen Bewusstsein angekommen ist. Kaum ein Regulierungsthema hat es in so kurzer Zeit von der ersten Idee in die gesetzliche Realität geschafft (Dapp u. a., 2016).

1.1 Zielsetzung und Abgrenzung

Anknüpfend an die zunehmende Relevanz von und Forderung nach Transparenz und Offenheit hat diese Arbeit den Anspruch, auf der einen Seite den wissenschaftlichen und fachlichen Diskurs über digitale Erzählstrukturen aufzuzeigen und aus informationstechnischen Gesichtspunkten zu analysieren. Auf der anderen Seite soll anhand einer konkreten und möglichst vollständigen Umsetzung eines Beispiels der reale Ablauf eines solchen Projektes dokumentiert und anfallende Fragen und Probleme erörtert werden. Durch den Anspruch an sowohl theoretischen Diskurs als auch praktische Umsetzung können einzelne Fragestellungen möglicherweise nicht in voller Tiefe erörtert werden und dienen somit lediglich als Einstieg in die weitere wissenschaftliche Betrachtung. Dieser zweigliedrige Ansatz wurde gewählt, um ein möglichst umfassendes Bild liefern zu können und die Abstraktion auf andere Fragestellungen zu erleichtern.

1.2 Aufbau der Arbeit

In Kapitel 2 wird der Journalismus und sein Wandel innerhalb der fortschreitenden Digitalisierung dargestellt. Aufbauend wird auf den noch relativ jungen Bereich des Datenjournalismus eingegangen und unterschiedliche Auffassungen und Zielsetzungen aufgezeigt. Der zweite Teil der Analyse beschäftigt sich mit dem Themenfeld Open-Data und Open-Government-Data. Eigenschaften und Anforderungen werden aufgezeigt sowie die Situation in Deutschland und speziell in Hamburg skizziert.

In Kapitel 3 wird ein Workflow für das Erstellen von datenjournalistischen Geschichten erarbeitet. Der im Data-Mining verbreitete KDD Prozess wird an die Anforderungen und Bedingungen des Journalismus angepasst und weiterentwickelt. Die fünf Schritte des Modells werden einzeln beschrieben und mögliche Ansätze zur Realisierung geliefert. Ein besonderer Fokus wird auf die Beschreibung der verschiedenen Verfahren zur Mustererkennung in Datensätzen gelegt.

Der zuvor entwickelte Ablauf wird in Kapitel 4 anhand eines konkreten Beispiels validiert. Die einzelnen Teilschritte und Vorgehensweise werden anhand eines Prognosesystems zur Auslastung der Parkhäuser in Hamburg veranschaulicht. Die Daten hierzu stammen aus dem zuvor beschrieben Open-Data Portal der Freien und Hansestadt Hamburg. Weiter werden verschiedene Klassifikationsalgorithmen an diesem Beispiel beschrieben und die erreichten Ergebnisse verglichen.

In der Schlussbetrachtung in Kapitel 5 werden die Ergebnisse dieser Arbeit zusammengefasst und ein Ausblick auf weitere Fragestellungen und mögliche weitere Untersuchungen vorgebracht.

2 Analyse

2.1 Journalismus

Journalismus ist in seiner Geschichte kein starrer Begriff, sondern befindet sich im stetigen Wandel. Journalistische Arbeit hat in demokratischen Gesellschaften einen hohen Stellenwert und trägt zur öffentlichen Meinungs- und Willensbildung bei. Sie wird häufig als „vierte Gewalt“ im Staat bezeichnet. Dieses spiegelt sich auch im Presse- und Rundfunkrecht wieder, in dem die „dienende Funktion“ betont wird, die nur erfüllt werden kann, wenn Artikel 5 des Grundgesetzes (Meinungsfreiheit (Autonomie)) gewahrt ist. Ihre besondere Aufgabe liegt in der Selbstbeobachtung der Gesellschaft und dem Schaffen eines Gesamtüberblicks über relevante und reale Geschehen (Neuberger und Kapern, 2013, 2-26).

Der *Deutsche Journalisten-Verband* beschreibt das Berufsfeld des/der Journalinst*in selbst wie folgt:

> „Journalistinnen und Journalisten haben die Aufgabe, Sachverhalte oder Vorgänge öffentlich zu machen, deren Kenntnis für die Gesellschaft von allgemeiner, politischer, wirtschaftlicher oder kultureller Bedeutung ist. Durch ein umfassendes Informationsangebot in allen publizistischen Medien schaffen Journalistinnen und Journalisten die Grundlage dafür, dass jede/r die in der Gesellschaft wirkenden Kräfte erkennen und am Prozess der politischen Meinungs- und Willensbildung teilnehmen kann. Dies sind Voraussetzungen für das Funktionieren des demokratischen Staates.“ (DJV, 2015, 2)

Journalismus hat den Anspruch, möglichst objektiv und autonom zu berichten. Unter Objektivität (oder Faktizität) ist hier ein Realitätsbezug, in Abgrenzung zum Fiktionalen zu verstehen. Er möchte sich nur mit tatsächlichen Ereignisse befassen und diese unverändert wiedergeben. Autonomie umfasst den Anspruch, weitgehend frei von politischen, ökonomischen und anderen Zwängen zu sein. Da Journalist*innen nicht bei jedem Ereignis selbst vor Ort sein können, sind sie auf verlässliche Quellen angewiesen (Neuberger und Kapern, 2013).

2.2 Wandel von Information und Journalismus

Die Erfindung des Buchdruckes durch Johannes Gensfleisch zu Gutenberg um 1450 war einer der größten Meilensteine in der Geschichte der Wissensarchivierung und -publizierung. Erstmals war es möglich einem breiten Publikum Information zugänglich zu machen. Seitdem wächst die Fülle an archivierten Informationen stetig. Eine weitere Ära wurde mit der Digitaltechnik eingeleitet. Durch sie wurde es möglich, Informationen immer schneller und kostengünstiger zu speichern und vereinfacht, sie einer breiten Masse zugänglich zu machen.

Im Jahr 2002 hat die Menschheit erstmals mehr Information in digitaler als in analoger Form gespeichert. Waren um die Jahrtausendwende noch 75 Prozent aller Informationen analog gespeichert (meist in Form von Videobändern), lagen im Jahr 2007 bereits 97 Prozent aller Informationen in digitalem Format vor (Woyteqicz, 2013).

Laut dem Cisco Visula Network Indesx (VNI) steigerte sich der weltweite Internet-Traffic von 100 GB pro Tag im Jahre 1992 innerhalb von 10 Jahr bereits auf 100GB pro Sekunde. Im Jahre 2014 wurden über 16.000 GB Daten pro Sekunde ausgetauscht. Für das Jahr 2019 prognostiziert Cisco eine weitere Verdreifachung auf über 51.000 GB pro Sekunde. Dies entspricht einem Datenvolumen von 58 Millionen DVDs pro Stunde. Siehe Tabelle 2.1.

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GBps
2007	2000 GBps
2014	16,144 GBps
2019	51,794GBps

Tabelle 2.1: Globaler Internetverkehr - Historisch (Quelle: Cisco VNI (2015))

2.2.1 Journalismus im digitalen Zeitalter

Durch die Digitalvernetzung verlieren traditionelle Medien zunehmend ihr Informationsmonopol, da jeder Mensch mit Zugang zum Internet selbst publizieren kann. Medien finden sich nicht mehr ausschließlich in der Rolle des „Gatekeepers“, sondern auch in der des „Gatewatchers“ wieder. Sie haben nicht mehr allein die Aufgabe zu entscheiden, welche Informationen publiziert werden, sondern müssen zudem auch eine Vielzahl an öffentlich zugänglichen Informationen sichten. Dies führt dazu, dass die Hierarchie zwischen Produzent*innen und Rezipient*innen immer weiter verschwindet. Medienhäuser müssen sich von der Vorstellung lösen, lediglich

auf exklusive Quellen zu setzen, sondern auch als Vermittlerinnen und Interpretinnen von offenen Quellen agieren (Neumüller und Kahn, 2015).

Auch die Wege, auf denen journalistische Inhalte und Informationen die Rezipient*innen erreichen, haben sich in den letzten Jahren verändert. Nachrichten werden zunehmend nicht mehr aktiv von einzelnen Quellen bezogen, sondern erreichen die Personen über Social-Media-Kanäle, in dem sie von Freund*innen, Bekannten oder Kolleg*innen geteilt werden.

Nach dem Medientheoretiker Clay Shirky besteht für den/die Nutzer*in das Problem nicht darin, über zu viele Informationen zu verfügen (*information overload*), sondern darin, jene Informationen herauszufiltern, die wirklich interessant sind (*filter failure*). Für diese Filterung braucht es sowohl die Expertise von Journalist*innen als auch die von Computerexpert*innen. Neumüller und Kahn sprechen sich für eine stärkere Kolaboration dieser beiden Berufsgruppe aus und verweisen auf den Begriff des *programmer-journalist* der u.a. von den amerikanischen Autoren Seth Lewis und Nikki Usher geprägt wurde (Neumüller und Kahn, 2015, 16).

2.2.2 Datenjournalismus

> *„Datenjournalismus stellt aus dem vorliegenden Material neue Zusammenhänge her. Alles andere ist Infografik."*
>
> — *Mitarbeiter des ORF nach Neumüller und Kahn (2015)*

Es ist in der Fachliteratur bislang keine einheitliche Definition von Datenjournalismus zu finden. Verallgemeinert kann er als eine Kombination aus Recherche-Form und Veröffentlichungsansatz gesehen werden. Häufig wird es als die berufliche Praxis von Journalist*innen verstanden, Daten zu sichten, aufzuarbeiten, mit statistischen Methoden zu analysieren und anschließend darzustellen (Neumüller und Kahn, 2015).

Das Arbeiten mit Daten ist für Journalist*innen jedoch nichts Neues. Die Nutzung des Internets als Recherchequelle und das Einbeziehen von Studien, Fotos und Infografiken in die Berichterstattung wird von ihnen bereits seit Jahrzehnten sowohl in der Presse und dem Rundfunk als in auch Online-Medien eingesetzt. Seit den 1990 Jahren werden Berichte mit Infografiken angereichert, um Zusammenhänge oder Tendenzen besser darstellen zu können. Interaktivität wird bereits seit dem Aufkommen von Adobe Flash eingesetzt und der Umgang mit Statistiken gehört bereits seit 1970 zum Alltag von US-Redakteur*innen (Kramp u. a., 2013).

Für den Einsatz dieser Methoden passt nach Matzat (2011) der englische Begriff des *data-driven-journalism*, also der datengetrieben Journalismus, besser, da Datensätze lediglich zur Unterstützung der Berichterstattung herangezogen werden.

> „Eine klickbare Karte ist zwar interaktiv, aber noch lange kein wirklicher Datenjournalismus, genauso wenig wie ein Balken- oder Tortendiagramm.“ (Matzat, 2011)

Datenjournalismus (*data journalism*) geht nach Matzat einen Schritt weiter. Daten fungieren hier nicht als Recherchequelle, sondern werden zum zentralen Gegenstand der Berichterstattung. Sie sind nicht Teil der Story, sondern diese entsteht erst aus den Daten. Datenjournalismus umfasst auf der eine Seite das Analysieren und Auswerten von Rohdaten, sowie das Visualisieren der Erkenntnisse. Auf der anderen Seite aber auch das Schaffen von größtmöglicher Transparenz über die herangezogenen Quellen. Diese sollten möglichst in ihrer Gesamtheit öffentlich und damit überprüfbar gemacht werden, welches das Vertrauen in die journalistische Arbeit stärken kann. Mit seinem offenen und transparenten Ansatz folgt der Datenjournalismus den Forderungen der Open-Data-Bewegung, die sich für die freie Verfügbar- und Nutzbarkeit von Informationen einsetzt (Haase, 2011, 17). Die Stärke von datenjournalistischer Arbeit liegt demnach, anders als bei den meisten traditionellen journalistischen Tätigkeiten, nicht allein bei dem Veröffentlichen exklusiver Informationen, sondern im verständlich und überprüfbar machen von bereits öffentlich Zugänglichem.

Auch Wissenschaftsjournalistin Eva Wolfangel (2015) betont in ihrem Artikel „Datenjournalismus: Selbst Leuchtturm-Projekte leuchten nicht“, dass Datamining in der journalistischen Arbeit nicht bei bunten Visualisierungen und interaktiven Karten aufhören darf. Auch viele hochgelobte und ausgezeichnete Projekte gehen ihrer Meinung nach momentan nicht über dieses hinaus. Der Grund hierfür ist nach Wolfangel, die noch fehlende Kompetenzenverteilung innerhalb der Redaktionen. Journalinst*innen sollten nicht versuchen zusätzlich gute Programmier*innen zu werden, sondern eine Ahnung davon haben, was mit Data-Mining und explorativer Datenanalyse möglich sei. Sie wollten sich jedoch vor allem auf ihre Expertise, also das saubere Recherchieren und das Finden und Erzählen von Geschichten, konzentrieren. Des Weiteren bräuchte es die enge Zusammenarbeit mit Informatiker*innen, Computerlinguist*innen, Statistiker*innen und Visualisierer*innen, die seriös und professionell den Daten ihre Geheimnisse entlocken können (Wolfangel, 2015). Prof. Emily Bell von der *Columbia School of Journalism* spricht sich sogar für ein eins zu eins Verhältnis zwischen Journalist*innen und Techniker*innen aus (Howard, 2014).

2.3 Open Data

Ein zentrales Element von Datenjournalismus ist der freie Zugang zu Informationen. Als Open-Data, oder *Offene Daten* werden Daten beschrieben, die frei und ohne Einschränkungen genutzt,

verändert und veröffentlicht werden dürfen. Die Open Knowledge Foundation Deutschland (2016) fasst die wichtigsten Eingenschaften von offenen Daten wie folgt zusammen:

> **Verfügbarkeit und Zugang:** Das Werk[1] soll als Ganzes verfügbar sein, zu Kosten, die nicht höher als die Reproduktionskosten sind, vorzugsweise zum gebührenfreien Download im Internet. Das Werk soll ebenso in einer zweckmäßigen und modifizierbaren Form verfügbar sein.
>
> **Wiederverwendung und Nachnutzung:** Die Daten müssen unter denjenigen Bedingungen bereitgestellt werden, die die Wiederverwendung, Nachnutzung und Verbindung mit anderen Datensätzen erlauben. Die Daten müssen maschinell lesbar sein.
>
> **Universelle Beteiligung:** Jede Person muss in der Lage sein, die Daten zu nutzen, wiederzuverwenden und nach zu nutzen. Es darf keine Diskriminierung gegen Handlungsfelder, Personen oder Gruppen vorliegen. Die Nachnutzung darf also nicht auf einzelne Bereiche begrenzt werden (z.B. nur in der Bildung), noch dürfen bestimmte Nutzungsarten (z.B. für kommerzielle Zwecke) ausgeschlossen sein.

Diese Eingenschaften können auf Wissensbestände aus allen Bereichen angewendet werden.

2.3.1 Open Government Data

Für den öffentlichen Sektor sind Informationen und Wissen von besonderer Bedeutung, da Abläufe und Entscheidungen in Politik und Verwaltung meist auf Grundlage von Informationen geplant bzw. gefällt werden.

Open-Government-Data (*Offene Verwaltungsdaten*) beschreibt den Ansatz, Daten des öffentlichen Sektors an Dritte zur Weiterverwendung zur Verfügung zu stellen. Ob die Daten als *offen* bezeichnet werden können, ist aus den, in Kapitel 2.3 aufgeführten, Anforderungen an Open-Data abzuleiten. Lucke und Geiger (2010) definieren in ihrem Gutachten über *"Frei verfügbare Daten des öffentlichen Sektors"* mit Verweis auf die *„Ten Principles for Opening Up Government Information“* der Sunlight Foundation (2010) offene Verwaltungsdaten wie folgt :

> „Offene Verwaltungsdaten sind jene Datenbestände des öffentlichen Sektors, die von Staat und Verwaltung im Interesse der Allgemeinheit ohne jedwede Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterverwendung frei zugänglich gemacht werden.“

[1] Werk ist hier als eine „übertragbare Wissenseinheit“ zu verstehen

Sie schließen explizit alle Informationen aus, „deren Veröffentlichungen nicht im Interesse öffentlicher Belange liegen, die geheim gehalten werden sollen beziehungsweise die personenbezogene Daten sowie Betriebs- und Geschäftsgeheimnisse beinhalten“ (Lucke und Geiger, 2010)

2.3.2 Situation in Deutschland

In Deutschland ist in den letzten Jahren ein steigendes Bewusstsein über die Notwendigkeit der Bereitstellung von offenen Daten sowohl auf bundes- als auch auf komunaler Ebene zu beobachten. Dies lässt sich u.a. daraus schließen, dass im Jahre 2012 von der Bundesregierung des Inneren beim *Frauenhofer-Institut* eine Studie zu *Open Government in Deutschland* in Auftrag gab (Klessmann u. a., 2012). Aus den Empfehlungen dieser Studie wurde das ebenenübergreifende Datenportal *GovData*[2] entwickelt, über das öffentliche Stellen aus Bund, Länder und Kommunen Daten aus der Verwaltung anbieten. Die rechtliche Grundlage regelt in Deutschland auf Bundesebene das *E-Government-Gesetz*, welches das Ziel hat, durch den Abbau bundesgesetzlicher Hindernisse die elektronische Kommunikation mit der Verwaltung zu erleichtern und das *Informationsfreiheitsgesetz (IFG)* welches jeder Person einen Rechtsanspruch auf Zugang zu amtlichen Informationen von Bundesbehörden einräumt. Auf Landesebende gibt es teilweise zusätzliche Gesetze (Siehe Kapitel 2.3.3).

Im *Open-Data-Index-2015* erreicht Deutschland jedoch lediglich Platz 26 und liegt demnach hinter den meisten anderen Industrienationen (Open Knowledge Foundation, 2015).

2.3.3 Hamburgisches Transparenzgesetz

Seid Oktober 2012 ist in der Freien und Hansestadt Hamburg *(FHH)* das *Hamburgische Transparenzgesetz (HmBTG)* in Kraft und löste somit das 2009 novellierte *Hamburgische Informationsfreiheitsgesetz (HmbIFG)* ab. Eine der wichtigsten Änderungen hierbei ist, dass nun eine aktive Veröffentlichungspflicht besteht. Der Zugang zu amtlichen Informationen wird nicht erst auf Anfrage eingeräumt, sondern muss unmittelbar nach der Entstehung erfolgen. Das Ziel des Gesetztes ist es, „über die bestehenden Informationsmöglichkeiten hinaus die demokratische Meinungs- und Willensbildung zu fördern und eine Kontrolle des staatlichen Handelns zu ermöglichen“ (§1 Gesetzeszweck - HmBTG). Veröffentlichungspflichtig sind alle Behörden, Senats- und Bezirksämter der FHH sowie Unternehmen, die öffentliche Aufgaben wahrnehmen und der Kontrolle der FHH unterliegen. Personenbezogene Daten werden in den meisten Fällen aus Datenschutzgründen unkenntlich gemacht (Freie und Hansestadt Hamburg, 2016).

[2] govdata.de

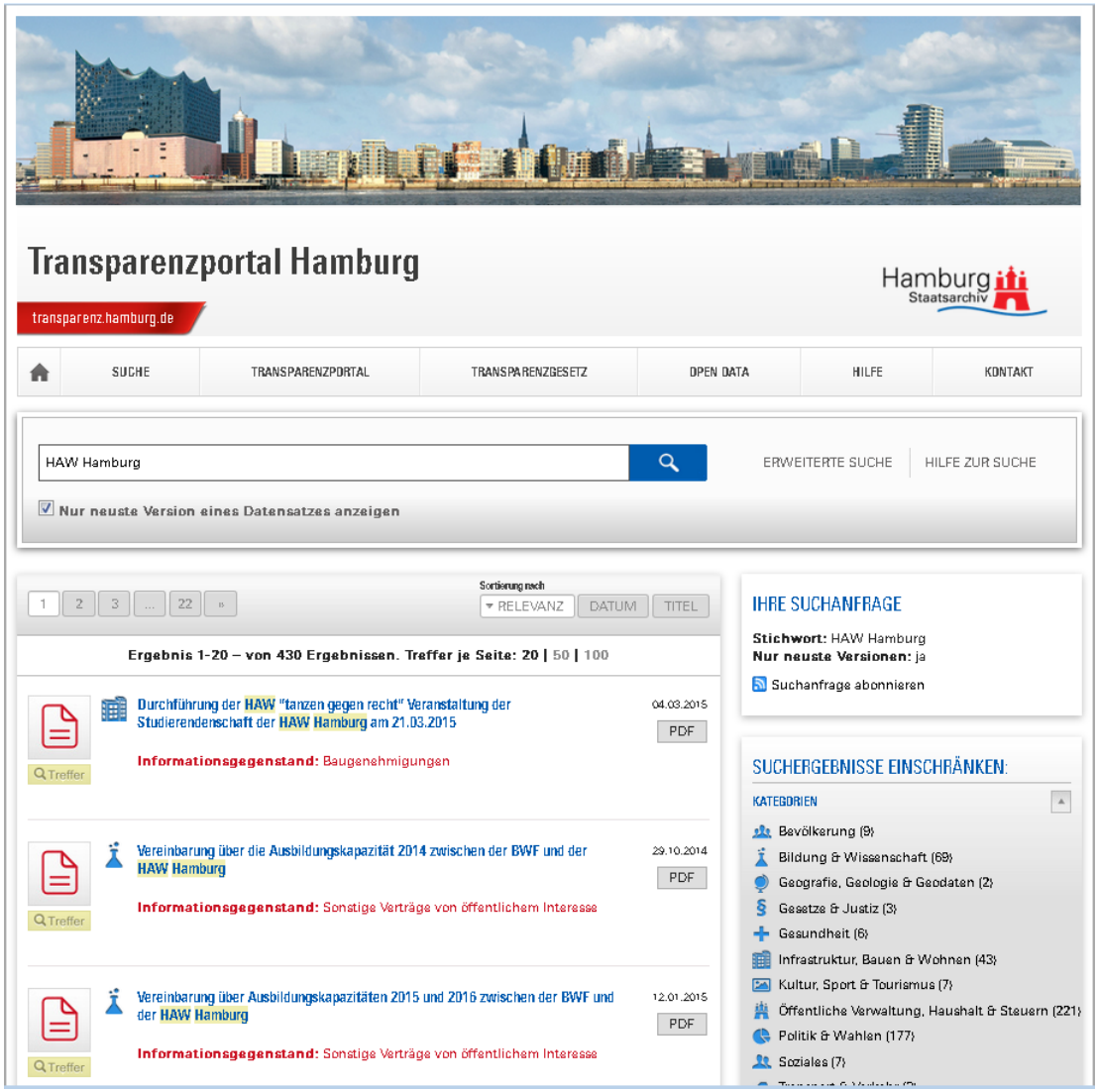


Abbildung 2.1: Transparenzportal - Screenshot

Das Transparenzportal[3] (Abbildung 2.1) ist der zentrale Punkt, in den alle Informationen, die unter das HmBTG fallen, gesammelt und bereitgestellt werden. Aus über 57 unterschiedlichen Liefersystemen werden Dokumente, Geodatenbestände, Mess- und Statistikdaten automatisiert eingelesen. Dokumente und Daten, die nicht bereits in einem dieser Systeme vorliegen, werden durch Sachbearbeiter*innen manuell in das System eingepflegt (Siehe Abbildung 2.2 (S. 11).

[3] http://transparenz.hamburg.de

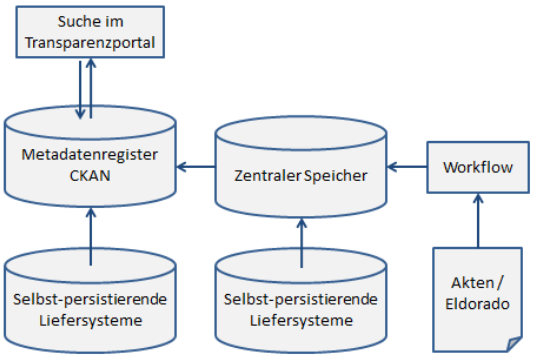


Abbildung 2.2: Tranzparenzportal - Aufbau (Quelle: transparenz.hamburg.de)

Die Metadaten aller Dateien werden von einem *Comprehensive Knowledge Archive Network (CKAN)* ausgelesen *(Harvesting)* und sind zentral per API und Webschnittstelle abruf- und durchsuchbar. CKAN wurde von der OKFN unter einer Open-Source-Lizenz entwickelt und hat sich zum weltweiten de-factor-Standard für Datenkatalogsoftware bei Open-Data durchgesetzt (GovData, 2016).

3 Workflow im Datenjournalismus

Damit Geschichten aus Daten gewonnen werden können, müssen diese zuvor systematisch analysiert und relevante Informationen extrahiert werden. Um diese komplexe Aufgabe erledigen zu können, braucht es einen strukturierten Workflow für Datenjournalist*innen. Aufgaben müssen in Teilschritte gegliedert werden, die nacheinander abgearbeitet werden können. Hierbei kann das, im Data-Mining verbreitete, Prozessmodell des *Knowledge Discovery in Databases (KDD)* zugrunde gelegt werden (Abbildung 3.1). Dieses wird als ein nicht-trivialer Prozess beschrieben, mit dem gültiges, neues und nutzvolles Wissen aus großen Datenmengen extrahiert werden kann. Er schließt nicht nur den eigentlichen Mining-Prozess an sich mit ein, sondern auch das Sammeln, die Datenvorverarbeitung sowie die Interpretation der Ergebnisse (Cleve und Lämmel, 2014).

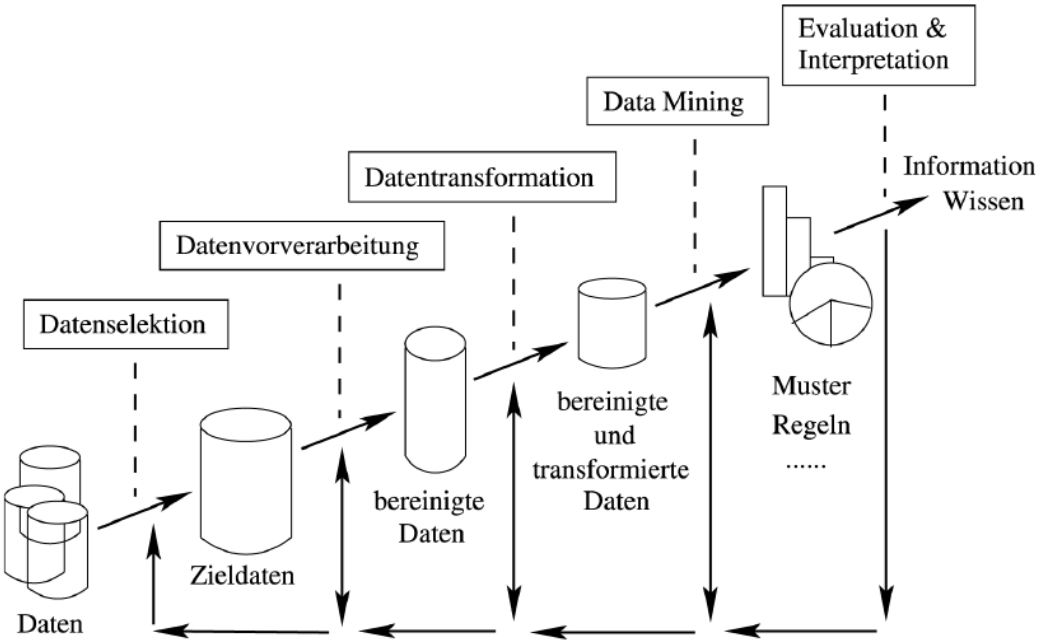


Abbildung 3.1: Klassischer KDD Prozess (Quelle: Cleve und Lämmel (2014))

Da sowohl die Datenbasis, als auch die eigentliche Fragestellung bei datenjournalistischer Arbeit in der Regel eine sehr große Bandbreite annehmen kann, bzw. zu Beginn des Prozesses noch nicht feststeht, muss das Vorgehensmodell an diese Anforderungen angepasst werden. Haase (2011) hat in seiner Arbeit über „Das Potential von Daten und ihr Einfluss auf den Journalismus“ im Jahr 2011 versucht diese Transformationen zu beschreiben. Auch Haase hält sich dabei an die Phasen des KDD Prozesses, geht jedoch nur gering auf die Anforderungen des Mining-Prozesses ein.

3.1 Recherchieren und Fokussieren

Im klassischen KDD-Prozess geht es im ersten Schritt darum, das Ziel der Anwendung zu definieren und ein Verständnis für dieses zu gewinnen. Verfügbare Daten werden gesichtet, bewertet und selektiert, sowie gegebenenfalls fehlende neu erhoben oder generiert (Ester und Sander, 2000).

3.1.1 Story

Der Impuls für die datenjournalistische Arbeit kann hierbei aus zwei Richtungen kommen. Entweder besteht bereits eine konkrete Fragestellung, welche durch das Analysieren von Daten beantwortet werden soll oder es existiert eine große Anzahl an Daten, die auf interessante Muster und Informationen hin untersucht werden soll. Im letzteren Fall entsteht die eigentliche Story oft erst während der Analyse. Auch Mischformen dieser beiden Ansätze sind denkbar. Bei dem Entdecken von Datensätze können Hypothesen über möglichen Mustern entstehen, die das Interesse an weiteren Analysen hervorbringt.

Daran anschließend muss eine Hintergrundrecherche erfolgen, um das Verständnis für die Daten zu steigern und neue potentielle Datenquellen zu finden. Anders als bei Prozessen innerhalb einer Firma ist im Journalismus die Bewertung auf Korrektheit und Unabhängigkeit der Informationsgeber*innen sehr viel wichtiger. Vor Beginn der eigentlichen Datenanalyse sollten die Zielvorgaben für das Projekt definiert und die Erfolgsaussichten mit Blick auf die vorhandenen Ressourcen analysiert werden.

3.2 Datenaggregation

Nach dem das Projektziel definiert wurde, müssen die benötigten Daten gesammelt und in strukturierter Form gespeichert werden. Bezieht sich das Projekt lediglich auf die Analyse einer konkret vorliegenden Datenbase, ist dieser Schritt sehr trivial. Häufig müssen Daten jedoch

erst aus mehreren Quellen extrahiert oder sogar selber erhoben werden. Auch sind verfügbare Informationen in vielen Fällen lediglich in Formaten wie PDF-Dokumenten vorhanden, welche nicht ohne weiteres maschinell ausgewertet werden können, so dass relevante Daten vorab extrahiert und gespeichert werden müssen.

Öffentliche Daten nutzen

Wie in Kapitel 2.3 (S. 8) beschrieben, werden gerade Statistik- und Verwaltungsdaten häufig als offene Daten angeboten und sind direkt über die entsprechende Portale und Schnittstellen durchsuch- und abrufbar. Diese Daten bieten in den meisten Fällen eine sehr hohe Güte und Benutzbarkeit. Nicht bereitgestellte staatliche Informationen können gegebenenfalls über das Informationsfreiheitsgesetz angefragt werden.

Daten selber erheben

Liegen keine verwertbaren Daten vor, können diese gegebenenfalls durch redaktionelle Arbeit selber aggregiert werden. Ein aktuelles Beispiel dafür ist das Datenprojekt „Gewalt gegen Flüchtlinge“ [1] in dem Mitarbeiter*innen von *DIE ZEIT* und *Zeit-Online* durch die eigene Erhebung und Bewertung von ausländer*innenfeindlichen Straftaten, unabhängige Daten erhalten haben, die den offiziellen Berichten des Staates entgegengestellt werden konnten.

Screen-Scraping Eine weitere Möglichkeit der Datenerhebung ist das strukturierte Extrahieren von Informationen aus Webseiten, die keine Programmierschnittstelle zu Verfügung stellen *(screen scraping)*. Dabei wird mit Hilfe eines Computerprogramms, *Wrapper* genannt, eine Internetseite heruntergeladen und anschließend die gewünschten Informationen separiert und gespeichert.

Diese Technik nutzte z.B. die Datenjournalismusagentur *OpenDataCity* im Jahr 2011 um die Verspätungen von Zügen für die Süddeutsche Zeitung auf einer interaktiven Karte darzustellen[2]. Da die Deutsche-Bahn keine strukturierten Daten zu den Verspätungen ausliefert, wurden diese über deren Internetauftritt ausgelesen (Süddeutsche Zeitung, 2012).

[1] http://www.zeit.de/politik/deutschland/2015-11/rechtsextremismus-fluechtlingsunterkuenfte-gewalt-gegen-fluechtlinge-justiz-taeter-urteile

[2] https://opendatacity.de/project/zugmonitor/

3.3 Vorverarbeitung

Ziel der Vorverarbeitung im KDD-Prozess ist es, die gesammelten Daten zu bereinigen und somit die Qualität dieser zu erhöhen. Daten aus unterschiedlichen Quellen müssen integriert, Inkonsistenzen, z.B. durch Messfehler, behoben und für fehlende Werte ein passender Umgang gefunden werden. Die Schwierigkeit hierbei ist, auf der einen Seite Anomalien zu entdecken (z.B. durch Analyse von „Ausreißern“ oder NULL-Werten), auf der Anderen die Wahl eines adäquaten Umgangs, z.B durch Ersetzen eines gemittelten Werts, dem Löschen des Datensatzes, o.ä., mit diesen (Ester und Sander, 2000).

Ein falscher Umgang mit fehlerhaften Daten würde unter Umständen die Berichterstattung verfälschen und die Glaubwürdigkeit des Mediums Datenjournalismus nachhaltig verringern. Wie bei jeder journalistischen Arbeit muss der/die Rezipient*in sich auf das fehlerfreie Arbeiten des Autors/der Autorin verlassen können.

3.4 Transformation

Die vorverarbeiteten Daten werden in dieser Phase in adäquate Repräsentationen umgewandelt. Das Ziel der Transformation hängt stark von dem verwendeten Data-Mining-Algorithmus ab. Manche Verfahren können beispielsweise nur auf nominale Datentypen angewendet werden, sodass metrische Werte in Intervalle zusammengefasst werden müssen *(Diskretisierung)*. Ein weiterer Transformationsschritt beinhaltet die Attribut-Selektion. Im Allgemeinen sind nicht alle Attribute für ein konkretes Verfahren relevant. Eine manuelle Vorauswahl der Attribute kann, bei genügend Anwendungswissen über die Daten sowie den Algorithmus, die Effizienz und Qualität des Algorithmus steigern (Ester und Sander, 2000). Da im journalistischen Bereich die konkrete Fragestellung häufig erst durch die Analyse der Daten entsteht, müssen die Schritte der Transformation, Mustererkennung und Visualisierung in der Regel häufig durchlaufen und auf neue Hypothesen angepasst werden.

3.5 Suche nach Mustern

In Daten können Informationen und Zusammenhänge stecken, die für den Menschen auf den ersten und zweiten Blick nicht erkennbar sind. Eine einfache Visualisierung und Filterung der Datenbestände kann helfen diese zu erkennen, reicht aber in vielen Fällen nicht aus, um neue Informationen zu generieren. Eine automatisierte Analyse kann weitere interessante Erkenntnisse liefern. Da häufig keine genaue Hypothese vorgegeben werden muss, kann eine

solche Analyse nicht nur vermutete Zusammenhänge beweisen, sondern ganz neue Ansätze liefern.

Abhängig von dem Ziel und der Art der Daten muss ein passender Algorithmus gewählt werden. Diese können in die folgenden Themenfelder eingeteilt werden.

Cluster-Analyse

Das Ziel von Klassenbildungsmethoden *(Clustering)* ist es, Informationen in Gruppen, sogenannten *Clustern* von Objekten einzuteilen, dessen Elemente untereinander möglichst ähnlich und zu Elementen der anderen Gruppen möglichst unähnlich sind. Objekte, die keiner spezifischen Gruppen zugeordnet werden können, werden als Ausreißer bezeichnet. In den meisten Fällen wird eine disjunkte Aufteilung der Datenbasis gesucht, d.h. jedes Objekt soll genau einem Cluster zugeordnet werden (*Partitionierung*). Kann ein Objekt auch Teil von mehreren Clustern sein, spricht man von *Fuzzy Clustering* oder *Clumping Methods*. Um die Ähnlichkeiten der Objekte zueinander zu bestimmen, gibt es verschiedene Ähnlichkeits- und Distanzmaße. Die Auswahl dieser ist hauptsächlich von der Skalierung der Merkmale abhängig (Cleve und Lämmel, 2014, 137-141) (Sharafi, 2013, 69-74).

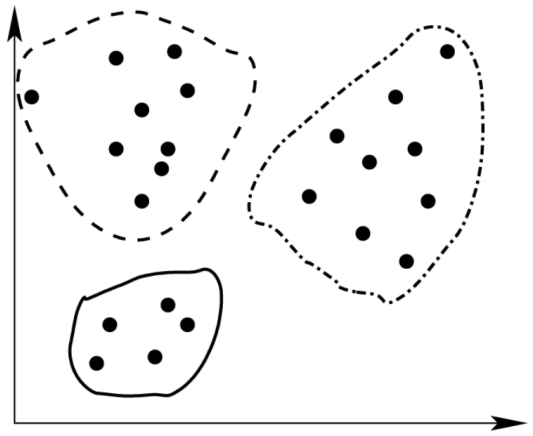


Abbildung 3.2: Clustering-Beispiel (Quelle: Cleve und Lämmel (2014))

Klassifikationsmethode

Das Ziel der Klassifikation besteht darin, Objekte anhand ihrer individuellen Merkmalskombination in Klassen einzuteilen. Anders als bei den Clustering-Verfahren bezieht sich diese

Einteilung nicht auf die gegebenen Werte, sondern auf eine bestimmte Ableitung aus diesen. Man spricht hier vom *überwachten Lernen.*

In der Lernphase werden aus der Datenbasis zufällig einige Objekte und ihre korrekte Klassenzuordnung als Trainingsmenge ausgewählt *(Trainingsdaten).* Anhand dieser Daten wird mittels eines Algorithmus ein Modell erstellt, welches anhand der Merkmalskombination die dazugehörige Klasse bestimmt. Dieses wird als *Klassifikator* bezeichnet. Es ist drauf zu achten, dass sich das Modell nicht zu sehr an die Trainingsdaten anpasst (*overfitting*), sondern flexibel genug bleibt, um auch unbekannte Daten korrekt klassifizieren zu können. Um die Brauchbarkeit des Klassifikators zu bestimmen, wird das Modell anschließend auf Daten angewandt, dessen korrekte Klassifikation bekannt ist, die aber nicht zum Trainieren des Modells benutzt wurden (*Testdaten*). Durch den Vergleich der prognostizierten und der realen Klassen kann so die Fehlerquote bestimmt werden.

Es wird unter anderem zwischen den multivariaten statischen Verfahren, den künstlichen neuronalen Netzen sowie den Entscheidungsbaumverfahren unterschieden (Cleve und Lämmel, 2014, 59-61) (Sharafi, 2013, 76-79).

Assoziationsanalyse

Bei der Assoziationsanalyse werden Beziehungen und Assoziationen zwischen Ausprägungen von Variablen gesucht. Diese wird auch als *Warenkorbanalyse (recommender engines)* bezeichnet, da sie häufig genutzt wird, um Kaufentscheidungen vorherzusagen. Die Beziehung zwischen zwei oder mehreren Items kann mittels Regeln in Form von *„Wer Produkt **A** kauft, kauft häufig auch Produkt **B**“* bzw. „$A \rightarrow B$“ dargestellt werden. Assoziationen sind dabei nicht immer korrekt, sondern zeigen lediglich eine hohe Wahrscheinlichkeit der Abhängigkeit auf (Cleve und Lämmel, 2014, 63-65) (Sharafi, 2013, 74-76).

Text-und Web-Mining

Text-Mining bezeichnet das Analysieren von un- oder schwachstrukturierten Textdaten. Das Ziel ist es, interessante Informationen herauszufiltern, um diese dann strukturiert weiter verarbeiten zu können. Hierzu werden zunächst irrelevante Wörter (*stop words*) herausgefiltert. Anschließend können Wörter auf ihren Wortstamm reduziert (*stemming*) und semantisch gleiche Begriffe zusammengefasst werden. Auf diese reduzierte Wortmenge (*bag of words*) kann nun die eigentliche Auswertung stattfinden, um so z.B. Verteilungen-, Häufigkeits-, oder Abhängigkeitsanalysen zu erstellen (Cleve und Lämmel, 2014, 65-67) (Sharafi, 2013, 79-93).

Durch *Sentiment Detection*, also der Stimmungserkennung, wird versucht eine positive oder negative Haltung in natürlichsprachlichen Texten zu erkennen.

Während des US-Wahlkampfes im Jahr 2012 hat der Mikroblogging-Dienst *Twitter* beispielsweise täglich mehrere hunderttausende „Tweets“ analysiert und nach positiven bzw. negativen Äußerungen zu den Kandidaten bewertet. So konnten sie eigenen Prognose über den Wahlkampf erstellen (Sharp, 2012). Die kalifornische Politikwissenschaftlerin Jennifer Ramos sagte im Jahr 2012 in einem Interview gegenüber *Zeit Online* : „Es scheint so, dass Twitter ein verlässliches Prognose-Instrument für die Präsidentschaftswahl gewesen ist.“(ZEIT ONLINE GmbH, 2012).

Ein weiteres Beispiel für Text-Mining mit journalistischem Hintergrund ist die Volltext Visualisierung der auf WikiLeaks veröffentlichten Geheimdokumente zum Irakkrieg (*Iraq War Logs*) von Jonathan Stray und Julia Burgess. Sie generierten ein Netz aus Schlüsselwörtern und deren Beziehungen untereinander (Stray, 2010). Siehe Abbildung: 3.3 (S. 19)

3.6 Evaluation und Visualisierung

In der letzten Phase müssen die Resultate, sprich die gefundenen Muster, dargestellt und beurteilt werden. Daten für sich sind nur Bits und Bytes und für eine Menschen nicht greifbar. Daher ist es wichtig, diese aufzuarbeiten, um sie verstehen zu können. Bereits eine Darstellung in Text- oder Tabellenform ist eine Form der Visualisierung. Es stellt sich also nicht die Frage, *ob* es sinnvoll ist die Daten zu visualisieren, sondern welche Art der Visualisierung die nützlichste ist. Dieses hängt stark von der Beschaffenheit der Daten ab. Eine Darstellung über eine einfache Tabelle hinaus ist jedoch in den meisten Fällen empfehlenswert, da es schwer ist nur mit ihnen einen kompletten Überblick zu gewinnen oder Muster zu erkennen (Aisch, 2016).

Um ein tiefes Verständnis für die vorliegenden Daten und deren Muster zu bekommen, ist eine ständige Iteration von Visualisierung, Analyse, Dokumentation und Anpassung notwendig (Abbildung: 3.4 (S. 20)). Anpassungen können dabei in jedem Abschnitt des beschreibenden Workflows notwendig sein.

3.7 Fazit

Der KDD Prozess scheint eine gute Grundlage für die Erstellung datengetriebener Erzählstrukturen zu bieten. Er muss jedoch in einigen Punkten angepasst oder erweitert werden. Die Recherche und Datenerhebung sind im journalistischen Kontext in der Regel aufwendiger als innerhalb einer Data-Warehouse Umgebung, da Ziel und Datenbasis häufig zu Beginn noch

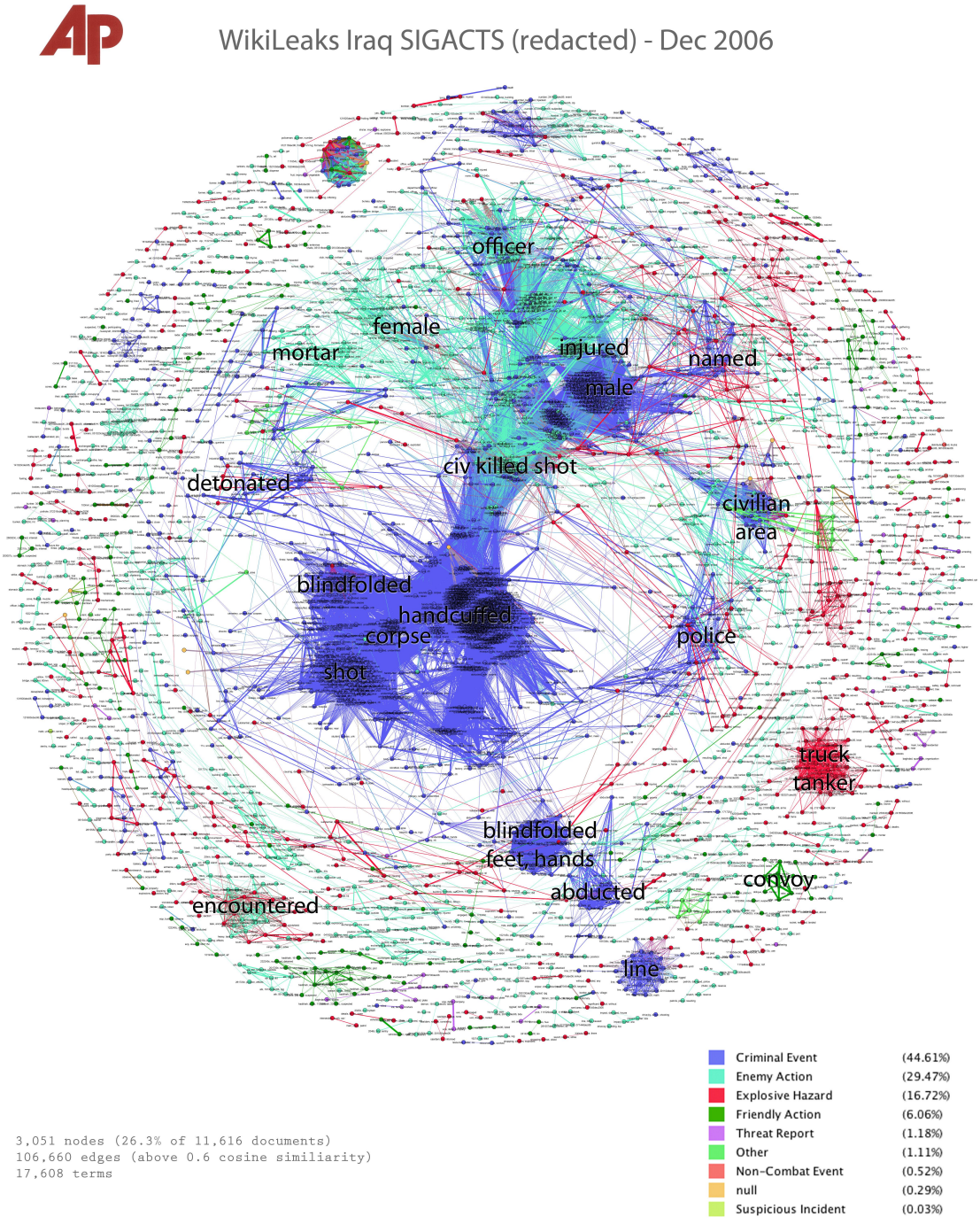


Abbildung 3.3: Visualisierung der Irakkrieg Dokumente (Quelle: Stray (2010))

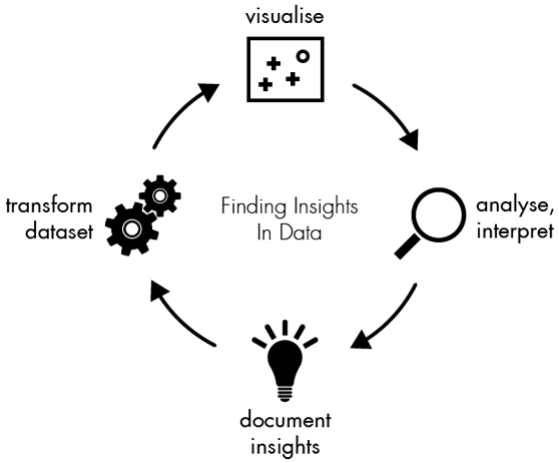


Abbildung 3.4: Verständnis über Daten (Quelle: Aisch (2016))

nicht vollständig definiert sind. Im folgenden Kapitel 4 werden die einzelnen Schritte anhand eines Beispiels weiter validiert.

4 Prognosemodell zur Parkhausauslastung

Um den beschriebenen Workflow zu validieren, werden die einzelnen Schritte in diesem Kapitel anhand eines Beispiels dargestellt und die Erfahrungen beim Erstellen einer datengetriebenen Geschichte aus informationstechnologischer Sicht dokumentiert. Auch wenn die Grundidee in Zusammenarbeit mit Journalist*innen entstanden ist, ist diese nur als Beispiel zur Verdeutlichung der technischen Umsetzung zu verstehen. Journalistische Kontexte werden lediglich beiläufig angeschnitten und diskutiert.

4.1 Die Idee

Am Anfang stand die Idee der Entwicklung eines Modells zur Analyse der Auslastung der hamburger Parkhäuser. Dieses soll den Nutzer*innen einen besseren Überblick über die aktuelle Parksituation verschaffen und so das zeitaufwendige Suchen nach freien Parkmöglichkeiten verringern. Des weiteren könnte es interessant sein, die Parkhausauslastung mit weiteren Informationen zu verknüpfen, um so Abhängigkeiten zu entdecken und Prognosemodelle zu entwickeln.

Die Ergebnisse könnten einerseits dem/der Nutzer*in auf einer interaktiven Applikation zur Verfügung gestellt werden, in der die aktuelle und die prognostizierte Auslastung abgerufen werden können. Für besondere Ereignisse, wie z.B. Großveranstaltungen, verkaufsoffene Sonntage, „dem ersten schönen Frühlingswochenende“, o.ä. könnten eigene Geschichten, z.B. mit Empfehlungen zur Parkhausnutzung entstehen.

Es ist auch denkbar, dass sich im Laufe der Analyse Thesen entwickeln, denen dann weiter nachgegangen werden können und so Geschichten entstehen, die im Vorfeld nicht abzusehen sind. Etwa, dass die Parksituation in bestimmten Bereichen generell sehr überlastet ist, sich ein neu gebautes Parkhaus aufgrund der niedrigen Belegung als Fehlinvestition herausstellt oder interessante Korrelationen zwischen Parkverhalten und externen Faktoren existieren, die vorher nicht bekannt waren.

Folgende Faktoren könnten die Belegung der Parkhäuser beeinflussen:

- Ort des Parkhauses

- Uhrzeit
- Wochentag
- Wetter
- Veranstaltungen (Dom, Hafengeburtstag, Fußballspiele, Messen,...)

4.2 Recherche und Selektion

In diesem Kapitel werden die verschieden Datenquellen analysiert.

4.2.1 Parkhausdaten

Über das Hamburger Transparenzportal lassen sich Daten über Parkhäuser im Stadtgebiet Hamburg aufrufen. Sie geben Information über

- Lage
- Öffnungszeiten
- Preise
- Stellplätze sowie tlw. Belegungsdaten (freie Stellplätze)
- tlw. Frauen- und Behindertenstellplätze sowie Einfahrtshöhe

(BWVI, 2015)

Die Informationen liegen in strukturierter Form vor und werden als GML[1] oder Textdatei ausgeliefert. Dabei werden die Daten nicht direkt im Portal selber vorgehalten, wie ich in einem Gespräch mit Dr. Lothar Hotz [2] (Universität Hamburg, Fachbereich Informatik und Mitglied des HITeC e.V.[3]) erfuhr. Im Portal wird lediglich ein Link zu den Informationen bereitgestellt, mit Metadaten angereichert und im Suchkatalog auffindbar gemacht.

Die einzelnen Attribute können in zwei Kategorien aufgeteilt werden:

Statische Attribute Dies sind beschreibende Informationen eines Parkhauses. Sie bleiben weitgehend stabil und werden nur alle zwei Jahre aktualisiert. Diese Veränderung wird im Rahmen dieser Arbeit vernachlässigt.

Variable Attribute Hierbei handelt es sich um Neartime-Daten, welche im 10 Minutentakt aktualisiert werden.

Zu beachten ist, dass nur in den seltensten Fällen alle Attribute geführt werden.

[1] Geography Markup Language
[2] http://kogs-www.informatik.uni-hamburg.de/~hotz/
[3] Hamburger Informatik Technologie-Center e.V. - hitec-hh.de

```
−<gml:featureMember>
 −<app:verkehr_parkhaeuser gml:id="APP_VERKEHR_PARKHAEUSER_10001">
    <app:id>10001</app:id>
    <app:name>Alsterhaus</app:name>
    <app:art>Parkhaus</app:art>
    <app:strasse>Jungfernstieg</app:strasse>
    <app:hausnr>16-20</app:hausnr>
    <app:einfahrt>Bei der Stadtwassermühle</app:einfahrt>
   −<app:preise>
      3,00 EUR/h|Tageshöchstsatz 30,00 EUR|Kundentarif (Alsterhaus)|Nachttarif
    </app:preise>
    <app:oeffnungszeit>täglich 24 Stunden</app:oeffnungszeit>
    <app:stellplaetze_gesamt>100</app:stellplaetze_gesamt>
    <app:bemerkung_verkehrsportal/>
    <app:frei>62</app:frei>
    <app:gesamt>100</app:gesamt>
    <app:auslastung>38</app:auslastung>
    <app:status>frei</app:status>
    <app:situation>frei</app:situation>
    <app:received>15.11.2015, 13:15</app:received>
    <app:pur>nein</app:pur>
    <app:punkt>multi</app:punkt>
    <app:datenherkunft>BWVI_V</app:datenherkunft>
   −<app:position>
     −<!--
        Inlined geometry 'APP_VERKEHR_PARKHAEUSER_10001_APP_POSITION'
      -->
     −<gml:Point gml:id="APP_VERKEHR_PARKHAEUSER_10001_APP_POSITION" srsName="EPSG:25832">
        <gml:pos>565703.765 5934210.671</gml:pos>
      </gml:Point>
    </app:position>
  </app:verkehr_parkhaeuser>
 </gml:featureMember>
```

Abbildung 4.1: Beispiel Eintrag eines Parkhauses (GML)

Name	Datentyp
id	Integer
name	String
art	String
strasse	String
hausnr	String
einfahrt	String
preis	String
oeffnungszeit	String
stellplaetze_gesamt	Integer
frauenst	Integer
behindertenst	Integer
einfahrtshoehe_in_meter	Decimal
bemerkung_verkehrsportal	String
gesamt	Integer
pur	String
punkt	String
datenherkunft	String
position	GeometryPropertyType

Tabelle 4.1: Statische Attribute der Parkhausdaten

Name	Datentyp
frei	Integer
auslastung	Decimal
status	String
received	String

Tabelle 4.2: Variable Attribute der Parkhausdaten

4.2.2 Wetterdaten

Es gibt eine Vielzahl meteorologischer Dienstleister im Internet. Für die Datenweiterverarbeitung ist ein eingeschränkter Wertebereich, z.B. durch Nutzung von Taxonomie oder Bildung von Wertintervallen, hilfreich (Cleve und Lämmel, 2014, S. 10). Aus diesem Grund ist der Anspruch an Granularität, Vollständig- und Genauigkeit der Wetterdaten eher gering. Weiter noch erleichtern bereits abstrahierte und zusammengefasste Werte die Datentransformation. Weiterer Ansprüche an den Dienst:

- stabile und einfache API [4]

[4] Application Programming Interface

- Ausgabe von strukturierten Daten (JSON[5], XML[6]) ohne viel Overhead
- Aktuelle Wetterdaten für die Stadt Hamburg
- Wettervorhersage (min. 1 Tag) für die Stadt Hamburg
- kostenfreie Nutzung
- möglichst offene Lizenz

Der Online-Dienst Openweathermap[7] erfüllt genau diese Anforderungen. Über einen einfachen API Aufruf via HTTP-Request lassen sich aktuelle, sowie prognostizierte Wetterinformationen für eine Stadt oder ein Gebiet abrufen und in XML, JSON oder HTML Format zurück geben (Beispielobjekte sieh Abbildungen 4.2 (S. 26) und 4.3 (S. 27)).

Der kostenlose Zugriff unterliegt zwar ein paar Einschränkungen (max. 60 Calls/minute, Prognosen nur für 5 Tage in die Zukunft, kein UV-Index, u.a.) erfüllt aber die Anforderungen in vollem Umfang. Die Daten von Openweathermap werden unter der „Creativ Commons Attribution-ShareAlike 2.0 Generic“ *(CC BY-SA 2.0)*[8] bereitgestellt, welche jegliches Verbreiten und Bearbeiten, auch für kommerzielle Zwecke, erlaubt. Lediglich eine Namensnennung und die Weitergabe unter gleichen Bedingungen/Lizenzen wird eingefordert. Die gelieferten Informationen können somit als *Open-Data* bezeichnet werden (Siehe Kapitel 2.3 (S. 8)).

4.2.3 Veranstaltungen

Großveranstaltungen mit vielen Besucher*innen wirken sich vermutlich punktuell stark auf die Frequentierung der Parkhäuser aus. Zudem wäre gerade eine Prognose der Auslastung bei bevorstehenden Veranstaltungen besonders interessant. Aus diesen Gründen wäre es wünschenswert diese in der Datenerhebung mit aufzunehmen und in der Auswertung zu berücksichtigen.

Veranstaltungshinweise könnten zum Beispiel auf der offiziellen Seite der Stadt Hamburg [9], der Hamburg Tourismus GmbH [10] oder verschiedener Tageszeitungen [11] [12] per *Screen Scraping* ausgelesen werden. Des Weiteren gibt es kostenpflichtige Anbieter, die Datenbanken mit

[5] JavaScript Object Notation
[6] Extensible Markup Language
[7] https://openweathermap.org
[8] https://creativecommons.org/licenses/by-sa/2.0/
[9] https://www.hamburg.de/veranstaltungen/
[10] http://www.hamburg-tourism.de
[11] http://veranstaltungen.abendblatt.de/hamburg/
[12] http://termine.mopo.de/

```
{
  ▾ coord: {
        lon: 10,
        lat: 53.55
    },
  ▾ weather: [
      ▾ {
            id: 500,
            main: "Rain",
            description: "light rain",
            icon: "10d"
        }
    ],
    base: "cmc stations",
  ▾ main: {
        temp: 4.39,
        pressure: 1008.7,
        humidity: 96,
        temp_min: 4.39,
        temp_max: 4.39,
        sea_level: 1013.3,
        grnd_level: 1008.7
    },
  ▾ wind: {
        speed: 7.32,
        deg: 267.001
    },
  ▾ rain: {
        "3h": 0.995
    },
  ▾ clouds: {
        all: 92
    },
    dt: 1455197433,
  ▾ sys: {
        message: 0.0034,
        country: "DE",
        sunrise: 1455173221,
        sunset: 1455207735
    },
    id: 2911298,
    name: "Hamburg",
    cod: 200
}
```


Abbildung 4.2: JSON-Objekt des aktuellen Wetters in Hamburg von openweahtermap.com

```
{
  ▾ city: {
        id: 2911298,
        name: "Hamburg",
      ▾ coord: {
            lon: 10,
            lat: 53.549999
        },
        country: "DE",
        population: 0,
      ▾ sys: {
            population: 0
        }
    },
    cod: "200",
    message: 0.0038,
    cnt: 40,
  ▾ list: [
      ▾ {
            dt: 1455202800,
          ▾ main: {
                temp: 4.39,
                temp_min: 4.39,
                temp_max: 4.39,
                pressure: 1008.7,
                sea_level: 1013.3,
                grnd_level: 1008.7,
                humidity: 96,
                temp_kf: 0
            },
          ▾ weather: [
              ▾ {
                    id: 500,
                    main: "Rain",
                    description: "light rain",
                    icon: "10d"
                }
            ],
          ▾ clouds: {
                all: 92
            },
          ▾ wind: {
                speed: 7.32,
                deg: 267.001
            },
          ▾ rain: {
                "3h": 0.995
            },
          ▾ sys: {
                pod: "d"
            },
            dt_txt: "2016-02-11 15:00:00"
        },
      ▾ {
            dt: 1455213600,
          ▾ main: {
                temp: 3.61,
                temp_min: 3.61,
                temp_max: 3.61,
                pressure: 1009.91,
                sea_level: 1014.54,
                grnd_level: 1009.91,
                humidity: 98,
                temp_kf: 0
            },
          ▾ weather: [
```


Abbildung 4.3: JSON-Objekt der Wetterprognose für Hamburg von openweahtermap.com (Auszug)

entsprechenden Schnittstellen zu Verfügung stellen, wie etwa die *Hamburg Tourismus GmbH* [13] oder das *OpenEventNetwork*[14]

Das automatische Erheben, zumindest eines Teils der Veranstaltungen, ist also technisch möglich. Schwierig wird jedoch die Interpretation und Abstraktion der konkreten Ereignisse. Würden alle Veranstaltungen gleich bewertet, um sie lediglich auf die Anzahl der gleichzeitig stattfinden Angebote zu reduzieren, würde dies zu Fehlinterpretationen führen. So würden zum Beispiel zwei Lesungen mit jeweils 15 Besucher*innen höher klassifiziert werden, als ein Großevent wie der Hafengeburtstag mit 1,5 Millionen Menschen. Wichtige beschreibende Attribute einer Veranstaltung wären demnach:

- Ort
- Dauer (Datum und Uhrzeit)
- Erwartete Besucher*innenzahl

Da es gerade über die Zahl der erwarteten Besucher*innen keine zugänglichen und auswertbaren Informationen gibt, ist eine automatisierte Interpretation von Veranstaltungen nicht möglich. Es wäre jedoch denkbar, die Informationen manuell nachzupflegen, um auch dieses Kriterium abdecken zu können.

4.3 Datenaggregation

Die in Kapitel 4.2 beschriebenen Datenquellen müssen nun analysiert und in einer einheitlichen, strukturierten Form für die Weiterverarbeitung gespeichert werden.

4.3.1 Parkhausdaten

Die Informationen zu den Parkhäusern werden über das Transparenzportal der Stadt Hamburg bereitgestellt. Es besteht jedoch kein Zugang zu historischen Daten, so dass diese in regelmäßigen Abständen abgerufen und in einer lokalen Datenbank gesammelt werden müssen. Hierzu wurde ein Java-Skript entwickelt, welches die Information abruft, und die zurückgelieferte XML-Datei in ein DOM-Object Trees umwandelt. Aus dieser Struktur können dann die Elemente *(Nodes)* mit den relevanten Informationen extrahiert und dauerhaft gespeichert werden. Da es sich, wie in Abschnitt 4.2.1 (S. 22) beschrieben, nur bei einem kleinen Teil um variable Attribute handelt, beschränkt sich die Auswahl der Nodes auf Grund der Datensparsamkeit nur auf diese. Die statischen Attribute werden lediglich einmalig erfasst und gespeichert.

[13] http://www.hamburg-tourism.de/service/veranstaltungsdatenbank/
[14] http://www.openeventnetwork.de/

4.3.2 Wetterdaten

Auch die Wetterinformationen müssen regelmäßig abgerufen und gespeichert werden. *Openweathermap* liefert einen relativ detaillierten Einblick in die aktuellen Wettermessungen. Ein Objekt umfasst Angaben zu Temperaturen, Luftdruck, Windeigenschaften, Wolkenstatus, Sonnenaufgang und vielem mehr (Siehe Abbildung 4.2 (S. 26)). Eine so granulare Auflistung ist für dieses Projekt nicht notwendig und würde möglicherweise die Mustersuche sogar erschweren. Daher wird bereits an dieser Stelle eine Selektion der zu Verfügung stehenden Informationen vorgenommen.

Als Abstraktion der Wettersituation bietet sich die Icon-ID an. In diesem werden die Wetterbedingungen bereits vom Anbieter in neun Kategorien unterteilt (Siehe Tabelle 4.3). Zu beachten ist jedoch, dass für Tages- und Nachtwerte jeweils ein unterschiedlicher Identifikator angehängt wird (*d* für Tag *(Day)*, *n* für Nacht *(Night)*). Um eine Vergleichbarkeit herzustellen müssen die Werte von dieser Information gesäubert werden.

Icon-ID Tag	Icon-ID Nacht	Beschreibung
01d	01n	clear sky
02d	02n	few clouds
03d	03n	few clouds
04d	04n	broken clouds
09d	09n	shower rain
10d	10n	rain
11d	11n	thunderstorm
13d	13n	snow
50d	50n	mist

Tabelle 4.3: Beschreibung der Icon-IDs von openwether.org

Um noch ein genaueres Bild zu erhalten, wird zusätzlich die Temperatur in Grad Celsius gespeichert. Mit diesen beiden Informationen wird vermutlich ein hinreichend präzises Bild der Wettersituation entstehen.

4.3.3 Veranstaltungen

Da, wie unter 4.2.3 (S. 25) beschrieben, Veranstaltungen zwar vermutlich von Relevanz sein werden, jedoch nur schwer automatisch eingelesen werden können, werden diese in dieser Arbeit lediglich in das Modell übernommen. In der konkreten Auswertung werden jedoch keine Veranstaltungen berücksichtigt.

- Startzeitpunkt *(date_from)*

- Endzeitpunkt *(date_to)*
- Veranstaltungsart *(description)*

4.3.4 Datenbank

Alle Informationen werden in einer relationalen Datenbank gespeichert. In dieser Arbeit wurde das seit 2010 von Oracle weiterentwickelte (vorher von *MySQL AB* und *Sun Microsystems*) Datenbankverwaltungssystem *MySQL* benutzt, da dieses unter einer freien Lizenz (*GPL*[15]) steht und auf einem Linux-Server betrieben werden kann. Weitere Anforderung wurden an die Datenbanklösung nicht gestellt. Um Redundanzen in der Datenhaltung zu vermeiden, wurde die Datenbank in mehren Tabellen aufgeteilt, wie in Abbildung 4.4 zusehen ist.

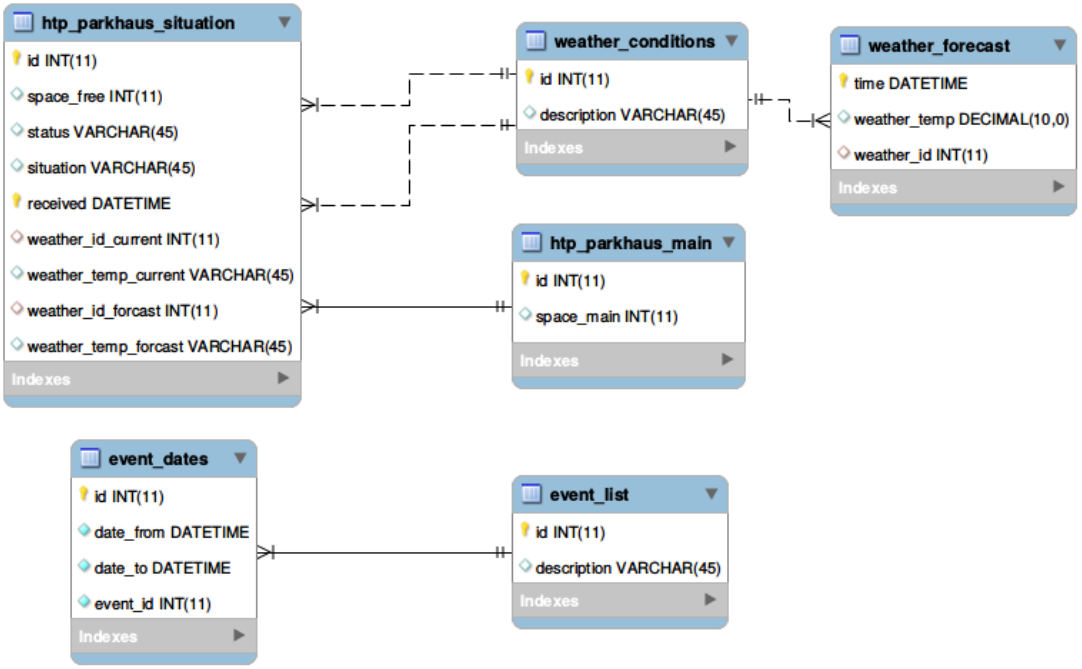


Abbildung 4.4: ER-Diagramm

4.4 Vorverarbeitung

Da unvollständige oder fehlerhafte Daten zu Fehlern im Data-Mining-Prozess führen könnten, ist es wichtig diese im Vorfeld zu identifizieren und zu beheben.

[15] GNU General Public License

4.4.1 Parkhausdaten

Wie bereits unter 4.3.1 (S. 28) aufgeführt, sind für das Mining-Verfahren lediglich die variablen Attribute relevant. Wobei der Wert *auslastung* von der Anzahl der freien Parkplätze (*frei*) und dem Öffnungsstatus des Parkhauses (*status*) abhängig ist. Somit bringt dieses Attribut keinen eigenen Mehrwert und kann vernachlässigt werden. Die Variable *status* nimmt entweder den Wert „frei“ oder „geschlossen“ an. Für eine prognostische Analyse sind nur die zugänglichen Parkhäuser relevant. Daher werden nur diese Parkhausdaten als gültige Werte betrachtet. Für andere Analysen, beispielsweise über die Häufigkeit von gesperrten Parkhäusern, müsste der Wertebereich hier anders definiert werden.

Gültige Wertebereiche können wie folgt definiert werden:

$$status == \text{“}frei\text{”} \tag{4.1}$$

$$frei \leq gesamt \tag{4.2}$$

$$received \leq now() \tag{4.3}$$

Durch die Definition aller Datensätze mit dem Status „geschlossenen“ als fehlerhaft (4.1), entsteht eine relativ hohe Fehlerquote von ca. 12%. Da die anderen Attribute dieser Datensätze jedoch auch keine relevanten Informationen beinhalten (ist ein Parkhaus geschlossen sind in der Regel auch alle Parkplätze frei), können diese Datensätze ohne Informationsverlust komplett ignoriert werden. Anforderung 4.2 und 4.3 werden in der aktuellen Datenbasis nie verletzt. Aus diesem Grund ist hier keine aufwendige Fehlerbehebung notwendig. Auch diese Datensätze werden in der weiteren Analyse ignoriert.

4.4.2 Wetterdaten

Bei den gesammelten Wetterdaten handelt es sich um analysierte und aufbereitete Informationen des Online-Dienstes *openweather.net*. Durch die begrenzten Ressourcen, bedingt durch den Rahmen dieser Arbeit, werden diese Daten als korrekt und vollständig betrachtet. Eine weitere Analyse dahingehend ist in Anbetracht des Umfangs der Arbeit nicht möglich.

4.4.3 Veranstaltungsdaten

Wie in 4.3.3 (S. 29) beschrieben, können Informationen zu Veranstaltungen nicht automatisiert ausgelesen werden, sondern müssten manuell eingetragen. Die Einträge unterliegen nur wenigen formalen Regeln. Es besteht keine Abhängigkeit der einzelnen Einträge untereinander, so

können auch keine Ausreißer o.ä. definiert werden. Lediglich für das Startdatum (*date_from*) und Enddatum (*date_to*) kann ein Wertebereich wie folgt definiert werden:

$$date\_from \leq date\_to \tag{4.4}$$

4.5 Transformation

In Kapitel 3.4 (S. 15) wurde bereits darauf eingegangen, dass die Art der Transformation stark von dem später eingesetzten Mining-Verfahren abhängt. Für die Analyse und Prognose der Parkhausauslastung wird hier ein Klassifikations-Algorithmus gewählt. Eine Assoziationsanalyse wäre aus dem Grund nicht sinnvoll, da nicht nach Zusammenhängen beliebiger Attribute gesucht werden soll (wie beispielsweise der zwischen Temperatur und Uhrzeit), sondern das Zielattribut, der Auslastungswert der Parkhäuser, bereits feststeht.

Als Datenbasis stehen die in der Tabelle 4.4 (S. 32) aufgelisteten Attribute zu Verfügung. Es handelt sich um nominale, ordinale, sowie metrische (intervall und absolut) Skalentypen.

Attribut	**Skalentyp**	**Beschreibung**
id	nominal	Identifikator des Parkhauses
space_free	metrisch, Absolutskala	Anzahl der freien Parkplätze
status	ordinal	Öffnungssituation
received	metrisch, Intervallskala	Zeitpunkt der Datenerhebung
space_main	metrisch, Absolutskala	Anzahl der Parkplätze
weather_id_current	nominal	Aktuelle Wettersituation
weather_temp_current	metrisch, Intervallskala	Aktuelle Temperatur
weather_id_forcast	nominal	Vorhergesagte Wettersituation
weather_temp_forcast	metrisch, Intervallskala	Vorhergesagte Temperatur
event_1	ordinal	Veranstaltungsgruppe 1
event_2	ordinal	Veranstaltungsgruppe 2

Tabelle 4.4: Datenbasis für die Analyse

4.5.1 Nominale Transformation

Für einige Verfahren der Klassifizierung dürfen die Merkmale nur nominale Ausprägungen besitzen. Vor diesem Hintergrund müssen alle in der Vorverarbeitung als metrische definierten Werte in nominale skalierende Merkmale umgewandelt werden. Dazu werden in den meisten Fällen Kategorien gebildet, in denen jede Untersuchungseinheit genau eingeteilt werden kann (Tabelle 4.5 (S. 33)). Aus dem Timestamp (*received*) werden hierbei zwei unterschiedliche

Merkmale generiert. Zum einen die Tageszeit (*Time*) und zum anderen der Wochentag als Nominalzahl (*Weekday*). Das Merkmal, dass für die Kategorisierung der Auslastung eines Parkhauses (*Usage*) genutzt wird, wird im Vorfeld aus der Datenbasis generiert ($Usage = 1 - \frac{space\_free}{space\_main}$).

Attribut	**Zeichenkette nominal**	**Intervall**
Usage	empty	0 - 25% belegt
	medium	25 - 50% belegt
	busy	50 - 75% belegt
	full	75 - 99,9% belegt
	closed	100% belegt
Forecast Temperature	verycold	< (-39) °C
	cold	(-39) - (-26) °C
	cooly	(-26) - (-13) °C
	minorycooly	(-13) - 0 °C
	comfortable	0 - 20 °C
	minorywarm	20 - 26 °C
	warm	26 - 32 °C
	hot	32 - 38 °C
	veryhot	> 38 °C
Current Temperature	verycold	< (-39) °C
	cold	(-39) - (-26) °C
	cooly	(-26) - (-13) °C
	minorycooly	(-13) - 0 °C
	comfortable	0 - 20 °C
	minorywarm	20 - 26 °C
	warm	26 - 32 °C
	hot	32 - 38 °C
	veryhot	> 38 °C
Weekday	1	Sonntag
	2	Montag
	3	Dienstag
	4	Mittwoch
	5	Donnerstag
	6	Freitag
	7	Samstag
Time	night	00 - 06 Uhr
	morning	06 - 12 Uhr
	afternoon	12 - 18 Uhr
	evening	18 - 24 Uhr

Tabelle 4.5: Transformation von metrischen zu nominalen Merkmalen

4.5.2 Merkmalanalyse

Eine Analyse über den Einflussfaktor der einzelnen Merkmale auf das Klassenattribut findet im besten Fall vor der Auswahl der Merkmale statt. Da in diesem Fall die Informationen erst gesammelt werden mussten, bevor sie untersucht werden konnten, findet der Schritt erst an dieser Stelle statt.

Unterschiede der einzelnen Parkhäuser

Betrachtet man den durchschnittlichen Füllungsgrad fällt auf, dass dieser zwischen den einzelnen Parkhäusern stark variiert (Abbildung 4.5 (S. 34)).

ID 10036 (Reeperbahn) ist im Durchschnitt zu 78% belegt, ID 10098 (Harburg) nur zu ca. 5%. ID 10018 ist auf Grund der geringen Datenmenge von nur 10 Datensätzen für eine Auswertung nicht repräsentativ.

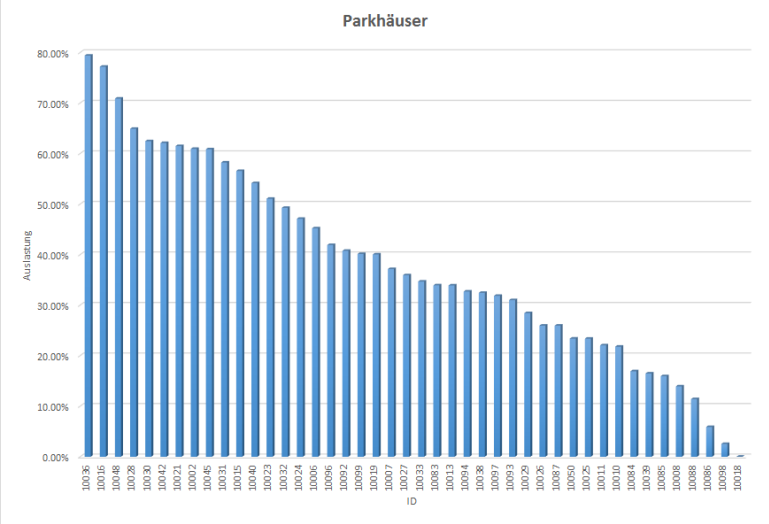


Abbildung 4.5: Durchschnittliche Auslastung der Parkhäuser

Wochentage

Der Einfluss der Wochentage hat, mit Ausnahme von Sonntag (1), auf den ersten Blick einen geringeren Einfluss als angenommen (Siehe Abbildung 4.6 (S. 35)).

Über alle Datensätze hinweg scheint es keinen signifikanten Unterschied der Auslastung zwischen Arbeitstagen (2-6) und Samstag (7) zu geben. Ein genauerer Blick in die Datenbasis zeigt jedoch, dass dies sehr abhängig von den einzelnen Parkhäusern ist. In der Einzelbe-

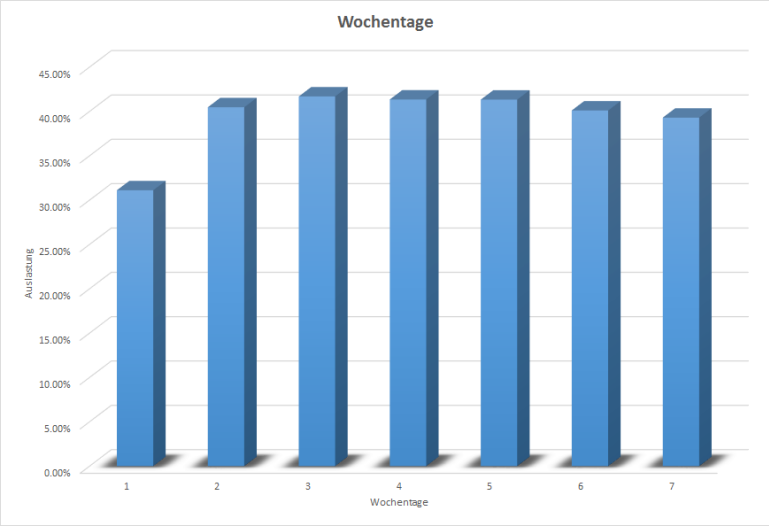


Abbildung 4.6: Durchschnittliche Auslastung nach Wochentagen

trachtung (Abbildung 4.7 (S. 35)) führt der Samstag an manchen Standorten zu einer höheren Auslastung (ID 1002; Jungfernstieg), an anderen wiederum zu einer geringeren Nutzung (ID 10026; Kunsthalle). Auf wieder Andere hat der Samstag, sowie andere Wochentage, keine nennenswerte Auswirkung (ID 10048; Altona(Mercado)).

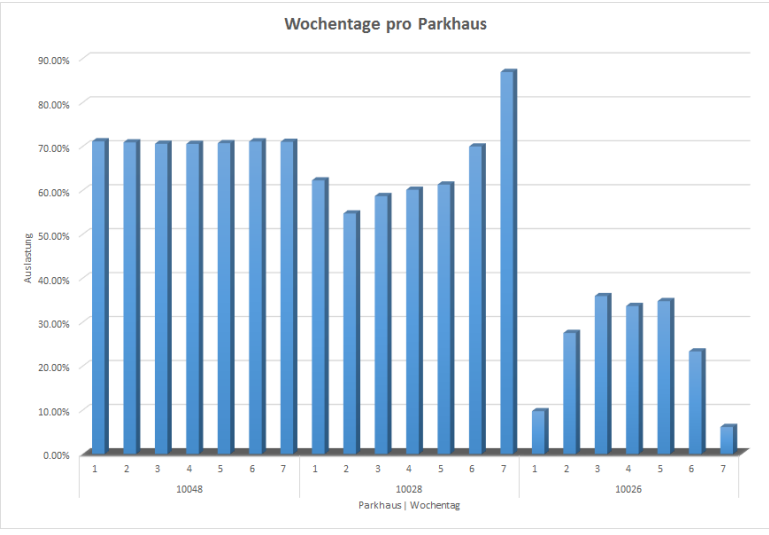


Abbildung 4.7: Durchschnittliche Auslastung einzelner Parkhäuser nach Wochentagen

Der Wochentag kann also für die Prognose einzelner Parkhäuser in der Tat ein entscheidendes Merkmal sein.

Uhr- und Tageszeit

Betrachtet man den Mittelwert der Auslastung über die Uhrzeiten, so ist zu erkennen, dass ab 9 Uhr die Nutzung von Parkhäusern ansteigt und um 13 Uhr ihren Höhepunkt erreicht. Ab ca. 22 Uhr ist nur noch wenig Veränderung zu erkennen.

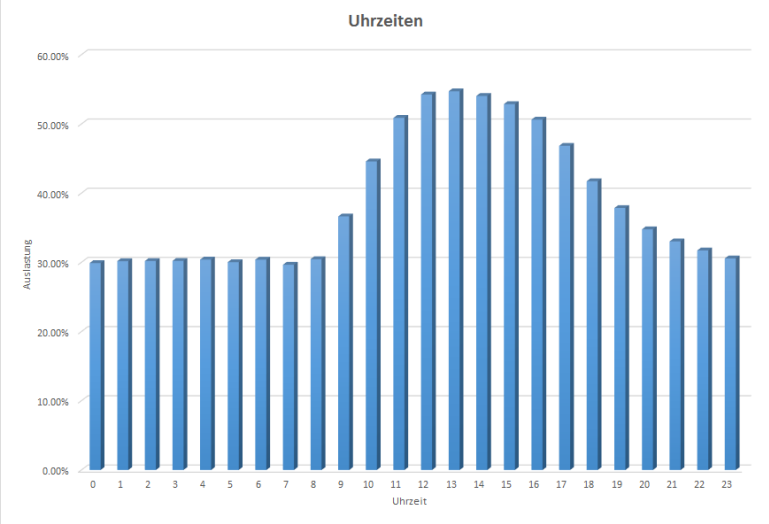


Abbildung 4.8: Durchschnittliche Auslastung nach Uhrzeit

Auch auf den nominalskalierten Werten ist dieses mit einem höheren Aufkommen am Nachmittag wiederzufinden (Abbildung 4.9), jedoch nicht mehr so klar erkennbar, wie in der detaillierten Betrachtung.

Wetter und Temperatur

Die realen Wetterbedingungen mit der höchsten Parkhausauslastungen (*shower rain, thunderstorm, few clouds*) haben im Durchschnitt eine Belegung von ca. 42%. Datensätze mit der Ausprägung *mist* weisen, mit knapp 36%, die geringste Auslastung auf. Somit ist lediglich ein Einfluss von ca. 6% auf die gesamten Daten zu beobachten.

Der Einfluss der Wettervorhersage ist mit ca. 3% sogar noch geringer. Interessant ist, dass im analysierten Zeitraum in den meisten Fällen (55%) Regen prognostiziert wurde (Siehe Abbildung 4.11 (S. 38)).

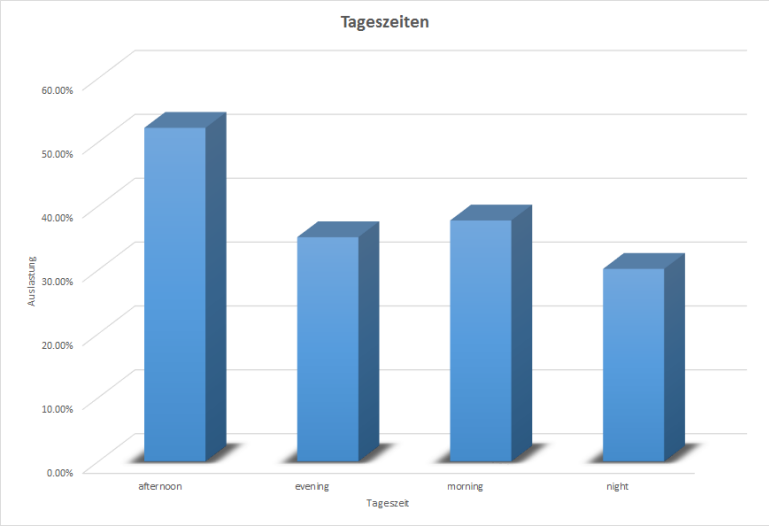


Abbildung 4.9: Durchschnittliche Auslastung nach Tageszeit

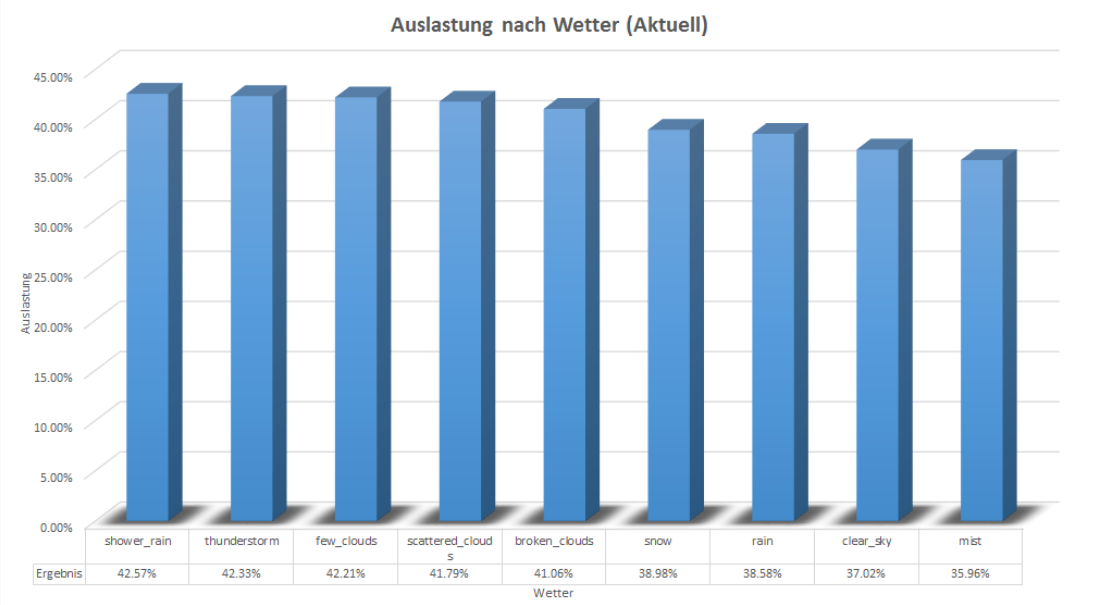


Abbildung 4.10: Durchschnittliche Auslastung nach Wetter (Aktuell)

Die Temperatur hat auf den ersten Blick eine etwas größere Auswirkung. Zwischen den Skalenwert *minorywarm* und *minorycooly* liegt ein Auslastungsunterschied von ca. 15%. Betrachtet man jedoch die Anzahl der Datensätze im Intervall *minorywarm*, fällt auf, dass diese mit lediglich 74 Eintragungen nicht repräsentativ sein dürften. Allgemein ist noch zu erwähnen,

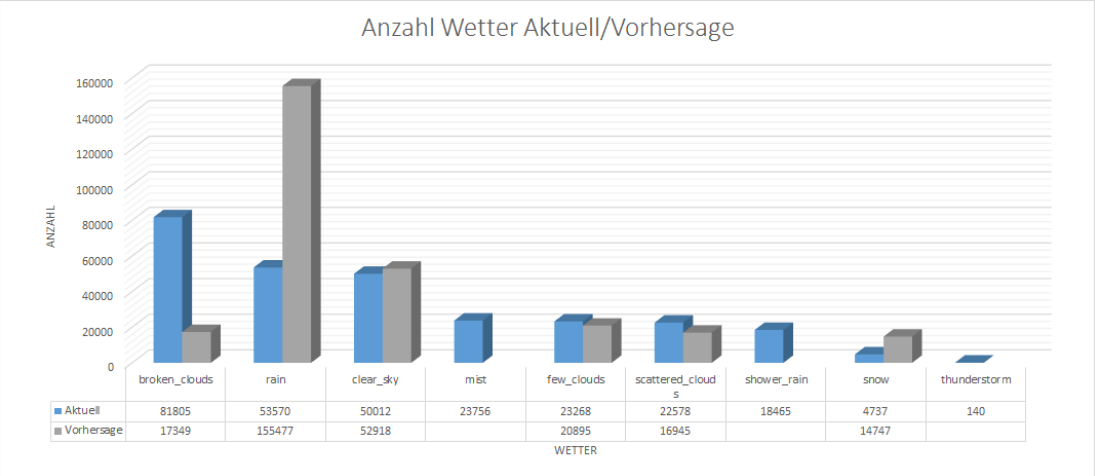


Abbildung 4.11: Vergleich: Anzahl Wetter Aktuell/Vorhersage

dass es in dem analysierten Zeitraum Temperaturschwankungen zwischen -8,58 °C bis 20,56 °C gab.

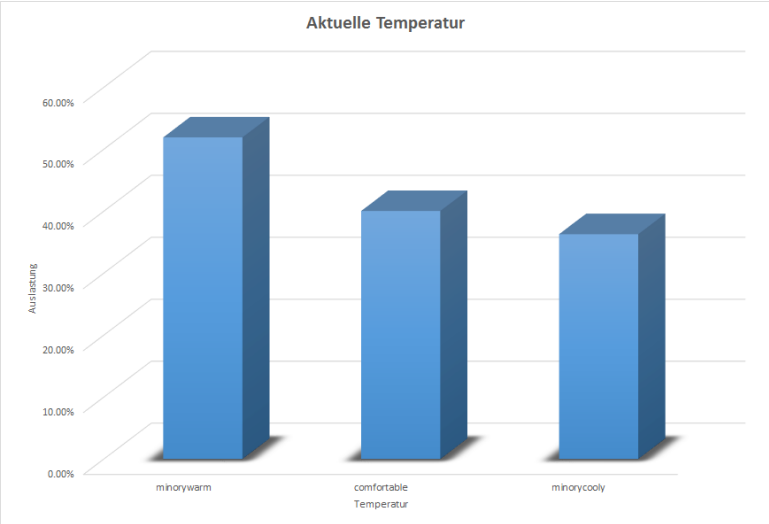


Abbildung 4.12: Auslastung nach Themperatur (Aktuell)

Fazit

In der erste Betrachtung der Merkmale fällt auf, dass die Parkhaus-ID mit Abstand das stärkste Unterscheidungsmerkmal sein wird. Den nächst größeren Einfluss auf die zusammengefasste

Datenbasis hat die Uhrzeit. Aufgrund der Ergebnisse erscheint eine Anpassung der Nominalskala der Tageszeit, wie in Tabelle 4.6 (S. 39) zu sehen, als sinnvoll.

Attribut	Zeichenkette nominal	Intervall
Time	night	22 - 09 Uhr
	morning	09 - 11 Uhr
	afternoon	11 - 16 Uhr
	evening	16 - 22 Uhr

Tabelle 4.6: Transformationsanpassung der Uhrzeit

Alle anderen Attribute zeigen für sich nur einen geringen Einfluss auf die Auslastung. Wie am Beispiel der Wochentage (Abbildung 4.7 (S. 35) aufgezeigt wurde, bedeutet dies jedoch nicht, dass diese nicht für die Auslastung einzelner Parkhäuser eine Bedeutung haben können. Dies wird im folgenden Kapitel 4.6 näher beschreiben.

4.6 Mustererkennung

Nachdem die Daten vorverarbeitet wurden, kann nun das eigentliche Mining, also das Suchen nach den Mustern, erfolgen. Wie bereits in Kapitel 4.5 (S. 32) erwähnt, werden hierfür Klassifikationsalgorithmen verwendet, um Modell zu finden, Datensätze einer der Zielklassen *(Auslastung)* zuordnen zu können. So können anschließend Prognosen für die Zukunft getroffen werden (Siehe auch Kapitel 3.5 (S. 16)). Um die Güte, also die Richtigkeit der Klassenzuteilung, überprüfen zu können, werden die Ausgangsdaten in Trainings- und Testdaten aufgeteilt (*Holdout*-Methode). Auf Basis der Trainingsdaten werden Modelle zur Klassifikation erzeugt und ihre Güte anschließend anhand der Testdaten bestimmt. Dieses Verfahren kann jedoch zu Problemen führen, wenn die zufällig bestimmten Trainingsdaten nicht repräsentativ für die Testdaten sind. Eine Lösung dieses Problems bietet die Kreuzvalidierung. Hier werden die Ausgangsdaten in mehrere gleichgroße Partitionen aufgeteilt. Nun wird eine dieser Teilmengen als Testmenge zurückgehalten, auf dem Rest wird das Modell trainiert und anschließend die Fehlerrate anhand der zurückgehaltenen Daten bestimmt. Dieses Verfahren wird nun mit immer einer anderen Partition als Testmenge wiederholt und der Mittelwert der jeweiligen Fehlerraten berechnet. So kann eine genauere Fehlerrate bestimmt werden als bei einem einfachen Validierung (Cleve und Lämmel, 2016, 235-242).

4.6.1 Verwendete Software

Für die Mustersuche und Evaluierung, sowie auch für die Realisierung der Datenvorverarbeitung und Transformation wurde die Software *RapidMiner Studio*[16] in der Version 6.4 verwendet (Siehe Abbildung 4.13 (S. 41)). RapidMiner wurde, zuerst unter dem Namen *YALE (Yet Another Learning Environment)*, vom Lehrstuhl für künstliche Intelligenz der Technischen Universität Dortmund entwickelt und wird seid dem Spinn-Off des Unternehmens *Rapid_I* ständig weiterentwickelt. Sie gilt als eines der am weitesten entwickelten und verbreiteten Data-Mining-Werkzeuge weltweit. RapidMiner bietet eine vollständige Integration der Lernverfahren und Attributevaluatoren der im akademischen Bereich sehr verbreiteten freien Software *Weka (Waikato Environment for Knowledge Analysis)*[17] (Sharafi, 2013, 95-96).

Die Software wird sowohl unter einer AGPL[18] als auch einer kommerziellen Lizenz [19] angeboten. Die kostenpflichtige Version bietet einige zusätzliche Features, wie z.B. eine größere Auswahl an einzubindenden Datenquellen (Kommerzieller Datenbanken (Oracle, MS SQL Server, etc.) Cloud-Sources (Twitter, Amazon S3, etc.), NoSql, u.a.) (RapidMiner, 2016).

Für diese Arbeit wurde mit einer akademischen Lizenz der kommerziellen Version im Rahmen des RapidMiner Academia Programms [20] gearbeitet.

4.6.2 Naive Bayes

Der *Naive-Bayes-Algorithmus* trifft Vorhersagen über die wahrscheinlichen Klassen. Hier wird kein Modell trainiert, sondern die Vorhersage direkt aus den Trainingsdaten geschlossen. Er geht dabei „naiv“ davon aus, dass es keine Abhängigkeiten der Attribute untereinander gibt (Cleve und Lämmel, 2016, 111-117).

Es wurde eine Güte von **63,40%** mit rein nominalen Attributen und **61,79%** mit Temperatur und Uhrzeit als metrischen Werten erreicht.

4.6.3 Entscheidungsbäume

Entscheidungsbäume erstellen ein Klassifikationsmodell anhand von geordneten und gerichteten Bäumen. Trainingsdaten werden solange rekursiv in disjunkte Teilmengen aufgeteilt,

[16] https://rapidminer.com
[17] http://www.cs.waikato.ac.nz/ml/weka/
[18] GNU AFFERO GENERAL PUBLIC LICENSE, Version 3 - https://opensource.org/licenses/AGPL-3.0
[19] https://rapidminer.com/wp-content/uploads/2016/05/RapidMiner_Clickwrap_20160520.pdf
[20] https://rapidminer.com/academia/

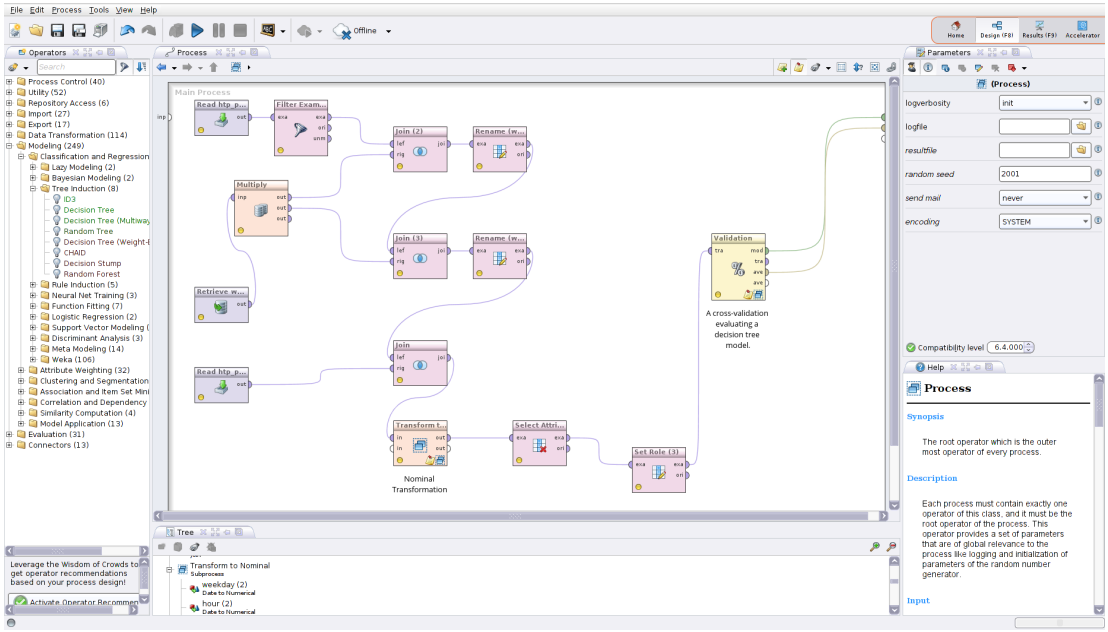


Abbildung 4.13: Rapdidminer - Prozessaufbau

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	41229	17874	7288	10942	53.31%
pred. busy	15297	47772	13110	2075	61.05%
pred. full	1566	4453	11742	1076	62.33%
pred. empty	20097	10387	4899	88188	71.37%
class recall	52.73%	59.35%	31.70%	86.22%	

Tabelle 4.7: Performancevektor: Naive Bayes mit Nominalwerten

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	41075	18520	6940	11201	52.84%
pred. busy	14884	43191	10646	2378	60.75%
pred. full	1962	5211	11337	174	60.68%
pred. empty	20268	13564	8116	88528	67.85%
class recall	52.53%	53.66%	30.61%	86.55%	

Tabelle 4.8: Performancevektor: Naive Bayes mit gemischten Werten

bis in jeder Teilmenge nur noch Objekte einer Klasse vorhanden sind. Findet keine weitere Verzweigung mehr statt, werden diese Knoten als Blatt bezeichnet.

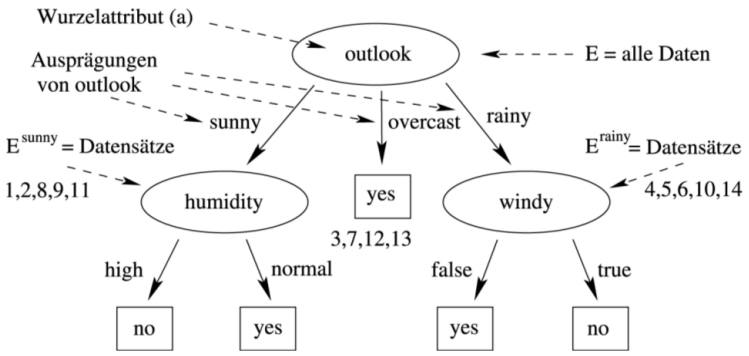


Abbildung 4.14: Entscheidungsbaum - Wetter-Beispiel (Quelle: Cleve und Lämmel (2014))

C4.5

Der *C4.5-Algorithmus* ist ein Nachfolger des *ID3* (*Iterative Dichotomiser 3*)-Algorithmus welcher von J. Ross Quinlan im Jahre 1979 erstmalig publiziert wurde. Mit ihm können möglichst kompakte Entscheidungsbäume generiert werden, indem aus den verbleibenden Attributen das mit dem höchsten Informationsgehalt (der kleinsten Entropie) in Bezug zur Trainingsmenge als nächster Baumknoten ausgewählt wird. Im Unterschied zum ID3 kann der C4.5 auch numerische Attribute verarbeiten (Cleve und Lämmel, 2016, S. 96-111) (Salzberg, 1994)).

Für diese Arbeit wurde die Implementierung von Weka (*W-J48*) im RapidMiner über die *Weka Extension*[21] benutzt. Alle Konfigurationseinstellungen wurden auf den Standarwerten gelassen, außer die Mindestanzahl pro Beispiele pro Blatt. Diese wurde auf 40 erhöht um ein repräsentativeres Bild zu erhalten. Mit den nominalen Werten konnte ein Modell erzeugt werden, welches **76,56%** der Testdaten der richtigen Zielklasse zuordnet.

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	49041	13018	2802	9418	66.02%
pred. busy	13730	56654	12834	2118	66.39%
pred. full	2146	5293	18513	645	69.61%
pred. empty	13272	5521	2890	90100	80.60%
class recall	62.72%	70.39%	49.98%	88.09%	

Tabelle 4.9: Performancevektor: Weka C4.5 mit Nominalwerten

[21] https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_weka

Da der C4.5-Algorithmus, wie oben beschrieben, auch mit numerischen Werten umgehen kann, wurde dieser zum Vergleich dieser auch mit den realen Werten für die Uhrzeit und der Temperatur getestet. Bei diesen wurde eine Güte von **71,92%** erreicht.

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	54338	10625	2329	8529	71.67%
pred. busy	12109	61124	11013	1810	71.03%
pred. full	1636	5114	21410	678	74.24%
pred. empty	10106	3623	2287	91264	85.07%
class recall	69.50%	75.94%	57.80%	89.23%	

Tabelle 4.10: Performancevektor: Weka C4.5 mit gemischten Werten

Decision Tree

RapidMiner bietet zudem einen eigenen Algorithmus zum Erzeugen von Entscheidungsbäumen, welcher vermutlich auch zum Großteil auf dem C4.5 beruht (im Nachfolgenden auch *R.M. Decision Tree* genannt). Die Ergebnisse sind mit einer richtigen Einteilung von **68,43%** bei rein nominalen Attributen und **72,03%** bei gemischten Werten vergleichbar, jedoch etwas schlechter als mit der Implementierung von Weka. Zu erwähnen ist, dass die Performance im Hinblick auf die Geschwindigkeit in der das Modell erzeugt werden konnte, hier wesentlich besser ist.

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	44795	12855	2510	12370	61.76%
pred. busy	15170	54670	14481	1418	63.76%
pred. full	4299	5937	16364	404	60.60%
pred. empty	13925	7024	3684	88089	78.15%
class recall	57.29%	67.92%	44.18%	86.12%	

Tabelle 4.11: Performancevektor: RapidMiner Decision Tree mit Nominalwerten

4.6.4 Neuronal Netze

Ein weiterer Ansatz zur Lösung von Klassifikationsaufgaben sind künstliche neuronale Netze. Dieses Konzept basiert auf den biologischen neuronalen Netzen, welche im Gehirn und Rückenmark gebildet werden. Viele einfache Nervenzellen werden miteinander verbunden und ergeben so ein komplexes Verhalten.

	true medium	true busy	true full	true empty	class precision
pred. medium	47082	9624	1838	10901	67.80%
pred. busy	15700	59011	13310	1224	66.12%
pred. full	4139	6484	18843	434	63.02%
pred. empty	11268	5367	3048	89722	82.01%
class recall	60.22%	73.32%	50.87%	87.72%	

Tabelle 4.12: Performancevektor: RapidMiner Decision Tree mit gemischten Werten

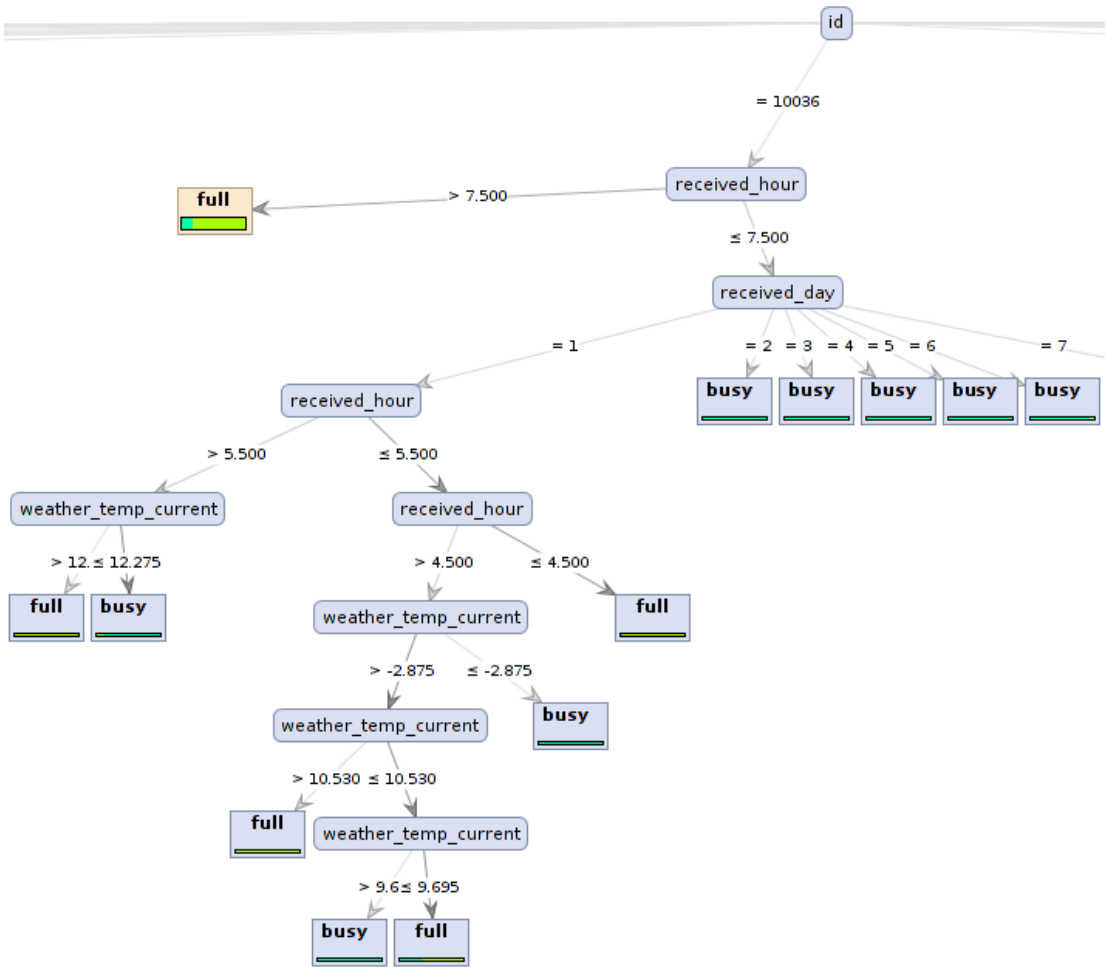


Abbildung 4.15: Modell: R.M. Decision Tree mit gemischten Werten (Ausschnitt)

Für die Klassifikation werden vorwärtsgerichtete neuronale Netze *(Pultilayer Preceptions (MLP))* eingesetzt. Diese bestehen aus einer Eingangsschicht *(Imput-Layer)*, einer beliebigen Anzahl von verdeckten Zwischenschichten *(Hidden-Layer)* sowie einer Ausgangsschicht

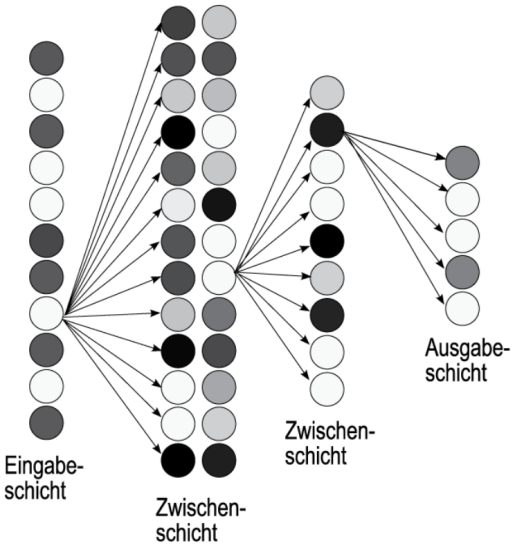


Abbildung 4.16: Vorwärtsgerichtetes neuronales Netz - Beispiel (Quelle: Cleve und Lämmel (2014))

(Output-Layer). Jede dieser Schichten besteht wiederum aus einer Menge von einzelnen künstlichen Neuronen (Cleve und Lämmel, 2016, 47-55, 117-129).

4.6.5 AutoMLP

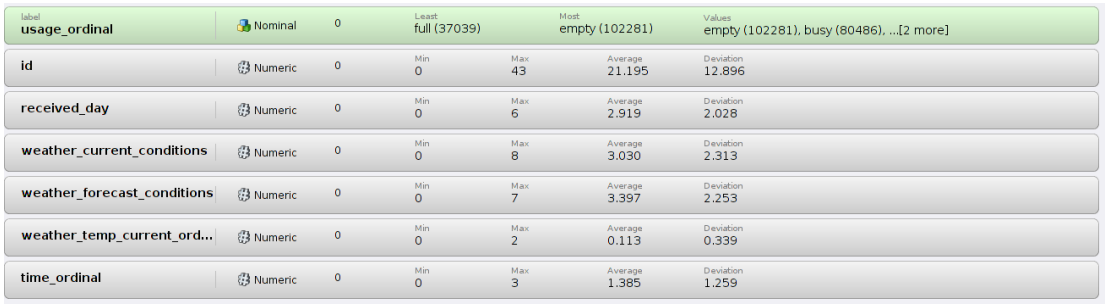
In dieser Arbeit wurde das Verfahren *AutoMLP* eingesetzt. Dies bestimmt die verschiedenen Faktoren, wie die Anzahl der Hidden-Layer und die Lernrate des neuronalen Netzes mit Hilfe von stochastischen und evolutionären Methoden automatisch (Breuel und Shafait, 2010).

Da neuronale Netze nur mit numerischen Werten arbeiten können, wurde allen Nominalattribute auf eine metrische Skala transformiert (Siehe Abbildung 4.17 (S. 46) und 4.18 (S. 46)).

Mit diesem Aufbau wurde eine Klassenzuordnung mit der Güte von **55,41%** erreicht.

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	29523	14397	5138	9157	50.71%
pred. busy	23378	52228	22243	13088	47.08%
pred. full	684	1838	3754	427	56.00%
pred. empty	24604	12023	5904	79609	65.18%
class recall	37.76%	64.89%	10.14%	77.83%	

Tabelle 4.13: Performancevektor: AutoMLP mit Intervallwerten



label			Least	Most	Values
usage_ordinal	Nominal	0	full (37039)	empty (102281)	empty (102281), busy (80486), ...[2 more]

			Min	Max	Average	Deviation
id	Numeric	0	0	43	21.195	12.896
received_day	Numeric	0	0	6	2.919	2.028
weather_current_conditions	Numeric	0	0	8	3.030	2.313
weather_forecast_conditions	Numeric	0	0	7	3.397	2.253
weather_temp_current_ord...	Numeric	0	0	2	0.113	0.339
time_ordinal	Numeric	0	0	3	1.385	1.259

Abbildung 4.17: Testdaten für AutoMLP - Übersicht

Row No.	usage_ordinal	id	received_day	weather_current_conditions	weather_forecast_conditions	weather_temp_current_ordinal	time_ordinal
164125	full	23	2	8	0	1	0
164126	full	23	2	8	0	1	0
164127	full	23	2	8	0	1	1
164128	full	23	2	8	0	1	1
164129	full	23	2	8	0	1	1
164130	full	23	2	1	0	1	1
164131	full	23	2	1	0	1	2
164132	full	23	2	1	0	1	2
164133	full	23	2	0	0	1	2
164134	full	23	2	0	0	1	2
164135	full	23	2	0	0	1	2
164136	full	23	2	0	0	1	2
164137	full	23	2	0	0	1	2
164138	busy	23	2	0	0	1	2
164139	busy	23	2	0	0	1	2
164140	busy	23	2	0	0	1	2
164141	busy	23	2	0	0	1	3
164142	busy	23	2	0	0	1	3
164143	busy	23	2	0	0	1	3
164144	busy	23	2	0	0	1	3
164145	full	23	2	0	0	1	3

Abbildung 4.18: Testdaten für AutoMLP (Ausschnitt)

4.6.6 Bewertung

Durch den Einsatz von Data-Mining-Verfahren konnten, auf Basis der vorliegenden Rohdaten, verschiedene Modelle zur Klassifizierung erstellt werden (Siehe Tabelle 4.14).

Das beste Ergebnis, der in dieser Arbeit getesteten Methoden, wurde durch die Entscheidungsbaumverfahren erreicht. Beide getesteten Algorithmen schnitten vergleichbar gut ab. Das beste Ergebnis lieferte die C4.5 Implementierung von Weka mit nominal skalierten Werten. Hier wurden 76,56% der getesteten Instanzen der richtigen Klasse zugeordnet. Im Vergleich würde eine naive zufällige Einteilung bei den vier Zielklassen zu einem Erfolg von 25% führe. Aus dieser Betrachtungsweise kann von einer signifikanten Verbesserung der Vorhersagemöglichkeit gesprochen werden.

Verfahren	Algorthmus	Skalentyp	Güte
Wahrscheinlichkeiten	Naive-Bayses	nominal	63,40%
		gemischt	61,79%
Entscheidungsbaum	C4.5	nominal	76,56%
		gemischt	71,92%
	R.M. Decision Tree	nominal	68,43%
		gemischt	72,03%
Neuronale Netze	AutoMLP	numerisch	55,41%

Tabelle 4.14: Vergleich der Mining-Verfahren

Unerwartet war das relativ schlechte Ergebnis der neuronalen Netze. Diese liefern häufig bessere Ergebnisse als transparente Methoden wie z.B. Entscheidungsbäume. Hierbei ist zu erwähnen, dass im Rahmen dieser Arbeit alle Verfahren lediglich in ihren Standardkonfigurationen benutzt wurden. Weitere Optimierungen wurden nicht vorgenommen. Es ist zu vermuten, dass die Ergebnisse durch gezielte Anpassungen weiter verbessert werden können. Die hier gelieferten Ergebnisse dienen somit lediglich als ersten Ansatzpunkt für weitere systematischen Analysen.

4.7 Visualisierung

Auf die tatsächliche Umsetzung der Visualisierung wird im Rahmen dieser Arbeit verzichtet. Stattdessen werden mögliche Ansätze skizziert.

Interaktive Karte Die Parkhäuser könnten mit ihren voraussichtlichen Auslastungen zu einer gewünschten Zeit auf einer interaktiven Karte dargestellt werden. Für Prognosen in die Nahe Zukunft, für die bereits Wetterprognosen verfügbar sind, könnte ein Modell zugrunde gelegt werden, welche diese miteinbezieht. Für Zeiten, die weiter in der Zukunft liegen, könnte wiederum auf ein Modell zurückgegriffen werden, welches nur die bereits feststehenden Attribute einbezieht (Siehe Abbildung 4.19 und 4.20)

Integration in die Routenplanug Die gefundenen Modelle könnten auch in eine Routenplanung integriert werden. So könnten vermutlich freie Parkmöglichkeiten schon vor Reisebeginn angezeigt werden und gegebenenfalls auf Park-and-Ride Angebote oder ähnliches hingewiesen werden.

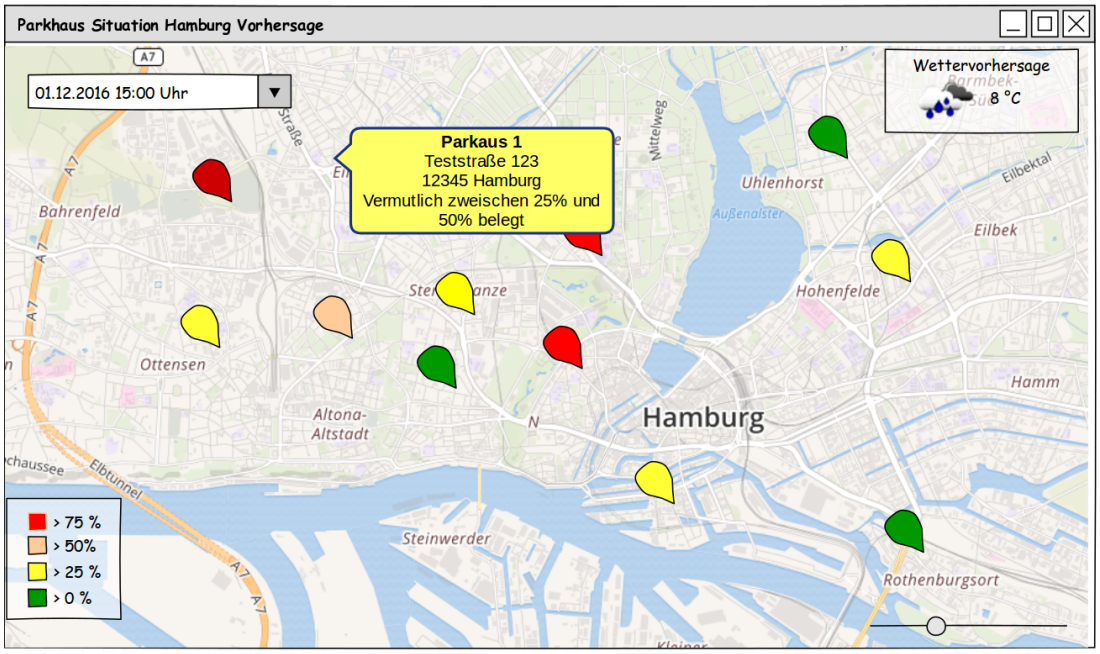


Abbildung 4.19: Mock-Up. Interaktive Karte

Abbildung 4.20: Prototyp. Interaktive Karte

Redaktionelles System Gefundene Modelle können auch intern in Redaktionen dafür genutzt werden, gezielte Empfehlungen zu geben. Werden systematisch Veranstaltungen in die Datenbase eingepflegt, könnten so Modelle für Ereignisse in der Zukunft erstellt werden.

4.8 Fazit und Bewertung der Ergebnisse

Der zuvor beschriebene Workflow konnte anhand des Beispiels eines Prognosemodelles zur Parkhausauslastung erfolgreich getestet werden. Die Datenerhebung gestaltete sich wie erwartet schwieriger als bei klassischen Mining-Prozessen innerhalb von Datawarehouse-Lösungen in Firmen, da nicht alle gewünschten Informationen in der gewünschten Form zugänglich waren, oder aus unterschiedlichen Quellen zusammengetragen werden mussten. Dies konnte jedoch zum Teil durch eigene Datenerhebung gelöst werden. Am Ende konnten Modelle geschaffen werden, die eine Prognose zur Parksituation zulassen und mögliche Darstellungsformen aufgezeigt werden. Es wäre nun die Aufgabe von Journalist*innen diese Ergebnisse publizistische einzubetten, zu bewerten und möglicherweise zu skandalisieren.

Alle hinzugezogenen Attribute wurden rein subjektiv ausgewählt. Es ist durchaus wahrscheinlich, dass entscheidende Faktoren in dieser Arbeit nicht berücksichtigt wurden.

Repräsentanz

Des Weiteren muss darauf hingewiesen werden, dass die zugrunde liegende Datenbase keine Repräsentanz für die Zukunft besitzt. Die hier ausgewerteten Daten umfassen lediglich den Zeitraum zwischen November 2015 und März 2016. Es liegt auf der Hand, dass, gerade im Blick auf Wetter- und Temperaturdaten, diese nicht für das ganze Jahr verallgemeinert werden können. Für eine Auswertung dieser Zusammenhänge müsste voraussichtlich eine Datenerhebung über mehrere Jahre hinweg stattfinden.

Zu Vergleichszwecken wurden zusätzliche Modelle allein aus den repräsentativen Attributen (*id, received_hour, received_day*) mittels Entscheidungsbaumverfahren generiert. Die jeweils erreichte Güte (R.M Decision Tree: **71,43%**, C4,5: **74,18%**) liegt lediglich ein paar Prozentpunkte unter den vorherigen Analysen.

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	45589	10141	1891	10284	67.14%
pred. busy	15824	58416	13526	1237	65.63%
pred. full	4118	6614	18502	421	62.39%
pred. empty	12658	5315	3120	90339	81.07%
class recall	58.31%	72.58%	49.95%	88.32%	

Tabelle 4.15: Performancevektor: RapidMiner Decision Tree ohne Wetter und Temperaturmerkmalen

	true medium	**true busy**	**true full**	**true empty**	**class precision**
pred. medium	51748	11712	2107	11245	67.37%
pred. busy	13876	60024	12448	1484	68.34%
pred. full	2206	4997	20116	379	72.63%
pred. empty	10359	3753	2368	89173	84.40%
class recall	66.18%	74.58%	54.31%	87.18%	

Tabelle 4.16: Performancevektor: C4.5 ohne Wetter- und Temperaturmerkmalen

Hieraus kann, wie bereits in der Merkmalanalyse 4.5.2 (S. 34) vermutet, gefolgert werden, dass Wetter- und Temperaturdaten in dem ausgewerteten Zeitraum lediglich einen geringen Einfluss auf die jeweilige Parkhausauslastung haben.

Kritik der Bewertung Da die Instanzklasse aus metrischen Werten auf eine Nominalskala transformiert wurde, ist eine hohe Fehlerrate bei Datensätzen am Rande des jeweiligen Klassenwertebereiches wahrscheinlich. Hat ein Datensatz beispielsweise eine reale Auslastung von 74% wird dieser der Klasse *busy* (50%-75%) zugeordnet. Wird nun die Klasse *full* (>75%) vorhergesagt, wird diese als Fehler gewertet. Real ist das Ergebnis jedoch nicht weit von dem vorhergesagten entfernt. Diese Problematik wurde durch die sehr grobe Einteilung in nur vier Instanzklassen möglichst klein gehalten, existiert jedoch weiterhin. Ca. 15% der analysierten Datensätze liegen im Schwellwertbereich von 3% zwischen zwei Klassen.

5 Schlussbetrachtung

Im Rahmen dieser Arbeit wurden Möglichkeiten und Vorgehensweisen von digitalen Erzählstrukturen untersucht. Dabei wurden anhand des Beispiels eines Prognosemodells zur Parkhausauslastung die Methoden des Data-Minings vorgestellt und deren Einsatz in der journalistischen Tätigkeit analysiert.

Zusammenfassung

Zunächst wurde auf den Journalismus und dessen Notwendigkeit für eine demokratische Gesellschaft eingegangen. Weiter wurde der Wandel von Informationen durch die Digitalisierung und dessen Auswirkungen auf die journalistische Arbeit beschrieben. Zwar ist der Umgang mit Daten und Informationen für Journalist*innen nichts Neues, das Potenzial von Datenjournalismus geht jedoch über das bloße Darstellen von Zahlen hinaus. Er kann als Ansatz gesehen werden, Geschichten aus Daten zu entwickeln und bietet innovative, neue Möglichkeiten der Publikationen.

Im weiteren Verlauf der Arbeit wurden die Eigenschaften und Anforderungen an Open-Data und Open-Government-Data beschrieben. Speziell wurde auf die Situation in Deutschland und in Hamburg eingegangen. Die Freie und Hansestadt Hamburg verfolgt mit dem *Hamburgischen Transparenzgesetz* im Bundesvergleich einen der offensten Ansätze im Umgang mit Verwaltungsdaten. Maschinenlesbare Informationen bieten die zwingende Grundlage für jede datengetriebene Geschichte.

Im nächsten Kapitel wurde ein Workflow für datenjournalistische Arbeit anhand des im Data-Mining angewendeten KDD-Verfahrens erarbeitet. Die einzelnen Schritte dieses Prozesses wurden erörtert und an die Anforderungen des Journalismus angepasst. Verschiedene Verfahren der Mustersuche wurde vorgestellt und ihre Vorgehensweise skizziert.

Im letzten Kapitel konnte dieser Workflow exemplarisch anhand eines Prognosemodells der Parkhausauslastung in Hamburg verifiziert, sowie Möglichkeiten und Probleme aufgezeigt werden. Zu Beginn wurden die verschiedenen Quellen identifiziert und die verfügbaren Daten analysiert. Aus diesen wurde dann im nächsten Schritt die Datengrundlage für die Auswertung geschaffen, indem Informationen aus den unterschiedlichen Quellen abgegriffen und historisch

gespeichert wurden. Im Rahmen der Vorverarbeitung wurden die gesammelten Daten auf ihre Güte geprüft und fehlerhafte Informationen identifiziert. Der nächste Verarbeitungsschritt umfasste die Skalentransformation in rein nominale Attribute, welche für die folgende Mustersuche notwendig war. Zunächst wurde jedoch der Einfluss der einzelnen Merkmale auf die zu prognostizierende Instanzklasse bestimmt. Hier zeigte sich, dass dieser stark variiert. Durch die anschließende Anwendung unterschiedlicher Klassifikationsalgorithmen konnten verschiedene Klassifikatoren erstellt und die Güte dieser Modell miteinander verglichen werden. In diesen Untersuchungen schnitten die entscheidungsbaumbasierten Verfahren am besten ab. Am Ende wurden mögliche Darstellungs- und Visualisierungsformen skizziert und anhand von Beispielen veranschaulicht.

Fazit

Daten werden in der Zukunft des Journalismus voraussichtlich eine wachsende Rolle spielen. Mit den Techniken aus Informatik und Datenverarbeitung können neue Ansätze geschaffen werden, Geschichten anders, verständlicher und spannender zu erzählen. Für die Zukunft des Datenjournalismus ist eine weitere Annäherung und interdisziplinäre Zusammenarbeit zwischen Journalist*innen und Informatiker*innen notwendig und wünschenswert. In der Vielzahl an Informationen, die Tag für Tag entstehen, stecken unzählige Zusammenhänge und Erkenntnisse, die mit bloßem Auge nicht zu begreifen sind. Mit Hilfe von computergestützten Verfahren können Algorithmen diese zu Tage tragen, wie anhand des Prognosesystems beispielhaft dargestellt werden konnte. Damit dies geschehen kann müssen Daten frei und strukturiert verfügbar sein. Dieses Verständnis von Transparenz und Offenheit hat sich in den letzten Jahren immer weiter verbreitet. Dennoch gibt es auch in Deutschland jede Menge Potenzial den freien Zugang weiter auszubauen. Es existieren bereits eine Vielzahl an zugänglichen Daten, nun ist es wichtig diese auch zu nutzen und ihnen ihre Informationen und Geheimnisse zu entlocken.

Ausblick

Die, im Rahmen dieser Arbeit neu geschaffene Datenbasis über die historischen Parkhausauslastung, bietet eine Vielzahl an neuen Auswertungs- und Analysemöglichkeiten. Durch die nun wachsende Anzahl an auswertbaren Informationen könnten in Zukunft repräsentativere und genauere Modelle zur Prädiktion möglich sein. Lediglich bei starken Veränderungen der Ausgangssituationen, z.B. durch einen rapiden Anstieg des Bezinpreises, die Einführung von Umweltzonen oder ähnlichem, müssten Modelle auf einer neuen Grundlage trainiert werden. Eine historische Datenhaltung bietet in diesen Fällen jedoch überhaupt erst die Möglichkeit,

Auswirkungen und Veränderungen erkennen und analysieren zu können.

Mit Blick auf die Prognose zur Parkhausauslastung könnten Untersuchungen mit einem größeren oder anderen Merkmalsspektrum weitere spannende Ansätze zur Prädiktion liefern. Veranstaltungen wurden in dieser Arbeit zwar in der Vorverarbeitung berücksichtigt, in der eigentlichen Analyse jedoch, aufgrund der fehlenden Datenbasis, nicht ausgewertet.

Mit weiteren Daten, z. B. aus der Parkraumbewirtschaftung zum Nutzungsverhalten von öffentlichen Parkplätzen könnten Prognosemodelle auf weitere Bereiche ausgedehnt werden. Bestehen Zusammenhänge zwischen dem Verkehrsaufkommen auf den Straßen, welches für Google bereits möglich ist zu prognostizieren, und der Parkhausauslastung? Welche weiteren Faktoren sind denkbar? Dies könnte ein möglicher Ansatzpunkt für weitere Untersuchungen sein.

Weitere Erfolge könnte die Kombination mehrerer Klassifikationsmethoden, sowie eine weitere Analyse und Anpassung der genutzten Algorithmen liefern.

Ein anderer Ansatzpunkt für zeitnahe Vorhersagen, welche beispielsweise für die Integration in Navigationssysteme interessant sein könnte, wären Zeitreihenanalysen. Hierbei werden, anders als bei der in dieser Arbeit betrachteten Klassifikationsmethoden, keine allgemeingültigen Modelle erstellt, sondern Trends und Entwicklungen anhand von zeitlich geordneten Datenpunkten vorhergesagt (vgl. Kreiss und Neuhaus (2006)).

Schwerpunkt weiterer Untersuchungen könnte die Generalisierung der hier beschriebenen Prozesse in ein softwaregestütztes Framework für Datenjournalist*innen sein. Es wird zwar davon ausgegangen, dass Datenanalyse nicht ohne Weiteres zu automatisieren ist, sondern immer auf fachliche Expertise aufbaut, es wäre jedoch denkbar Systeme zu schaffen, die die Arbeit, gerade bei kleineren datengetriebenen Geschichten, erleichtern oder zum besseren Austauschen zwischen den unterschiedlichen Disziplinen beitragen. Hierfür wären weitere Analysen mit interdisziplinärem Anspruch notwendig.

Literaturverzeichnis

[Aisch 2016] Aisch, Gregor: *Using Data Visualization to Find Insights in Data - The Data Journalism Handbook*. 2016. – URL http://datajournalismhandbook.org/1.0/en/understanding_data_7.html. – Zugriffsdatum: 15.01.2016

[Breuel und Shafait 2010] Breuel, Thomas ; Shafait, Faisal: AutoMLP: Simple, Effective, Fully Automated Learning Rate and Size Adjustment. In: *The Learning Workshop*, URL http://madm.dfki.de/publication&pubid=4950. – Zugriffsdatum: 07.05.2016, 2010

[BWVI 2015] BWVI: *Parkäuser Hamburg - MetaVer*. Aug 2015. – URL http://metaver.de/trefferanzeige?docuuid=0EEE1494-36DD-410C-B5A3-7531DC457014. – Zugriffsdatum: 15.11.2015

[Cisco VNI 2015] Cisco VNI: *The Zettabyte Era-Trends and Analysis*. 2015. – URL http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.pdf. – Zugriffsdatum: 28.11.2015

[Cleve und Lämmel 2014] Cleve, Jürgen ; Lämmel, Uwe: *Data Mining*. 1. Aufl. Walter de Gruyter GmbH, 2014. – ISBN 978-3-486-71391-6

[Cleve und Lämmel 2016] Cleve, Jürgen ; Lämmel, Uwe: *Data Mining*. 2. Aufl. Walter de Gruyter GmbH, 2016. – ISBN 978-3-11-045675-2

[Dapp u. a. 2016] Dapp, Marcus M. ; Balta, Dian ; Palmetshofer, Walter ; Krcmar, Helmut: *Open Data. The Benefits*. Konrad-Adenauer-Stiftung e.V., 2016. – URL http://www.kas.de/wf/doc/kas_44906-544-1-30.pdf. – Zugriffsdatum: 16.07.2016

[DJV 2015] DJV: *DJV Wissen 4. Berufsbild Journalistin - Journalist*. Feb 2015. – URL http://www.djv.de/fileadmin/user_upload/Infos_PDFs/Flyer_Broschuren/wissen4_Berufsbild.pdf. – Zugriffsdatum: 23.11.2015

[Ester und Sander 2000] Ester, Martin ; Sander, Jörg: *Knowledge Discovery in Databases - Techniken und Anwendungen.* 1. Aufl. Springer-Verlag, 2000. – ISBN 978-3-540-67328-6

[Freie und Hansestadt Hamburg 2016] Freie und Hansestadt Hamburg: *Transparenzportal FAQ.* 2016. – URL http://transparenz.hamburg.de/fragen-und-antworten. – Zugriffsdatum: 13.06.2016

[GovData 2016] GovData: *Metadatenstruktur für Daten in Deutschland.* 2016. – URL https://www.govdata.de/metadatenschema. – Zugriffsdatum: 30.06.2016

[Haase 2011] Haase, Alexander: *Stories in Data. Das Potential von Daten und ihr Einfluss auf den Journalismus.* 2011. – URL http://hsmw.bsz-bw.de/frontdoor/index/index/year/2011/docId/1524. – Zugriffsdatum: 14.12.2015

[Howard 2014] Howard, Alecander B.: *The Art and Science of Data-driven Journalism.* 2014. – URL http://towcenter.org/wp-content/uploads/2014/05/Tow-Center-Data-Driven-Journalism.pdf. – Zugriffsdatum: 30.06.2016

[Klessmann u. a. 2012] Klessmann, Jens ; Denker, Philipp ; Schieferdecker, Ina ; Schulz, Sönke E.: *Open Government Data Deutschland, Eine Studie zu Open Government in Deutschland im Auftrag des Bundesministerium des Innern.* Bundesministerium des Innern, 2012. – URL https://www.bmi.bund.de/SharedDocs/Downloads/DE/Themen/OED_Verwaltung/ModerneVerwaltung/opengovernment.pdf. – Zugriffsdatum: 13.06.2016

[Kramp u. a. 2013] Kramp, Leif ; Novy, Leonard ; Ballwieser, Dennis ; Wenzlaff, Karsten ; Kramp, Leif ; Novy, Leonard ; Ballwieser, Dennis ; Wenzlaff, Karsten: *Journalismus in der digitalen Moderne - Einsichten - Ansichten - Aussichten.* 1. Aufl. Berlin Heidelberg New York : Springer-Verlag, 2013. – ISBN 978-3-658-01144-4

[Kreiss und Neuhaus 2006] Kreiss, Jens-Peter ; Neuhaus, Georg: *Einführung in die Zeitreihenanalyse.* Springer, 2006. – ISBN 978-3-540-25628-1

[Lucke und Geiger 2010] Lucke, Jörn von ; Geiger, Christian P.: *Open Government Data. Frei verfügbare Daten des öffentlichen Sektors.* Zeppelin University, 2010. – URL https://www.zu.de/institute/togi/assets/pdf/TICC-101203-OpenGovernmentData-V1.pdf. – Zugriffsdatum: 21.01.2016

[Matzat 2011] Matzat, Lorenz: *Datenjournalismus*. Aug 2011. – URL http://www.bpb.de/gesellschaft/medien/opendata/64069/datenjournalismus. – Zugriffsdatum: 20.11.2015

[Neuberger und Kapern 2013] Neuberger, Christoph ; Kapern, Peter: *Grundlagen des Journalismus*. 1. Aufl. Berlin Heidelberg New York : Springer-Verlag, 2013. – ISBN 978-3-531-94191-2

[Neumüller und Kahn 2015] Neumüller, Fritz ; Kahn, Eram: *Perlentauer im Datenmeer. Sieben Thesen zu Datenjornalismus im New News Process*. 2015. – URL journal.kommunikation-media.at

[Open Knowledge Foundation 2015] Open Knowledge Foundation: *Global Open Data Index*. 2015. – URL http://index.okfn.org/place/. – Zugriffsdatum: 27.01.2016

[Open Knowledge Foundation Deutschland 2016] Open Knowledge Foundation Deutschland: *Offene Daten | Open Knowledge Foundation Deutschland*. 2016. – URL https://okfn.de/themen/offene-daten/. – Zugriffsdatum: 2016-1-18

[RapidMiner 2016] RapidMiner: *Compare RapidMiner Studio Editions*. 2016. – URL https://rapidminer.com/products/comparison/. – Zugriffsdatum: 03.06.2016

[Salzberg 1994] Salzberg, Steven L.: *C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993*. 1994. – URL http://dx.doi.org/10.1007/BF00993309

[Sharafi 2013] Sharafi, Armin: *Knowledge Discovery in Databases - Eine Analyse des Änderungsmanagements in der Produktentwicklung*. Wiesbaden : Springer Gabler, 2013. – ISBN 978-3-658-02002-6

[Sharp 2012] Sharp, Adam: *Election Night 2012*. 2012. – URL https://blog.twitter.com/2012/election-night-2012. – Zugriffsdatum: 01.04.2016

[Stray 2010] Stray, Jonathan: *A full-text visualization of the Iraq War Logs*. 2010. – URL http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs. – Zugriffsdatum: 11.01.2016

[Süddeutsche Zeitung 2012] Süddeutsche Zeitung: *Werkstattbericht - Wie der Zugmonitor entstanden ist*. 2012. – URL http://www.sueddeutsche.de/kolumne/werkstattbericht-wie-der-zugmonitor-entstanden-ist-1.1303418. – Zugriffsdatum: 20.06.2016

[Sunlight Foundation 2010] Sunlight Foundation: *Ten Principles for Opening Up Government Information*. 2010. – URL https://sunlightfoundation.com/policy/documents/ten-open-data-principles/. – Zugriffsdatum: 19.01.2016

[Wolfangel 2015] Wolfangel, Eva: *Datenjournalismus. Selbst Leuchtturm-Projekte leuchten nicht*. 2015. – URL http://www.meta-magazin.org/2015/11/12/datenjournalismus-selbst-die-leuchttuerme-visualisieren-bloss. – Zugriffsdatum: 30.06.2016

[Woyteqicz 2013] Woyteqicz, Daniela: *Mit Daten Geschichten erzählen: Das Potential von Datenjounalismus im World Wide Web*. 2013. – URL http://www.fbi.fh-koeln.de/institut/papers/kabi/volltexte/band071.pdf. – Zugriffsdatum: 28.11.2015

[ZEIT ONLINE GmbH 2012] ZEIT ONLINE GmbH: *US-Wahl: Twitter wusste, wer die Wahl gewinnt*. 2012. – URL http://www.zeit.de/digital/internet/2012-11/twitter-obama-wahl. – Zugriffsdatum: 11.01.2016

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 1. August 2016 Simon Dreyer