HTW Chur
Hochschule für Technik und Wirtschaft
University of Applied Sciences

Churer Schriften
zur Informationswissenschaft
Herausgegeben von
Wolfgang Semar

Arbeitsbereich
Informationswissenschaft

Schrift 91

Customer Engagement Analytics: Clustering User Navigation Behaviour

Sven Lenz

Chur 2017

Churer Schriften zur Informationswissenschaft

Herausgegeben von Wolfgang Semar

Schrift 91

Customer Engagement Analytics: Clustering User Navigation Behaviour

Sven Lenz

Diese Publikation entstand im Rahmen einer Thesis zum Master of Science FHO in Business Administration, Major Information and Datamanagement.

Referent: Prof. Dr. habil. Wolfgang Semar

Korreferent: Prof. Dr. habil. Albert Weichselbraun

Verlag: Arbeitsbereich Informationswissenschaft

ISSN: 1660-945X

Chur, November 2017

Abstract

Durch die Digitalisierung der Geschäftsprozesse stehen Anwendern von Informationssystemen immer mehr Services zur Verfügung, die ihre Tätigkeiten effizienter und effektiver werden lassen. Ihnen soll ein informationeller Mehrwert geboten werden, so dass Kundenloyalität aufgebaut und eine Kundenabwanderung verhindert wird. Die Thesis beschäftigt sich damit, wie aus den Daten, die bei der Interaktion eines Benutzers mit einem Informationssystem entstehen, ein Kundenverständnis erlangt werden kann. Anhand eines Smart Data Systems werden die Benutzer in Gruppen segmentiert. Durch eine Online-Umfrage werden die Persönlichkeitsprofile und Clickstream-Sequenzen aufgezeichnet und eine Segmentierung mit Hilfe eines Divisive Hierarchical Clustering durchgeführt. Die Ergebnisse zeigen auf, dass innerhalb der segmentierten Gruppen keine Übereinstimmung auf ähnliche Persönlichkeitsprofile basierend auf den Big Five Faktoren gefunden werden konnte. Indessen kann eine Effizienz- und Effektivitätswirkung für den Benutzer erreicht werden, indem Navigationsstrukturen und Produktplatzierungen für die Cluster optimiert werden.

Schlagwörter. Smart Data Discovery, Web Usage Mining, Clustering, Persönlichkeitsprofile, Big Five

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AWS	Amazon Web Services
bzw.	beziehungsweise
bzgl.	bezüglich
B5T®	Big-Five-Persönlichkeitstest von Dr. Satow
CSV	Comma Separated Value
d.h.	das heisst
DHC	Divisive Hierarchical Clustering
EC2	Elastic Cloud Compute
EBS	Elastic Block Storage
EM	Expectation Maximum
etc.	et cetera
ETL	Extraktion, Transformation, Laden
evtl.	eventuell
HDFS	Hadoop Distributed File System
i.d.R.	in der Regel
IP	Internet Protocol
IR	Information Retrieval
IT	Information Technology
JSON	JavaScript Object Notation
PLSA	Probabilistic Latent Semantic Analysis
resp.	respektive
SII	Schweizerisches Institut für Informationswissenschaft
sog.	so genannt
u.a.	unter anderem / unter anderen
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
usw.	und so weiter
v.a.	vor allem
vgl.	vergleiche
vs.	versus
z.B.	zum Beispiel

1 Einführung in die Thematik

Die Digitalisierung der Geschäftsprozesse hat in den letzten Jahren in diversen Unternehmen Einzug gehalten. Auch in der U-NICA Systems AG, die hauptsächlich im Bereich des Marken- und Produktschutzes tätig ist, wurde erkannt, dass die Verknüpfung der analogen und digitalen Welt wichtig ist, um innovative Schutzmöglichkeiten für Markenprodukte anbieten zu können. Im Zuge der Digitalisierung wurde die Applikation scryptoTrace entwickelt, die schnell und zuverlässig anhand von versteckten digitalen Merkmalen auf einer Verpackung erkennen kann, um was für ein Produkt es sich handelt, und ob das Produkt authentisch ist. Durch die zunehmende Verbreitung der Digitalisierung und der Möglichkeiten der Konsumenten beim Kauf von Produkten mit dem Internet verbunden zu sein, führt dies zu der Idee, dass sich die Applikation vom reinen Expertensystem zu einer Endkonsumenten-Lösung entwickeln soll.

Die Rede ist von einem Customer Engagement, das dem Konsumenten durch eine Smartphone-Applikation einen informationellen Mehrwert bietet. Basierend auf den in scryptoTrace implementierten Bildverarbeitungsalgorithmen wird eine neue Smartphone-Applikation scryptoSight entwickelt, die einem Konsumenten mittels eines Scanvorgangs der Verpackung weitere Daten zu diesem Produkt anzeigt. Die Daten über das Produkt werden dem User in einer hierarchischen Form präsentiert, und bieten dem Nutzer die Möglichkeit, innerhalb der Applikation durch die verschiedenen Themenbereiche zu navigieren.

Durch Gespräche mit den Businesspartnern zeigt sich nun vermehrt das Bedürfnis, dass die mit der Applikation gesammelten Daten so ausgewertet werden können, dass einerseits ein Mehrwert für die Konsumenten entsteht, andererseits auch die Markenanbieter einen Nutzen aus der Applikation ziehen können. Das nachfolgende Kapitel erläutert das Forschungsproblem sowie die Forschungsfragen, die aus diesem Kundenbedürfnis abgeleitet werden können.

1.1 Aufbau der Arbeit

Das Kapitel 1 bietet eine Einführung in die Thematik und beschreibt das gewählte Vorgehen für die Umsetzung der Forschung. Basierend auf der Konzeptualisierung der Thematik, die einen ersten Einblick in die Themengebiete der Masterarbeit gibt und dazu dient, grundlegende Konzepte aus der Literatur zu erarbeiten, werden diese Konzepte visualisiert und die relevanten Suchbegriffe für die anschliessende vertiefte Literaturanalyse ausgewählt. Sämtliche Ergebnisse aus der Literaturanalyse sind in die drei nachfolgenden Hauptkapitel integriert, die jeweils eine der drei Forschungsfragen bearbeiten. Grundsätzlich sind diese Kapitel nach einem ähnlichen Schema aufgebaut. Die Einleitung bietet einen Kontext zu dem

thematischen Kapitel und stellt einen Bezug zu einer der Forschungsfragen her. Danach werden die in der Literatur gefundenen theoretischen Erkenntnisse in Kapitel vorgestellt und die relevanten Aspekte für die Beantwortung der Forschungsfrage ausgearbeitet. Sämtliche Forschungsergebnisse werden gegen Prüfraster aus der Literatur validiert. In Kapitel 3 wird das vorgeschlagene System durch einen Anforderungskatalog an moderne Datenmanagementsysteme geprüft. Die Stabilität der potentiellen Algorithmen mittels einer relativen Validierung wird in Kapitel 3 miteinander verglichen. Des Weiteren werden in Kapitel 4 zwei unterschiedliche Gold Standards erarbeitet, um die verschiedenen Modellierungen mittels einer externen Validierung vergleichen zu können, um so die optimale Modellierung für das Forschungsproblem zu erarbeiten. Jedes der drei Kapitel schliesst mit einem Fazit und nimmt dabei Bezug auf die Beantwortung der im Kapitel behandelten Forschungsfrage.

In der abschliessenden Diskussion in Kapitel 5 werden nochmals sämtliche Erkenntnisse zur Beantwortung der drei Forschungsfragen zusammengetragen und unter einer kritischen Reflexion präsentiert. Die Diskussion zeigt auf, wo die Stärken und Schwächen der Masterarbeit liegen, und wie die Schwächen in zukünftigen Forschungen angegangen werden können.

1.2 Konzeptualisierung der Thematik

Durch die Konzeptualisierung der Thematik sollen die für diese Arbeit wichtigen und relevanten Konzepte aufgezeigt werden, so dass basierend auf diesen Konzepten eine vertiefte Analyse der Literatur vorgenommen werden kann. Als Einstieg in die Konzeptualisierung dient die Theorie der informationellen Mehrwerte (Kuhlen, 1995) sowie das Standardwerk „Grundlagen der praktischen Information und Dokumentation“ von Kuhlen et al. (2014). Da sich das Forschungsproblem zudem über mehrere Bereiche hinwegzieht und sowohl technologische, informationswissenschaftliche als auch sozialwissenschaftliche Aspekte betrachtet werden müssen, fällt die Wahl auf die Datenbanken Library and Information Sciences Abstracts (LISA) sowie Library, Information Science & Technology Abstracts (LISTA) für eine vertiefte Literaturanalyse. Als zusätzliche Quelle wird Google Scholar genutzt, um weitere Definitionen und Ausführungen zur recherchierten Literatur ausfindig zu machen.

1.2.1 Informationelle Mehrwerte

In der Einleitung ist die Rede von einem informationellen Mehrwert, der die neue Smartphone-Applikation einem Konsumenten bzw. einem Benutzer dieser Applikation bieten soll. Gemäss Kuhlen (1995, S. 82) kann dann von einem realen informationellen Mehrwert gesprochen werden, wenn das Informationssystem vom Nutzer akzeptiert wird und eine systembezogene Mehrwertleistung entsteht. Damit ein informationeller Mehrwert geschaffen

werden kann, müssen Wissensobjekte vorhanden sein, die mittels einer Informationsarbeit aus dem Wissen in eine Information für den Nutzer überführt werden können. Diese Wissensobjekte bleiben über die Zeit nicht unverändert und müssen von dem Informationssystem für erneute Informationsarbeit wieder aktualisiert zur Verfügung gestellt werden.

In der Theorie werden drei Entstehungsgründe für informationelle Mehrwerte definiert. Grundlage für den Aufbau von Informationssystemen ist die (1) Wissensrekonstruktion, bei der Wissensobjekte wie z.B. Artikel oder Bücher auf informationelle Ressourcen (Datenbanken) abgebildet werden. Der Zugriff auf ein Informationssystem geschieht über verschiedene Formen der (2) Informationserarbeitung, die wiederum einen Mehrwert darstellen, in dem die informationellen Ressourcen auf Relevanzinformation abgebildet werden, d.h. als Information, die für ein gegebenes Problem relevant ist. Ob diese Relevanzinformation nun von einem Nutzer verwendet wird, hängt von der (3) Informationsaufbereitung ab. Die semantischen Eigenschaften der Informationseinheiten alleine reichen nicht aus, sondern es muss ein informationeller Mehrwert durch eine Veredelung bzw. eine Anpassung an die speziellen Kundenwünsche geschaffen werden.

Kuhlen (1995, S. 85) zeigt in seinem Mikromodell der Informationsarbeit, dass eine Informationsverwaltung für diese drei Schritte der Mehrwerterzeugung eine zentrale Bedeutung hat. Die Informationsverwaltung ist zuständig dafür, dass alle Zwischenstufen der bearbeiteten Information abgelegt werden können. Abbildung 1 illustriert das von Kuhlen vorgestellte Mikromodell der Informationsarbeit und zeigt auf, wie bereits im Jahr 1995 erkannt worden ist, dass die Informationsverwaltung eine zentrale Rolle im Prozess der Informationsarbeit einnimmt. Der Prozess rund um die Informationsverwaltung ist iterativ. Während sämtlichen Iterationsstufen werden die Relevanzinformation (Abgeleitet aus der informationellen Ressource), die aufbereitete Information und die Handlungsinformation in der zentralen Informationsverwaltung abgelegt.

Abschliessend definiert Kuhlen (1995) den Prozess der Informationsverarbeitung, in dem entweder durch die kognitive Leistung des Menschen "die Teilmenge selektiert wird, die als Handlungsinformation für die aktuelle Problemlösungssituation tatsächlich gebraucht wird" (S. 89) oder, dass aktuell immer mehr Informationssysteme am Entstehen sind, die die Informationsverarbeitung mit dem Ziel der direkten Problemlösung übernehmen.

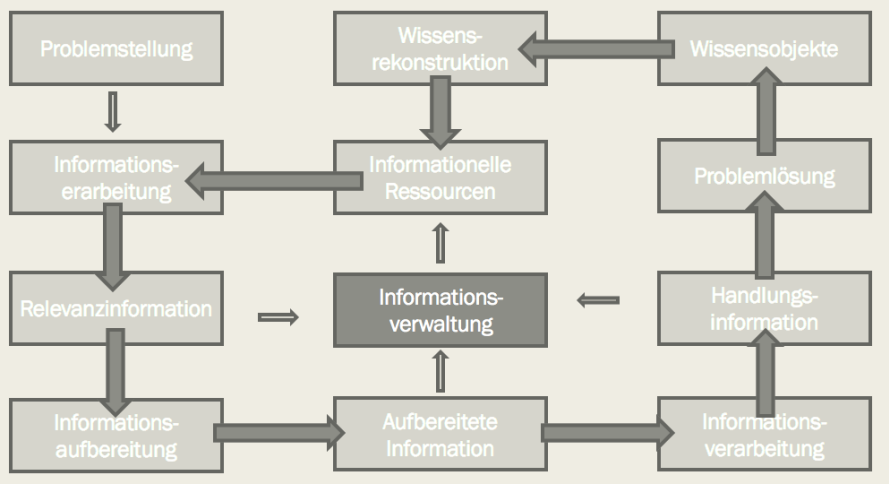

Abbildung 1: Mikromodell der Informationsarbeit (eigene Darstellung in Anlehnung an Kuhlen)

1.2.2 Datenmanagement

Wie aus der Theorie der informationellen Mehrwerte entnommen werden kann, ist die Informationsverwaltung, heute auch oft Datenmanagement genannt, zentral für die Mehrwerterzeugung. Das Customer Engagement hat zum Ziel, dass eine grosse Zahl an Benutzern mit der Applikation erreicht werden kann und daher kann mit einer sehr umfangreichen Datenmenge gerechnet werden. Ein gutes Datenmanagement scheint für den weiteren Verlauf der Arbeit somit von zentraler Bedeutung zu sein. Gemäss Klump und Bertelmann (2014, S. 575) lassen sich Datenvolumen in die beiden Bereiche Big Data und Small Data einteilen. Von Big Data wird dabei ab einigen Gigabyte Daten innerhalb der Datenbank gesprochen. Eine Kategorisierung von Big Data und Small Data von O'Neal (2012) zeigt, dass dann von Big Data gesprochen werden kann, wenn unter anderem Log Daten ausgewertet werden, und die Verarbeitung in Real Time erfolgt. Als Small Data Kategorisiert O'Neal z.B. Web Transactions, die in einer Batch oder quasi Real Time Verarbeitung erfolgen. Neben Small Data taucht auch immer öfters der Begriff Smart Data auf, bei dem es darum geht, die Datenmengen zu verstehen und über das Wissen zu verfügen, an die wirklich nützlichen Daten heranzukommen (Heuring, 2015). Diese Transformation von Big Data nach Smart Data deckt sich mit dem von Kuhlen vorgestellten Mehrwert der Wissensrekonstruktion, welche Wissensobjekte wie z.B. die Interaktionsdaten eines Nutzers mit einem System in informationelle Ressourcen transformiert.

1.2.3 Analyse von Daten

Auffindbare und zur Verfügung stehende Daten sind die wesentlichen Aspekte, damit eine Integration und Analyse grösserer Datenbestände ermöglicht wird (Klump & Bertelmann, 2014, S. 879). Für die Analyse der Daten nach relevanter Information in einem bestimmten

Kontext helfen u.a. die Ansätze der Wissensorganisation, semantische Technologien sowie linguistische Methoden. Eine weitere Kategorie von Ansätzen versucht mit Hilfe von Personalisierung, die Relevanz und Nützlichkeit der einem Benutzer gelieferten Informationsmenge zu verbessern (Reimer, 2014, S. 238).

Wie in den Grundlagen zu den informationellen Mehrwerten aufgezeigt werden konnte, liegt diesen Ansätzen jeweils eine Informationsarbeit zugrunde, die als Ergebnis dem Anwender einen informationellen Mehrwert liefert. Gemäss Kuhlen (1995, S. 91) kann zwischen drei Typen von informationellen Mehrwerten unterschieden werden. Produktbezogene informationelle Mehrwerte, organisationsbezogene informationelle Mehrwerte und wirkungsbezogene individuelle informationelle Mehrwerte. Die letzte Kategorisierung ist in dieser Arbeit von Bedeutung, da diese Mehrwerte für einen Nutzer eine Effizienz- und Effektivitätswirkung bringen. Nutzer können durch die geschaffenen Mehrwerte bisherige Tätigkeiten schneller durchführen und ihre Ziele besser erreichen. Durch ästhetische, emotionale Komfortmehrwerte wird die Akzeptanz beim Umgang mit einem technischen System erhöht. Mit einer Variabilität bei der Erstellung von informationellen Leistungen kann flexibel auf unterschiedliche Informationsverhalten reagiert werden. Um diesen wirkungsbezogenen individuellen informationellen Mehrwert für einen Nutzer zu erreichen, ist es notwendig, dass mittels der Informationserarbeitung eine Analyse der Daten nach relevanter Information in einem bestimmten Kontext durchgeführt wird. Diese relevante Information wird anschliessend in einer pragmatischen Informationsaufbereitung den unterschiedlichen Bedürfnissen und dem unterschiedlichen Informationsverhalten der Nutzer angepasst. Diese Anpassungen beruhen auf individuellen oder stereotypen Benutzermodellen.

Wurde früher noch der Einsatz einer wissensbasierten Komponente zur Benutzermodellierung verwendet, kommen heute vor allem Empfehlungssysteme zum Einsatz, die personalisierte Information selektieren (Informationserarbeitung) und anbieten (Informationsaufbereitung). Dabei können die Empfehlungen aufgrund von angeklickten und heruntergeladenen Artikeln (inhaltsbasierte Filterung) sowie von anderen bzgl. ihres Profils ähnlichen Benutzern ermittelt werden (kollaborative Filterung) (Reimer, 2014, S. 238). Hybride Ansätze der beiden Vorgehensweisen sind ebenfalls gebräuchlich. Im Rahmen dieser Arbeit erscheint der Ansatz der kollaborativen Filterung als sehr relevant, da diese voraussetzt, dass Benutzer eines Systems miteinander verglichen werden können.

Für die Berechnung der Ähnlichkeiten zwischen Empfehlungsobjekten oder zwischen Quellen können verschiedene Datenquellen verwendet werden. Unter anderem die Benutzerdaten (Demografische Daten, Interessensprofil), Benutzungsdaten (Interaktionen der Benutzer im Informationssystem), Kontextdaten (Kontext, in dem sich der User befindet) sowie Daten über Merkmale der Empfehlungsobjekte (Reimer, 2014, S. 240). Letztere sind

für die kollaborative Filterung nicht relevant, da diese die Benutzer und nicht die Empfehlungsobjekte betrachtet.

Für die Bestimmung ähnlicher Nutzer kann (1) ein benutzerbezogenes Verfahren eingesetzt werden, das Benutzer bestimmt, die dem aktuellen Nutzer ähnlich sind und anschliessend dessen Präferenzen weiterempfiehlt, (2) mittels einem elementbasierten Verfahren ähnliche Objekte anhand von Benutzerpräferenzen bestimmt werden, oder (3) anhand einem modellbasierten Ansatz zunächst ein generalisiertes Modell erstellt wird, mit dessen Hilfe dann Empfehlungen abgeleitet werden können (Reimer, 2014, S. 243).

Während die Ansätze des benutzerbezogenen und elementbasierten Verfahrens auf der gesamten Datenmenge basieren und somit mit steigender Anzahl an Datensätzen ineffizienter werden, unterliegt der dritte Ansatz einem Benutzermodell, das die Benutzer in ähnliche Gruppen segmentiert, so dass die Benutzer innerhalb der Gruppen ähnlichen Präferenzen unterliegen. Die Erstellung von ähnlichen Gruppen nach gemeinsamen Eigenschaften aus Elementen ohne einer vorgegebenen Struktur nennt sich Clustering und kann dem Bereich des Data Mining zugeordnet werden (Mandl, 2014, S. 183). Als Analogie zum Data Mining entstand der Begriff Web Mining und bezieht sich im speziellen auf Daten aus dem Internet (Mandl, 2014, S. 184). Die Entwicklung des Data Mining ist auf das maschinelle Lernen zurückzuführen und beruht auf der Anwendung eines induktiven Verfahrens, um neue Erkenntnisse aus Daten gewinnen zu können.

1.2.4 Verständnis über das Benutzerverhalten erlangen

Während im Standardwerk „Grundlagen der praktischen Information und Dokumentation“ (Kuhlen, Semar, & Strauch, 2014) der Informationswissenschaft im Bereich des Data Mining noch keine Ansätze zur Erkennung des Benutzerverhaltens dokumentiert sind, zeigen Entwicklungen im Bereich des Information Retrieval, dass die beteiligten Akteure mit ihrem kognitiven Hintergrund vermehrt in den Mittelpunkt gestellt werden. Es ist eine Umorientierung weg von der technisch, systemgetriebenen Sicht hin zu einem die Eigenschaften und Spezifika eines Benutzers berücksichtigenden System zu beobachten (Womser-Hacker C. , 2014, S. 336). Gemäss Womser-Hacker (2014, S. 336) befasst sich ein solches System mit der Analyse des Benutzerverhaltens im Umgang mit Information und Wissen und kann als Kognitives Retrieval bezeichnet werden. Mit der Analyse der gewonnenen Erkenntnisse soll die Schnittstelle zwischen Mensch und Maschine verbessert werden.

Während eines dreitägigen strategischen Workshops in Lorne (SWIRL 2012) sind 45 Forscher aus dem Themengebiet des Information Retrieval zusammengekommen und haben Herausforderungen und Möglichkeiten in diesem Forschungsfeld diskutiert (Allan, Croft, Moffat, & Mark, 2012). Gemäss Allan et al. (2012, S. 16) besteht trotz dem weit-

verbreiteten Einverständnis, dass das Verstehen eines Benutzers essentiell für die Erstellung, Verbesserung und Evaluation eines IR System ist, immer noch eine grosse Lücke zwischen der Studie von Benutzern und der Studie von IR Algorithmen. Die Ergebnisse dieser Konferenz zeigen, dass Wissen um den Benutzer in einem gewissen Kontext eine immer wichtigere Rolle für das Information-Retrieval darstellt.

Um Verständnis über das Benutzerverhalten zu erlangen und dadurch das Information Retrieval zu verbessern, schlagen Allan et al. (2012, S. 16) vor, dass die Benutzer vor, während und nach der Nutzung mit einem Informationssystem studiert und unter einer Vielzahl von Methoden untersucht werden. Als Basis dazu dienen gut dokumentierte Forschungsdaten, die für weitere Untersuchungen zur Verfügung gestellt werden. Es werden zwei Vorschläge unterbreitet, wie ein solches Programm aussehen könnte. Mittels der ersten Variante soll eine kontrollierte Überwachung von Personen, die sich in Interaktion mit einem System befinden beobachtet werden. Die zweite vorgeschlagene Variante zielt nicht auf die individuellen Aktionen einer Person, sondern ist ein Logging von (Such-)Interaktionen in einem grösseren Massstab, so dass die Verteilungen der Interaktionsmuster betrachtet werden können. Als Ergebnis einer solchen Forschung schlagen Allan et al. (2012, S. 16) vor, dass eine Sammlung von Datensätzen und eine Infrastruktur entsteht, mittels der die IR Community das Benutzerverhalten untersuchen kann. Die aufgezeigten Ansätze und Erkenntnisse aus dem kognitiven Information Retrieval können als Leitfaden für diese Arbeit genutzt werden, um das Benutzerverhalten genauer zu analysieren.

1.2.5 Operationalisierung

Die Konzeptualisierung zeigt auf, dass eine Notwendigkeit besteht, innerhalb der Informationsverwaltung aus einem vermeintlich grossen Datenbestand (Big Data) mittels einer geeigneten Wissensrekonstruktion ein Smart Data Set bereitzustellen, das besser für eine anschliessende Analyse geeignet ist. Besser bedeutet, dass das Datenset so aufbereitet ist, dass es ohne grössere Aufwände im Folgeprozess der Analyse und Auswertung des Benutzerverhaltens genutzt werden kann. Dazu werden die Smart Data mittels einem kollaborativen Filtering nach ähnlichen Interaktionsmustern innerhalb einer Benutzersession durchsucht. Die Ähnlichkeiten werden mittels einem Clustering über sämtliche Navigationsmuster ermittelt und daraus ein Benutzermodell erstellt, das für die anschliessende Informationsaufbereitung verwendet werden kann. Durch die anschliessende Beobachtung der weiteren Interaktionen der Benutzer, die einem gemeinsamen Modell zugeteilt sind, lassen sich weitere Adaptionen vornehmen. Diese Adaptionen werden wieder in der Informationsverwaltung als neue informationelle Ressourcen abgespeichert und können in einer nächsten Iteration als Input für eine neue Analyse verwendet werden.

Zusammenfassend lassen sich die aus der Konzeptualisierung gewonnenen Erkenntnisse, wie in Abbildung 2 dargestellt, operationalisieren. Die aufgedeckten Terme und Konzepte sollen während der weiteren Literatursuche dazu dienen, tieferen Einblick in die Thematik zu erhalten und sämtliche für das Forschungsproblem relevanten Theorien aufzufinden.

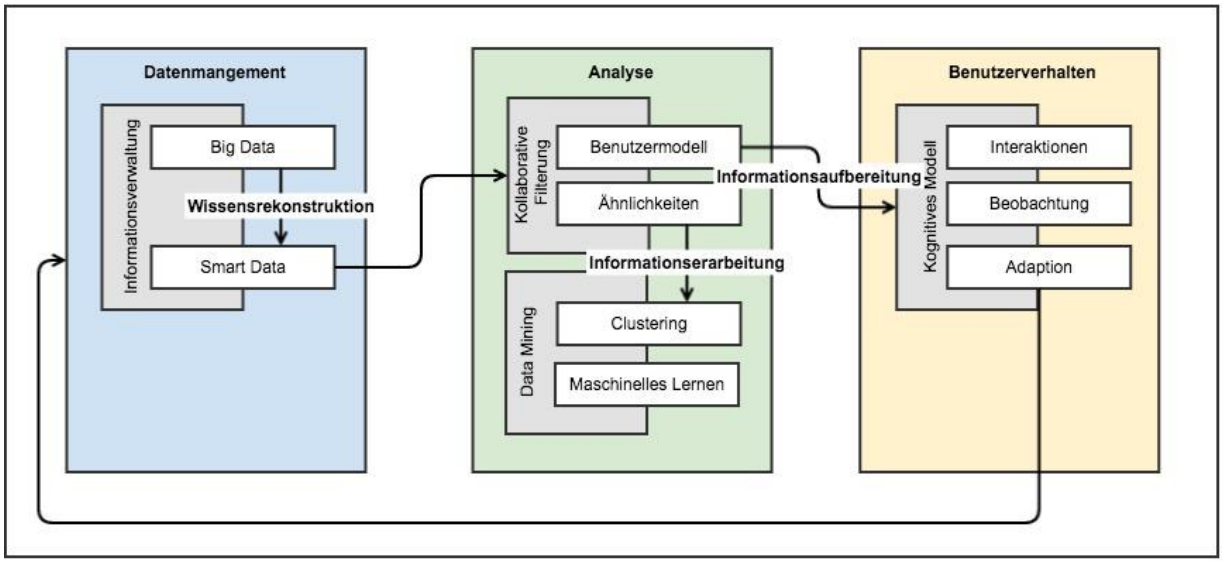

Abbildung 2: Konzeptualisierung (eigene Darstellung)

1.3 Forschungsagenda der Masterarbeit

Für das Customer Engagement soll eine analytische Datenauswertung erarbeitet werden, die die Userbasis der Applikation in Gruppen von ähnlichen Profilen segmentieren kann, um aufbauend auf den Gruppen gezieltere Marketingkampagnen durchzuführen, Produkte basierend auf ähnlichen Präferenzen vorzuschlagen oder mit gezielten Rabatten User anzusprechen.

Diese Arbeit beschäftigt sich mit dieser Customer Engagement Analytics und zeigt auf, wie ein solches System implementiert werden kann, und welche Besonderheiten im Umgang mit den anfallenden Daten beachtet werden müssen. Die Implementierung des Systems beinhaltet die Umsetzung eines Prototyps zur Datenerhebung und Datenauswertung, um die Erkenntnisse aus der Theorie in der Praxis austesten zu können.

1.3.1 Übersicht der Forschungsfragen

Die Forschungsfragen der Arbeit basieren auf den Erkenntnissen aus dem Forschungsproblem und der Konzeptualisierung der Thematik, die in Kapitel 1.1 beschrieben wird.

Forschungsfrage I

Wie muss eine Datenmanagement Plattform für die Informationsverwaltung des Benutzerverhaltens konzipiert werden?

Die Forschung dreht sich um die Frage, wie das Benutzerverhalten gespeichert, weiterverarbeitet und wiederverwendet werden kann. Dafür wird ein System beschrieben, das unter den genannten Faktoren eingesetzt werden kann, um die während der Forschung und in einer späteren Anwendung anfallenden Daten zu speichern.

Forschungsfrage II

Mit welchem Algorithmus lässt sich das aufgezeichnete Benutzerverhalten in verschiedene Gruppen segmentieren?

Mit der zweiten Forschungsfrage soll ein Data-Mining Algorithmus gefunden werden, mittels dem sich das Benutzerverhalten in unterschiedliche Gruppen segmentieren lässt. Sämtliche Algorithmen zu implementieren und testen würde den Rahmen dieser Arbeit überschreiten. Es werden geeignete Algorithmen vorgestellt, evaluiert und die Auswahl des definitiven Algorithmus theoretisch argumentiert sowie mittels einer relativen Validierung miteinander verglichen. Der schlussendlich gewählte Algorithmus wird in einem Prototyp implementiert und mittels den aus der Umfrage erhobenen Testdaten verifiziert.

Forschungsfrage III

Kann durch die Beobachtung der Benutzer vor und während der Interaktion mit einem Informationssystem Gruppierungen mit ähnlichen psychologischen Profilen erstellt werden?

Diese Frage soll beantworten, ob mittels der Beobachtung eines Benutzers vor und während der Interaktion mit einem System eine Aussage über den Kontext eines Benutzers getroffen werden kann. Mittels eines interaktiven psychometrischen Fragebogens soll die Validierung des Clustering-Algorithmus verifiziert werden. Die Validierung soll Erkenntnisse dazu liefern, inwiefern die mit dem Algorithmus aus Forschungsfrage II segmentierten Benutzer einem einheitlichen oder einem zufälligen psychologischen Profil unterliegen. Für die Validierung wird eine externe Validierung verwendet, indem prototypische Profile als Gold Standard genutzt werden, um die Precision sowie den Recall der anhand des Clickstream segmentierten Gruppen bestimmen zu können.

1.3.2 Meilensteine des experimentellen Forschungsdesigns

Abbildung 3: Forschungsdesign (eigene Darstellung)

Grundsätzlich handelt es sich bei dieser Arbeit um eine quantitative Forschung. Anhand des Forschungsproblems ist ersichtlich, dass die Arbeit dem Bereich Big Data zugeordnet werden kann, weshalb für diese Arbeit ein quantitativer Ansatz einem qualitativen vorgezogen wird. Das Forschungsdesign der Arbeit kann der experimentellen Forschung zugeordnet werden. Für die Beantwortung der Forschungsfragen wird in der Arbeit zuerst ein Literaturreview durchgeführt, mittels dem die theoretischen und konzeptuellen Grundlagen erarbeitet werden. Das Literaturreview wird den aktuellen Stand der Forschung im Bereich des Forschungsproblems ausleuchten und aufzeigen, welche Erkenntnisse bereits gemacht worden sind. Diese aus der Literaturanalyse gewonnene Information dient als Ansatzpunkt für die Definition der Forschungsfragen sowie für die Erarbeitung der Theorie, die die Forschungsfragen beantwortet. Die weiteren Untersuchungen werden unter Verwendung eines experimentellen Forschungsdesigns weitergeführt.

Der Prototyp wird in einer kontrollierten Umgebung implementiert, die nach Bedarf auch beeinflusst werden kann. Anhand dieses Prototyps werden die aufgearbeiteten Analysen unter Einbezug von Testdaten ausgetestet und optimiert. Dieser Arbeitsschritt bedarf eines hohen Masses an Flexibilität und Kreativität; deshalb die Wahl einer agilen Vorgehensweise während diesem Schritt der Arbeit.

Die Verifizierung des Experimentes ist in zwei Phasen eingeteilt. In einem ersten Schritt wird der Prototyp auf einem Subset der erhobenen Daten getestet. Anhand dieser Daten sollen Reliabilität und Validität der Algorithmen getestet werden und ein finaler Algorithmus für den zweiten Schritt ausgewählt werden. In einem zweiten Schritt werden weitere Tests mit der gesamten Datenmenge durchgeführt. Für die Datenerhebung wird im Prozessschritt Quantitative Research eine Umfrage erstellt. Der Aufbau dieser Umfrage ist in Kapitel 4.2.1 Quantitative Research und die Auswahl der Teilnehmer in Kapitel 4.2.2 Sampling beschrieben. Die erhobenen Daten werden in einem anschliessenden Schritt analysiert und zur finalen Diskussion, die die Beantwortung der Forschungsfragen beinhaltet aufbereitet.

1.3.3 Abgrenzung

Mit der Erhebung und Speicherung von personenbezogenen Daten unterliegen diese dem Datenschutz und es bedarf somit spezieller rechtlicher Aspekte wie z.B. einer Datenschutzerklärung, die transparent Auskunft über die Datenerhebung und -nutzung gibt. In dieser Arbeit wird auf diese Thematik nicht weiter eingegangen; diese muss separat abgeklärt bzw. definiert werden. Sämtliche in dieser Arbeit erhobenen Daten werden nur in anonymisierter Form verwendet und nicht an Dritte für eine Weiterverarbeitung zur Verfügung gestellt.

Der in dieser Masterthesis erarbeitete Prototyp ist basierend auf Grundkonzepten aus den Ergebnissen der Forschungsfrage I implementiert. Der Prototyp ist jedoch so einfach wie möglich konzipiert. D.h. die Datenerhebung und -speicherung wird so einfach wie nötig gestaltet, um möglichst schnell Ergebnisse erzielen zu können, ohne die gesamte Datenmanagement Plattform implementieren zu müssen. Beispielsweise wird auf eine Orchestrierungskomponente verzichtet und die Arbeit der Orchestrierung wird manuell durchgeführt. Die Lösung ist geeignet für den Einsatz basierend auf wenigen Benutzersessions. Für eine Skalierung der Lösung muss der Prototyp jedoch weiter ausgearbeitet werden.

2 Konzipierung einer Datenmanagement Plattform

Wie in der Konzeptualisierung in Kapitel 1.2 aufgezeigt werden konnte, befasst sich diese Arbeit mit Daten, die zum Zeitpunkt der Analyse einem Machine Learning Prozess unterzogen werden. Gemäss H.J. Miller (zit. in Kitchin, 2014, S. 103) ist Machine Learning nicht einfach eine automatisierte Wissenschaft, die auf Knopfdruck angewendet werden kann, sondern erfordert Domänenexpertise. Um einen Systemdesignvorschlag für das Forschungsproblem zu erarbeiten, ist es daher notwendig, die Domäne, in der sich das Problem befindet zu kennen. Anhand der Domäne lassen sich die Benutzer des Systems definieren und dadurch auch die Daten kategorisieren, die der Benutzer im System hinterlässt und durch ein System verarbeitet werden können. Anhand dieser Erkenntnisse wird abschliessend zur Beantwortung der Forschungsfrage I eine Systemarchitektur erarbeitet.

2.1 Eingrenzung der Forschungsdomäne

In diesem Kapitel wird die Domäne erarbeitet, in der das Forschungsproblem anzugliedern ist. Basierend auf der Konzeptualisierung werden die operationalisierten Begriffe weiter verfeinert und mittels einer Trendanalyse in den heutigen Kontext gebracht.

Web Usage Mining

Ausgehend von dem in Kapitel 1.2.2 vorgestellten Begriff Data Mining und der Analogie Web Mining soll die Domäne noch enger gefasst werden können. Eine erste Unterteilung von Web Mining findet sich von Cooley et al. (1997) und unterteilt das Gebiet in Web Content Mining und Web Usage Mining. Auch Kosala & Blockeel (2000) haben erkannt, dass der Begriff Web Mining zu unklar ist, weshalb sie eine Kategorisierung des Begriffs in die drei Unterkategorien Web Content Mining, Web Structure Mining und Web Usage Mining vorgenommen haben. Während im Web Content Mining hauptsächlich die Inhalte von Webseiten und im Web Structure Mining die Topologien der Webseiten untersucht werden, behandelt das Web Usage Mining die Analyse der Nutzung von Webseiten. Cooley et al. (1997) definieren Web Usage Mining als automatisierte Erkundung von Benutzer-interaktionen mit einem Webserver. Kosala & Blockeel (2000) sehen die Einsatzgebiete von Web Usage Mining speziell im Bereich des Marketings und der Benutzermodellierung, um den Aufbau von Seiten durch Adaption an Benutzermodelle zu verbessern.

Trendanalyse

Da der Begriff Web Usage Mining (Cooley, Mobasher, & Srivastava, 1997) als Domäne bereits ein wenig in die Jahre gekommen ist, wird an dieser Stelle die Trendentwicklung der

letzten Jahre betrachtet, um daraus einen Rückschluss auf die heutige und aktuelle zuordnungsbare Domäne zu ziehen. Dazu kann ein Blick auf den Hype Cycle der Emerging Technologies von Gartner geworfen werden, der jährlich erscheint und die aktuellsten Trends sowie deren Verlauf über die letzten Jahre aufzeigt.

Der aktuellste Hype Cycle von Gartner (siehe Abbildung 4) beinhaltet eine neue Emerging Technology in der Phase „Innovation Trigger“, die für das Forschungsproblem geeignet zu sein scheint. Die Rede ist von Smart Data Discovery, welche als eine Next-Generation Data Discovery Fertigkeit eingestuft wird (Sallam & Parenteau, 2015). Smart Data Discovery hat das Ziel, möglichst einfach auch einem nicht ausgebildeten Data Scientist Zugang zu fortgeschrittenen Analysen zu gewähren.

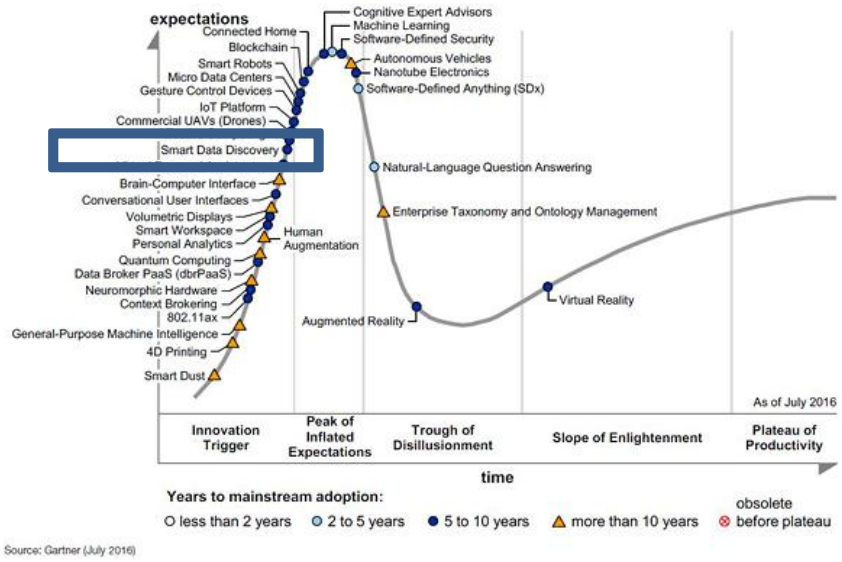

Abbildung 4: Hype Cycle (Gartner 2016)

Smart Data Discovery

Smart Data Discovery Applikationen unterstützen ihre Benutzer in den drei Phasen der Aufbereitung der Daten, dem Ermitteln von Patterns in den Daten und der Vermittlung und Operationalisierung der Ergebnisse (Sallam & Parenteau, 2015). Das klassische Web Usage Mining beinhaltet die Phasen des Preprocessing, der Knowledge Discovery und der Pattern Analysis (Cooley, Mobasher, & Srivastava, 1999). Eine Gegenüberstellung dieser Phasen in der Tabelle 1 zeigt, dass sich diese Phasen decken und Smart Data Discovery somit durchaus als moderne Variante des Web Usage Mining angesehen werden kann.

Phasen	Web Usage Mining	Smart Data Discovery
Phase 1	Preprocessing	Aufbereitung der Daten
Phase 2	Knowledge Discovery	Ermitteln von Patterns in den Daten
Phase 3	Pattern Analysis	Vermitteln und Operationalisierung der Ergebnisse

Tabelle 1: Gegenüberstellung der Data Mining Phasen

In einem Whitepaper von Aspire Systems werden die Vorzüge von Smart Data Discovery im Bereich des Retail aufgezeigt, im Bereich, in dem mit dem Endkunden in Kontakt getreten wird (Ghosh, 2015). Gemäss Gosh (2015) können mittels einem Smart Data Discovery Model aktionsrelevante Einsichten von Real Time Data genutzt werden, um den Wert der digitalen Strategien zu maximieren. Die Rede ist von einer Real Time basierten 360 Grad Ansicht des Kunden als Treiber für diese Wertsteigerung. Über diese Real Time Ansicht des Kunden wird ermöglicht, ein Kundenverständnis über das Kaufverhalten, die Zeit sowie den Kanal zu erlangen. Mittels diesen Erkenntnissen über einen Kunden können Kundenprofile erstellt werden, die in Modellen zusammengefasst und basierend auf diesen eine Segmentierung von Kundengruppen vorgenommen wird. Anhand dieser Kundengruppen kann dem Benutzer nun eine massgeschneiderte Customer Experience angeboten werden, welche Kundenloyalität aufbauen und so eine Kundenabwanderung verhindern soll.

2.2 Aufbereitung der Daten (Preprocessing)

Wie im vorangegangen Kapitel 2.1 aufgezeigt wurde, basiert „Smart Data Discovery“ (im Kontext des Forschungsproblems) auf einer Modellierung des Benutzers anhand von Daten, die aus den Interkationen eines Nutzers mit einem System entstehen. Um ein solches Modell zu erstellen, mittels dem Vorhersagen darüber getroffen werden können, was ein Individuum unter einem anderen Umstand und an einem anderen Ort tun könnte, ist es notwendig, die Daten zu sammeln und angemessen zu strukturieren (Kitchin, 2014, S. 43). Aus der Strukturierung der Daten entstehen sogenannte abgeleitete Daten (Derived Data) sowie individuelle oder flächendeckende Profile der Nutzer, die für die Erstellung des Vorhersagemodells genutzt werden können. Für die Beantwortung der Forschungsfrage ist es daher notwendig zu wissen, welche Daten vom System aufzubereiten sind und in welche Struktur diese gebracht werden müssen, um diese in einem späteren Schritt auf gemeinsame Muster hin zu untersuchen.

Welche Art von Daten aufgezeichnet werden, wird gemäss Kitchin (2014, S. 28) geprägt durch das Sampling Frame (wo die Daten anfallen), die eingesetzte Technologie und Plattform, dem Kontext, der den Daten zugrundeliegenden Ontologie sowie dem regulatorischen Umfeld. Daher wird in den folgenden Kapiteln ein grundlegendes Verständnis über die Kategorisierung der Daten für dieses Forschungsproblem aufbereitet, um darauf aufbauend die Informationsarchitektur von Online-Services zu analysieren. Anhand der Informationsarchitektur lassen sich dann Metriken für die Modellierung von Benutzermodellen aus den Navigationspfaden ableiten. Basierend auf diesen Metriken kann ein Schema abgeleitet werden, das definiert, welche Daten überhaupt aufbereitet werden müssen. Sämtliche in diesen Kapiteln aufgezeigten theoretischen Grundlagen dienen als Basis für die Systemarchitektur in Kapitel 2.3.2.

2.2.1 Kategorisierung der Daten

Mittels der Kategorisierung der Daten wird ein grundlegendes Verständnis der Daten erarbeitet, welche bei der Beobachtung des Benutzerverhaltens während der Interaktion mit einem System anfallen können. Gemäss Kitchin (2014, S. 3) können Daten nach den folgenden fünf Kriterien kategorisiert werden:

1. Form der Daten: Qualitativ oder Quantitativ.
2. Strukturierung der Daten: Strukturiert, Semi-Strukturiert und Unstrukturiert.
3. Quelle der Daten: Captured, Derived, Exhaust, Transient.
4. Produzent der Daten: Primär, Sekundär oder Tertiär.
5. Typ der Daten: Indexierend, Attribut, Metadatei.

Anhand dieser fünf Kriterien lassen sich für das Forschungsproblem die folgenden Einschätzungen an das System ableiten:

Kriterium	**Einschätzung Forschungsproblem**
Form	Im vorliegenden Forschungsproblem muss davon ausgegangen werden, dass nicht nur mit quantitativen Daten gearbeitet werden kann, weshalb der Einsatz von Machine Learning für die Weiterverarbeitung genutzt werden muss.
Strukturierung	Von Vorteil für eine anschliessende Analyse ist, wenn die gesammelten Daten in strukturierte oder zumindest semi-strukturierter Form vorliegen.
Quelle	Mit der Aufzeichnung des Clickstreams eines Benutzers wird versucht, aus den „Exhaust Data“ abgeleitete Daten zu erstellen,

	um diese anschliessend strategisch wertvoll einzusetzen.
Produzent	Die Daten werden in Form von Primärdaten erhoben, da nicht vorgesehen ist, diese an weitere Forscher zur Verfügung zu stellen.
Typ	Bei der Überführung der Exhaust Data in Derived Data muss darauf geachtet werden, das indexierende Daten mit aufgenommen werden, die eine Identifikation der Daten zulassen. Mittels Metadaten sollen die Daten so beschrieben werden, dass in der anschliessenden Analyse ein Verständnis der Daten ermöglicht wird.

Tabelle 2: Kategorisierung der Daten

2.2.2 Informationsarchitektur

Um zu verstehen, welche Daten aus den Interaktionen eines Benutzers mit einem System gesammelt werden können, ist es notwendig, die Informationsarchitektur eines Systems zu kennen. Der Begriff Informationsarchitektur wurde von Richard S. Wurman eingeführt und beschreibt, wie Daten so als Information dargestellt werden, dass sie von einem Benutzer des Systems verstanden werden können (Wurman, 2000). Die Informationsarchitektur beschreibt, wie die Information in einem System gruppiert ist, wie sie dem Nutzer präsentiert wird und wie die Navigation durch das System strukturiert ist (Nyman, 2013). Informationsarchitektur ist eine Kombination aus Organisation, Verschlagwortung, Suche und Navigation innerhalb einer Website oder eines Intranets (Morville & Rosenfeld, 2006). Nyman (2013) kategorisiert die Architektur von Webseiten in vier verschiedene Designs: (1) Matrix Design, (2) Hierarchisches Design, (3) Tunnel Design und (4) Hybrides Design.

Eine Website im Matrix Design ist so aufgebaut, dass ein Benutzer jederzeit auf jede Seite der Webseite zugreifen kann. Der Benutzer kann so frei seine individuellen Interessen verfolgen. Das Hierarchische Design verfolgt einen Top-Down Ansatz, der zuerst Information von genereller Natur präsentiert und diese immer spezifischer gestaltet, je tiefer sich der Benutzer in der Navigationsstruktur bzw. im Navigationsbaum befindet. Das Tunnel Design ist das Gegenteil des Matrix Designs (Nyman, 2013) und lässt dem Benutzer gar keine Wahl in der Navigation. Sämtliche Seiten sind sequentiell hintereinandergestellt und es gibt im System somit nur einen möglichen Navigationspfad. Unter dem Hybriden Design versteht Nyman (2013), dass die Module der Website unterschiedlich aufgebaut sein können. Während einige Teile z.B. einem hierarchischen Design unterstellt sind, kann der Prozess beim Kauf eines Produktes aus dem Warenkorb dem Tunnel Design folgen.

Aus diesen vier verschiedenen Informationsarchitekturen lässt sich schliessen, dass das hierarchische Design am besten dazu geeignet ist, um eine Profilierung der Nutzer durch Interaktionen mit einem System vorzunehmen. Während das Matrix Design dem Nutzer zu viele Freiheiten lässt und die Anzahl unterschiedlicher Navigationspfade somit sehr hoch wird, ergibt sich für das Tunnel Design für jeden Nutzer denselben Navigationspfad. Das hierarchische Design befindet sich in Bezug auf die Navigationsfreiheit zwischen den anderen beiden Designs und erlaubt eine gewisse Individualität, schränkt diese jedoch auch ein, so dass die Anzahl an Navigationsmöglichkeiten nicht unbegrenzt sein können.

2.2.3 Metriken für die Modellierung eines Benutzermodells

Belk et al. (2013) haben die Modellierung von kognitiven Modellen untersucht und Metriken für das Tracking des Navigationsverhalten von Benutzern aufgezeigt. Benutzermodelle können gemäss Belk et al. (2013) aus expliziten oder auch impliziten Daten des Benutzers erhoben werden. Die Generierung eines Benutzermodells aus den expliziten Daten ist die einfachere, weil diese Daten direkt übernommen werden können. Beispielsweise könnte der Nutzer in seinen expliziten Daten seine Vorzüge bekannt geben, und das System adaptiert sich auf diese. Intelligentere Verfahren greifen auf die impliziten Daten zurück, um dann mittels Data Mining und Machine Learning Ansätzen Gemeinsamkeiten in den Navigationsmustern zu finden und daraus ein Benutzermodell abzuleiten.

Unter expliziten Benutzerdaten werden Daten verstanden, die durch den Benutzer zur Verfügung gestellt werden. Dies sind unter anderem demographische Merkmale des Benutzers wie z.B. das Alter, Geschlecht oder der Beruf sowie persönliche Interessen und Vorlieben (Belk, Papatheocharous, Germanakos, & Samaras, 2013). Erhoben werden können diese Daten über eine Profilseite des Benutzers und können beispielsweise Teil des Registrierungsprozesses sein. Ein Nachteil dieser expliziten Daten ist, dass diese durch den Benutzer nicht oder ungenau ausgefüllt werden können.

Als implizite Benutzerdaten werden dynamisch generierte Daten verstanden, wie z.B. das Navigationsverhalten des Benutzers. Das Navigationsverhalten eines Nutzers entspricht einem Clickstream (siehe Abbildung 5), also einer seriellen Abfolge von Interaktionen mit dem System.

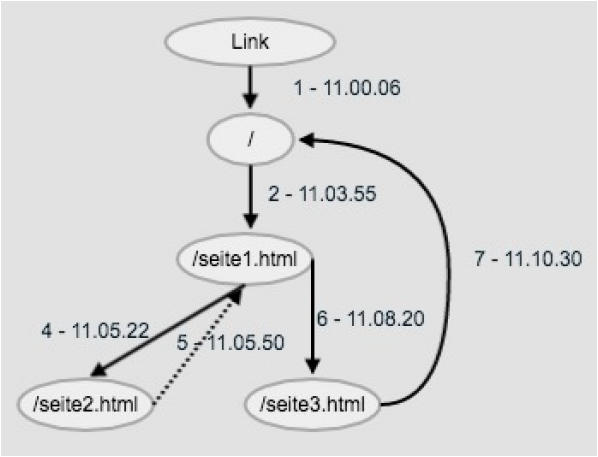

Abbildung 5: Clickstream (eigene Darstellung)

Die implizite Datenerhebung bezeichnen Belk et al (2013) als einen für den Benutzer transparenten Mechanismus, der ihn nicht in seiner Tätigkeit stört und somit auch keinen Zusatzaufwand vom Benutzer benötigt. Diese Daten können aus zwei Quellen erhoben werden. Möglich ist die nachträgliche Erhebung der Daten aus der „Browsing History“ des Benutzers, also aus den aufgezeichneten Daten, die auf einem Proxy oder in einem lokalen Cache des Benutzers über längere Zeit gespeichert werden. Ein Vorteil dieses Ansatzes ist sicherlich, dass die Daten rückwirkend analysiert werden können und keine spezielle Software für die Erhebung der Daten benötigt wird. Eine zweite Möglichkeit ist, die Erhebung der Daten direkt auf dem Client-System vorzunehmen. Belk et al (2013) bezeichnen dies als den Einsatz eines Agenten, der beispielsweise als Browser Plugin implementiert werden kann. Der Vorteil dieses zweiten Ansatzes ist, dass die Daten direkt bei der Entstehung erhoben werden können und nicht erst nachträglich aus einem Speicher heraus extrahiert werden müssen. Gemäss Kitchin (2014, S. 76) kommen Systeme heute beim Aufzeichnen und Verarbeiten von Daten an ihre Grenzen. Als Lösung für dieses Problem sieht Kitchin (2014, S. 77) eine Datenvorsortierung, die den Umfang der Daten vor der Übertragung und der Verarbeitung reduziert. Dies ist somit der grosse Vorteil der Datenerhebung direkt auf dem Client-System. Es kann bereits früh im Prozess eine Vorselektion der relevanten Daten vorgenommen werden und dadurch das Preprocessing der Daten effizienter gestaltet werden. Wird eine Real-Time Ansicht eines Benutzers benötigt, ist diese Effizienz notwendig, um mit einer geeigneten Geschwindigkeit auf die Interaktionen des Benutzers reagieren zu können. Ein weiteres Problem, das durch den Einsatz eines Client-System gelöst wird, ist dasjenige des Cachings. Gemäss Cooley et al. (1999) unterliegt das Web Usage Mining basierend auf Server Log Files dem Problem, dass sowohl der Browser als auch ein Proxy Seiten zwischenspeichern kann. Somit können nicht alle durch einen Benutzer durchgeführten Schritte anhand eines Server Log Files reproduziert werden. Als Lösung für das

Problem setzen Cooley et al. (1999) auf ein sogenanntes „Path Completion“, das während dem Preprocessing versucht, Rückschlüsse auf Lücken zu finden und diese bestmöglich anhand von „Referrer“-Daten schliesst. Dieses grosse Problem kann beim Einsatz des Client-System umgangen werden, da die Benutzernavigationen direkt im Browser aufgezeichnet und an das Zielsystem gesendet werden. Das Caching-Problem kann so elegant umgangen werden. Als Nachteil des zweiten Ansatzes kann hier zusätzliche Software aufgeführt werden, die bei diesem Vorgehen benötigt wird.

Julian und van Oostendorp (2006) haben in ihrer Studie „Individual differences and behavioral metrics involved in modeling web navigation“ Metriken ausgearbeitet, die bei der Modellierung der Navigation in einem webbasierten System angewendet werden können und auf impliziten Daten basieren. Dabei unterscheiden sie zwischen den zwei Metriken „First-Order“ und „Second-Order“. First-Order Metriken sind Daten, die direkt aus den Rohdaten bzw. den Captured Data abgeleitet werden können. Die Second-Order Metriken sind gemäss dem Forscherpaar lineare Kombinationen der „First-Order“ Metriken. Insgesamt haben Juvina und van Oostendorp (2006) 19 First-Order Metriken definiert. Betrachtet man die einzelnen Metriken, wird schnell ersichtlich, dass es notwendig ist einen Identifier jeder Seite inklusive dem genauen Zeitpunkt des Besuchs zu speichern, um daraus später die Pfadlänge, Pfaddichte oder auch die durchschnittliche Betrachtungszeit ableiten zu können.

Cooley et al. (1999) haben einen Architekturvorschlag für ein WEBMINER System erarbeitet, das ähnliche Datenerhebungen voraussetzt, wie es für die von Juvina und van Oostendorp (2006) vorgestellten Metriken notwendig ist. Cooley et al. (1999) beschreiben die Phase des Preprozessing, das Inputdaten so aufbereitet, dass sie für das anschliessende Knowledge Discovery genutzt werden können. Das Preprozessing beinhaltet unter anderem die Schritte der Benutzer-, Session- und Transaktionsidentifikation. Die Benutzeridentifikation ist kein einfaches Unterfangen, da durch den Einsatz von Proxies nicht jeder Benutzer einwandfrei anhand der IP-Adresse identifiziert werden kann (Cooley, Mobasher, & Srivastava, 1999). Daher müssen für die Benutzeridentifikation neben der IP-Adresse noch ein weiteres Merkmal, ein sogenannter Fussabdruck des Systems (Browser, Operation System), der der Benutzer nutzt, mit einbezogen werden. Da die Sammlung und Auswertung von Navigationsdaten ein dauerhafter Prozess in der Real-Time Modellierung eines Benutzers ist, müssen die Sessions der einzelnen Benutzer ebenfalls identifiziert werden können. Eine einfache Identifizierung kann über den Zeitabstand zwischen zwei Aufrufen ermittelt werden (Cooley, Mobasher, & Srivastava, 1999). Gängiger ist heutzutage jedoch der Einsatz von Cookies, die einen Session Identifier abspeichern. Die Transaktionsidentifizierung dient dem Zweck, später aussagekräftige Cluster bilden zu können. Daher ist es notwendig, Transaktionen zu identifizieren, die in die Gruppierung mit aufgenommen werden können (Cooley, Mobasher, & Srivastava, 1999). Neben der URI als eindeutigen Identifier für eine Webseite müssen in heutigen System noch weitere Interaktionsmöglichkeiten in Betracht gezogen werden. Viele

Online-Systeme werden vermehrt als One-Page Applikationen konzipiert. Dies bedeutet, dass vom Server nur eine Seite geladen wird und danach sämtliche Interaktionen eines Nutzers mit dem System innerhalb dieser Seite durchgeführt werden und sich demzufolge die Navigationspfade nicht alleine aus der URI ableiten lassen. Daher ist es notwendig, solche Transaktionsinteraktionen bei der Datenaufbereitung ebenfalls zu berücksichtigen. Mögliche Aktionen sind dabei unter anderem das Absenden von Formularen, das Ein- und Ausblenden von Elementen auf der Seite sowie die Navigation innerhalb dieser One-Page Struktur. Des Weiteren werden in neueren Applikationen auch vermehrt Social- und Multimediainhalte genutzt. Ein Nutzer kann z.B. seine Emotionen in Form von Likes oder Favoriten ausdrücken oder auch Inhalte wie Video- oder Audiodateien nutzen.

Zusammenfassend zeigt die Tabelle 3 sämtliche impliziten Daten, die für die Modellierung eines Benutzermodells berücksichtigt werden sollen.

Gruppe	**Bezeichner**	**Erklärung**
Benutzeridentifikation	ipAddress	Zugriffs-IP Adresse des Benutzers
	userAgent	Eindeutige Kennzeichnung des eingesetzten Browsers sowie des Betriebssystems, welches genutzt wird.
Sessionidentifikation	sessionId	Eindeutiger Session Identifier des Benutzers, um eine Gruppierung der Nutzerinteraktionen zu einer bestimmten Session einordnen zu können. Umgeht zudem das Problem, wenn sich mehrere Nutzer einen Proxy Server teilen, und dadurch dieselbe IP Adresse haben. Gespeichert werden kann diese Information in einem Cookie.
Transaktionsidentifikation	location	Eindeutige Kennzeichnung jeder Seite, welche innerhalb des System besucht werden kann. Dazu wird die URI der Seite verwendet.
	timestamp	Genauer Zeitpunkt, wann eine Seite auf dem Client aufgerufen worden ist.
	duration	Aufenthaltsdauer auf einer bestimmten Seite. Durch den Einsatz von einem Client-Seitigen System für die Erhebung von Daten, kann die Aufenthaltsdauer direkt auf dem Client berechnet werden und gestaltet das Preprocessing effizienter.

	event	Spezielle Interaktionen auf einer Seite. Diese sind applikationsspezifisch und müssen daher individuell definiert werden. Mögliche Werte sind: Like, Favorite, BackToOverview, Scroll, PlayVideo, PauseVideo, StopVideo, PlayAudio, ... etc.

Tabelle 3: Clickstream Schema

2.3 Systemdesign einer Datenmanagement Plattform

In den kommenden Kapiteln wird der Aufbau einer Datenmanagement Plattform mit den erwähnten Komponenten aufgezeigt und die Einsatzbarkeit mittels einem Fragenkatalog verifiziert.

2.3.1 Hauptkomponenten einer Datenmanagement Plattform

Bei der Sammlung von Clickstream Daten entsteht ein grosses Datenvolumen, das in kurzer Kadenz gespeichert werden muss. Unter diesem Aspekt sind konventionelle relationale Datenbanken nicht mehr einsetzbar und haben Probleme, mit diesen erhöhten Anforderungen klarzukommen (Kitchin, 2014). Gemäss Driscoll (2012) sind NoSQL Datenbanken besser dazu geeignet, Daten zu speichern, die noch nicht einer vorbestimmten Relation unterliegen. Ein Beispiel für eine solche NoSQL Datenbank im Big Data Bereich ist das Open-Source Framework Hadoop, mittels dem MapReduce Funktionalitäten angeboten werden (Kitchin, 2014, S. 87).

Dass Hadoop auch für das Sammeln und Analysieren von Clickstream Daten geeignet ist, zeigen Grover et al. (2015) anhand eines Architekturdesigns basierend auf Hadoop. Die Autoren argumentieren, dass auf einer aktiven Webseite täglich Gigabyte an Logdaten generiert werden können, und das Speichern und Analysieren von solch immensen Daten ein robustes und verteiltes System benötigt. Alles Charakteristiken, für welche Grover et al. (2015) Hadoop als sehr geeignet einstufen.

Die Designüberlegungen von Grover et al. (2015) beinhalten fünf Bereiche, die es zu evaluieren gilt: Storage, Ingestion, Processing, Analyse und Orchestration. Unter Storage werden die Überlegungen rund um das Speichersystem verstanden (Datenformat, Datenmodelle, Kompression etc.). Als Ingestion wird die Phase bezeichnet, in welcher die Rohdaten entgegengenommen werden, um anschliessend im Speichersystem für die anschliessende Verarbeitung und Analyse abgelegt zu werden. Die Orchestrierung automatisiert die Koordination der einzelnen Schritte der Ingestion, Prozessing und Analyse untereinander. Adersberger (2016) nutzt in seinem Architekturvorschlag für eine Clickstream

Applikation ebenfalls die von Grover et al. (2015) vorgeschlagenen Phasen der Ingestion und des Processing, führt jedoch als erste Phase die Collection, also die eigentliche Sammlung der Rohdaten auf. Ein Clickstream-Architektur Vorschlag von Rajagopalan (2016) listet ähnliche Phasen auf, die in einem Architekturdesign berücksichtigt werden müssen. Rajagopalan (2016) sieht die fünf Bereiche Data, Aquire, Organise, Analyse und Decide.

Im Rahmen der Forschungsfrage I sind vor allem die Bereiche Collection, Ingestion und Storage relevant, während Processing und Analyzing Folgeprozesse sind, die auf den ersteren Aufbauen und die Orchestrierung die einzelnen Prozesschritte zusammenhält. Dieser Designvorschlag deckt sich auch mit dem von Belk et al (2013) vorgeschlagenen Agenten, der genutzt werden kann, um die impliziten Daten in strukturierter Form zu sammeln (Collection) und an eine Schnittstelle zu senden (Ingestion), die diese Daten dann in einem Datenbehälter (Storage) für eine spätere Benutzung abspeichert (siehe auch Kapitel 2.2).

2.3.2 Systemarchitektur

Wie in Kapitel 2.3.1 beschrieben, soll das Design die drei Bereiche Collection, Ingestion und Storage beinhalten, um Clickstream Daten sammeln, aufnehmen und speichern zu können.

Der von Belk et al. (2013) vorgeschlagene Einsatz eines Browser Plugin für die Sammlung von Daten hat den Nachteil, dass der Benutzer dieses Plugin lokal auf dem Rechner installieren muss. Ein System, das keine Installationen vom Benutzer benötigt, ist ein besserer Designansatz für die Implementation des Agenten. Daher wird in diesem Systemdesign ein javascriptbasierter Agent vorgeschlagen, der ohne jegliche Interaktionen seitens des Benutzers eingesetzt werden kann.

Ein Framework, das die Verwendung eines javascript Agenten berücksichtigt, ist der Divolte Collector[1]. Das Framework wurde speziell für den Einsatz bei der Sammlung, Entgegennahme und Abspeicherung von Clickstream Daten entworfen. Die Abbildung 6 verschafft einen Überblick, aus welchen Komponenten das System besteht.

[1] http://divolte.io/

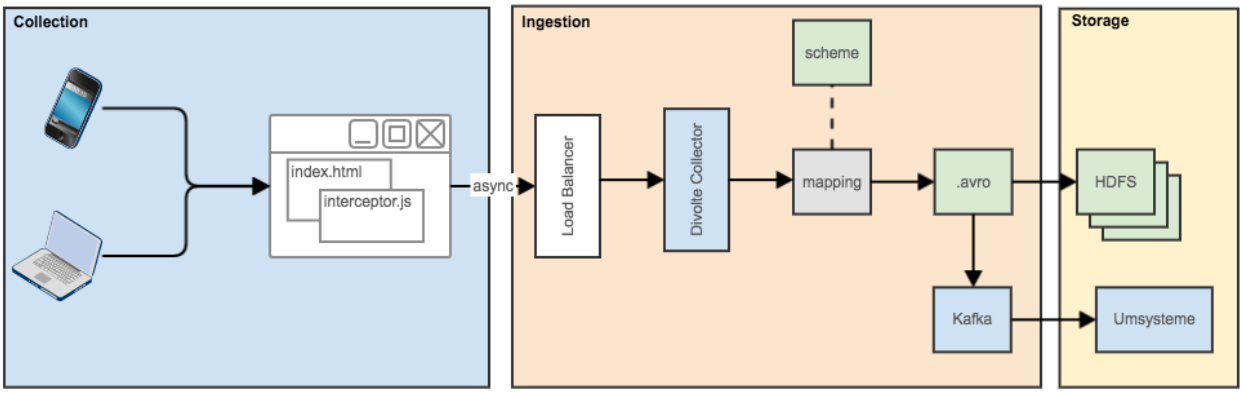

Abbildung 6: Systemarchitektur (eigene Darstellung)

Collection

Die Phase der Datensammlung basiert auf einem Javascript-Interceptor, der sämtliche Click- oder Custom-Events, die in einem Browser generiert werden, asynchron (im Hintergrund) an einen Server sendet, auf dem ein Divolte Collector läuft. Der Collector behindert somit nicht den Arbeitsfluss eines Benutzers und hat zudem den Vorteil, dass die Clickstream Daten durch die Sammlung am Ort des Entstehens auch in Real-Time verteilt werden können und dadurch eine schnellere Analyse der Daten und somit ein direkteres Profiling des Benutzers zulassen.

Ingestion

Für die Phase der Datenentgegennahme ist der Divolte Collector zuständig. Der Collector sammelt die empfangenen Rohdaten und bildet diese auf ein vordefiniertes Schema basierend auf dem Datenformat Avro ab. Während dieser sogenannten Mapping-Phase besteht zudem die Möglichkeit, weitere Sekundärdaten zu den empfangenen Rohdaten zu linken. Dieser Schritt wird als sogenannte Datenanreicherung oder Data Linking bezeichnet. Diese Datenanreicherung hat einen verstärkenden Effekt auf die Daten und gewährt dadurch Einsichten, die durch ein alleinstehendes Datenset nicht ermöglicht werden (Crampton, et al., 2012). Im Kontext der Clickstream Daten kann beispielsweise ein Geolocation Service genutzt werden, um mittels der IP-Adresse genauere Geodaten zu ermitteln und mit den Clickstream Daten abzuspeichern.

Während der Ingestion kann der Divolte Collector enorm unter Last stehen, da er zentral für die Sammlung der Daten zuständig ist und somit als Single Point of Entry angesehen werden kann. Um zu verhindern, dass die gesamte Ingestion nicht mehr läuft, da der Divolte Collector die Abarbeitung der Clickstream Daten nicht genügend schnell vornehmen kann, muss im Systemdesign ein Load Balancer vorgesehen werden, der mehrere Divolte Collectors ansprechen kann. Der Divolte Collector ist zustandslos und folgedessen müssen

die einzelnen Collector voneinander keine Kenntnisse haben. Der Einsatz des Load Balancer erlaubt, eine hochverfügbare und skalierbare Ingestion zur Verfügung zu stellen.

Storage

Die im Avro-Datenformat gespeicherten Daten werden daraufhin vom Divolte Collector im Hadoop Filesystem HDFS abgespeichert. HDFS ist ein geeignetes Dateisystem, weil das nachfolgende Processing meist auf Batch-Transformationen über mehrere Datenrecords ablaufen wird. HDFS wird von Grover et al. (2015) als Dateiformat vorgeschlagen, da dies effizient Batch-Operationen in einem grossen Umfang zulässt. Auch im späteren Schritt der Analyse, in der Anfragen aggregiert werden und Analysen auf grossen Datensätzen durchgeführt werden, sehen Grover et al. (2015) HDFS als geeignet an. Als Datenformat schlagen Grover et al. (2015) ein spaltenbasiertes Format vor, weil die Analysen meist auf einem eingeschränkten Spaltenbereich durchgeführt werden. Das vom Divolte Collector verwendete Avro-Format ist ein solches spaltenbasiertes Format und eignet sich daher als Dateiformat für eine spätere Analyse. Avro ist ein Serialisierungsformat und wird dazu genutzt, um Daten von einem Prozess zu einem anderen zu transferieren oder Daten in einem Filesystem wie z.B. HDFS abzuspeichern (Grover, Malaska, Seidman, & Shapira, 2015). Die Daten selbst werden in Binärform im Avro abgespeichert. Dies macht die Dateien sehr kompakt und mithilfe von Markern in den Dateien können Map Reduce Operationen die grossen Files schnell und effizient verarbeiten.

Eine weitere Überlegung, die im Design des Speichers berücksichtigt werden muss, ist, wie lange die Daten gespeichert werden sollen. Grover et al. (2015) wählen beim Design von Hadoop Applikationen meist eine Speicherung der Daten auf unbestimmte Zeit, anstelle diese nach der Prozessierung wieder zu löschen. Dadurch können die Daten einfach nochmals verwendet werden, falls während dem ETL Prozess ein Fehler auftreten sollte. Auch können die Daten zu einem späteren Zeitpunkt unter einem anderen Aspekt analysiert werden und so Einblicke ermöglichen, an die nach einer ersten Verarbeitung gar nicht gedacht worden ist. Nicht zuletzt ist das Aufheben der Daten auch für Audit-Zwecke geeignet, sollte zu einem späteren Zeitpunkt eine Reproduktion notwendig sein. Aus den genannten Gründen wird im vorliegenden Systemdesign kein Löschen der gesammelten Daten vorgenommen, und die Avro-Files auf unbegrenzte Zeit im HDFS gespeichert.

2.3.3 Verifikation anhand Fragebogen für moderne Datenmanagementkonzepte

Obwohl Big Data in aller Munde ist, finden sich in der Literatur noch wenig wissenschaftliche Artikel und Texte zum Aufbau und der Verifikation von Big Data Architekturen. Die Industrie 4.0 wurde zwar ausgerufen, jedoch hat ein Grossteil der Unternehmen noch keine Infrastruktur aufgebaut, um das Ziel einer informationsbetriebenen Wirtschaft im Zeitalter von

Big Data zu erreichen (Linked Enterprise Data Service, 2016). Isele und Arnd (2016) haben sich der Herausforderung angenommen und sieben qualitativ-konzeptionelle Fragen formuliert, die ein modernes Datenmanagementkonzept beantworten muss. In diesem Kapitel wird auf die sieben Fragen eingegangen und aufgezeigt, ob die Systemarchitektur aus Kapitel 2.3.2 diese Fragen klärt.

Frage 1

Wie kann die Rückverfolgbarkeit von Daten sichergestellt werden, um sowohl Authentizität als auch Richtigkeit und Kontext der Daten zu überprüfen?

Durch die iterative Informationsarbeit (siehe Kapitel 1.2.1) werden einmal in der Informationsverwaltung abgespeicherte Ressourcen immer wieder verändert. Ein starkes Qualitätsmanagement umfasst daher die Definition von eindeutigen Identifier in den Metadaten sowie der Nachverfolgbarkeit der Daten zu den Ursprungssystemen (Isele & Arnd, 2016). Der Divolte Collector bietet mit dem Mapping die Möglichkeit, die notwendigen Attribute in den Metadaten aufzunehmen. Die notwendigen Identifier wie z.B. die IP-Adresse oder ein Session Identifier werden wie bereits in Kapitel 2.2.3 beschrieben im Clickstream mitgesendet und müssen somit nur noch auf das Avro-File abgebildet werden. Zusätzlich kann an dieser Stelle noch ein weiteres Attribut hinzugenommen werden, das den Kanal eindeutig identifiziert. Über diesen Kanal können die aufgezeichneten Clickstream-Daten zu ihrem Ursprungssystem zurückverfolgt werden. Im vorliegenden Fall kann der Kanal aus dem User Agent gebildet werden.

Frage 2

Wie können Inhalte inventarisiert werden, um eine sinnvolle Ablage und schnelle Wiederauffindbarkeit zu gewährleisten?

Neue Business Intelligence entsteht nur, wenn grosse und teilweise heterogene Datensätze miteinander verknüpft werden können (Isele & Arnd, 2016). Den im HDFS abgespeicherten Daten liegt das Avro-Schema zugrunde, das gewährleistet, dass die Inhalte inventarisiert abgespeichert werden. Die Daten selbst sind im Binärformat abgespeichert und mit Map-Reduce Markern versehen, so dass eine schnelle Wiederauffindbarkeit garantiert ist. Ebenfalls konnte bereits in Kapitel 2.3.2 aufgezeigt werden, dass das HDFS als Ablage genutzt werden kann, um verschiedene Datensätze miteinander zu verlinken.

Frage 3

Was sind kritische Datenelemente und wie können diese kontrolliert werden?

Als kritische Datenelemente werden Daten angesehen, welche die sinngebende Funktionsweise der Anwendung sicherstellen, und deren strukturellen Beziehungen untereinander durch zentrale Taxonomien und Ontologien erfasst werden (Isele & Arnd, 2016). Im vorliegenden Forschungsproblem sind die Benutzungsdaten und die Interaktionen der

Benutzer mit dem Informationssystem die kritischen Datenelemente. Die Kontrolle dieser Daten erfolgt über den Divolte Mapper mit dem Abgleich der Daten auf das zugrundeliegende Schema, wodurch eine konsistente und transparente Verwaltung der Daten ermöglicht wird.

Frage 4

Welche Standards sind notwendig und dabei flexibel genug, um auf individuelle und sich ändernde Anforderungen reagieren zu können?

Das Avro Format, das vom Divolte Collector genutzt wird, hat ein sogenanntes Schema Evolution Konzept berücksichtigt. Datenmodelle können sich über die Zeit verändern und Avro Dateien unterstützen dies, in dem einerseits beim Lesen oder Schreiben zwingend Schemas vorausgesetzt werden, anderseits jedoch nicht dasselbe verwendet werden muss. Das Avro Framework kümmert sich darum, die fehlenden, zusätzlichen oder modifizierten Felder richtig zu interpretieren. Mittels dieser Schema Evolution kann somit eine Rückwärtskompatibilität gewährleistet werden und ermöglicht ein robustes und entkoppeltes System für die Zukunft.

Frage 5 & 6

Wie kann eine unternehmensweite Ontologie zur einheitlichen Abbildung von Datenelementen definiert und umgesetzt werden?

Wie kann eine Harmonisierung von Daten über mehrere verteilte Datenbanken erreicht werden?

Die Fragen 5 und 6 befassen sich mit einer unternehmensweiten Ontologie und können somit zur Verifizierung der Systemarchitektur nicht beigezogen werden, da dies nicht im Rahmen der Forschungsfrage I oder des Forschungsproblems erarbeitet wird. Jedoch verhindert das System in keinem Fall, dass eine unternehmensweite Ontologie und eine Harmonisierung der Daten über mehrere Datenbanken erreicht werden kann.

Frage 7

Wie wird mit externen, freien Datenquellen umgegangen?

Die Verknüpfung und Anreicherung von externen Datenquellen kann ebenfalls über den Divolte Collector (siehe Kapitel 2.3.2) oder zu einem späteren Zeitpunkt in der Phase des Processing basierend auf den Daten im HDFS durchgeführt werden.

2.4 Zusammenfassung des Kapitels

In Kapitel 2.1 konnte aufgezeigt werden, wie die aus der Konzeptualisierung entdeckten Begriffe Machine Learning und Data Mining in den Kontext des Forschungsproblems gebracht werden können. Web Usage Mining beschreibt den Prozess der Analyse von

Benutzerinteraktionen mit einem Onlinesystem. Die Aktualität der Problematik ist immer noch relevant, wie das Aufkommen von Smart Data Discovery im Hype Cylce von Gartner zeigt. Parallelen zwischen Smart Data Discovery und Web Usage Mining konnten aufgezeigt werden, und das Bedürfnis nach einem einfach einsetzbaren System, das die Benutzermodellierung unterstützt, aufgedeckt werden. Abschliessend lässt sich die Domäne „Smart Data Discovery“ mit den Hauptaspekten der Einfachheit der Benutzerbarkeit des Systems sowie der Einsetzbarkeit bei der Real Time Personalisierung als für das Forschungsproblem geeignet einstufen.

Ein Smart Data Discovery System besteht aus den Phasen der Aufbereitung der Daten, dem Ermitteln von Patterns innerhalb der Daten sowie der Vermittlung der Ergebnisse. Zur Beantwortung der Forschungsfrage I ist die erste Phase der Aufbereitung, auch Preprocessing genannt, von besonderer Relevanz, da dies die Daten sind, die durch das Informationssystem gespeichert und für die spätere Verarbeitung in den Phasen 2 und 3 zur Verfügung gestellt werden müssen. Diese können sich sogar, wie bereits in der Theorie der informationellen Mehrwerten aufgezeigt, über die Zeit verändern. Das Kapitel 2.2 zeigt auf, welches Datenmanagement Konzept der Systemarchitektur zu Grunde liegen muss, um die Daten in einer strukturierten oder semi-strukturierten Form abzuspeichern und diese im Anschluss mittels Machine Learning Ansätzen weiterverarbeiten zu können. Als Grundlage für die Datenerhebung eignen sich Online-Services, die nach einer hierarchischen Informationsarchitektur gestaltet sind, da mit diesem Designansatz analysierbare Clickstreams entstehen können. Als Basis für die Clickstreamauswertung werden implizite Daten direkt auf dem Client erhoben und in einer strukturierten Form an das Datenmanagement-System gesendet. Dadurch wird der Benutzer nicht in seinem Arbeitsprozess gestört, und Probleme wie z.B. das Caching durch den Browser oder Proxies, können umgangen werden. Für die Strukturierung der Daten soll das Clickstream Schema (siehe Tabelle 3) eingesetzt werden, das in einer späteren Analyse die Identifikation des Benutzers, der Session sowie der einzelnen Transaktionen ermöglicht.

Das Kapitel 2.3.2 beantwortet die Forschungsfrage I, in dem eine Architektur für eine Datenmanagement Infrastruktur basierend auf dem Divolte Framework vorgestellt und mittels einem Fragenkatalog verifiziert wird. Es kann aufgezeigt werden, dass die von Isele und Arnd formulierten Herausforderungen an ein modernes Datenmanagement mit dem Systemvorschlag beantwortet werden können. Die Systemarchitektur eignet sich als Grundlage für eine Datenmanagement Plattform für die Informationsverwaltung des Benutzerverhaltens, womit die Forschungsfrage I dieser Arbeit beantwortet werden konnte.

3 Evaluation eines Clickstream Clustering Algorithmus

Im vorliegenden Kapitel werden die theoretischen Erkenntnisse aus der Literatur zur Beantwortung der Forschungsfrage II aufgezeigt. Wie bereits in der Bestimmung der Domäne des Forschungsproblems aufgezeigt werden konnte, kann das Problem als ein Web Usage Mining bzw. ein Smart Data Discovery eingestuft werden (siehe 2.1). Dieser Prozess kann in die drei Phasen Preprocessing, Knowledge Discovery (bzw. auch Pattern Discovery genannt) und Pattern Analysis unterteilt werden. Für die Identifizierung von ähnlichem Interaktionsverhalten mit einem Informationssystem ist die zweite Phase das Knowledge Discovery, zuständig. Diese Phase basiert auf der Anwendung von Statistiken und Data Mining Methoden, die als Resultat ein nutzbares Set an Patterns retournieren, mittels welchem wiederum das Benutzerverhalten eingestuft werden kann (Eirinaki & Vazirgiannis, 2003). Dabei kommen in den Studien über Web Usage Mining verschiedene Techniken zum Einsatz. Eirinaki & Vazirgiannis (2003) sehen die Assoziationsanalyse, die Sequenzmuster-analyse, das Clustering sowie die Klassifizierung als solche Data Mining Techniken. Dabei können die Klassifizierung sowie das Clustering als sehr effektive Methoden eingestuft werden (Joan & Venifa, 2012). Gemäss Wang et al. (2013) kommen auch weitere Methoden wie das graphbasierte Markov Chain Clustering sowie baumbasierte Modelle zum Einsatz. Somit stehen für die Lösung des Problems viele verschiedene Algorithmen zur Verfügung, die genutzt werden können. Jedoch muss darauf geachtet werden, dass die Wahl des Algorithmus vom Ziel der Miningaufgabe abhängig ist.

Aus diesem Grund wird in den nachfolgenden Kapiteln zuerst eine Kategorisierung der Web Mining Techniken vorgestellt (siehe Kapitel 3.1). Anhand dieser Kategorisierung können die Algorithmen, die für die Beantwortung der Forschungsfrage II in Frage kommen, einge-schränkt werden. Eine wichtige erste Komponente, die jedem Data Mining Algorithmus zugrunde liegt, ist die Repräsentation des Models, bzw. die Sprache, die genutzt wird, um entdeckbare Patterns zu beschreiben (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). In Kapitel 3.2 wird ein solches Clickstream Model definiert und basierend auf diesem Model und der Kategorisierung eine Evaluation eines geeigneten Algorithmus vorgenommen.

3.1 Anwendungsgebiete im Web Usage Mining und ihre Algorithmen

Wie im vorherigen Kapitel aufgezeigt werden konnte, gibt es in der Literatur verschiedene Lösungsansätze im Bereich des Web Usage Mining. Die Wahl des Algorithmus muss gemäss Mobasher (2007) von der eigentlichen Miningaufgabe abhängig gemacht werden, da nicht jeder Algorithmus auf ein beliebiges Miningproblem angewendet werden kann. In der Tabelle 4 werden die von Mobasher (2007) vorgeschlagenen Einsatzgebiete und Beschrei-

bungen der gängigsten Pattern Discovery und Analysetechniken aufgelistet sowie einige in der Literatur gefundene Algorithmen der jeweiligen Einsatzgebiete zusammengetragen. Die Tabelle hilft dabei, den für das Forschungsproblem geeigneten Algorithmus zu finden.

3.1.1 Kategorisierung der Web Usage Mining Techniken

Anwendungsgebiete nach Mobasher (2007)	Beschreibung nach Mobasher (2007)	Algorithmen aus der Literatur
Session und Besucher Analyse	- Statistische Analyse auf Pre-Processed Data - Daten werden mittels vorbestimmten Einheiten (wie z.B. Tagen, Sessions, Besucher oder Domänen) aggregiert. - Aus diesen Aggregationen können Reports gebildet werden, die z.B. die meist frequentierten Seiten beinhalten, die durchschnittlichen Ansichts-zeiten einer Seite, die durchschnittliche Pfadlänge oder andere aggregierte Messgrössen. - Vorteil: Es können schnell und einfach nützliche Messgrössen über die Seite sowie die Benutzer erhoben werden. - Ein Nachteil dieses Anwendungsfeldes ist die fehlende Tiefe der Analyse.	- Aggregation (Mobasher, 2007) - Online Analytical Processing (Mobasher, 2007)
Cluster Analyse und Besucher Segmen-tierung	- Data Mining Technik, bei der es um das Gruppieren von Items geht, die dieselbe Charakteristik aufweisen. - Es kann zwischen dem User Clustering und dem Page Clustering unter-schieden werden. - Das Clustering von Benutzern tendiert dazu Gruppen zu kreieren, die ein ähnliches Navigationsverhalten aufweisen. - Sehr nützlich bei der demographischen Segmentierung der Benutzer für Marketingaktivitäten oder um personalisierten Webinhalt an Nutzer mit denselben Interessen anbieten zu können. - Zum Einsatz kommen klassische partitionierende oder hierarchische	- k-Means (Heer & Chi, 2002; Shahabi, Zarkesh, Adibi, & Shah, 1997) - k-Medoids (Bandari, Xiang, & Leskovec, 2017) - Divisive Hierarchical Clustering (Wang, Zhang, Tang, & Zhao, 2016) - Markov Chain Modelle (Sadagopan & Li, 2008; Lu,

	Clusteringalgorithmen, die die Gruppierungen der Cluster basierend auf einem Distanzmass oder Ähnlichkeitsmass berechnen. - Auch stochastische Methoden werden unter anderem für das Clustering, vor allem aber für die Modellierung von Benutzern eingesetzt.	Dunham, & Meng, 2005) - Probabilistic Latent Semantic Analysis (PLSA) (Jin, Zhou, & Mobasher, 2004)
Assoziations- und Korrelationsanalyse	- Dienen dem Auffinden von Gruppen von Items oder Seiten, die gemeinsam aufgerufen oder gekauft worden sind. - Eine der Hauptanwendungen der Assoziationsanalyse ist im Mining von Regeln um, basierend auf diesen, Vorschläge für weitere Produkte machen zu können. - Vorteile: Mit dem Mining von Regeln kann beispielsweise auch überprüft werden, ob Werbekampagnen den gewünschten Erfolg haben (in dem eine Regel von der Werbung -> Produkt ermittelt wird). Die Navigationsstruktur von Seiten kann verbessert werden, wenn erkannt wird, dass es keinen direkten Link zwischen A und B hat, es jedoch eine Regel A->B ergibt. - Nachteile: Assosatzionsregeln basierte Recommender Systeme haben den Nachteil, dass keine Empfehlungen gemacht werden können, wenn das Datenset zu zerstreut ist. Dies kommt im Web Usage Mining und in kollaborativen Filtering Applikationen oft vor, da ein Benutzer nur wenige Seiten/Produkte des gesamten Angebots besucht. Umgangen werden kann das Problem in dem die Dimensionalität verkleinert wird oder ein Ranking der Regeln vorgenommen wird.	- Apriori (Kumar & Rukmani, 2010) - D-Apriori (Kousalya, Pradeepa, & Saravanan, 2013)

Analyse von Sequenz-und Navigationspatterns	- Mit der Analyse von Sequenz- und Navigationspattern werden Patterns innerhalb der Session gesucht, in welchen die Items zeitlich geordnet sind. - Mittels dieser Web Usage Mining Technik können anhand der Items in der laufenden Session Vorhersagen darüber getroffen werden, was ein Benutzer als nächstes aufrufen könnte. Dies kann bei der Trendanalyse, Change Point Detection oder Similarity Analyse genutzt werden. - Modelliert werden diese Benutzerinteraktionen auf der Webseite mittels Markov Modellen, wobei jede Seitenansicht/Aktivität als ein Status angesehen werden kann und die Transitionswahrscheinlichkeit zwischen diesen Stati der Wahrscheinlichkeit entspricht, dass ein User von Status A nach B wechselt.	- Markov Chain Modelle (Sadagopan & Li, 2008; Lu, Dunham, & Meng, 2005)
Klassifizierung	- Technik, bei der ein Item in eine von mehreren vordefinierten Klassen zugeteilt wird. - Die Klassifizierung wird mittels Supervised Learning Algorithmen trainiert und dazu eingesetzt, um Modelle von Nutzern anhand von vordefinierten Metriken zu erstellen. - Viele der kollaborativen Filtering Methoden in den heutigen Recommender Systemen nutzen die Klassifizierung mittels k-nearest Neighbor um User Rating oder die Kaufbereitschaft vorhersagen zu können.	- Decision Tree,Naive Bayesian Classifier, K-nearest neighbor classifiers (Mobasher, 2007) - Support Vector Machines (Sitaraman, 2014; Wang, Konolige, Wilson, & Zhao, 2013)

Tabelle 4: Web Usage Mining Algorithmen

3.1.2 Auswahl des Anwendungsgebiets

Wie dem Forschungsproblem entnommen werden kann, geht es in dieser Masterarbeit darum, die Benutzer von Applikationen in ähnliche Profile zu segmentieren, um aufbauend auf den Gruppen gezielte Marketing- oder Verkaufsmassnahmen durchführen zu können. Aus der Tabelle 4 wird ersichtlich, dass das Anwendungsgebiet der „Cluster Analyse und Besucher Segmentierung“ am besten dafür geeignet ist. Im nachfolgenden Kapitel muss also ein Clickstream Modell definiert werden, das entweder von einem partitionierenden oder hierarchischen Clusteringverfahren, Markov Chain Modellen oder Probabilistic Latent Semantic genutzt werden kann.

3.2 Clickstream Modellierung

Eine wichtige Komponente bei der Wahl des Algorithmus ist, wie die Daten, die der Algorithmus verarbeitet, modelliert werden sollen. Das Verfahren kann die Reihenfolge der Aktionen in Betracht ziehen oder ignorieren (Bandari, Xiang, & Leskovec, 2017). Viele Forschungen im Bereich des Clickstream Mining nutzen unterschiedliche Verfahren und es zeigt sich, dass je nach Gebiet ein anderes Verfahren angewendet werden muss. Beispielsweise kann ein rollenbasiertes Recommender-System basierend auf der Assoziationsanalyse eingesetzt werden, um in Real-Time eine Warenkorbanalyse durchzuführen und passende weitere Produkte vorzuschlagen. Geht es jedoch um die Analyse von den besuchten Seiten, um beispielsweise das Layout der Webseite für den Nutzer zu optimieren, muss gemäss Zaiane et al. (2003) ein Distanzmass eingesetzt werden, das sowohl die strukturelle Information an sich (Reihenfolge der Elemente) sowie auch die zweidimensionale Sequenz (gegeben durch die Zeit) berücksichtigt. Dank der Kombination dieser zwei Faktoren kann eine Auskunft darüber erlangt werden, wie gross das Interesse des Benutzers an der Seite war und durch das Distanzmass kann nach ähnlichen Interessen gruppiert werden. Auch Mobasher (2007) sieht ein Set von besuchten Seiten (eine Transaktion) als semantisch aussagekräftige Entität, die für das Mining Task genutzt werden kann. Konzeptuell kann jede Transaktion als eine Sequenz von geordneten Paaren angesehen werden, bestehend aus der Aktion (Besuch einer Seite) sowie einer Gewichtung, die die Wichtigkeit der Aktion repräsentiert (Mobasher, 2007). Dabei ist die Gewichtung gemäss Mobasher (2007) in den meisten Web Usage Mining Anwendungen entweder Binär (Pageview existiert / existiert nicht), oder sie misst die Aufenthaltsdauer auf der Seite während der Transaktion. Diese Beispiele zeigen, dass es für die Wahl des Algorithmus entscheidend ist, welches Clickstream Modell aus den Daten im Schritt des Pre-Processing generiert wird. Im folgenden Kapitel werden daher Clickstream Modelle vorgestellt, die verschiedene Clickstream Features (wie Seite, Zeit, Reihenfolge etc.) N-Dimensional abbilden können.

3.2.1 Zweidimensionale Modelle

Gemäss Mobasher (2007) können in vielen der Data Mining Anwendungen die Clickstream Modelle als eine User-Pageview Matrix (siehe Tabelle 5) dargestellt werden.

Benutzer	**Seiten**		
	A	B	C
Benutzer 1	20	0	23
Benutzer 2	5	7	13
Benutzer 3	100	20	5

Tabelle 5: User-Pageview Matrix

Die Matrix besteht aus den zwei Dimensionen Benutzer sowie Seiten und als Wert kann die Zeit eingetragen werden, die der Benutzer auf der Seite verbracht hat. Auf diese Tabelle kann nun eine vielfältige Anzahl an unsupervised Techniken wie z.B. das Clustering angewendet werden (Mobasher, 2007).

Eine solche User-Pageview Matrix kann sehr umfangreich werden, da sämtliche Seiten in der Matrix aufgeführt werden. Fu et al. (2000) schlagen daher einen „Generalization-based Clustering“ Ansatz vor, der darauf basiert, dass Unterseiten sowie deren Besuchszeiten zu ihren Oberseiten generalisiert werden und dadurch die Dimensionalität stark vereinfacht werden kann. Eine einzelne generalisierte Session kann dann als Vektor in der Form von (Session-ID, t1, t2, ..., tn) abgebildet werden, wobei die Seiten-IDs nicht Bestandteil sein müssen, da sämtliche Sessions auf denselben generalisierten Sessions basieren.

3.2.2 n-dimensionale Modelle

Einen weitaus komplexeren Ansatz nutzen Heer und Chi (2002) mit ihrem Multi-Modal Clustering (MMC) Ansatz, der zusätzliche Datenmerkmale aus dem Inhalt und den Strukturen sowie der Reihenfolge der einzelnen Seiten hinzuzieht. Der MMC Ansatz ist eine Technik, die genutzt werden kann, um mehrere Datenmerkmale (sogenannte Modularitäten) zu nutzen, um Cluster zu produzieren. In der vorliegenden Studie wurden die folgenden Datenmerkmale genutzt, um den Modularity Vektor zu bilden:

- Inhalt: TF-IDF gewichteter Vektor, der sämtliche Wörter der Seiten beinhaltet
- URL: TF-IDF gewichteter Vektor, der sämtliche Tokens der URL beinhaltet
- Inlink/Outlink: ein Vektor, der eingehende Seiten sowie ein Vektor, der ausgehende Seiten beinhaltet
- User-Session: Vektor der die Sessiontransaktion in Reihenfolge beinhaltet.

Für die Abbildung der Werte des User-Session Vektors haben Heer und Chi (2002) mit verschiedenen Pfadgewichten experimentiert. Nachfolgend die Liste der einzelnen Gewichte:

- Uniform: Jede Seite erhält dieselbe Gewichtung.
- TF-IDF: Jede Seite erhält eine TF-IDF Gewichtung (Session = Dokument, zugegriffene Seiten = Terme der Dokumente).
- Lineare Reihenfolge: Die Reihenfolge vom Zugriff auf die Seiten wird für die Gewichtung genutzt.
- View Time: Jede Seite ist gewichtet nach der Zeit, die der User auf der Seite verbracht hat.
- Sowie diverse Kombinationen dieser Schemata.

Anhand dieser Multi-Modalen Vektoren und dem Cosine-Distanzmass konnten Heer und Chi (2002) unter Anwendung eines Bisecting K-Means (Kombination von KMeans und hierarchischem Clustering) Clickstreams gruppieren. Von Heer und Chi (2002) durchgeführte Experimente mit Uni- und Multi-Modalen Vorgehen zeigen vor allem zwei interessante Resultate auf:

1. Das Multi-Modale Vorgehen zeigt keinen signifikanten Unterschied zu einem Uni-Modalen Ansatz. Jedoch kann der Multi-Modale Ansatz ein wenig robuster eingestuft werden, da er auf mehrere Datenmerkmale zurückgreifen kann und dadurch auch funktioniert, wenn der Inhalt einer Seite z.B. nur aus Bildern besteht.
2. Die Gewichtung der Seiten mit der View Time hat eines der besten Ergebnisse geliefert, während TF-IDF und Uniform ziemlich schlecht abgeschnitten haben.

3.2.3 Repräsentation der Modellierung

Als Fazit aus dem Vergleich der verschiedenen Clickstream Modelle lässt sich schlussfolgern, dass einfache Clickstream Modelle unter Berücksichtigung des rohen Clickstream Pfads und der View Time sehr gute Resultate liefern. Komplexere Modelle mit mehr als diesen zwei Dimensionen sind ein wenig robuster, jedoch wiegt sich das nicht auf gegenüber den höheren Kosten (Zeit, CPU, ...), die für das Verarbeiten der zusätzlichen Dimensionen anfallen. Aus diesem Grund wird für die Beantwortung der Forschungsfrage II ein Algorithmus angewendet, der als Basis ein zweidimensionales Clickstream Modell verwendet, mit dem die Aktionssequenz (in Reihenfolge) sowie die Zeitintervalle zwischen den Aktionen erhoben werden.

Im Vergleich zu der Frage der Dimensionalität der Modellierung von Klicksequenzen ist die Wahl einer geeigneten Datenstruktur gemäss Riesen und Bunke (2010) ebenfalls ein wichtiger Schritt in jedem Pattern Recognition System. Daher werden vor der Gegenüber-

stellung der Algorithmen die zugrundeliegenden Repräsentationsmöglichkeiten mit ihren Vor- und Nachteilen aufgezeigt.

Grundlegend stellt sich neben der Dimensionalität der Modellierung auch die Frage, wie das Modell repräsentiert werden soll, da die Repräsentation massgeblich für die Wahl des Algorithmus verantwortlich ist. Die Wahl einer geeigneten Datenstruktur ist laut Riesen und Bunke (2010) der erste und ein wichtiger Schritt in jedem Pattern Recognition System. Riesen und Bunke (2010) unterscheiden zwischen zwei verschiedenen Arten der Encodierung der Daten: (1) mittels einem statistischen Vorgehen, in dem Objekte und Patterns durch sogenannte Feature-Vektoren abgebildet werden oder (2) durch ein strukturelles Vorgehen, das die Objekte und Patterns in symbolischen Datenstrukturen wie Strings, Bäumen oder Graphen abbildet (wobei Strings und Bäume einfach spezielle Formen eines Graphen darstellen). Beide mögliche Repräsentationen haben ihre Nachteile. Vektoren repräsentieren immer ein vordefiniertes Set an Features und daher müssen die Vektoren unabhängig von der Komplexität des Objektes immer dieselbe Grösse haben. Zudem gibt es keine direkte Möglichkeit, Beziehungen, die zwischen Teilen von Patterns bestehen zu beschreiben. Diese Nachteile werden beim Einsatz von Graphen eliminiert, jedoch haben diese im Gegenzug zu den Vektoren das Problem, dass die Komplexität der Algorithmen signifikant ansteigen kann (Riesen & Bunke, 2010).

Eine Lösung für das Problem ist stellt der Brückenschlag zwischen dem statistischen und strukturellen Vorgehen dar, indem der Graph in einem Vektorraum dargestellt wird, und die Partitionierung des Graphen mittels „Dissimilarites“ durchgeführt wird (Riesen & Bunke, 2010).

3.3 Auswahl des Algorithmus

Wie in der Tabelle 4: Web Usage Mining aufgezeigt, wird ein Algorithmus aus der Cluster Analyse und der Benutzer Segmentierung benötigt. Zu den beliebten Algorithmen aus dem Bereich des Clustering von zeitintervallbasierten Daten gehören unter anderem partitionierende, hierarchische oder modelbasiertes Verfahren (Zhang, Liu, Du, & Lv, 2011). In den nachfolgenden Unterkapiteln werden die in der Literaturanalyse gefundenen Algorithmen zum Thema Clickstream-Clustering den drei Bereichen partitionierend, hierarchisch und modellbasiert zugeteilt. Für jeden Bereich werden die allgemeinen Vor- und Nachteile erläutert und danach wird im Detail auf die zuvor aufgelisteten Kriterien für jeden Algorithmus eingegangen. Abschliessend werden diese Algorithmen miteinander verglichen und ein Kandidat für die Implementation im Prototypen ausgewählt.

3.3.1 Definition des Evaluierungsrasters

In den vorangegangenen Kapiteln konnte aufgezeigt werden, dass die Modellierung der Clickstream-Sequenz sowie die Repräsentation der Clickstream-Sequenz eine wichtige Rolle bei der Wahl eines Algorithmus haben. Daher wird die Modellierung als Kriterium für die Umsetzbarkeit einbezogen. Abgesehen von der Repräsentation können bei der Evaluierung eines Models herkömmlicherweise auch quantitative Statements oder Fit Functions genutzt werden, um eine Aussage darüber zu treffen, inwieweit ein bestimmtes Pattern das Ziel des Knowledge Discovery Prozess erfüllt (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Als quantitatives Statement wird dazu oft die Genauigkeit des Algorithmus für den Vergleich genutzt. Dieses Kriterium soll bei der Auswahl des Algorithmus ebenfalls berücksichtigt werden. Weitere Kriterien sind gemäss Fayyad et al. (1996) die Neuheit des Ansatzes, die Nützlichkeit für das vorliegende Forschungsproblem sowie die Verständlichkeit des Algorithmus bzw. der Modellierung. Auch diese Kriterien werden für den Vergleich der in der Literatur aufgefundenen Lösungsansätze für ein Clickstream-Clustering basierend auf hierarchischen, partitionierenden oder modelbasierten Algorithmen herangezogen.

Die Auswertung der objektiven Kriterien Genauigkeit, Neuheit und Modellierung basieren auf theoretisch begründeten Ergebnissen aus der Literaturanalyse. Die Nützlichkeit sowie die Verständlichkeit basiert auf einer subjektiven Einschätzung für die Umsetzbarkeit des Algorithmus im Rahmen des Prototypings dieser Masterarbeit.

3.3.2 Partitionierende Algorithmen

Partitionierende Algorithmen werden in der Regel dazu verwendet, um einen Datensatz in K Partitionen aufzuteilen, wobei jede Partition dabei ein Cluster repräsentiert (Soni & Ganatra, 2012). Zu den partitionierenden Algorithmen zählen unter anderem k-Means und k-Medoid, die für das Clustering jeweils einen Repräsentanten pro Cluster auswählen und in einem iterativen Vorgehen die Cluster um diese Repräsentanten gebildet werden. Bei k-Means wird dazu der Mittelwert des Clusters als Repräsentant gewählt, bei k-Medoid ein mittleres Objekt als Repräsentant verwendet.

Gemäss Sisodia et al. (2012) sind die Stärken der partitionierenden Algorithmen, dass sie sehr einfach zu implementieren sind und gleichzeitig auch gut skalieren. Als Nachteil wird aufgeführt, dass die Effektivität in hochdimensionalen Räumen degradiert, die Cluster selbst schlecht beschrieben werden, die Clusteranzahl im Vorhinein bekannt sein muss und die Verfahren sehr anfällig sind auf Noise (Patterns, die nicht dazugehören) und Ausreisser. Da die Anzahl Cluster im Vorhinein bekannt sein muss, sind diese Algorithmen nicht deterministisch (Manning, Raghavan, & Schütze, 2008).

Die Tabelle 6 und Tabelle 7 zeigen zwei unterschiedliche Verfahren im Clickstream Clustering basierend auf dem K-Means und K-Medoid Clustering Algorithmus auf.

Algorithmus	**K-Means**
Genauigkeit	Niedrig, es konnte aufgezeigt werden, dass k-Means eine Fehlerrate von 10-30% besitzt (Shahabi, Zarkesh, Adibi, & Shah, 1997; Heer & Chi, 2002; Wang, Zhang, Tang, & Zhao, 2016; Heer & Chi, 2002).
Neuheit	Sowohl im Data Mining als auch im Clickstream Clustering keine Neuheit. Der Algorithmus wird in vielen Forschungen angewendet, da er so einfach zu implementieren ist.
Modellierung	Für die Modellierung kann wie von Heer und Chi (2002) aufgezeigt der Raw Path als Vektor verwendet werden. Die Vektoren beschreiben die Anzahl an Seitenaufrufen, die im Pfad vorkommen. Die Dimension der Zeitintervalle wird mittels der Gewichtung der Pfade modelliert (jede Seite der Session erhält die Zeit als Gewicht). Folglich wären zwar beide Dimensionen präsent, jedoch sind die Seitenaufrufe nicht in der Reihenfolge des Auftretens abgebildet.
Nützlichkeit	Das Vorgehen von Bandari et al. (2017) unterstützt die Dimension der Zeitintervalle nicht, weshalb es nur bedingt einsetzbar ist. Dieses Vorgehen wird jedoch in vielen Forschungen genutzt, da der Algorithmus leicht zu implementieren ist.
Verständlichkeit	Leicht verständlich und viele Standardimplementationen[2] verfügbar. Kann auch direkt auf den Raw-Input Path angewendet werden (Heer & Chi, 2002), wodurch das Pre-Processing weniger komplex ausfällt.

Tabelle 6: K-Means Verfahren

Algorithmus	**K-Medoid**
Genauigkeit	Der Algorithmus erreicht eine Vorhersagegenauigkeit von über 85% (Bandari, Xiang, & Leskovec, 2017).
Neuheit	Die Literaturrecherche zeigt zwar, dass K-Medoid als Algorithmus zur Umsetzung von Clickstream-Clustering vorgeschlagen wird [siehe auch (Hahsler & Dunham, 2010)], jedoch findet sich eine

[2] z.B. https://github.com/algorithmfoundry/Foundry oder https://mahout.apache.org/users/clustering/k-means-clustering.html

	konkrete Anwendung nur von Bandari et al. (2017).
Modellierung	Für die Modellierung einer User-Session werden nur die Events innerhalb der Session verwendet. Zeitintervalle werden keine genutzt. In dem Verfahren werden alle Events, die in weniger als 5% der Sessions verwendet werden, herausgefiltert und basierend auf den restlichen Events eine TF-IDF Gewichtung vorgenommen und jede Session als Vektor dargestellt. Wie in Kapitel 3.2 aufgezeigt, werden diese Vektoren sehr umfangreich. Daher nutzen Bandari et al. (2017) eine Hauptkomponenten-analyse, um die Dimensionen zu reduzieren und wenden erst danach das K-Medoids Clusteringverfahren an.
Nützlichkeit	Das Vorgehen von Bandari et al. (2017) unterstützt die Dimen-sion der Zeitintervalle nicht, weshalb es nur bedingt einsetzbar ist.
Verständlichkeit	Leicht verständlich, jedoch ist die Komplexität leicht erhöht, da eine Reduktion der Dimensionen vorgenommen wird. Die Reduk-tion könnte weggelassen werden, indes sind keine Angaben darüber bekannt, wie sich das auf die Genauigkeit auswirken würde.

Tabelle 7: K-Medoids Verfahren

3.3.3 Hierarchische Algorithmen

Die hierarchischen Algorithmen verfolgen den Ansatz, dass ein Datenset aus N Objekten in eine Hierarchie von Gruppen unterteilt wird, die in einer Baumstruktur, einem sogenannten Dendogramm, dargestellt werden können (Soni & Ganatra, 2012). Bei den hierarchischen Algorithmen kann zwischen agglomerativen (bottom-up) und divisiven (top-down) Verfahren unterschieden werden.

Die Vorteile der hierarchischen Algorithmen sind, dass der Output des Clusterings viel informativer ist als die unstrukturierten Cluster, die bei den partitionierenden Verfahren entstehen. Zudem sind die Algorithmen deterministisch und benötigen keine Vorkenntnisse über die Anzahl an zu bildenden Cluster (Manning, Raghavan, & Schütze, 2008). Gemäss Manning et al. (2008) gehen diese Vorteile jedoch auf Kosten der Komplexität der Algorithmen, die mindestens eine quadratische Komplexität haben (im Vergleich zu K-Means oder EM, die lineare Komplexität aufweisen).

Die Tabelle 8 und Tabelle 9 zeigen zwei unterschiedliche Verfahren im Clickstream Cluste-ring basierend auf einem divisiven und agglomerativen Algorithmus.

Algorithmus	**Divisive Hierarchical Clustering (DHC)**
Genauigkeit	Hoch, es konnte eine Genauigkeit von 93% und ein Recall von 94% nachgewiesen werden (Wang, Zhang, Tang, & Zhao, 2016).
Neuheit	Der Algorithmus inkl. dem zugrundeliegenden Modell von Wang et al. (2016) ist verglichen mit den anderen in der Literaturrecherche gefundenen Ansätzen neu im Clickstream Clustering.
Modellierung	Wang et al. (2016) modellieren die Clickstream Sequenz als Similarity Graphen, wobei jede Node einem User entspricht und die Kanten gewichtet nach der Similarity zwischen zwei Clickstreams sind. Der Vergleich wird dabei auf Subsequenzen des Clickstreams, die in Reihenfolge und mit Zeitintervallen zur Verfügung stehen, vorgenommen. Mittels einem Iterativen Feature Pruning unter dem Einsatz vom Divisive Hierarchical Clustering wird der Similarty Graph in detailliertere Subgraphen unterteilt.
Nützlichkeit	Mit dem vorgestellten Verfahren konnte erfolgreich zwischen fake und realen Benutzern anhand dem Navigationsverhalten unterschieden werden. Wang et al. (2016) schätzen, dass ihr Verfahren auch in anderen Bereichen erfolgreich eingesetzt werden kann.
Verständlichkeit	Der Algorithmus ist gut dokumentiert, und es wird eine Python-Implementation[3] des Algorithmus zur Verfügung gestellt. Der Similarity Graph beinhaltet von Beginn an sämtliche Datensätze und wird iterativ durch das Feature Pruning weiter geteilt. Das Verfahren ist somit divisiv.

Tabelle 8: Divisives Verfahren

Algorithmus	**Balanced Interative Reducing and Clustering using Hierarchies (BIRCH)**
Genauigkeit	Keine Angaben zur Genauigkeit verfügbar. Es wird lediglich erwähnt, dass das Clustering mit zunehmender Dimensionalität degradiert (Fu, Sandhu, & Shih, 2000).
Neuheit	Der Einsatz von BIRCH Algorithmus im Bereich des Clickstream Clustering konnte nur in der Forschung von Fu et al. (2000) gefunden werden.

[3] http://sandlab.cs.ucsb.edu/clickstream/

Modellierung	Das Verfahren von Fu et al. (2000) basiert darauf, dass Seitenhierarchien generalisiert werden können, indem Unterseiten einer Session zu einer bestimmten Hauptseite in der Hierarchie generalisiert werden. Zu jeder Generalisierung wird die auf der Seite verbrachte Zeit mitgespeichert. Wurde eine Seite mehrmals besucht, wird die Zeit in der Generalisierung aufaddiert. Durch die Generalisierung kann die Dimensionalität des Session-Vektors verkleinert werden, was wiederum dem BIRCH Algorithmus zugutekommt, weil zu dünnbesetzte Vektoren dem Clustering Algorithmus Probleme bereiten.
Nützlichkeit	Der BIRCH Algorithmus liefert gute Ergebnisse mit geringen Dimensionalitäten und skaliert auch gut mit grossen Datenmengen. Jedoch geht die Reihenfolge, in der die Seiten aufgerufen werden durch die Generalisierung und Speicherung im Vektor verloren.
Verständlichkeit	Die Generalisierung ist einfach beschrieben und anwendbar. Für die Implementation muss mit dem BIRCH ein Cluster Feature Tree gebildet werden, in dem die Nodes eine nach der anderen eingefügt werden. Das Verfahren kann zu den agglomerativen Verfahren gezählt werden.

Tabelle 9: Agglomerative Verfahren

3.3.4 Modellbasierte Algorithmen

Streng genommen können auch die Probabilistic (Mixture Model) Based Algorithmen den partitionierenden Algorithmen zugeordnet werden (Soni & Ganatra, 2012). Anstelle der Wahl eines Repräsentanten wie beim K-Means wird bei den modellbasierten Algorithmen jedoch basierend auf vorgegebenen Modellannahmen die beste Zerlegung in einzelne Cluster, basierend auf paarweiser Berechnung von Ähnlichkeitsmassen, in einem iterativen Vorgehen ermittelt (Soni & Ganatra, 2012). Modellbasiertes Clustering geht davon aus, dass die Daten durch ein Modell kreiert worden sind und versucht, das Originalmodell aus den vorhandenen Daten zurückzugewinnen (Manning, Raghavan, & Schütze, 2008).

Der Vorteil von modellbasierten Algorithmen ist die Flexibilität. Während partitionierende und hierarchische Algorithmen nur Mutmassungen über die Daten machen, können modellbasierte Verfahren auf die Verteilung der zugrundeliegenden Daten adaptiert werden. Mit dem zugrundeliegenden probabilistischen Modell ist die Berechnung der Anzahl an Cluster nur noch ein statistisches Problem, was ein grosser Vorteil gegenüber den heuristischen Clustering Algorithmen ist, die die Anzahl Cluster nicht selbst bestimmen können (Yeung, Murua, Raftery, & Ruzzo, 2001). Ein grosser Nachteil gegenüber den anderen Verfahren ist, dass klassische model-basierte Clusteringtechniken schlecht mit höheren Dimensionen zurechtkommen (Bouveyron & Brunet, 2013).

Die Tabelle 10 und Tabelle 11 zeigen zwei unterschiedliche Verfahren im Clickstream Clustering basierend auf Markov Modellen und einem Probabilistic Latent Semantic Analysis (PLSA) Verfahren.

Algorithmus	**Expectation-Maximum basierend auf Markov Modellen (MMC)**
Genauigkeit der Vorhersage	Die Genauigkeit der Modelle ist nachweislich abhängig von der Komplexität der Markov Modelle (Pirolli & Pitkow, 1999). Je komplexer die Modelle sind (je höher die Ordnung ist), desto ungenauer werden die Ergebnisse.
Neuheit	Modellbasiertes probabilistisches Clustering ist im generellen (nicht im Web-Umfeld) seit langem bekannt und weit angewendet, wird jedoch erst von Cadez et al. (2003) im Bereich des Clickstream Clustering für die Segmentierung von Benutzerverhalten vorgeschlagen.
Modellierung	Jede User Session wird von Cadez et al. (2003) in eine Sequenz konvertiert, wobei diese Sequenz einer geordneten Liste von diskreten Symbolen entspricht. Jedes dieser Symbole repräsentiert dabei eine mögliche Kategorie von Webseite, die vom User aufgerufen worden ist. Für das Clustering wird ein first-order Markov Model genutzt, das die Reihenfolge der aufgerufenen Seiten berücksichtigt. Das Clustering selbst wird mit einem Expectation Maximum Algorithmus durchgeführt, einem iterativen Algorithmus, der das lokale Maximum für die Parameter der Modelle findet.
Nützlichkeit	Die meisten der gefundenen Modelle wurden für Vorhersagen und nicht für Gruppierung eingesetzt. Das vorliegende Beispiel nutzt den EM Algorithmus, um diese Modelle zu clustern. Mit dem Einsatz des first-order Markov Model wird zwar die Reihenfolge wiedergegeben, indes sind die Zeitintervalle zwischen den Aufrufen nicht abgebildet.
Verständlichkeit	Wie von Borges und Mark in ihrem Literatur-Review aufgezeigt wird, handelt es sich um ein komplexes Thema, da unterschiedliche Modellierungen zu unterschiedlichen Laufzeiten und Genauigkeiten führen (Borges & Mark, 2004). Auch laut Cadez et al. (2003) gibt es einige Aspekte der Daten (Dynamik, Heterogenität, Umfang), die nicht-trivial sind und daher die Umsetzung der Daten in Modelle erschweren. Ein Set von Tools bietet z.B. das Package „Clickstream“[4] für R.

Tabelle 10: Markov Modelle

[4] https://cran.r-project.org/web/packages/clickstream/clickstream.pdf

Algorithmus	Expectation-Maximum basierend auf PLSA
Genauigkeit der Vorhersage	Basierend auf der Genauigkeit von Produktvorschlägen konnten Jin et al. (2004) aufzeigen, dass die Genauigkeit besser ausfällt als bei Experimenten mit dem k-Means Algorithmus.
Neuheit	PLSA wurde von Hofmann (2001) als neue statistische Methode eingeführt und wird im Zusammenhang mit Web Usage Mining nur von Jin et al. (2004) im Bereich der kollaborativen Filterung angewendet.
Modellierung	Die Websession eines Users kann gemäss Jin et al. (2004) als eine mxn Session-Seitenview Binärmatrix modelliert werden. Dabei entsprechen die Werte den Gewichtungen und repräsentieren entweder in binärer Form die Existenz oder Inexistenz eines Seitenaufrufes oder können auch die Anzahl Aufrufe oder Aufenthaltsdauer wiederspiegeln. PLSA ist ein latentes Variablenmodell, das diese Matrix nutzt, um einen Zusammenhang aus den beobachtbaren Variablen und den dahinterliegenden latenten Variablen zu beschreiben. Der Output (die Wahrscheinlichkeiten von gemeinsamen Vorkommen von Session und Seitenaufruf) wird dann mittels dem Expectation Maximum Algorithmus geclustert.
Nützlichkeit	Durch eine komplexere Modellierung können zwar die Anzahl Seitenaufrufe oder die Aufenthaltsdauer auf den Seiten modelliert werden, jedoch geht die Reihenfolge sowie einer der beiden Werte verloren. Ein Vorteil des PLSA Ansatzes ist, dass ein User nicht zwingend nur einem Cluster zugeteilt werden muss, da durch das PLSA Verfahren eine Zugehörigkeit in mehrere Cluster erlaubt wird.
Verständlichkeit	Wie auch beim Markov Model Clustering wird bei der PLSA neben der Modellierung ein Algorithmus wie z.B. Expectation-Maximum für die Bestimmung der Ähnlichkeiten benötigt. Im Gegensatz zum Markov Model ist die Modellierung indessen einfacher.

Tabelle 11: PLSA

3.3.5 Gegenüberstellung

Die folgende Tabelle 12 ist eine zusammenfassende Auswertung der Kriterien der sechs Algorithmen aus den drei Bereichen partitionierend, hierarchisch, modellbasiert. Für die Bewertung wird die folgende Skala pro Kriterium angewendet:

- Genauigkeit: (-) = tief, (0) = mittel, (+) = hoch
- Neuheit: (-) = selten angewendet, (0) = neutral, (+) = viel angewendet
- Modellierung: (-) = komplexe Modellierung, (0) = neutral, (+) = einfache Modellierung
- Nützlichkeit (Dimensionalität Forschungsproblem): (-) = schlecht, (0) = neutral, (+) = erreicht
- Verständlichkeit: (-) = schwer verständlich, (0) = neutral, (+) = einfach verständlich

Algorithmus	Genauigkeit	Neuheit	Modellierung	Nützlichkeit	Verständlichkeit
k-Means	-	+	+	+	+
k-Medoids	0	-	+	-	0
DHC	+	-	0	+	0
BIRCH	0	-	0	0	0
MMC	0	0	-	-	-
PLSA	+	-	+	0	0

Tabelle 12: Vergleich der Algorithmen

Aus der Gegenüberstellung lässt sich ableiten, dass für die Beantwortung der Forschungsfrage das Divisive Hierarchical Clustering (DHC) Verfahren von Wang et al. (2016) am besten geeignet ist, da es das einzige Verfahren ist, das sowohl die Reihenfolge der Events als auch die Zeitintervalle zwischen den Events berücksichtigt. Als hierarchisches Verfahren entsteht ohne benötigte Vorkenntnisse über die Anzahl an Cluster ein leicht verständlich strukturierter Output auf Kosten einer erhöhten Laufzeitkomplexität.

Neben dem DHC wurde auch das PLSA Verfahren aus dem modellbasierten Bereich erfolgreich im Clickstream Clustering eingesetzt. Das Verfahren überzeugt vor allem mit der sehr guten Performance, die es gegenüber dem viel häufiger eingesetzten K-Means Clustering (Hou, 2015) aufweist sowie dem Umstand, dass PLSA gegenüber den traditionellen Clustering Algorithmen den Vorteil hat, dass ein User in mehrere Segmente eingeteilt werden kann (Wu, et al., 2009).

Ebenfalls häufig zum Einsatz kommt der K-Means Algorithmus, da dieser im Vergleich zu den anderen Algorithmen sehr einfach zu implementieren ist und mit grossen Datenmengen sehr gut skaliert. Der K-Means Algorithmus aus dem Bereich der partitionierenden Algorithmen ist jedoch sehr anfällig auf Noise und Ausreisser sowie die Anzahl an Cluster muss im Vorhinein bekannt sein.

3.3.6 Testversuche

Nach dem Vergleich steht aus jedem der drei Bereiche jeweils ein Algorithmus zur Verfügung, der auf das Forschungsproblem angewendet werden könnte. Um die Genauigkeit dieser drei Algorithmen besser miteinander vergleichen zu können, wurde im Rahmen dieser Masterarbeit anhand von 50 User Sessions, die jeweils mehr als 30 Events beinhalten, eine Segmentierung in drei Cluster durchgeführt. Für jeden dieser drei Algorithmen wurde das Clustering 10-mal unter denselben Parametern neu initialisiert. Eine detaillierte Auflistung des Testsetups sowie der resultierenden Cluster findet sich im Anhang in Kapitel 7.1.

Die Validierung von Clusterings kann in drei Kategorien eingeteilt werden. Die interne, externe sowie die relative Validierung (Dalton, Ballarin, & Brun, 2009). Bei der internen Validierung werden keine externen Daten herangezogen, um ein Qualitätsmass der einzelnen Cluster zu ermitteln. Da die drei Clusteringverfahren stark voneinander abweichen, kann diese Art der Validierung für den Vergleich der Algorithmen nicht verwendet werden. Die externe Validierung hingegen nutzt bekannte Klassen, um das Ergebnis gegenüber diesen Klassen zu vergleichen. Auf den erhobenen Clickstream-Daten sind keine solchen Klassen verfügbar, weshalb auch dieser Ansatz nicht zur Anwendung kommt. Die relative Validierung kann als Mischform angesehen werden und wird angewendet, um verschiedene Datenteile, Algorithmen und Parameter untereinander zu vergleichen (Plöhn, 2014). Für die Auswertung der erstellten Testcluster wird diese relative Validierung angewendet. Ausgewertet wird die Reproduzierbarkeit der Clusterzugehörigkeit einzelner User in diesen 10 Clusterdurchgängen sowie die Clusterstabilität, wenn die Menge der zu gruppierenden Clickstream-Sessions verändert wird.

Zusammenfassend lässt sich aus diesem Test die folgende Reproduzierbarkeit der Clusterzugehörigkeit ableiten:

Während beim PLSA mit EM 68% der Clusterzuteilungen der 50 Sessions übereinstimmen, sind es beim K-Means 77% Übereinstimmungen. DHC im Vergleich erreicht bei jeder Segmentierung eine 100% Übereinstimmung.

Ein Vergleich der Clusterzuteilungen der drei Algorithmen untereinander ergibt, dass DHC/K-Means mit 24% die ähnlichsten Zuteilungen haben, während DHC/PLSA 21% und PLSA/K-Means nur 17% gemeinsame Zuteilungen aufweisen. Die geringen Übereinstimmungen erklären sich durch die völlig unterschiedlichen Modellierungen der Clickstreams, wobei PLSA und K-Means auf derselben User/Event Matrix mit Anzahl Vorkommen der Events basieren. Hier hätte eine grössere Übereinstimmung erwartet werden können.

Ein weiterer Testversuch mit dem DHC hat ergeben, dass das Clustering eine Stabilität von über 87% aufweist, wenn die Anzahl an User Sessions auf die Hälfte reduziert werden. Im Vergleich PLSA mit 76% und K-Means, die nur noch 24% Übereinstimmung aufweisen.

Insgesamt lässt sich daraus schliessen, dass Divisive Hierarchical Clustering auf dem gegebenen Datenset eine grössere Reproduzierbarkeit sowie Stabilität als die anderen beiden Algorithmen belegen. K-Means reproduziert die Cluster gegenüber PLSA besser, hat jedoch eine tiefere Stabilität bei der Veränderung des Umfangs des Testest als PLSA.

3.4 Zusammenfassung des Kapitels

In diesem Kapitel konnte Anhand der von Mobasher (2007) eingeführten Kategorisierung von Web Usage Mining Algorithmen das Forschungsproblem dem Bereich der „Cluster Analyse und Besucher Segmentierung“ zugeordnet werden. Anhand dieser Zuordnung konnte eine erste Eingrenzung von Web Usage Mining Algorithmen vorgenommen werden. Konkret handelt es sich um partitionierende, hierarchische oder modellbasierte Clustering Algorithmen, die in diesem Bereich angewendet werden können. Mittels einer Literaturrecherche dieser Clustering-Bereiche und der Eingrenzung dieser auf den Bereich des Clickstream Clusterings, konnten sechs unterschiedliche Verfahren aufgezeigt und miteinander Verglichen werden (siehe Tabelle 6 - Tabelle 12).

Anhand von selbst durchgeführten Tests auf Clickstream-Daten von 50 Usersessions mit jeweils einem Vertreter jedes Bereichs – K-Means als Vertreter der partitionierenden, DHC als Vertreter der hierarchischen sowie PLSA als Vertreter der modellbasierten Algorithmen – konnte ermittelt werden, dass das Divisive Hierarchical Clustering die besten Ergebnisse in Bezug auf die Reproduzierbarkeit sowie Stabilität umfasst.

Aus diesen Gründen wird für eine Integration in den Prototypen und die anschliessende Beantwortung der Forschungsfrage III der von Wang et al. (2016) vorgeschlagene Algorithmus DHC basierend auf einem Similarity Graphen für die Segmentierung der User ausgewählt.

4 Analyse von Benutzerinteraktionen

Gemäss von Fayyad et al. (1996) besteht ein Data-Mining Algorithmus aus den drei Komponenten (1) Repräsentation des Models, (2) Evaluation des Models und (3) Suche nach der besten Optimierung. In Kapitel 3 konnte die Repräsentation des Clickstream Model beschrieben werden sowie der Divisive Hierarchical Clustering Ansatz von Wang et al. (2016) als geeigneter Algorithmus für die Segmentierung der Benutzer nach ähnlichem Interaktionsverhalten ausgewählt werden. Dieses Kapitel beschäftigt sich mit den weiteren zwei Komponenten eines Data-Mining Algorithmus. Es wird aufgezeigt, wie anhand von Studien aus der Psychologie prototypische Profile definiert werden können. Die Evaluation des Models entspricht quantitativen Statements oder Fit Functions, mit denen gemessen werden kann, wie gut die aufgefundenen Patterns des Data-Mining auf das Ziel der Aufgabe passen (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Die Genauigkeit des Algorithmus als quantitatives Statement wurde bereits in Kapitel 3.3 behandelt. In diesem Kapitel wird das prototypische Profil als Fit Function angesehen, um aufzugzeigen, ob die anhand des Algorithmus aufgedeckten Patterns diesen Profilen entsprechen oder nicht. Diese prototypischen Profile werden in Kapitel 4.1 aus den theoretischen Grundlagen der Psychologie erstellt.

Als Grundlage für die Evaluation dieses Modells werden ein Datenset von Persönlichkeitsmerkmalen, die auf ein prototypisches Profil zurückschliessen lassen sowie Clickstream-Sequenzen von demselben Benutzer benötigt. Das Kapitel 4.2 zeigt auf, wie diese Daten basierend auf den Grundlagen der Datenmanagement Plattform aus Kapitel 2 erhoben werden.

Um eine externe Validierung der Clusterergebnisse zu ermöglichen, werden in Kapitel 4.3.2 zwei unterschiedliche Gold Standard Clusterings ermittelt. Das erste Set basiert auf einem Clustering anhand der prototypischen Profile aus der Literatur, das zweite Set wird direkt aus den Umfrageergebnissen ermittelt. Diese zwei Gold Standard Sets bieten die Grundlage, um die dritte von Fayyad et al. (1996) beschriebene Data-Mining Komponente umzusetzen. Dabei handelt es sich um die Phase der Optimierung des Data-Minings. Dazu werden in Kapitel 4.3.3 verschiedene Parametrierungen und Modellierungen der Daten getestet und die Cluster-Ergebnisse auf eine Übereinstimmung auf eines der Gold Standard Sets geprüft.

Aus dem Nachweis einer Optimierung, so dass eine Übereinstimmung der Clickstream-Clusterergebnisse zu einem der beiden Gold Standard Cluterings entsteht, kann schlussendlich die Forschungsfrage III beantwortet werden.

4.1 Modellierung der Persönlichkeit

Die Big Five Faktoren haben in den letzten Jahren das Interesse in der Wirtschaft geweckt, da führende Wirtschafspsychologen aufzeigen konnten, dass die Personalauswahl, statt in einem Assessment Center auch genauso gut mittels einem 20-minütigen Big-Five-Persönlichkeitstest durchgeführt werden konnte (Satow, 2012). Der Big Five Test überzeugt vor allem dadurch, dass er trotz einer sehr hohen Reliabilität (Testgenauigkeit) kürzer ist als andere verfügbare Tests und durch die Einfachheit auch sehr gut für den Einsatz als Online Test geeignet ist (2012). Aus diesen Gründen wurde der Big Five Test für die Modellierung der Persönlichkeiten in dieser Masterarbeit ausgewählt.

In den nachfolgenden Kapiteln wird auf die Big Five Faktoren eingegangen. Ausserdem werden Fragebogen evaluiert, die für die Erhebung von Persönlichkeitsprofilen basierend auf diesen Faktoren verwendet werden können. Des Weiteren wird aufgezeigt, wie verschiedene Persönlichkeiten anhand von den Resultaten eines solchen Big Five Test miteinander verglichen werden können.

4.1.1 Big Five

Für die Modellierung der Persönlichkeit wird in dieser Masterarbeit auf die Big Five Faktoren zurückgegriffen. Die Big Five beschreiben fünf unterschiedliche Persönlichkeitsfaktoren. Den Ursprung haben die Big Five in einer empirischen Faktorenanalyse, in der die Teilnehmer sich in Bezug auf viele unterschiedlichen Eigenschaften selbst beurteilen mussten. Die Faktorenanalyse wurde daraufhin genutzt, um die vielen Items in den Persönlichkeitsintervallen auf möglichst wenige und voneinander unabhängige Faktoren zu reduzieren (Asendorpf & Neyer, 2012, S. 106). Die aus der Faktorenanalyse ermittelten Eigenschaftsdimensionen konnten von unabhängigen Forschergruppen bestätigt werden (Asendorpf & Neyer, 2012, S. 107) und die Faktoren wurden schlussendlich durch Goldberg (1981) unter dem Begriff „Big Five“ zusammengefasst. Zudem sind die fünf Faktoren auch in anderen Persönlichkeitsinventaren auffindbar, die unter anderen Methodiken konstruiert worden sind. Dazu gehört z.B. gemäss Asendorpf & Neyer (2012, S. 108) auch der NEO-FFI von Costa und McCrae (1989).

Die Big Five werden wie folgt charakterisiert:

1. ***O****penness:* Die Dimension der Offenheit gegenüber neuen Erfahrungen bezieht sich auf die intellektuelle Neugier, das Gefühl für Kunst und Kreativität und korreliert positiv mit der Intelligenz und v.a. der Bildung (Asendorpf & Neyer, 2012, S. 107). Zudem können Menschen mit einem hohen Wert in dieser Dimension mit besonders hoher Toleranz, Neugier und Interesse verbunden werden (Satow, 2012).

2. ***C**onscientiousness:* Die Dimension der Gewissenhaftigkeit bezieht sich auf die Ordentlichkeit, Beharrlichkeit und Zuverlässigkeit (Asendorpf & Neyer, 2012, S. 107). Personen mit einem hohen Gewissenhaftigkeitswert sind zudem pflichtbewusst, ordnungsliebend und gehen prinzipiell systematisch und genau vor (Satow, 2012).
3. ***E**xtraversion:* Bezieht sich auf die Geselligkeit, Ungehemmtheit und Aktivität (Asendorpf & Neyer, 2012, S. 107). Extrovertierte Personen sind nach aussen orientiert und werden dadurch gesprächig, gesellig und abenteuerlustig wahrgenommen (Satow, 2012).
4. ***A**greeableness:* Die Dimension der Verträglichkeit bezieht sich auf die Freundlichkeit, Hilfsbereitschaft und Wärme im Umgang mit anderen Personen (Asendorpf & Neyer, 2012, S. 107). Menschen mit einem hohen Wert in dieser Dimension sind oft gute Team-Player, bemühen sich um andere und sind allgemein beliebt (Satow, 2012).
5. ***N**euroticism:* Der Neurotizismus bezieht sich auf die Nervosität, Ängstlichkeit und die Gefühlsschwankungen (Asendorpf & Neyer, 2012, S. 107). Personen, die einen hohen Wert in dieser Dimension haben, sind gemäss Satow (2012) oft angespannt, zweifeln und können weniger gut mit Stress umgehen.

Die Kürzel der englischen Dimensionen ergeben den Namen OCEAN, unter dem der Test auch bekannt geworden ist und mit dessen Hilfe sich die Big Five gut merken lassen (Asendorpf & Neyer, 2012, S. 107).

4.1.2 Fragebogen

Heutzutage können online diverse Fragebögen zur Erfassung der Big Five gefunden werden. Asendorpf und Neyer (2012, S. 108) verweisen z.B. auf den Big-Five-Inventar (BFI) von Lang et al. (2001) mit 7-10 Items pro Faktor und dessen Kurzversion (BFI-S) von Gerlitz und Schupp (2005) mit nur drei Faktoren pro Dimension. In der vorliegenden Arbeit wird der Big-Five-Persönlichkeitstest (B5T®) von Dr. Lars Satow verwendet. Der Test wurde bereits mehr als 50'000 Mal durchgeführt und ist als Paper-Pencil, Excel sowie als Online Version verfügbar und darf für nichtkommerzielle Forschungs- und Unterrichtszwecke verwendet werden (Satow, 2012). Der Test umfasst die Erhebung von 72 Items und erfasst nicht nur die fünf grundlegenden Persönlichkeitsdimensionen, sondern auch drei Grundmotive ‚Bedürfnis nach Anerkennung und Leistung', ‚Bedürfnis nach Einfluss und Macht' und ‚Bedürfnis nach Sicherheit und Ruhe' sowie eine Kontrolle von Testverfälschungen durch positive Selbstdarstellungen (Satow, 2012). Die Masterthesis konzentriert sich zwar nur auf die Big Five Faktoren, jedoch können so die anderen drei Faktoren mitgesammelt werden und für eine spätere Analyse verwendet werden. Von Vorteil ist auch, dass mittels der Testverfälschung nicht seriöse Clickverhalten ausgeschlossen werden können.

4.1.3 Persönlichkeitstypen

Persönlichkeitstypen können auf zwei unterschiedliche Arten klassifiziert werden. Die erste Variante ist, dass zwei unterschiedliche Beurteiler anhand von „streng gehandhabten" Regeln unabhängig voneinander die Personen klassifizieren, so dass eine intersubjektive Objektivität erreicht wird (Asendorpf & Neyer, 2012, S. 82). Diese intersubjektive Objektivität wird dann erreicht, wenn die beiden Klassifizierungen für fast alle Personen identisch sind. Die zweite Variante beruht auf der automatisierten Einteilung der Personen in Extremgruppen, also in unterschiedliche Gruppen mit möglichst hohen oder niedrigen Ausprägungen in einer bis mehreren der Eigenschaftsvariablen (Asendorpf & Neyer, 2012, S. 111). Die zweite Variante hat den Vorteil, dass nicht zwei Beurteiler benötigt werden, da die Beurteilung maschinenunterstützt durchgeführt werden kann. Asendorpf et al. zeigen auf, dass mittels einer Clusteranalyse der Big Five Faktoren Persönlichkeitstypen ermittelt werden können (2001). Mittels dem statistischen Verfahren der Clusteranalyse lassen sich die Teilnehmer aus einer Stichprobe in Gruppen ähnlicher Profile einteilen, dem sogenannten prototypischen Profil. Gemäss Herzberg und Roth (2006) ist eine Stichprobengrösse unter 500 Teilnehmern nicht ausreichend, um eine stabile Clusterlösung zu finden. Bei der Wahl der Stichprobe für eine solche Studie ist es daher wichtig, dass eine heterogene Stichprobe mit genügend Personen verwendet wird. Clusteranalysen sind nachweislich sensibel gegenüber Stichprobengrössen und -zusammensetzung (Aldenderfer & Blashfield, 1996) und daher kann es bei homogenen Stichproben zu einer Verzerrung der Clusterlösung und somit auch zu einer Verzerrung der Persönlichkeitstypen führen (Pöschl, 2010). Mittels diesem prototypischen Profil können anschliessend andere Personen einem ähnlichen Typ zugeordnet werden, in dem sie demjenigen prototypischen Profil zugwiesen werden, das die kleinste Distanz zum Profil der Person aufweist (Asendorpf & Neyer, 2012, S. 113). Für die Bestimmung der kleinsten Distanz wird bei der Clusteranalyse das sogenannte Ähnlichkeitsmass verwendet, das in diesem Anwendungsgebiet gemäss Asendorpf und Neyer (2012) der euklidischen Distanz entsprechen sollte und definiert wird als „die Wurzel aus der Summe der quadrierten Differenzen in den einzelnen Eigenschaften" (S. 112).

In ihrer Studie können Asendorpf et al. (2001) belegen, dass mittels der Clusteranalyse Persönlichkeiten in drei Hauptprototypen beschrieben werden können. Jeder dieser Prototypen wird durch ein eigenes Dimensionspattern der Big Five Dimensionen ausgezeichnet und somit durch das mittlere Profil des Clusters charakterisiert. Die ermittelten Prototypen bezeichneten sie als stressresistente (Resilients), überkontrollierte (Overcontrollers) und unterkontrollierte (Undercontrollers) Persönlichkeiten. Asendorpf et al. (2001) beschreiben die drei Persönlichkeitstypen wie folgt:

- Der resilente Persönlichkeitstyp wird durch eine sehr hohe Stressresistenz und Belastbarkeit ausgezeichnet.
- Der überkontrollierte Persönlichkeitstyp charakterisiert sich durch übermässige Emotions- und Motivationskontrolle.
- Der unterkontrollierte Persönlichkeitstyp charakterisieret sich im Gegensatz zu dem überkontrollierten Persönlichkeitstyp durch mangelnde Emotions- und Motivationskontrolle.

Die Bildung dieser drei Typen wurde durch Block (1971) geprägt und in weiteren Studien wie derjenigen von Robins et al. (1998) mit einer Stichprobe von 300 jugendlichen Teilnehmern und in derjenigen von Asendorpf et al. (2001) mit einer weitaus heterogeneren Stichprobe von über 2000 Teilnehmern belegt.

Gemäss Asendorpf et al. (2001) zeichnen sich die drei Persönlichkeitstypen durch die nachfolgenden Big Five Prototypen aus:

- *Resilients:* Die stressresistenten Persönlichkeiten zeichnen sich durch einen sehr sozialen (durchschnittlichen) Wert aus in allen fünf Big Five Faktoren ausser dem Neurotizismus, indem sie einen tiefen Wert aufweisen.
- *Overcontrolled:* Die überkontrollierten Persönlichkeiten sind introvertiert und neurotisch veranlagt, was sich durch einen hohen Neurotizismus- und einen tiefen Extraversionswert wiederspiegelt. Die restlichen Big Five Faktoren bewegen sich um den Mittelwert.
- *Undercontrolled:* Die unterkontrollierten Persönlichkeiten werden als nicht sehr verträglich und nicht sehr gewissenhaft wahrgenommen.

Die Abbildung 7 veranschaulicht das prototypische Profil dieser drei Persönlichkeitstypen nach Asendorpf et al. (2001). Durchschnittliche Werte werden mit einer 5 erreicht, höhere Werte entsprechen einem ausgeprägten Big Five Faktor, tiefere Werte einem weniger ausgeprägten Big Five Faktor.

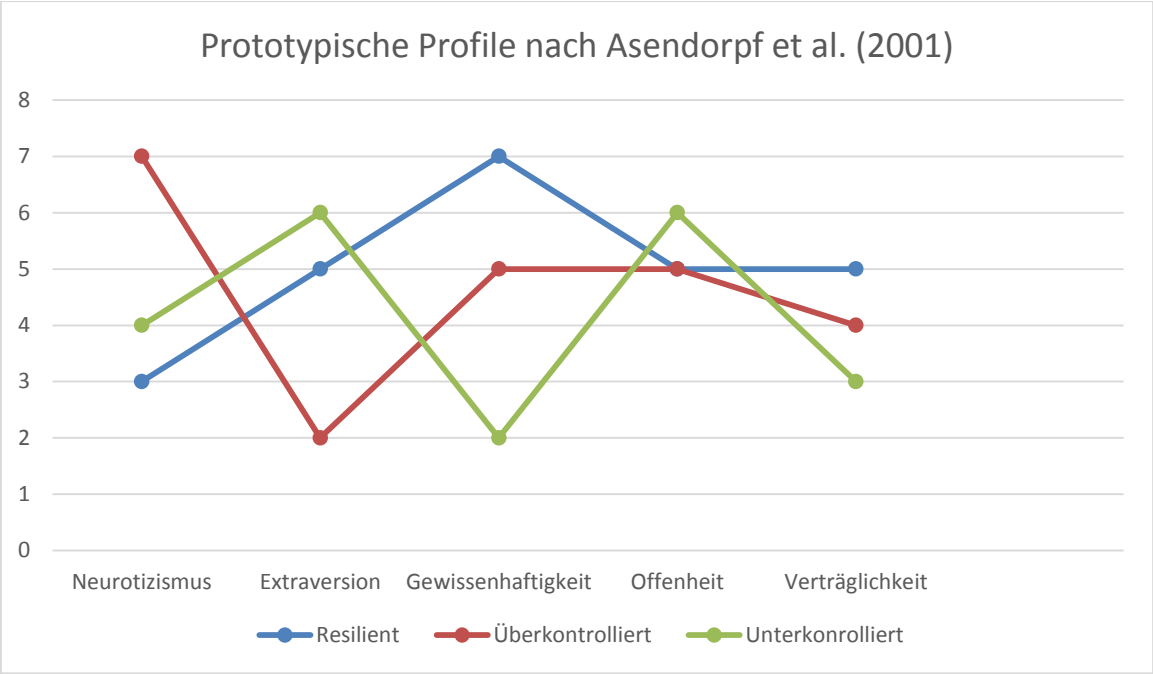

Abbildung 7: Persönlichkeitstypen nach Asendorpf et al. (eigene Darstellung)

In der Studie von Asendorpf et al. (2001) wird aufgezeigt, dass das resiliente prototypische Profil dem grössten Cluster entspricht (49%), gefolgt von der unterkontrollierten Gruppe (28%) und schlussendlich der überkontrollierten Gruppe (23%). Jedoch ist zu erwähnen, dass Asendorpf et al. (2001) vier verschiedene Studien durchgeführt haben, und die Werte der überkontrollierten und unterkontrollierten Gruppe in den jeweiligen Studien unterschiedlich waren, weshalb die Prozentangaben gemittelte Werte über die vier Studien sind.

Herzberg und Roth (2006) konnten aufzeigen, dass eine Dreiclusterlösung basierend auf den drei Typen Resilient, Überkontrolliert und Unterkontrolliert auch in anderen Studien verwendet wurden, die drei prototypischen Profile indes nicht konsistent ausfallen. Die Dreiclusterlösung wird von Herzberg und Roth (2006) als „weit von der Perfektion entfernt“ gekennzeichnet. In einer von Herzberg und Roth (2006) durchgeführten Umfrage mit einer Stichprobe von über 1900 deutschen Erwachsenen konnten sie aufzeigen, dass die Dreiclusterlösung in eine Fünfclusterlösung mit guter Replizierbarkeit unterteilt werden kann.

Neben den bekannten Prototypen Resilient (16%), Überkontrolliert (12%) und Unterkontrolliert (24%) wurde noch der selbstbewusste Prototyp (22%) mit hohen Offenheits- und Extraversionswerten sowie der zurückhaltende Prototyp (26%) mit tiefen Offenheitswerten der Clusterlösung hinzugefügt (siehe auch Abbildung 8).

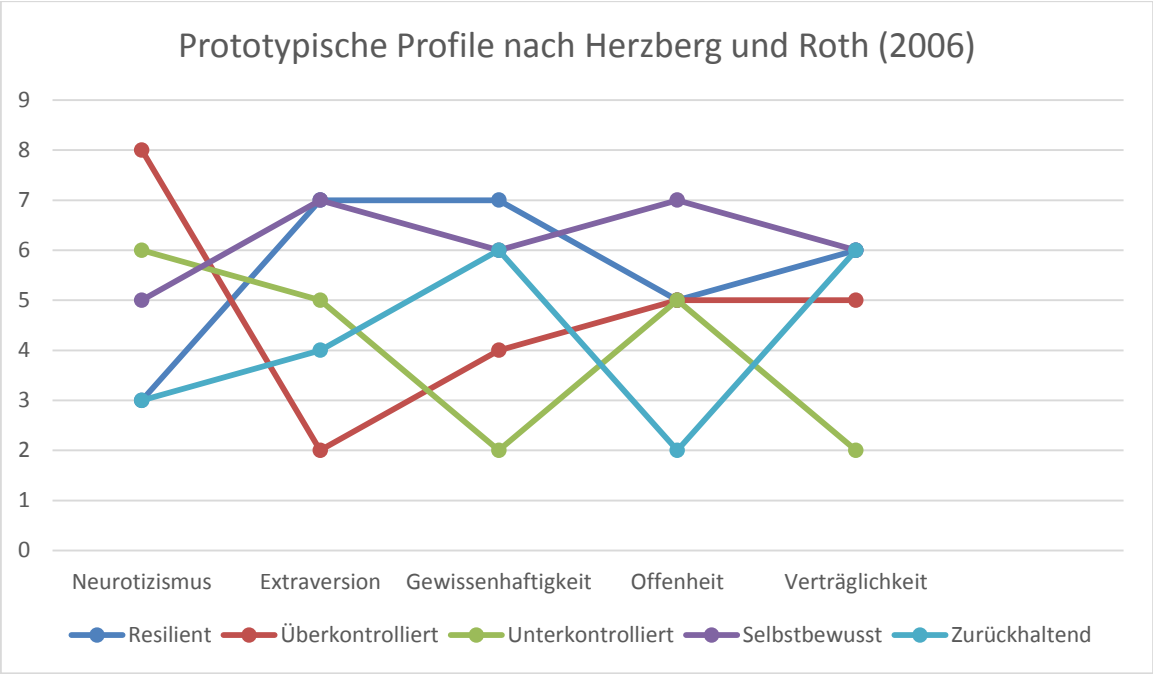

Abbildung 8: Persönlichkeitstypen nach Herzberg und Roth (eigene Darstellung)

4.1.4 Zusammenfassung des Kapitels

In Kapitel 4.1 konnte aufgezeigt werden, dass die Big Five grundlegende Persönlichkeitsfaktoren sind, die durch diverse psychologische Studien nachgewiesen werden konnten. Der Big Five Test ist ein sehr einfach zu bedienender Test, der ohne Supervision durchgeführt und somit ohne Probleme über eine Onlineumfrage an viele Personen verteilt werden kann. Diese Attribute sprechen für den Einsatz dieses Tests für die Beantwortung der Forschungsfrage III, da möglichst viele Personen online nach ihrer Persönlichkeit befragt sowie deren Klickverhalten aufgezeichnet werden soll. Dank der Zurverfügungstellung des B5T von Dr. Satow steht für den Testaufbau in Kapitel 4.2 ein Fragebogen mit 72 Items, eine Excel-Auswertungshilfe sowie eine Test- und Skalendokumentation zur Verfügung, um die gesammelten Ergebnisse des B5T Tests auszuwerten.

Für die Auswertung der Persönlichkeiten kann auf ein sogenanntes prototypisches Profil zurückgegriffen werden. Das prototypische Profil beschreibt eine Gruppe von Charakterzügen, welche aus grossen Stichproben durch eine Clusteranalyse ermittelt werden konnte. In Kapitel 4.1.3 wurden zwei Modelle mit prototypische Profilen vorgestellt, die einmal die Personen in ein Dreiercluster und einmal in ein Fünfercluster gruppieren. Mittels diesen zwei Modellen können die Ergebnisse aus dem Clickstreamclustering aus Kapitel 4.3.3 mit den Modellen verglichen werden, um eine Aussage darüber zu treffen, ob aus dem Navigationsverhalten von Benutzern auf die Persönlichkeit zurückgeschlossen werden kann. Der Vergleich mit diesen Modellen wird benötigt, da für ein repräsentatives Clustering der Persönlichkeiten eine Stichprobe mit über 500 Teilnehmern benötigt wird. Eine solch grosse Stichprobe konnte im Rahmen dieser Masterarbeit nicht erhoben werden.

4.2 Testaufbau

Das vorliegende Kapitel beschreibt den Testaufbau für die Beantwortung der Forschungsfrage III. Mittels dem Testaufbau soll untersucht werden können, ob das Benutzerverhalten durch eine Umfrage, bestehend aus zwei Teilen, bestimmt werden kann. Neben der Beschreibung des Testaufbaus sowie des Samplings wird in diesem Kapitel auch noch die Architektur der Umfrage aufgezeigt, die für die Datenerhebung und -auswertung genutzt worden ist.

4.2.1 Quantitative Research

Für die quantitative Forschung werden die von Allan et al. (2012, S. 16) vorgeschlagenen Methoden zur Beobachtung der Benutzung eines Informationssystems durch einen User eingesetzt. Dazu wird in einer ersten Phase eine kontrollierte Überwachung von Personen, die sich in Interaktion mit einem System befinden, durchgeführt. Dem Benutzer wird genau beschrieben, welche Aufgaben er als Teilnehmer der Studie durchzuführen hat. Zudem werden demografische Beschreibungen der Teilnehmer sowie allenfalls Vorkenntnisse des Benutzers beim Einsatz mit dergleichen Systemen erhoben. Wichtig ist, dass ein Instrument genutzt wird, das die Evaluation der Nutzbarkeit der erhobenen Daten belegen kann. Als Instrument in dieser Arbeit wird der von Dr. Satow (2011) erstellte Big-Five-Persönlichkeitstest eingesetzt. Mittels dem Test kann eine befragte Person in die fünf grundlegenden Persönlichkeitsdimensionen der Psychologie eingestuft werden (siehe auch Kapitel 4.1.1) und erfasst zudem die Grundmotive 'Bedürfnis nach Anerkennung und Leistung', 'Bedürfnis nach Einfluss und Macht' und 'Bedürfnis nach Sicherheit und Ruhe' (Satow, 2017). Das Testmanual mit den Normen sowie der Fragebogen in der Excel-Version wurden von Dr. Satow für die Durchführung dieser Arbeit zur Verfügung gestellt.

Ein Test basierend auf dem Five Factor Model wurde bereits von Heinström erfolgreich eingesetzt, die mit einem klassischen Fragebogen bestehend aus 60 Fragen zur Persönlichkeit und 70 Fragen zum Informationsverhalten nachweisen konnte, dass drei unterschiedliche Muster von Informationsverhalten existieren (Heinström, 2003).

Mittels einem Logging von Interaktionen innerhalb einer Session wird in einem zweiten Schritt der Umfrage das Interaktionsverhalten des Benutzers protokolliert. Gespeichert werden implizite Indikatoren wie Klicken, Verweilen auf einer Seite oder das Zurückkehren zu einer bestimmten Seite und limitierte Kontextinformation des Nutzers wie z.B. der Ort sowie der Zeitraum der Nutzung durch einen Benutzer. Eine ausführliche Beschreibung der impliziten Daten, die gesammelt wurden, ist in Kapitel 2.2.3 zu finden.

Gemäss Womser-Hacker und Mandl (2014) etablieren sich bei der Analyse des Informationssuchverhaltens immer mehr Mischmethoden aus qualitativen und quantitativen

Betrachtungen, die synergetisch ineinandergreifen und voneinander profitieren. Mit der Kombination dieser zwei Vorgehensweisen während der Umfrage kann einerseits ein psychologisches Profil gebildet, und andererseits können die gesammelten Log-Daten der einzelnen Nutzer geclustert werden.

Für das Logging wird eine in die Umfrage integrierte HTML Applikation genutzt, die denselben Informationsgehalt anzeigt wie später die Produkte innerhalb der Applikation. Ein Benutzer sollte dabei möglichst ohne Vorgabe bestimmter Tasks sein natürliches Informationssuchverhalten zeigen können. Für die Umfrage wurden vier unterschiedliche Produkte ausgewählt, die möglichst ein breites Spektrum an Personen ansprechen soll und trotzdem genügend unterschiedlich voneinander sein sollen, um unterschiedliche Navigationsverhalten zu erzeugen. Daher wurde bei der Auswahl der Produkte auf vier unterschiedliche Produktklassen gesetzt: Food (Chiquita Banana), Automobil (Audi A3), Spiele (Nintendo Wii) sowie Möbel (Billy von Ikea). Innerhalb der einzelnen Produkte wurden jeweils dieselben vier unterschiedlichen Subkategorien abgebildet: Beschreibung über das Unternehmen, Herkunft des Produktes, Rabatte und Spezifikationen bzw. Inhaltsstoffe zum Produkt. Auch hier war das Ziel, dass die Benutzer des Systems möglichst gut differenzieren zu können, in dem die Kategorien sehr unterschiedlich voneinander sind. Der Benutzer soll dazu bewogen werden, eine Priorisierung über die Geschichte, den ökologischen Aspekt, Sondervorzügen oder Zahlen zum Produkt vorzunehmen und diese in seinem Navigationsverhalten widerzuspiegeln.

Wie bereits in Kapitel 2.2.3 aufgezeigt werden konnte, werden heute auch vermehrt Social- und Multimediainhalte genutzt. Daher wurde in der Umfrage jeweils ein Video zum Produkt integriert und ein Favoriten-Button hinzugefügt, über den ein Produkt für einen späteren Schnellzugriff markiert werden kann. Sämtliche dieser unterschiedlichen Produkte, Subkategorien und sozial- sowie medialen Interaktionen dienen dazu, die individuelle Persönlichkeit eines jeden Teilnehmers anzusprechen, um einen möglichst unvoreingenommenen und individuellen Clickstream zu erzeugen.

Sämtliche Daten werden anonym erhoben und nicht an dritte Personen weitergereicht. Ein entsprechender Vermerk zur Datensicherheit und -nutzung wird in die Umfrage integriert. Da psychologische Tests generell nur von psychologisch geschultem Personal ausgewertet werden sollten (Satow, 2017), wird dem Befragten keine Auswertung des Persönlichkeitsprofiles am Ende der Befragung angezeigt.

4.2.2 Sampling

Generell sollen mit der Umfrage so viele Personen wie möglich erreicht werden. Daher werden nur wenige Limitierungen für die Auswahl der Befragten gesetzt. Bei Jugendlichen in

der Entwicklungsphase hat sich die Persönlichkeit noch nicht gefestigt (Satow, 2017), so dass die Ergebnisse nicht sehr aussagekräftig sind. Daher werden für den Test nur Teilnehmer berücksichtigt, die mindestens 16 Jahre alt sind. Zudem steht der Test nur in deutscher Sprache zur Verfügung, weshalb die Umfrage auch nicht in einer anderen Sprache durchgeführt werden kann. Auf eine Übersetzung wird verzichtet, da unklar ist, ob dies die Ergebnisse nicht verfälschen würde.

Die Verteilung der Umfrage wird mittels eines Schneeballsystems durchgeführt. Die Umfrage soll sich über Kollegen und Familie weiterverteilen, um möglichst viele Personen in der zur Verfügung stehenden Zeit zu erreichen. Bevor die Umfrage weitergeleitet wird, wird mit einer kleinen Auswahl von 3-5 Testpersonen ein Pretest durchgeführt. Der Pretest wird unter denselben Bedingungen wie die spätere Umfrage durchgeführt und soll Verbesserungspotential an der Art der Umfrage sowie den einzelnen Fragen aufdecken.

4.2.3 Architektur des Umfragetools

Dieses Unterkapitel erläutert, wie die die Architektur der Umfragetools gemäss den Grundlagen aus dem Kapitel 2 umgesetzt worden ist. Die Architektur richtet sich nach den Phasen der Datensammlung (Collection), Datenaufnahme (Ingestion), Speicherung der Daten (Storage), Verarbeitung der Daten (Processing) und der Analyse der Daten (Analyzing). Wie in Kapitel 2 dargelegt, kann für die Verbindung der einzelnen Phasen untereinander eine Orchestrierungsinstanz verwendet werden kann. Um den Aufwand für die Implementation der Architektur zur Erhebung und Auswertung der Umfragedaten gering zu halten, wurde auf eine solche Orchestrierungsinstanz verzichtet, und die einzelnen Phasenübergänge manuell durchgeführt.

Abgrenzung

Da keine automatisierte Orchestrierung implementiert wird, wird das in Kapitel 2.3.2 aufgezeigte Dateisystem HDFS nur für die Sammlung der Clickstreamdaten eingesetzt. Für sämtliche weiteren Dateien wird ein normales Dateisystem eingesetzt, weil dieses im manuellen Einsatz einfacher zu handhaben ist als ein HDFS, das sämtliche Dateien in einer internen Struktur abspeichert.

Komponenten und Frameworks

Für die Sammlung der Daten auf dem Client, wird wie bereits in Kapitel 4.2.1 beschrieben, eine HTML Applikation eingesetzt. Als Webserver für die Verteilung der HTML Applikation wird ein nginx Server[5] verwendet, da sich dieser durch seinen geringen Bedarf an CPU und Hauptspeicher auszeichnet.

[5] https://nginx.org/en/

Die HTML Applikation basiert auf einem AngularJS Client[6], der mit einem Spark Server[7] sowie einem Divolte Collector[8] kommuniziert. AngularJS eignet sich sehr gut, um dynamische Seiten zu generieren und das Default HTML Vokabular mit eigenen Tags (sogenannte Komponenten) zu erweiterten.

Die Umsetzung des B5T wird mittels dem Framework survey.js vorgenommen[9]. Das Framework bietet online einen visuellen Editor, um eine Umfrage vorzubereiten und sämtliche Fragen und Antworten in einem JSON File zu exportieren. Mittels einer AngularJS Komponente kann das JSON dann einfach im Client hinzugeladen werden, und das Framework übernimmt die Visualisierung sowie die Prozessführung durch die Umfrage.

Betriebskonzept

Der Betrieb der Applikation wird über die Amazon Web Services[10] sichergestellt (siehe Abbildung 9). Sämtliche für die Umfrage benötigten Komponenten sind auf einer Elastic Cloud Compute (EC2) Instanz installiert. Die EC2 Instanz kann mit verschiedenen Betriebssystemen ausgestattet werden. Für die vorliegende Masterarbeit wird das von Amazon angebotene Linux den Anforderungen der Komponenten gerecht. Für die Speicherung der anfallenden Daten wird der EC2 Instanz ein acht Gigabyte grosser Elastic Block Store (ECB) zugewiesen. Eine Security Group stellt sicher, dass der eingehende Datenverkehr nur über den Default-Port 80 vorgenommen werden kann. Ein den Komponenten vorgelagerter nginx Server wird als Reverse Proxy eingesetzt, um den eingehenden Datenverkehr auf die verschiedenen Komponenten (HTML Ressourcen, Spark Server, Divolte Collector) zu verteilen.

[6] https://angularjs.org/
[7] https://spark.apache.org/
[8] http://divolte.io/
[9] http://surveyjs.org/
[10] https://aws.amazon.com

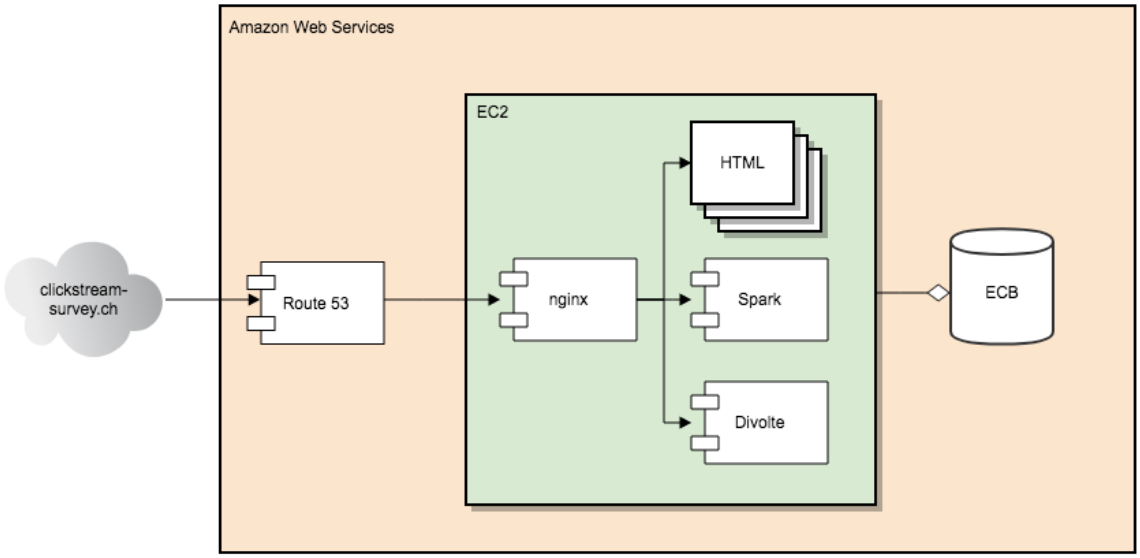

Abbildung 9: Betriebskonzept (eigene Darstellung)

Mittels dem Service „Route 53“ von Amazon, wird die IP Adresse der EC2 Instanz an die Domäne http://clickstream-survey.ch gebunden, die für die Durchführung der Umfrage registriert worden ist.

Collection und Ingestion

Das Sammeln und Aufnehmen der Umfrage- und Clickstreamdaten entspricht der Informationserarbeitung, die Kuhlen (1995) in seinem Mikromodell der Informationsarbeit definiert. Die informationellen Ressourcen sind im vorliegenden Fall die Umfrage- und Clickstreamdaten, die durch die beiden Informationserarbeitungsschritte 1 & 2 (siehe Abbildung 10) in Relevanzinformation übergeführt werden um, anschliessend in der Informationsverwaltung (Storage) abgespeichert zu werden. Eine Beschreibung der Informationserarbeitungsschritte findet sich in der anschliessenden Aufzählung.

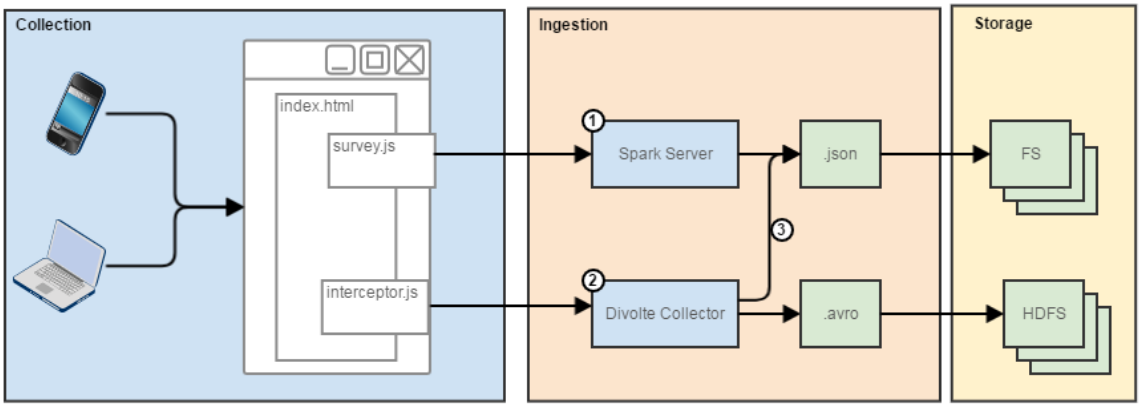

Abbildung 10: Collection und Ingestion (eigene Darstellung)

Informationserarbeitung

1. Das Ergebnis einer Umfrage wird als JSON aus dem survey.js Framework exportiert und mittels einem HTTP POST Request an den Spark Server gesendet. Der Spark Server stellt sicher, dass sämtliche benötigten Antworten vorhanden sind und speichert die empfangenen JSON Daten in einem separaten File auf dem Dateisystem (FS) des EBS-Volume. Für jede Umfrage wird ein neuer Ordner auf dem Dateisystem erstellt und erhält den Session-Identifier als Ordnernamen, um später sämtliche Files einer Session schnell wieder aufzufinden.
2. Das Clickverhalten eines Benutzers wird vom ersten Aufruf der Seite bis zum Ende der Teilnahme an der Umfrage gesammelt und asynchron (im Hintergrund) an den Divolte Collector gesendet. Der Divolte Collector mapped die im JSON Format empfangenen Daten gemäss dem in Kapitel 2.2.3 beschriebenen Schema (Tabelle 3: Clickstream Schema) und speichert die Avro Files im HDFS, das sich ebenfalls auf dem EBS-Volume befindet.
3. Um das anschliessende Processing im Prototypen einfacher zu gestalten, werden sämtliche Events direkt während der Ingestion neben dem Avro Format auch im JSON Format zusammen mit den bereits erhobenen Events abgespeichert.

Processing

Sämtliche Prozessschritte der Abbildung 11 entsprechen der Informationsaufbereitung aus dem Mikromodell der Informationsarbeit (Kuhlen, 1995). Die während der Ingestion abgespeicherten Relevanzinformation wird in dieser Phase für die anschliessende Analyse der Daten aufbereitet und wiederum in der zentralen Informationsverwaltung abgespeichert. Eine Beschreibung der Informationsaufbereitungsschritte findet sich in der anschliessenden Aufzählung.

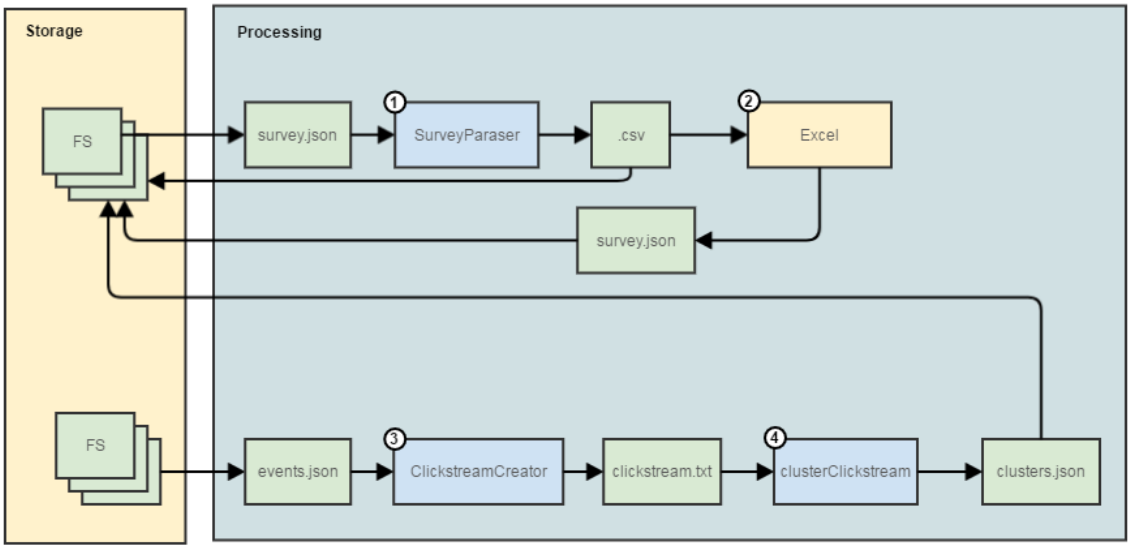

Abbildung 11: Processing (eigene Darstellung)

Informationsaufbereitung

1. Für die Auswertung und Analyse der B5T Daten müssen die im JSON-Format empfangenen und gespeicherten Umfragedaten in ein neues Format überführt werden, damit diese im Excel genutzt werden können. Dazu wird eine Java-Applikation eingesetzt, die die Umfragedaten vom JSON-Format in ein CSV-Format umwandelt und im Session-Ordner abspeichert.
2. Der Schritt zwei der Informationsaufbereitung der B5T Daten für die Erstellung des Benutzerprofils entspricht einem manuellen Schritt. Die CSV Daten werden in das Excel geladen und das Persönlichkeitsprofil des Teilnehmers wird im ursprünglichen Umfrage JSON ergänzt.
3. Die abgespeicherten Events im JSON Format werden im dritten Informations-aufbereitungsschritt von einer Java Applikation ausgelesen und sämtliche User-Events in ein zentrales Clickstream-File überführt, das vom Divisive Hierachical Clustering Algorithmus genutzt wird. Verschiedene Möglichkeiten zur Modellierung können genutzt werden. Ein Vergleich der Modellierungen findet sich in Kapitel 4.3.3. Dieses File wird nur temporär benötigt und daher nicht in der Informationsverwaltung abgelegt. Eine Verbesserung des Prototyps wäre hier anstelle der Events im JSON-Format die im HDFS abgelegten Avro Files zu nutzen, um von sämtlichen in Kapitel 2.3.2 vorgestellten Vorteilen profitieren zu können.
4. Die Clickstream-Datei wird im vierten Informationsaufbereitungsschritt vom Clustering-Algorithmus verarbeitet und der daraus resultierende Clusteroutput, welcher sämtliche Cluster beinhaltet im Dateisystem für die spätere Analyse abge-speichert. Im Rahmen des Prototyps wird der Output als Listen von Benutzer-IDs pro Cluster in einem zentralen Dokument abgespeichert.

Analyzing

Die Prozessschritte der Abbildung 12 entsprechen der Informationsverarbeitung aus dem Mikromodell der Informationsarbeit (Kuhlen, 1995). Die während der Informationsaufbereitung abgespeicherten Ressourcen werden während der Informationsverarbeitung aus der Informationsverwaltung gelesen und in Handlungsinformation überführt, die der Problemlösung (der Beantwortung der Forschungsfrage III) dienen sollen. Als grundlegendes Format für die Handlungsinformation wurde ein CSV Format gewählt, da die Daten sich so einfach im Excel visualisieren lassen. Eine Beschreibung der Informationsverarbeitungs-schritte findet sich in der anschliessenden Aufzählung.

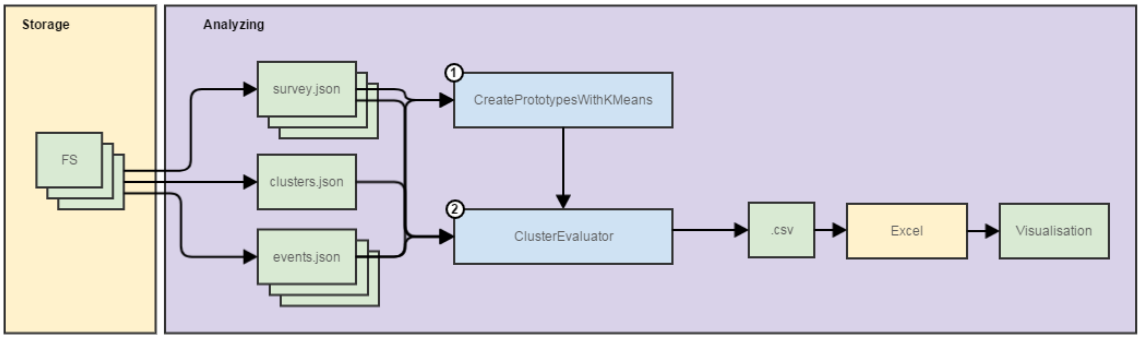

Abbildung 12: Analyzing (eigene Darstellung)

Informationsverarbeitung

1. Sämtliche Umfrageergebnisse werden mithilfe eines Java-Programms unter Anwendung der euklidischen Distanz den prototypischen Profilen aus Kapitel 4.1.3 zugeordnet. Die Clusterergebnisse werden als Input für eine künftige Analyse im ClusterEvaluator abgespeichert. Die Zuteilung der Umfrageergebnisse zu den prototypischen Profilen soll als Vergleichswert dienen, um einerseits die in der Literatur gefundenen Stereotypen zu prüfen sowie um die Clusterergebnisse der Benutzerinteraktionen aus dem Schritt zwei zu verifizieren.
2. Mittels dem ClusterEvaluator werden die Ergebnisse aus dem Clickstream-Clustering mit den zuvor aufbereiteten Daten zusammengeführt und der F1-Score sowie weitere Werte zu den Clusterergebnissen in einer CSV-Datei für die anschliessende Analyse gespeichert.

Auf die Ergebnisse der Informationsverarbeitung wird im folgenden Kapitel genauer eingegangen.

4.3 Analyse der Daten

4.3.1 Statistik

Insgesamt sind 126 Rückmeldungen auf die Onlineumfrage eingegangen. Diese unterteilen sich in 55 Frauen und 71 Männer.

Die zwei nachfolgenden Tabellen zeigen sowohl die Altersverteilung sowie die digitale Versiertheit im Umgang mit elektronischen Geräten zwischen Frauen, Männern und im Totalen.

Alter	Frauen	Männer	Total
Jünger als 20	0	0	0
20 bis 30 Jahre	20	13	33
31 bis 40 Jahre	27	33	60
41 bis 50 Jahre	3	15	18
Älter als 50 Jahre	5	10	15

Tabelle 13: Altersstatistik

Erfahrung	Frauen	Männer	Total
Keine Erfahrung	0	0	0
Tiefe Erfahrung	5	1	6
Mittlere Erfahrung	20	18	38
Hohe Erfahrung	30	52	82

Tabelle 14: Digitale Versiertheit

Es zeigt sich, dass eine homogene Gruppe mit der Umfrage erreicht werden konnte, wobei ein Grossteil der Teilnehmer sich im Alter zwischen 31 und 40 Jahren befindet und eine mittlere bis grosse Erfahrung im Umgang mit elektronischen Geräten vorweist.

Die Tabelle 15 zeigt die Klickrate (Anzahl Klicks/Sekunde) der unterschiedlichen Kategorien.

Gruppe	Klicks	Zeit	Klickrate
Frauen	25	106s	4.3
Männer	30	175s	5.8
20 bis 30 Jahre	27	186s	**6.8**
31 bis 40 Jahre	27	111s	4.1
41 bis 50 Jahre	30	157s	5.2
Älter als 50 Jahre	25	172s	**6.9**
Tiefe Erfahrung	34	128s	**3.7**
Mittlere Erfahrung	24	147s	6.1
Hohe Erfahrung	28	144s	5.1

Tabelle 15: Klickrate

Die Auswertung der Klickrate zeigt erstaunlicherweise, dass diejenigen User mit der tiefsten Erfahrung im Umgang mit elektronischen Geräten die schnellste Klickrate haben, während die zwei Altersgruppen ‚20 bis 30 Jahre' sowie ‚älter als 50 Jahre' die langsamsten Klickraten aufweisen. Eine schnelle Klickrate bedeutet, dass wenig Zeit zwischen den einzelnen Klicks vergeht, während bei einer hohen Klickrate längere Zeit zwischen zwei Klicks vergeht.

4.3.2 Validierung der Persönlichkeitsprototypen

In der Validierung der Persönlichkeitsprototypen wird belegt, inwiefern die drei in Kapitel 4.1 vorgestellten Persönlichkeitsprototypen Resilient, Überkontrolliert und Unterkontrolliert nach Asendorpf et al. (2001) und die fünf Persönlichkeitsprotoypen Resilient, Überkontrolliert, Unterkontrolliert, Selbstbewusst und Zurückhaltend nach Herzberg und Roth (2006) in den erhobenen Big Five Profilen der einzelnen Teilnehmer wiederentdeckt werden können. Dazu wird eine Diskriminanzfunktion verwendet, die mithilfe der euklidischen Distanz eine Zuteilung der Big Five Profile auf die Prototypen vornimmt. Die z-Score Werte der Prototypen von Asendorpf et al. (2001) und Herzberg und Roth (2006) werden auf die Skala des B5T® Test von Dr. Satow umgerechnet und mittels der euklidischen Distanz die geringste Distanz zu einem der Prototypen berechnet. Mit diesem Vorgehen der Validierung kann ein Vergleich der erhobenen Profile vorgenommen werden, ohne dass die Ergebnisse von der Sensitivität der Sample-Grösse sowie Zusammensetzung abhängig sind (Herzberg & Roth, 2006).

Für den Prototypen von Asendorpf et al. (2001) ergeben sich die folgenden drei Vektoren:

Resilient	**new int**[] { 3, 5, 7, 5, 5 }
Überkontrolliert	**new int**[] { 7, 2, 5, 5, 4 }
Unterkontrolliert	**new int**[] { 4, 6, 2, 6, 3 }

Tabelle 16: Vektoren Asendorpf et al. (2001)

Und für den Prototypen von Herzberg und Roth (2006) ergeben sich die folgenden fünf Vektoren:

Resilient	**new int**[] { 3, 7, 7, 5, 6 }
Überkontrolliert	**new int**[] { 8, 2, 4, 5, 5}
Unterkontrolliert	**new int**[] { 6, 5, 2, 5, 2 }
Selbstbewusst	**new int**[] { 5, 7, 6, 7, 6 }
Zurückhaltend	**new int**[] { 3, 4, 6, 2, 6 }

Tabelle 17: Vektoren Herzberg und Roth (2006)

Die in den Tabelle 16 und Tabelle 17 dargestellten Vektoren entsprechen den Faktoren: *{Neurotizismus, Extraversion, Gewissenhaftigkeit, Offenheit, Verträglichkeit}*

Die Zuteilung der einzelnen Profile zu diesen Vektoren wurde mithilfe eines Java Programms StatDump[11] umgesetzt. Die Ergebnisse der Zuteilungen sind in Tabelle 18 und Tabelle 19 aufgelistet. Den Tabellen kann entnommen werden, dass die prozentuale Zuteilungsrate, die in der Literatur gefunden wird, nicht mit den Ergebnissen aus dieser Zuteilung korreliert. Während die Zuteilung in die drei Prototypen nach Asendorpf et al. (2001) noch eine Zuteilung auf sämtliche Cluster zulässt, zeigt sich bei der Zuteilung zu den Prototypen von Herzberg und Roth (2006), dass den beiden Prototypen Resilient und Überkontrolliert keine (bzw. fast keine) Probanden zugeteilt werden.

Diese Unterschiede können darauf hindeuten, dass das erhobene Sample nicht die gewünschte Heterogenität erreicht hat, jedoch merken Herzberg und Roth (2006) auch an, dass der Vergleich mittels einer Diskriminanzfunktion für andere Kulturen als in Deutschland noch empirisch erhoben werden sollten. Allgemein könne eine Abhängigkeit von kulturellen Einflüssen jedoch ausgeschlossen werden, da den Big Five Faktoren gültige interkulturelle Vergleiche nachgewiesen werden konnte.

Profil	Zuteilungen	Anzahl	%	% Literatur
Resilient	1, 10, 12, 20, 26, 32, 41, 46, 50, 54, 59, 62, 63, 81, 92, 93, 102, 110, 111, 112, 117, 120	22	17	49
Über-kontrolliert	2, 4, 15, 16, 19, 21, 22, 23, 27, 29, 30, 35, 36, 38, 39, 43, 44, 45, 47, 57, 61, 64, 66, 67, 69, 71, 72, 78, 82, 84, 85, 89, 90, 91, 94, 96, 97, 101, 103, 105, 106, 113, 114, 118, 124	45	36	23
Unter-kontrolliert	3, 5, 6, 7, 8, 9, 11, 13, 14, 17, 18, 24, 25, 28, 31, 33, 34, 37, 40, 42, 48, 49, 51, 52, 53, 55, 56, 58, 60, 65, 68, 70, 73, 74, 75, 76, 77, 79, 80, 83, 86, 87, 88, 95, 98, 99, 100, 104, 107, 108, 109, 115, 116, 119, 121, 122, 123, 125, 126	59	47	28

Tabelle 18: Prototypzuteilung Asendorpf et al. (2001)

[11] https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/utils/StatDump.java

Profil	Zuteilungen	Anzahl	%	% Literatur
Resilient	-	0	0	16
Über-kontrolliert	30, 35	2	1	12
Unter-kontrolliert	3, 4, 5, 9, 11, 12, 14, 17, 18, 22, 23, 25, 27, 28, 31, 32, 33, 34, 37, 38, 39, 40, 42, 45, 48, 49, 51, 52, 53, 55, 56, 57, 58, 59, 60, 65, 67, 68, 70, 71, 73, 74, 75, 76, 77, 78, 79, 80, 82, 83, 84, 85, 86, 87, 88, 90, 91, 93, 97, 98, 99, 100, 104, 105, 112, 113, 115, 116, 119, 121, 122, 123, 125, 126	74	59	24
Selbst-bewusst	7, 8, 10, 20, 24, 29, 66, 81, 92, 101, 102, 103, 107, 110, 111, 117, 120, 124	19	15	22
Zurück-haltend	1, 2, 6, 13, 15, 16, 19, 21, 26, 36, 41, 43, 44, 46, 47, 50, 54, 61, 62, 63, 64, 69, 72, 89, 94, 95, 96, 106, 108, 109, 114, 118	32	25	26

Tabelle 19: Prototypzuteilung Herzberg und Roth (2006)

Da die Zuteilung auf die Prototypen von Herzberg und Roth (2006) schlechter ausfällt als diejenige von Asendorpf et al. (2001) und allgemein in der Literatur viel mehr Prototypen basierend auf drei Clustern zu finden sind (siehe auch Vergleich von Herzberg und Roth (2006)), werden für die Auswertung des Clickstreams die drei Prototypen von Asendorpf et al. (2001) verwendet. Da jedoch auch in dieser Zuteilung der Vergleich nicht wie gewünscht ausgefallen ist, werden neben diesen Prototypen noch drei Referenz-Prototypen für diese Studie erstellt. Dazu werden die erhobenen Profile mittels dem K-Means Algorithmus unter Verwendung der euklidischen Distanz in drei Cluster aufgeteilt. Verwendet wird dazu eine K-Means Library der Cognitive Foundry[12] und die Implementation der Prototypenerstellung findet sich in der Java Applikation CreatePrototypesWithKMeans[13] (siehe auch Kapitel 7.2). Für die Ermittlung der Vektoren aus Tabelle 20 wurden ausreichend Iterationen durchgeführt und die einzelnen Ergebnisse gemittelt, so dass das finale Ergebnis nicht von der Random-Initialisierung der drei Cluster abhängig ist. Mehrere Kontrolldurchgänge bestätigen das Ergebnis.

[12] https://github.com/algorithmfoundry/Foundry

[13] https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/kmeans/CreatePrototypesWithKMeans.java

Prototyp 1	**new int**[] { 6, 6, 5, 4, 4 }
Prototyp 2	**new int**[] { 5, 5, 4, 4, 3 }
Prototyp 3	**new int**[] { 5, 5, 3, 4, 3 }

Tabelle 20: Vektoren Referenz Prototypen

Ein Vergleich dieser drei Prototypen zeigt, dass Prototyp 1 sich durch erhöhte Neurotizismus sowie Extraversions Werte auszeichnet. Während der Prototyp 2 ziemlich durchschnittliche Werte und lediglich eine etwas tiefere Verträglichkeit aufweist, hat der Prototyp 3 auch tiefe Werte in der Gewissenhaftigkeit.

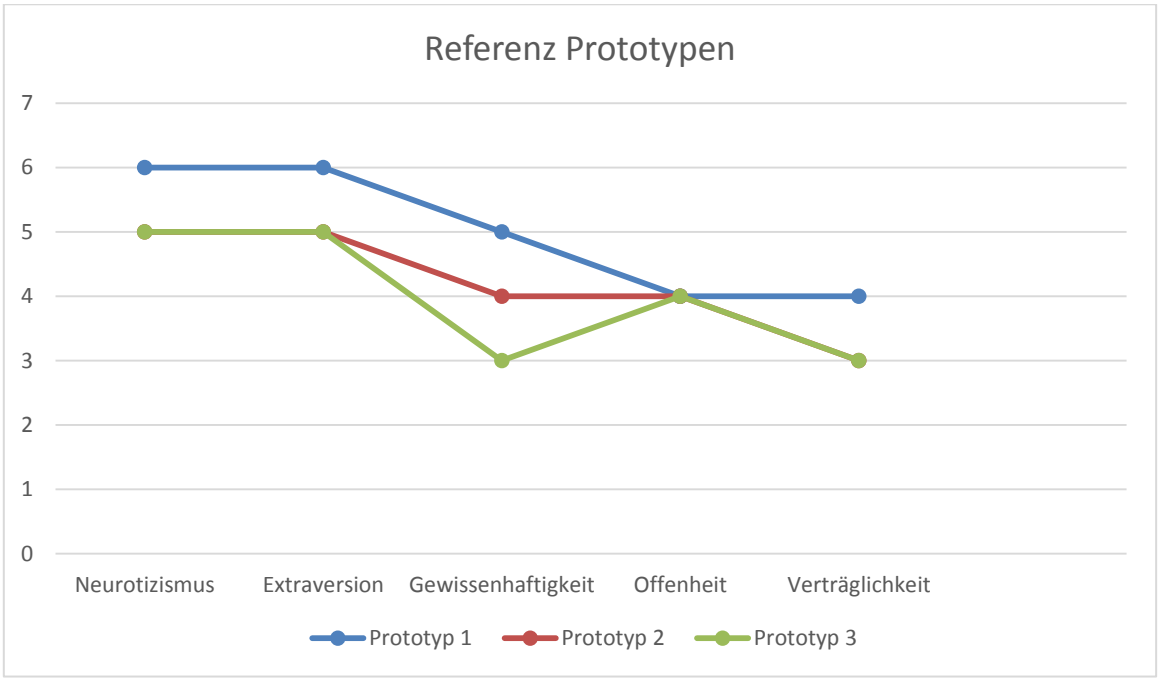

Abbildung 13: Referenz Prototypen (eigene Darstellung)

Ein Vergleich dieser Referenz Prototypen mit den Prototypen aus Abbildung 7 und Abbildung 8 zeigt auf, dass die Prototypen nicht mit denjenigen aus der Literatur übereinstimmen. Einzig der Prototyp 3 scheint gewisse Ähnlichkeiten mit den unterkontrollierten Prototypen aus der Literatur zu haben.

4.3.3 Clickstream Clustering

Im diesem Kapitel werden die erhobenen Clickstream Daten der einzelnen Benutzer mit dem in Kapitel 3.3 ausgewählten Divisive Hierarchical Clustering Algorithmus geclustert, und die einzelnen Zuordnungen der Profile gegenüber den Referenz Prototypen sowie den Prototypen von Asendorpf et al. (2001) geprüft. Dazu werden verschiedene Modellierungen

verwendet, um die Ergebnisse zu optimieren. Anschliessend an die Optimierung wird eine Validierung der Ergebnisse und eine Interpretation der Clusterergebnisse vorgenommen.

Modellierungen zur Optimierung der Ergebnisse

Wie bereits in der Einleitung erwähnt, muss ein Algorithmus in einer letzten Phase optimiert werden, um die besten Ergebnisse für die Data-Mining Aufgabe zu erreichen (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Für den in dieser Arbeit verwendeten DHC Algorithmus bedeutet dies, dass die Modellierung der Navigationssequenzen unterschiedlich vorgenommen werden kann und durch das Feature Pruning jeweils anders geartete Unterbäume entstehen. Auch der beim Algorithmus definierbare Parameter Threshold hat einen Einfluss auf die Ergebnisse der Cluster, da dieser steuert, inwiefern die Gruppen noch weiter unterteilt werden können. In den nachfolgenden Tests wurde der Threshold jeweils so getunt, dass möglichst drei Hauptgruppen entstehen, ohne dass eine grosse Anzahl an Ausreissern resultieren.

Für die Variation der Modellierung der Clickstreams wurden Modelle erstellt, die sich einerseits in der Abstraktionsstufe der Events unterscheiden, andererseits in den Zeitintervallen zwischen den Events. Insgesamt wird zwischen drei Abstraktionsstufen und drei Zeitintervallstufen unterschieden. Die Abkürzungen in Klammern ergeben den Namen der Modellierung.

Abstraktionsstufen:

- High Level (HL). Höchste Abstrahierung. Die Events werden auf die grundlegendsten Aktionen wie OPEN, VIDEO oder BACK zurückgeführt. Dabei ist nicht ersichtlich, welches Produkt oder Detail der Benutzer gerade angeklickt hat, oder wie er mit dem Video interagiert.
- Medium Level (ML). Im Gegensatz zu HL wird bei ML jeweils das Produkt oder Detail zu dem Event hinzugefügt. Auch beim Video ist z.B. ersichtlich, ob es auf einem Produkt oder auf einem Detail angeklickt worden ist.
- Low Level (LL). Tiefstes Level der Abstraktion. Aus jedem Event ist ersichtlich, zu welchem Produkt und Detail er gehört. Während im ML Ansatz vom Detail nicht auf das Produkt geschlossen werden kann, ist dies im LL Ansatz möglich. Zudem wird unterschieden, ob der Benutzer ein Video abspielt, stoppt oder zu Ende ansieht.

Zeitintervalle:

- Sekunden (S): Die Zeit, die zwischen den Events liegt wird in Sekunden angegeben.
- Millisekunden (M): Die Zeit, die zwischen den Events liegt wird in Millisekunden angegeben.
- Normalisierte Zeiteinheit (N): Die Zeit, die zwischen den Events liegt wird in einer normalisierten Zeiteinheit angegeben. Für die Normalisierung wurde die Fibonacci-Reihe verwendet, um eine gute Trennung zwischen grossen Intervallen und kleinen

Intervallen zu erreichen. Sämtliche Events über 30 Sekunden werden dabei demselben Bucket zugeteilt.

Als weitere Variation wurden zudem jeweils zwei Durchgänge mit demselben Clickstream durchgeführt. Die beiden Durchgänge unterscheiden sich wie folgt:

1. Alle Events werden berücksichtigt.
2. Der erste Open-Event, der bei jedem Benutzer derselbe ist, wird aus der Sequenz entfernt, um sicherzustellen, dass diese Gemeinsamkeit zwischen allen Sessions nicht zu einer Überanpassung führt.

Die Tabelle 21 zeigt Beispiele für sämtliche Modellierungsarten, die aus diesen unterschiedlichen Variationen entstehen können:

Modellierung	**Beispiel**
HLS1	open(1)open(1)open(1)video(8)back(37)
HLS2	open(1)open(1)video(8) back(37)
HLM1	open(1)open(1)open(1)video(8566)video(13707) back(36951)
HLM2	open(1)open(1)video(8566)back(36951)
HLN1	open(1)open(1)open(1)video(13)back(55)
HLN2	open(1)open(1)video(13)back(55)
MLS1	openProduct(1)openProduct0(1)openDetail2(1)videoDetail2(8) backDetail2(37)
MLS2	openProduct0(1)openDetail2(1)videoDetail2(8)backDetail2(37)
MLM1	openProduct(1)openProduct0(1)openDetail2(1)videoDetail2(8566)backDetail2(3096)
MLM2	openProduct0(1)openDetail2(1)videoDetail2(8566) backDetail2(3096)
MLN1	openProduct(1)openProduct0(1)openDetail2(1)videoDetail2(13) backDetail2(55)
MLN2	openProduct0(1)openDetail2(1)videoDetail2(13)backDetail2(55)
LLS1	open(1)open0(1)open02(1)playingVideo02(8)back00(36)
LLS2	open0(1)open02(1)playingVideo02(13)back00(36)
LLM1	open(1)open0(1)open02(1)playingVideo02(8566)back00(36951)
LLM2	open0(1)open02(1)playingVideo02(8566)back00(36951)

LLN1	open(1)open0(1)open02(1)playingVideo02(13) backToProduct03(55)
LLN2	open0(1)open02(1)playingVideo02(13)backToProduct03(55)

Tabelle 21: Modellierungsarten

Die Modellierungen der einzelnen Clickstreams wurden mit der Java Applikation ClickstreamConverter[14] vorgenommen. Die daraus resultierenden Inputfiles für das Clustering finden sich im Results-Verzeichnis[15] im Git Projekt.

Validierung der Modelle

Für die Überprüfung der aus dem DHC resultierenden Cluster wird das F-Measure als externes Validierungsmass verwendet, da dieses sowohl die Precision der Clusterzuteilungen als auch den Recall berücksichtigt. Recall, Precision und das F-Mass sind etablierte Masse, die für die Evaluation einer Klassifizierung verwendet werden können (Mandl, 2014). Obwohl die Ergebnisse durch Clustering und nicht durch eine Klassifizierung entstanden sind, wird in diesem Absatz evaluiert, wie gut die Klassifizierung gegenüber einem Gold Standard Testset (den Prototypen) ist. Optimal wäre, wenn sowohl Precision als auch Recall Werte 100% erreichen würden. Sowohl Precision als auch Recall sind wichtig, um eine möglichst hohe Clusterstabilität zu erreichen. Denn wenn die Precision hoch ist, jedoch ein tiefer Recall erreicht wird, bedeutet dies automatisch, dass ein anderes Cluster einen schlechteren Precision Wert aufweisen wird. Für die Berechnung des F1-Wert der resultierenden Cluster mit den Prototypen wird die Java Applikation CompareToPrototypes[16] verwendet. Dabei werden die drei Cluster aus dem Clustering in voller Kombinatorik mit den jeweiligen Prototypen verglichen und das Resultat mit dem höchsten durchschnittlichen F1-Wert ausgewählt.

Die Ergebnisse der Clusterings mit den Referenz Prototypen als Vergleich können der Tabelle 22 entnommen werden. Der Vergleich der Clustering zu den Prototypen von Asendorpf et al. ist in der Tabelle 23 dokumentiert. Ein detaillierter Clusteroutput zu sämtlichen Durchgängen ist im Dokument Evaluation[17] hinterlegt.

[14] https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/utils/ClickstreamConverter.java

[15] https://github.com/svenlenz/clickstream-survey/tree/master/results/clickstreams

[16] https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/utils/CompareToPrototypes.java

[17] https://github.com/svenlenz/clickstream-survey/blob/master/results/clickstreams/evaluation.docx

Modus	Threshold / Ausreisser	Cluster 1			Cluster 2			Cluster 3			AVG F-Measure
		Fm	Pr	Re	Fm	Pr	Re	Fm	Pr	Re	
HLS1	0.18 / 0	0.386	0.654	0.27	0.4	0.313	0.556	0.381	0.334	0.445	0.388
HLS2	0.18 / 0	0.383	0.654	0.27	0.4	0.313	0.556	0.381	0.334	0.445	0.388
HLM1	0.12 / 0	0.404	0.513	0.334	0.193	0.313	0.139	0.355	0.247	0.63	0.317
HLM2	0.18 / 0	0.404	0.513	0.334	0.193	0.313	0.139	0.355	0.247	0.63	0.317
HLN1	0.14 / 0	0.51	0.628	0.429	0.416	0.324	0.584	0.178	0.223	0.149	0.368
HLN2	0.14 / 0	0.496	0.62	0.413	0.219	0.316	0.167	0.37	0.262	0.63	0.361
MLS1	0.5 / 3	0.286	0.786	0.175	0.372	0.296	0.5	0.347	0.271	0.482	0.335
MLS2	0.5 / 3	0.297	0.667	0.191	0.379	0.306	0.5	0.411	0.327	0.556	0.363
MLM1	0.5 / 3	0.54	0.54	0.54	0.086	0.182	0.056	0.422	0.327	0.593	0.349
MLM2	0.15 / 3	0.366	0.567	0.27	0.219	0.316	0.167	0.377	0.257	0.704	0.321
MLN1	0.5 / 2	0.286	0.786	0.175	0.405	0.318	0.556	0.46	0.362	0.63	0.384
MLN2	0.5 / 2	0.472	0.582	0.397	0.282	0.286	0.278	0.439	0.348	0.593	0.398
LLS1	0.5 / 2	0.47	0.52	0.429	0.325	0.316	0.334	0.296	0.265	0.334	0.363
LLS2	0.25 / 0	0.329	0.275	0.408	0.381	0.334	0.445	0.416	0.553	0.334	0.376
LLM1	0.25 / 1	0.509	0.526	0.493	0.348	0.364	0.334	0.334	0.304	0.371	0.397
LLM2	0.5 / 4	0.477	0.596	0.397	0.348	0.364	0.334	0.352	0.277	0.482	0.392
LLN1	0.4 / 1	0.471	0.616	0.381	0.4	0.364	0.445	0.406	0.334	0.519	0.426
LLN2	0.6 / 1	0.539	0.523	0.556	0.381	0.334	0.445	0.055	0.1	0.038	0.325

Tabelle 22: Clusterergebnisse Referenz Prototypen

Modus	Threshold / Ausreisser	Cluster 1			Cluster 2			Cluster 3			AVG F-Measure
		Fm	Pr	Re	Fm	Pr	Re	Fm	Pr	Re	
HLS1	0.18 / 0	0.167	0.154	0.182	0.371	0.417	0.334	0.488	0.469	0.509	0.342
HLS2	0.18 / 0	0.167	0.154	0.182	0.371	0.417	0.334	0.488	0.469	0.509	0.342
HLM1	0.12 / 0	0.158	0.188	0.137	0.457	0.377	0.578	0.4	0.488	0.339	0.339
HLM2	0.18 / 0	0.158	0.188	0.137	0.457	0.377	0.578	0.4	0.488	0.339	0.339
HLN1	0.14 / 0	0.15	0.167	0.137	0.41	0.419	0.4	0.501	0.477	0.526	0.354
HLN2	0.14 / 0	0.147	0.158	0.137	0.391	0.405	0.378	0.484	0.462	0.509	0.341
MLS1	0.5 / 3	0.223	0.286	0.182	0.474	0.459	0.489	0.567	0.558	0.577	0.421
MLS2	0.5 / 3	0.56	0.56	0.56	0.462	0.457	0.467	0.25	0.278	0.228	0.424
MLM1	0.5 / 3	0.061	0.091	0.046	0.447	0.429	0.467	0.558	0.54	0.577	0.355
MLM2	0.15 / 3	0.196	0.211	0.182	0.555	0.446	0.734	0.405	0.6	0.306	0.385
MLN1	0.5 / 2	0.278	0.358	0.228	0.435	0.426	0.445	0.591	0.572	0.611	**0.435**
MLN2	0.5 / 2	0.281	0.229	0.364	0.44	0.435	0.445	0.393	0.466	0.339	0.371
LLS1	0.5 / 2	0.3	0.237	0.41	0.279	0.324	0.245	0.487	0.52	0.458	0.355
LLS2	0.25 / 0	0.258	0.188	0.41	0.448	0.475	0.423	0.454	0.579	0.373	0.386
LLM1	0.25 / 1	0.219	0.182	0.273	0.359	0.425	0.312	0.526	0.526	0.526	0.368
LLM2	0.5 / 4	0.348	0.256	0.546	0.359	0.425	0.312	0.456	0.548	0.39	0.388
LLN1	0.4 / 1	0.263	0.206	0.364	0.414	0.429	0.4	0.525	0.614	0.458	0.401
LLN2	0.6 / 1	0.063	0.1	0.046	0.447	0.374	0.556	0.486	0.542	0.441	0.332

Tabelle 23: Clusterergebnisse Prototypen nach Asendorpf et al. (2001)

Interpretation des Clustering

Die besten Ergebnisse werden mit der LLN1 Modellierung (F-Score 0.426) für die Referenz Prototypen und der MLN1 Modellierung (F-Score 0.435) für die Prototypen nach Asendorpf et al. (2001) erreicht. Das F-Mass kann Werte zwischen 0 und 1 annehmen, woraus bereits ersichtlich wird, dass ein Wert mit 0.4 keine optimalen Lösung für das gesuchte Problem ist. Um diese Ergebnisse noch detaillierter betrachten zu können, zeigen Tabelle 24 und Tabelle 25 die Verteilungsraten der Clickstream-Clusterergebnisse gegenüber den Einteilungen, wenn die Big Five Profile geclustert werden. Aus den Tabellen lässt sich ableiten, dass in sämtlichen Clickstream-Clusterergebnissen Profile aus unterschiedlichen Prototypen präsent sind. Ein Pattern, das auf bestimmte Persönlichkeitsprofile innerhalb eines Clusters schliessen lässt, ist aus diesen Ergebnissen nicht ableitbar, da die Cluster zu ähnlich ausfallen und keine klaren Tendenzen auszumachen sind.

Cluster	Anzahl	Prototyp 1	Prototyp 2	Prototyp 3
Cluster 1	44	52%	12%	36%
Cluster 2	42	35%	35%	30%
Cluster 3	39	61%	20%	18%

Tabelle 24: Verteilungsrate LLN1 gegenüber Referenz Prototypen

Cluster	Anzahl	Resilient	Überkontrolliert	Unterkontrolliert
Cluster 1	63	12%	30%	57%
Cluster 2	47	17%	42%	40%
Cluster 3	14	35%	42%	21%

Tabelle 25: Verteilungsrate MLN1 gegenüber Prototypen nach Asendorpf

Auch ein visueller Vergleich der gemittelten Big Five Faktoren in Abbildung 14 und Abbildung 15 zeigt keine markanten Unterschiede innerhalb der jeweiligen Cluster. Während beim LLN1 die Cluster 1 und 3 genau gleich ausfallen und lediglich das Cluster 2 sich durch einen tieferen Neurotizismuswert differenziert, sind die Unterschiede beim MLN1 ein wenig grösser, da sich das Cluster 3 allgemein mit höheren Werten als Cluster 1 und 2 auszeichnet.

Auch wenn die Unterschiede der Persönlichkeitsprofile nicht so hoch ausfallen, kann anhand des Clusterings trotzdem ein Merkmal identifiziert werden, mittels dem zumindest ein Recommendersystem basierend auf Interessengruppen von Produkten aufgebaut werden

könnte. Tabelle 42 illustriert für den MLN1 Ansatz die Anzahl Klicks, die pro Produkt innerhalb eines Clusters getätigt worden sind. Aus dieser Analyse ergibt sich, dass Cluster Gruppe 1 die meisten Klicks auf Produkt 4, Gruppe 2 auf Produkt 2 und Gruppe 3 auf Produkt 3 gemacht haben. Daraus lassen sich nun Regeln für weitere Produktvorschläge ableiten. Beispielweise kann für das Cluster 1 das Produkt 3 vorgeschlagen werden, für das Cluster 2 die Produkte 2 und 3,sowie für das Cluster 3 die Produkte 1 und 4.

4.4 Zusammenfassung des Kapitels

In diesem Kapitel konnte aufgezeigt werden, wie sich das Clickstream Clustering mit der anschliessenden Analyse der Benutzerprofile innerhalb der segmentierten Gruppen zur Beantwortung der Forschungsfrage III umsetzten lässt. Es werden Teile aus der Datenmanagementplattform aus Forschungsfrage I und der Algorithmus aus Forschungsfrage II genutzt, um zwei weitere wichtige Komponenten jedes Data-Mining Algorithmus beantworten zu können: die Evaluation des Modells sowie die Suche nach der besten Optimierung.

Für die Evaluation des Modells wurden zwei verschiedene Gruppen von prototypischen Profilen erstellt. Die erste Gruppe von prototypischen Profilen wurde in der Literatur von Asendorpf et al. eingeführt (2001), die zweite Gruppe an prototypischen Profilen wurde aus den erhobenen Daten abgeleitet und als Referenz Prototypen gekennzeichnet. Mittels diesen zwei Gruppen von Prototypen konnten zwei unterschiedliche Gold Standard Clusterings basierend auf den Persönlichkeitstest-Resultaten aufgebaut werden. Mittels diesen Gold Standard Clusterings, die unabhängig vom Clickstream der jeweiligen User gebildet worden sind, konnten die effektiven Clickstream-Clusterergebnisse gegen diese Standards verglichen werden.

Für den Vergleich wurden unterschiedlichste Modellierungsarten der Inputdaten miteinander verglichen, um das bestmögliche Ergebnis zu erhalten. Dabei wurden verschiedene Abstraktionsstufen der Events sowie unterschiedliche Arten der Zeitintervalls-Einheiten miteinander verglichen. Dank dem F-Mass konnten die besten Clusterergebnisse dieser unterschiedlichen Modellierungen ausgewählt werden. Für die Prototypen von Asendorpf et al. (2001) erwies sich das MLN1 (Medium Level Abstraktion mit normalisierter Zeiteinheit) als bestes Clusterergebnis, während bei den Referenz Prototypen das LLN1 (Low Level Abstraktion mit normalisierter Zeiteinheit) am besten abgeschnitten hat.

Ein Vergleich der Persönlichkeitsprofile dieser zwei Ergebnisse offenbart, dass das MLN1 differenziertere Profile innerhalb der ermittelten Cluster liefert als der LLN1 Ansatz. Indessen zeigt sich auch, dass die gemittelten Persönlichkeitsprofile der einzelnen Cluster weit von den ursprünglichen Prototypen abweichen und darum anhand des Clickstream-Clustering

nicht auf die für die Evaluation ausgewählten Persönlichkeitsprototypen zurückgeführt werden kann.

Demzufolge kann draus geschlossen werden, dass das in dieser Arbeit vorgeschlagene Vorgehen zur Segmentierung von Benutzern in Gruppen von ähnlichen Persönlichkeitsprofilen zu homogen ausgefallen ist, und nur bedingt für den Einsatz für spezifische Marketingaktivtäten geeignet ist. Hingegen konnte aufgezeigt werden, dass innerhalb der Cluster klare Präferenzen bzgl. der Produktauswahl bestehen. Folgedessen könnte das hier beschriebene Vorgehen genutzt werden, um personalisierte Webinhalte (Produktvorschläge) basierend auf weiteren beliebten Produkten innerhalb eines Clusters vorzunehmen.

5 Zusammenfassung und Diskussion der Ergebnisse

Mit dieser Arbeit konnte ein vertiefter Einblick in die Auswertung von Benutzerinteraktionen mit einem Informationssystem mittels der Definition einer modernen Datenmanagement-plattform, einer Evaluation eines Clickstream-Algorithmus sowie einer Auswertung der resultierenden Cluster gewonnen werden. Die Beantwortung der drei aufgestellten Forschungsfragen dient schlussendlich dazu, dem Benutzer des Informationssystems einen wirkungsbezogenen individuellen informationellen Mehrwert zu liefern; er kann bisherige Tätigkeiten schneller durchführen und seine Ziele können besser erreicht werden. Das Profiling des Benutzers dient dabei als Treiber für diese Wertsteigerung, so dass durch eine Segmentierung der Benutzermodelle eine massgeschneiderte Customer Experience angeboten werden kann und dadurch Kundenloyalität aufgebaut wird. Aus Sicht des Anbieters eines Informationssystems entsteht ein Vendor Lock-In Effekt, der sicherstellt, dass durch die Kundenloyalität eine Kundenabwanderung verhindert werden kann.

Die drei Hauptkapitel 2, 3 und 4 haben bereits im Fazit jeweils einen Bezug zu der jeweiligen Forschungsfrage aufgezeigt. In der Diskussion werden diese Antworten nochmals aufge-arbeitet und kritisch reflektiert, um einen Ausblick auf weitere Forschungsrichtungen in diesem Bereich geben zu können.

5.1 Konzeption einer Datenmanagement Plattform

Mittels der Forschungsfrage I wird die Grundlage einer Datenmanagement-Plattform geschaffen, die notwendig ist, um ein Profiling von Benutzern überhaupt vornehmen zu können. Die Forschungsfrage kann dem Bereich des Machine Learning und Data Mining zugeordnet werden und wird in der Literatur als Web Usage Mining oder neuerdings auch als Smart Data Discovery bezeichnet. Dieser Bereich des Data Mining behandelt den Prozess der Analyse von Benutzerinteraktionen mit einem Informationssystem und besteht aus den drei Phasen (1) Aufbereitung der Daten, (2) Ermitteln von Patterns in den Daten und (3) Vermitteln und Operationalisierung der Ergebnisse. Für die Forschungsfrage I ist die erste Phase von besonderer Relevanz, während die Phasen 2 und 3 im Rahmen der Forschungsfrage III von besonderer Bedeutung sind. Für die Aufbereitung der Daten wird eine Systemarchitektur, bestehend aus den drei Bereichen Collection (Datenerhebung), Ingestion (Datenaufnahme) und Storage (Datenspeicherung) konzipiert. Die Collection der Daten wird bereits auf dem Client vorgenommen, wodurch die Performance der Daten-aufnahme erhöht werden kann und somit eine Verarbeitung von Logfiles, wie es in vielen Artikeln aus der Literatur vorgeschlagen wird, hinfällig wird. Mittels der Definition eines

Clickstream Schema können die Click-Events, die für eine spätere Suche nach Patterns benötigt werden, bereits auf dem Client semantisch definiert und abgespeichert werden.

Das Gesamtkonzept behandelt erfolgreich die von Isele und Arnd formulierten Herausforderungen an ein modernes Datenmanagement. Die Systemarchitektur eignet sich daher als Grundlage für eine Datenmanagement Plattform für die Informationsverwaltung des Benutzerverhaltens, womit die Forschungsfrage I dieser Arbeit beantwortet werden konnte.

Die konzipierte Datenmanagement Plattform konnte in dieser Arbeit in einem Prototyp umgesetzt und zur Datenerhebung, Datenaufnahme und der anschliessenden Analyse der gespeicherten Daten genutzt werden. Rückwirkend lässt sich aus den mit dem Protoypen gemachten Erfahrungen bestätigen, dass die Datenerhebung direkt auf dem Client eine geeignete Methode ist. Durch die Definition des Clickstream-Schema konnten die Events, die in der Analyse benötigt wurden, einfach am Ort des Entstehens erhoben werden und während der Analyse wird nur ein geringer Aufwand benötigt, um diese Clickstream-Events in ein auswertbares Format überzuführen. Ein gutes Beispiel für die Einfachheit der Lösung ist die Erhebung des Zeitintervalls, die auf dem Client direkt zur Verfügung steht; basierend auf einem Serverlog müsste dieser rückwirkend berechnet werden. Zudem stehen alle im Clickstream-Schema definierten Events zur Verfügung, während das Serverlog nur Daten von Anfragen beinhaltet, die auch beim Server eingetroffen sind. Mit den heutigen modernen javascript Webapplikationen, bei denen eine grosser Teil der Logik auf dem Client liegt und die Kommunikation mit dem Server immer geringer wird, ist dieser Ansatz sicherlich besser geeignet. Jedoch verbirgt sich dahinter auch der Nachteil, dass nur Events zur Verfügung stehen, die im Clickstream Schema definiert sind. Werden neue Anforderungen bekannt, können diese zwar im Schema ergänzt werden, indessen können rückwirkend keine Auswertungen auf diesen neuen Anforderungen durchgeführt werden, da diese Events nicht gespeichert worden sind. Natürlich wäre denkbar, dass in so einem Fall versucht werden kann, die Daten aus einer anderen Quelle wie z.B. einem Serverlog oder einer Datenbank zu ermitteln und die gespeicherten Daten mit den neuen Anforderungen zu ergänzen.

Die meisten Daten konnten mit dem Divolte.io Framework erhoben werden, indes gab es einige Teilnehmer, die zu Beginn der Umfrage mitteilten, dass die Umfrage nicht nutzbar sei. Es hat sich herausgestellt, dass das Divolte Collector Skript von gewissen Browsern wegen einem aktiven AdBlocker geblockt war. Um die Erhebung der Umfrage nicht unnötig zu verzögern, konnte das Problem schnell mit einem selbstgeschriebenen Collector-Skript behoben werden, das nicht von einem AdBlocker erkannt wird und die Events als Backup in einem JSON File auf dem Server abspeichert. Für die Datenauswertung wurden dann schlussendlich diese JSON Files ausgewertet und dadurch rund 20% mehr Ergebnisse für die anschliessende Auswertung zur Verfügung gestanden sind. Der Nachteil besteht darin,

dass die im HDFS abgespeicherten Avro-Files in der Theorie eine sehr gute Performance und Flexibilität versprechen, dies jedoch im Prototypen nicht ausgetestet werden konnte.

Eine Komponente, die in der Literatur erwähnt, jedoch im Prototypen nicht berücksichtigt wurde ist die Orchestrierung. Die Orchestrierungskomponente wäre dafür zuständig, die einzelnen Bereiche automatisiert miteinander zu verbinden, so dass möglichst wenig manueller Aufwand notwendig ist. Im Nachhinein wird sichtbar, dass bereits mit Rund 140 auswertbaren Ergebnissen der manuelle Aufwand sehr viel Zeit in Anspruch nehmen kann: Sei dies vom Kopieren der Clickstream-Events vom Server auf eine Plattform, wo die Analyse durchgeführt wird, dem Übertrag der Umfrage-Daten vom JSON in ein Excel zur Auswertung der Persönlichkeiten oder dem sequentiellen Ablauf von verschiedenen Applikationen zur Analyse der erhobenen Daten. Gemäss der Definition ist eine Smart Data Applikation möglichst einfach von einem Benutzer zu bedienen. Demzufolge kann ohne diese Komponente noch nicht wirklich von einer Smart Data Applikation gesprochen werden.

Bei einer weiterführenden Clickstream-Forschung basierend auf dieser Datenmanagement Plattform wäre der Einsatz einer Orchestrierungskomponente die erste Priorität, um dadurch die Auswertungen und Analysen viel effizienter durchführen zu können. Dies bedeutet, dass initial ein wenig Mehraufwand geleistet werden muss, dafür in den späteren Phasen der Ermittlung von Patterns sowie der Vermittlung der Ergebnisse vielseitigere Ansätze verfolgt werden können, da nicht sämtliche Schritte manuell durchgeführt werden müssen. Die Auswertungen und Analysen würden sich durch den Einsatz dieser Orchestrierungskomponente viel einfacher und effizienter gestalten und dadurch könnte eine Smart Data Applikation erreicht werden, die vom Benutzer einfach zu bedienen ist.

5.2 Algorithmus zur Segmentierung von Benutzerverhalten

Mittels der zweiten Forschungsfrage konnte ein Data-Mining Algorithmus evaluiert werden, der, basierend auf der Datenbasis aus der Datenmanagement Plattform aus Forschungsfrage I, eine Segmentierung der Benutzer nach unterschiedlichem Interaktionsverhalten mit einem Informationssystem vornimmt.

Der Bereich des Web Usage Mining ist ein gut erforschter Teil des Data Mining und es haben sich bereits viele unterschiedliche Algorithmen etabliert. Daher wurden die zur Verfügung stehenden Algorithmen anhand einer Kategorisierung nach den Anwendungsgebieten in die Bereiche der Session und Besucher Analyse, Cluster Analyse und Besucher Segmentierung, Assoziations- und Korrelationsanalyse, Analyse von Sequenz-und Navigationspatterns sowie Klassifizierung unterteilt. Für die Beantwortung der Forschungsfrage II und somit auch des gesamten Forschungsproblems, fällt die Wahl auf den Anwendungsbereich der Cluster Analyse und Besucher Segmentierung, einer Data Mining Technik,

bei der es um das Zusammengruppieren von Items geht, die dieselbe Charakteristik aufweisen. Dieses Anwendungsgebiet des Web Usage Mining ist sehr nützlich bei der demographischen Segmentierung der Benutzer für Marketingaktivitäten, oder um personalisierte Webinhalte an Nutzer mit denselben Interessen anbieten zu können. Dies entspricht den informationellen Mehrwerten, die mit der Beantwortung des gesamten Forschungsproblems resultieren sollen.

Im Anwendungsbereich der Cluster Analyse und Besucher Segmentierung kommen hauptsächlich partitionierende, hierarchische oder modellbasierte Clustering Algorithmen zum Einsatz. Für die Auswahl eines geeigneten Algorithmus wurden daher verschiedene Kriterien aufgestellt, die der Algorithmus erfüllen soll. Die beiden für die Arbeit am wichtigsten eingestuften Kriterien sind die Präzision des Algorithmus sowie die Modellierung der Daten, die vom Algorithmus verwendet werden können. Aus dem Vergleich von verschiedenen Clickstream Modellen lässt sich schliessen, dass einfache Modelle unter Berücksichtigung eines geordneten Clickstream Pfads und der View Time sehr gute Resultate geliefert werden können. Komplexere Modelle mit mehr als diesen zwei Dimensionen erweisen sich zwar ein wenig robuster, dazu müssen für diese Modelle zusätzliche Kosten wie Verarbeitungszeit oder erhöhte CPU-Nutzung miteinberechnet werden. Neben diesen zwei Hauptkriterien spielten bei der Auswahl auch subjektive Kriterien wie die Verständlichkeit und Umsetzbarkeit der Algorithmen eine grosse Rolle. Anhand einer Literaturrecherche dieser Clustering-Bereiche und der Eingrenzung dieser auf den Bereich des Web Usage Mining sowie das Clickstream Clustering, konnten sechs unterschiedliche Verfahren aufgezeigt und unter den gegebenen Kriterien miteinander Verglichen werden (siehe Tabelle 6 - Tabelle 12).

Aus jedem Bereich wurde jeweils ein Verfahren zur genaueren Untersuchung ausgewählt. K-Means wurde als Vertreter der partitionierenden Algorithmen ausgewählt, da dies der meistverwendete und zitierte Algorithmus war, der in der Literaturrecherche gefunden werden konnte. Im Bereich der hierarchischen Algorithmen wurde ein Divisive Hierarchical Clustering basierend auf einem Similiarty Graphen ausgewählt, da der Algorithmus die meisten der aufgestellten Kriterien am besten erfüllt hat. Sowohl die Modellierung als auch die Präzision deckten die Anforderungen und es konnte eine einfach zu verwendende Implementation eines DHC Algorithmus gefunden werden, so dass auch mit einer erfolgreichen Umsetzbarkeit in dieser Arbeit gerechnet werden konnte. Die Probabilist Latent Semantic Analysis wurde als Vertreter der modellbasierten Algorithmen ausgewählt, weil dieses Verfahren leichter umsetzbar war als der Einsatz von Markov Chain Modellen und in der Literatur gute Ergebnisse mit dem PLSA nachgewiesen werden konnten.

Für den Vergleich dieser drei Algorithmen wurde anhand eines Datenset, bestehend aus 50 Clickstreamsessions, die Reproduzierbarkeit der Cluster sowie die Stabilität getestet. Der

DHC Algorithmus ist der einzige der dreien, der bei jedem Clusteringdurchgang dieselben Zuweisungen liefern kann. Sowohl beim K-Means als auch PLSA variierten die Clusterergebnisse zwischen verschiedenen Durchgängen teilweise ziemlich stark. Für die Bestimmung der Reproduzierbarkeit wurde das Datenset, bestehend aus 50 Sessions, einmal halbiert und einmal um 50 weitere Sessions erweitert und geprüft, wie viele der ursprünglichen Zuteilungen noch in denselben Clustern waren. Auch in diesem Vergleich schnitt das DHC besser ab, als die anderen beiden Verfahren.

Da das Divisive Hierarchical Clustering sowohl in Bezug auf Reproduzierbarkeit der Cluster als auch in Bezug auf die Stabilität den anderen beiden Verfahren überlegen ist, konnte als Beantwortung der Forschungsfrage II dieser Algorithmus bestimmt werden.

Das DHC Verfahren konnte in dieser Arbeit in einem Prototyp eingesetzt werden, um die Forschungsfrage III beantworten zu können. Rückwirkend ist es schwierig zu beurteilen, wie gross der Einfluss der Wahl dieses Algorithmus sowie der Modellierung auf die Ergebnisse der Forschungsfrage III und folgedessen auch auf die Ergebnisse des gesamten Forschungsproblems sind. Insgesamt konnte anhand der Theorie zwar aufgezeigt werden, welche Algorithmen theoretisch am besten geeignet sind und anhand der Reproduzierbarkeit sowie Stabilität konnte auch ein finaler Algorithmus ausgewählt werden. Es ist durchaus auch denkbar, dass z.B. ein Verfahren basierend auf einer Markov Modellierung mindestens gleich gute Ergebnisse gebracht hätte. Die Einschränkung auf die drei Algorithmen K-Means, PLSA, und DHC wurde sicherlich auch durch subjektive Einflüsse geprägt, da für diese drei Algorithmen einfach einsetzbare Projekte gefunden werden konnten, so dass der Aufwand für die Modellierung sowie Integration des Prototypen möglichst gering gehalten werden konnte.

Des Weiteren könnten die Tests auf Reproduzierbarkeit und Stabilität noch weiter verbessert werden, in dem für jeden der Algorithmen noch eine Phase der Optimierung durchgeführt würde. Beispielsweise könnten für den K-Means Algorithmus verschiedene Initialisierungsarten der Zentroiden oder auch unterschiedliche Distanzmasse getestet werden, um die Reproduzierbarkeit sowie Stabilität verbessern zu können. Auch der PLSA Ansatz war von einem eindeutigen Wahrscheinlichkeitswert über die drei Cluster abhängig. In der Tat gab es aber auch Lösungen, die eine eindeutige Zuweisung nicht ermöglichen und die Wahrscheinlichkeit einer Zuteilung bei allen Clustern gleich gross war. In diesem Falle wurde einfach das erste Cluster als finales Cluster ausgewählt. Auch hier wäre infolge anderer Modellierungen oder Inputparameter evtl. noch Optimierung der Eindeutigkeit möglich gewesen.

In einer zukünftigen Forschung wäre es daher empfehlenswert, wenn der Fokus auf die Optimierung dieser drei Algorithmen gelegt würde oder sogar noch weitere Algorithmen in

einen direkten Vergleich mit einbezogen würden. Um den Vergleich noch besser zu gestalten, wäre es sinnvoll, eine Gold Standard Datenset zu erarbeiten, mittels dem die Zuteilung der einzelnen Algorithmen besser geprüft werden könnte. Hier stellt sich die Schwierigkeit, wie ein solches Gold Standard Datenset erstellt werden kann, da die Resultate erst durch das Data Mining ersichtlich werden. Eine Forschung in diese Richtung würde die Auswahl geeigneter Algorithmen stark vereinfachen.

5.3 Psychologische Profile in Clickstream Benutzersegmenten

Mit der Forschungsfrage III wird aufgezeigt, wie ein Benutzer vor und während der Interaktion mit einem Informationssystem beobachtet werden kann, um die gesammelten Daten anschliessend so auszuwerten, dass das Clustering von Clickstream-Sessions in ähnliche Benutzerprofile erfolgt.

Für die Beantwortung dieser Forschungsfrage wurde in einem ersten Schritt die Beobachtung des Benutzers vor der Interaktion mit dem Informationssystem umgesetzt. Dazu wurde der B5T® von Dr. Satow ausgewählt. Dabei handelt es sich um einen psychometrischen Fragebogen, dank diesem von jedem Teilnehmer der Studie ein psychologisches Profil, bestehend aus den fünf Faktoren Neurotizismus, Extraversion, Gewissenhaftigkeit, Offenheit und Verträglichkeit erstellt werden kann. Der Test wurde ausgewählt, weil er sehr einfach ohne Supervision durchgeführt werden kann und sich daher gut für den Einsatz in einer Onlineumfrage eignet. Die Onlineumfrage wurde in den Gesamtprototypen integriert, und die Daten wurden auf derselben Datenmanagement Plattform gespeichert, auf der auch die Clickstream Daten gespeichert worden sind.

Der zweite Schritt für die Beantwortung dieser Forschungsfrage bestand darin, den Teilnehmern der Studie ein Informationssystem zur Verfügung zu stellen, in welchem die Click-Events der Teilnehmer beobachtet bzw. gespeichert werden können. Dazu wurde eine HTML Applikation basierend auf dem AngularJS Framework entwickelt, die es ermöglicht, die in Forschungsfrage I erarbeitete Datenmanagement Plattform im Prototypen zu integrieren. Der Prototyp beinhaltet vier Produkte, die möglichst unterschiedlicher Natur sind, so dass sich jeder Benutzer von mindestens einem der Produkte angesprochen fühlt. Innerhalb der Produkte konnten unterschiedliche Inhalte wie z.B. geschichtliche Aspekte über das Unternehmen, Fakten über das Produkt selbst oder auch mediale Inhalte wie eingebettete Videos über das Produkt angesehen werden. Auch hier wurden die Inhalte so gewählt, dass ein möglichst breites Eventspektrum jedem Benutzer angeboten wird und sich die einzelnen Clickstream-Sequenzen durch die Breite an zur Verfügung stehenden Events noch besser segmentieren lassen.

Basierend auf diesen zwei Beobachtungen wurde die abschliessende Evaluierung durchgeführt. Um die Qualität der Clusterergebnisse beurteilen zu können, wurden zwei unterschiedliche Gold Standard Clustergruppen erarbeitet. Die Grundlage dieser Clustergruppen basiert darauf, dass sich unterschiedliche Persönlichkeiten zu prototypischen Profilen zuordnen lassen. In der Literatur gibt es mehrere Definitionen solcher prototypischen Profile. Für die Erstellung des ersten Gold Standards wurden die von Asendorpf et al. (2001) vorgeschlagenen drei prototypischen Profile verwendet, und mittels der euklidischen Distanz sämtliche in der Umfrage erhobenen psychologischen Profile einem dieser drei Profile zugeteilt. Ein zweites Gold Standard Set bestehend aus drei prototypischen Profilen wurde direkt aus den psychologischen Profilen ermittelt, indem unter Einsatz eines K-Means Algorithmus sämtliche Profile geclustert wurden. Als Distanzmass wurde die Euklidische Distanz gewählt und für die Modellierung wurden die Skalenwerte der Big Five Faktoren als Vektoren abgebildet. Die Wahl von zwei Gold Standards wurde getroffen, da es in diesem Forschungsbereich noch keinen etablierten Gold Standard gibt und auch in der Literatur der Psychologie keine Einigkeit darüber herrscht, wie die prototypischen Profile basierend auf den Big Five Faktoren definiert werden können. In der Literatur werden unter anderem sogar fünfer Gruppen aufgezeigt. In dieser Arbeit wurden jedoch bewusst Dreiergruppen gewählt, da zu wenige Umfrageergebnisse zur Verfügung standen, um erfolgreich ein Clustering in fünf Gruppen durchzuführen. Mit der Wahl eines Gold Standards aus der Literatur und eines Gold Standards, der aus den Daten selbst abgeleitet werden konnte, wurden die Chancen erhöht, einen für das Clickstream Clustering validen Gold Standard zu definieren. Anhand dieser Gold Standard Clusterings, die unabhängig vom Clickstream erstellt worden sind, konnten die effektiven Clickstream-Clusterergebnisse messbar gemacht werden. Dazu kam das F-Mass zum Einsatz, dank diesem die Clusterergebnisse auf ihre Präzision sowie Reliabilität gegenüber den Gold Standards verglichen werden konnten.

Zu jeder Data-Mining Aufgabe gehört auch die Phase der Optimierung, in der die Algorithmen und die Inputparameter auf das Data-Mining Problem hin optimiert werden. Für den zugrundeliegenden DHC Algorithmus wurden diverse Modellierungsarten der Clickstreamsequenzen in Bezug auf die Abstraktionsstufen der Events sowie unterschiedlichen Zeitintervalls-Einheiten miteinander verglichen. Durch den Vergleich des F-Mass der einzelnen Ergebnisse konnte aufgezeigt werden, dass eine Medium Level Abstraktion mit einer normalisierten Zeiteinheit basierend auf den prototypischen Profilen von Asendorpf et al. (2001) die besten Ergebnisse liefert.

Es verdeutlicht auch, dass das F-Mass der besten gefundenen Modellierung immer noch sehr tief ausfällt (0.435 bei einer Skala von 0.0 bis 1.0) und prototypischen Profile, die sich durch die Clusterings ergeben, weit von den ursprünglichen prototypischen Profilen abweichen. Die einzelnen Cluster fallen zu homogen aus und daher lässt sich die Forschungs-

frage mit den vorliegenden Ergebnissen nicht abschliessend beantworten, da die Cluster nur bedingt für spezifische Marketingaktivitäten nutzbar sind. Einzig die Individualisierung der Inhalte könnte mit dem aufgezeigten Vorgehen angegangen werden, da aus den Clustergruppen klare Präferenzen hinsichtlich der Produktauswahl ausgemacht werden konnte. Das Vorgehen könnte somit genutzt werden, um personalisierte Produktvorschläge basierend auf weiteren beliebten Produkten innerhalb eines Clusters vorzunehmen.

Mit dieser Forschungsfrage sollte aufgezeigt werden, ob durch die Beobachtung eines Benutzers vor und während der Interaktion mit einem Informationssystem und der Anwendung eines Clickstream Clustering ähnliche Persönlichkeiten pro segmentierte Gruppe nachgewiesen werden können. Aus den Ergebnissen ist ersichtlich, dass dies nicht erreicht werden konnte, da die Clusterings zu homogen ausgefallen sind. Lediglich der Rückschluss auf Produktpräferenzen innerhalb der Cluster konnte nachgewiesen werden, was dem gesamten Forschungsproblem zugutekommt, da dadurch wenigstens ein informationeller Mehrwert für den Benutzer geschaffen werden kann.

Erschwerend für die Erforschung des Clustering von Persönlichkeitsprofilen war, dass keine Gold Standards für die Evaluierung der Ergebnisse zur Verfügung standen, und daher diese für die vorliegende Arbeit erarbeitet werden mussten. Da selbst in der Literatur keine Einigkeit über die prototypischen Profile besteht, ist es durchaus denkbar, dass die beiden erarbeiteten Gold Standard Sets nicht optimal sind. Das Erarbeiten dieser Sets hat sich jedoch bewährt, da dadurch zusammen mit dem F-Mass eine Grundlage geschaffen werden konnte, um die einzelnen Clusterergebnisse überhaupt messbar zu gestalten.

Der in dieser Arbeit vorgestellte Ansatz zur Erarbeitung eines Gold Standard im Clickstream Clustering hat sich bewährt und daher sollte in einer zukünftigen Forschung der Fokus auf die Erarbeitung eines definitiven Gold Standards gelegt werden. Um dies zu erreichen, müssten genügend Umfrageergebnisse erhoben werden, damit die selbst definierten Prototypen eine statistische Relevanz erhalten. Dazu werden mindestens 500 auswertbare Ergebnisse benötigt, von Vorteil wäre, wenn noch mehr zur Verfügung stehen würden.

Ein weiterer Forschungsaspekt wäre, dass anstelle des Big Five Test ein anderer psychometrischer Fragebogen verwendet wird. Dadurch könnte aufgezeigt werden, ob durch die Messung von anderen psychometrischen Faktoren ein Clustering in ähnliche Profile erfolgreicher wäre.

Eine Schwäche des gewählten Vorgehens könnte sein, dass die Clickstream Daten zu gering ausgefallen sind. Rund die Hälfte der Teilnehmer hat 60 oder weniger Clicks durchgeführt. In einer zukünftigen Forschung wäre es angebracht, dass eine umfangreichere Applikation für die Beobachtung der Interaktionen verwendet wird und jeder Benutzer über eine längere Zeit beobachtet wird. Dadurch könnten genügend Daten gewonnen werden, um

während der Phase der Optimierung noch unterschiedlichere Modellierungen miteinander vergleichen zu können und so das am besten geeignete Vorgehen zu finden, das ein gutes bis sehr gutes F-Mass gegenüber dem Gold Standard aufweist.

5.4 Schlussbemerkung

Auch wenn die Forschungsfrage III mit den in dieser Arbeit erhobenen Persönlichkeitsprofilen und Clickstream Daten nicht beantwortet werden konnte, diente diese Forschungsfrage dazu, die zuvor erarbeiteten theoretischen Erkenntnisse der Forschungsfragen I und II in einen praktischen Kontext überzuführen und auszutesten. Die Gliederung der Arbeit in diese drei Forschungsfragen hat sich somit bewährt, denn erst mit den Erkenntnissen aus dem Praxisbezug konnten die Schwächen aus den theoretischen Erkenntnissen aufgezeigt werden. Trotz den aufgezeigten Mängeln zeigt die Arbeit indessen auch, dass viel Potential in der Datenmanagement Plattform steckt und der Algorithmus mit einer grösseren Datenbasis sowie einer erneuten Optimierung erfolgreich eingesetzt werden könnte.

Beim Forschungsproblem dieser Arbeit ging es schlussendlich darum, dem Benutzer einen wirkungsbezogenen individuellen informationellen Mehrwert zu liefern, so dass bisherige Tätigkeiten vom Benutzer schneller durchgeführt und Ziele besser erreicht werden können. Mit der in dieser Arbeit vorgestellten Datenmanagement Plattform und dem Algorithmus zum Clustering der Clickstream Daten kann ein System erarbeitet werden, das dem Benutzer diejenigen Produkte vorschlägt, die andere Nutzer aus seiner Gruppe ebenfalls angeklickt haben. Auch die Navigationsstruktur liesse sich individualisieren, in dem die im Cluster häufig verwendeten Bereiche innerhalb eines Produktes prominenter platziert werden. Der Benutzer könnte dadurch schneller seine gewünschten Daten abrufen und durch die Produktvorschläge weniger Zeit mit der Suche nach geeigneten Produkten verbringen. Das ganze Konzept lässt sich natürlich auch in andere Bereiche übertragen. Zum Beispiel könnten die Navigationsstrukturen innerhalb eines Internet Banking so verbessert werden, dass ein Trader schneller zum Kauf oder Verkauf bei Börsengeschäften kommt.

Mit weiteren Forschungen in diesem Bereich könnte der Umfang der informationellen Mehrwerte, die ein Benutzer durch dieses Clickstream Clustering erhält sicherlich noch weiter ausgebaut werden. Obwohl es noch viel Potenzial hat, kann das Forschungsproblem trotzdem als beantwortet betrachtet werden.

6 Literaturverzeichnis

Adersberger, J. (2016). Clickstream Analysis with Spark. Spark Summit New York.

Aldenderfer, M., & Blashfield, R. K. (1996). Cluster Analysis.

Allan, J., Croft, B., Moffat, A., & Mark, S. (2012). Frontiers, Challenges, and Opportunities for Information Retrieval. *Report from SWIRL 2012.* Lorne.

Asendorpf, J., & Neyer, F. (2012). *Psychologie der Persönlichkeit* (5 Ausg.). Springer.

Asendorpf, J., Borkenau, P., Ostendorf, F., & van Aken, M. (Mai 2001). Carving personality description at its joints: Confirmation of three replicable personality prototypes for both children and adults. *European Journal of Personality* .

Bandari, D., Xiang, S., & Leskovec, J. (2017). Categorizing User Sessions at Pinterest. *KDD 2017*, (S. 9). Halifax.

Belk, M., Papatheocharous, E., Germanakos, P., & Samaras, G. (2013). Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques. *Journal of Systems and Software*.

Block, J. (1971). *Lives through time.* Berkeley, CA: Bancroft books.

Borges, J., & Mark, L. (2004). *A Dynamic Clustering-Based Markov Model for Web Usage.* Abgerufen am 15. Mai 2017 von arXiv:cs/0406032.

Bouveyron, C., & Brunet, C. (2013). Model-Based Clustering of High-Dimensional Data : A review. *Computational Statistics and Data Analysis, 71*, 52-78.

Burton, B., & Smith, S. (2015). *Gartner Hype Cycles 2016.* Abgerufen am 28. April 2017 von http://www.gartner.com/technology/research/hype-cycles/

Cadez, I. V., Heckerman, D., Meek, C., Smyth, P., & White, S. (2003). Model-Based Clustering and Visualization of Navigation. (Springer, Hrsg.) *Data Mining and Knowledge Discovery, 7*(4), S. 399-424.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. *IEEE.*

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*.

Cooper, H. (1988). Organizing knowledge syntheses: a taxonomy of literature reviews. *Knowledge in Society, 1*(1), S. 104-126.

Costa, P., & McCrae, R. (1989). NEO PI/FFI Manual Supplement for Use with the NEO Personality Inventory and the NEO Five-Factor Inventory.

Crampton, J., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M., & Zook, &. M. (2012). Beyond the geotag: situating ‘big data’ and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 130-139.

Dalton, L., Ballarin, V., & Brun, M. (2009). Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics. *Current Genomics, 10*(6), S. 430-445.

Driscoll, K. (2012). From Punched Cards to "Big Data": A Social History of Database Populism. *Futures of Communication*.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. S. 1-27.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. In U. Fayyad, G. Piatetsky-Shapiro, & P. Smyth, *Advances in Knowledge Discovery and Data Mining* (S. 1-34). Menlo Park, CA, USA: American Association for Artificial Intelligence.

Fu, Y., Sandhu, K., & Shih, M.-Y. (2000). A Generalization-Based Approach to Clustering of Web Usage Sessions. *WEBKDD '99 Revised Papers from the International Workshop on Web Usage Analysis and User Profiling.* Springer-Verlag.

Gerlitz, J.-Y., & Schupp, J. (2005). *Zur Erhebung der Big-Five-basierten Persönlichkeits-merkmale im SOEP.* Berlin: Deutsches Institut für Wirtschaftsforschung.

Ghosh, S. (2015). *aspire SYSTEMS.* Abgerufen am 28. April 2017 von http://www.aspiresys.com/WhitePapers/Smart-Data-Discovery-in-Retail.pdf?pdf=smart-data-discovery-whitepaper

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of Personality and social psychology, 1*, S. 141–165.

Grover, M., Malaska, T., Seidman, J., & Shapira, G. (2015). *Hadoop Application Architectures Designing Real-World Big Data Applications.* O'Reilly Media.

Hahsler, M., & Dunham, M. H. (2010). rEMM: Extensible Markov Model for Data Stream Clustering in R. *Journal of statistical software*.

Heer, J., & Chi, E. H. (2002). Separating the swarm: categorization methods for user sessions on the web. *Proceeding CHI '02 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 243-250). New York: ACM.

Heinström, J. (2003). Five personality dimensions and their influence on information behaviour. *Information Research, 3*.

Herzberg, P., & Roth, M. (2006). Beyond Resilients, Undercontrollers, and Overcontrollers? An Extension of Personality Prototype Research. *European Journal of Personality*.

Heuring, W. (2015). *Siemens*. Abgerufen am 12. März 2017 von https://www.siemens.com/innovation/de/home/pictures-of-the-future/digitalisierung-und-software/von-big-data-zu-smart-data-warum-big-data-smart-data-werden-muss.html

Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review, 34*(1), S. 135-174.

Hoerl, R. W., Snee, R. D., & De Veaux, R. D. (2014). Applying statistical thinking to ‘Big Data’ problems. *WIREs Comput Stat, 6*, S. 222–232.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. (Springer, Hrsg.) *Machine Learning, 1*, S. 177–196.

Hou, W. (2015). K-means vs. GMM & PLSA.

Isele, R., & Arnd, N. (2016). Mit semantischer Datenverwaltung Big Data in den Griff bekommen. *Wirtschaftsinformatik & Management*(4).

Jin, X., Zhou, Y., & Mobasher, B. (2004). Web usage mining based on probabilistic latent semantic analysis. *KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* New York: ACM .

Joan, J. M., & Venifa, M. G. (2012). User Profile Tracking by Web Usage Mining in Cloud Computing. *Procedia Engineering*.

Juvina, I., & van Oostendorp, H. (2006). Individual differences and behavioral metrics involved in modeling web navigation.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* SAGE.

Klump, J., & Bertelmann, R. (2014). Forschungsdaten. In R. Kuhlen, W. Semar, & D. Strauch, *Grundlagen der praktischen Information und Dokumentation.* Berlin: De Gruyter Saur.

Kosala, R., & Blockeel, H. (2000). Web Mining Research: A Survey. *SIGKDD Explorations, 2*.

Kousalya, R., Pradeepa, S., & Saravanan, V. (2013). Web Usage Mining Using D-Apriori And DFP Algorithm. *International Journal of Scientific & Engineering Research, 4*.

Kuhlen, R. (1995). *Informationsmarkt : Chancen und Risiken der Kommerzialisierung von Wissen* (Bd. XXVIII). Konstanz: Universitätsverlag.

Kuhlen, R., Semar, W., & Strauch, D. (2014). *Grundlagen der praktischen Information und Dokumentation* (6 Ausg.). Berlin: De Gruyter.

Kumar, B., & Rukmani, K. (2010). Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms. *International Journal of Advanced Networking and Applications*(1).

Lang, F., Lüdtke, O., & Asendorpf, J. (2001). Testgüte und psychome- trische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen. *Diagnostica*(47), S. 111–121.

Lemieux, V. L., Gormly, B., & Rowledge, L. (2014). Meeting Big Data challenges with visual analytics. *Records Management Journal, 24*(2), S. 122-141.

Linked Enterprise Data Service. (2016). Abgerufen am 20. Mai 2017 von http://www.leds-projekt.de/de/aktuelles/2016/semantische-Datenverwaltung-im-Griff.html

Lu, L., Dunham, M., & Meng, Y. (2005). Mining Significant Usage Patterns from Clickstream Data. *WebKDD 2005: Advances in Web Mining and Web Usage Analysis .* Berlin: Springer.

Mandl, T. (2014). Text Mining und Data Mining. In R. Kuhlen, W. Semar, & D. d. Strauch, *Grundlagen der praktischen Information und Dokumentation.* Berlin: De Gruyter Saur.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Mobasher, B. (2007). Chapter 12: Web Usage Mining in Data Collection and PreProcessing. *The 13th ACM SIGKKD International Conference on Knowledge Discovery and Data Mining .*

Morville, P., & Rosenfeld, L. (2006). *Information Architecture for the world wide web: designing large-scale web sites.* O'Reilly.

Nyman, M. (2013). Navigation behavior analysis and user profling based on automatically collected site data.

O'Neal, K. (2012). *BeyeNetwork*. Abgerufen am 12. März 2017 von http://www.b-eye-network.com/blogs/oneal/archives/2012/09/whats_the_big_d.php

Pöschl, S. (2010). *Die Handhabung mobiler Erreichbarkeit.* Kohlhammer.

Pirolli, P. L., & Pitkow, J. E. (1999). Distributions of Surfers' Paths through the WWW. *World Wide Web, 2*(1), S. 29-45.

Plöhn, P. (2014). Kategorisierung von Indizes zur Clustervalidierung. *Studienarbeit*. Darmstadt.

Rajagopalan, N. (2016). *Ness*. Abgerufen am 13. Mai 2017 von https://www.ness.com/big-data-analytics-with-clickstream/

Reimer, U. (2014). Empfehlungssysteme. In R. Kuhlen, W. Semar, & D. Strauch, *Grundlagen der praktischen Information und Dokumentation.* Berlin: De Gruyter Saur.

Riesen, K., & Bunke, H. (2010). *Graph Classification and Clustering Based on Vector Space Embedding.* River Edge, NJ, USA: World Scientific Publishing Co.

Robins, R., John, O., Caspi, A., & Stouthamer-Loeber, M. (1998). Resilient, Overcontrolled, and Undercontrolled Boys: Three Replicable Personality Types. *Journal of Personality and Social Psychology*, S. 151-171.

Sadagopan, N., & Li, J. (2008). *WWW '08 Proceedings of the 17th international conference on World Wide Web* (S. 885-894). Beijing, China: ACM.

Sallam, R., & Parenteau, J. (2015). *Gartner*. Abgerufen am 2. Mai 2017 von https://www.gartner.com/doc/3084217/smart-data-discovery-enable-new

Satow, L. (2011). Big-Five-Persönlichkeitstest (B5T) [Testbeschreibung und Items]. (Z. f. Dokumentation, Hrsg.)

Satow, L. (2012). Big-Five-Persönlichkeitstest (B5T): Test- und Skalendokumentation. Online im Internet: URL: http://www.drsatow.de. Abgerufen am 18. März 2017 von „Satow, L. (2012). Big-Five-Persönlichkeitstest (B5T): Test- und Skalendokumentation. Online im Internet: URL: http://www.drsatow.de.

Satow, L. (2017). Abgerufen am 18. März 2017 von http://www.drsatow.de/tests/fragen-und-antworten-zur-test-verwendung.html

Shahabi, C., Zarkesh, A. M., Adibi, J., & Shah, V. (1997). Knowledge Discovery from Users Web-Page Navigation. *Proceeding RIDE '97 Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97) High Performance Database Management for Large-Scale Applications.* IEEE Computer Society Washington.

Sisodia, D., Singh, L., Sisodia, S., & Saxena, K. (2012). Clustering Techniques: A Brief Survey of Different Clustering Algorithms . *International Journal of Latest Trends in Engineering and Technology*.

Sitaraman, G. (2014). Inferring Big 5 Personality from Online Social Networks. Washington.

Soni, N., & Ganatra, A. (2012). Categorization of Several Clustering Algorithms from Different Perspective: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*.

vom Brocke, J. S. (2009). Reconstructing the giant: on the importance of rigour in documenting the literature search process. *paper presented at the 17th European Conference on Information Systems (ECIS), 8*(10).

Wang, G., Konolige, T., Wilson, C., & Zhao, B. Y. (2013). You are how you click: clickstream analysis for Sybil detection. *Proceedings of the 22nd USENIX conference on Security.*

Wang, G., Zhang, X., Tang, S., & Zhao, B. Y. (2016). Analysis, Unsupervised Clickstream Clustering for User Behavior. *The 2016 CHI Conference.*

Womser-Hacker, C. (2014). Kognitives Information Retrieval. In R. Kuhlen, W. Semar, & D. Strauch, *Grundlagen der praktischen Information und Dokumentation.* Berlin: De Gruyter Saur.

Womser-Hacker, Christa, & Mandl, T. (2014). Information Seeking Behaviour. In R. Kuhlen, W. Semar, & D. Strauch, *Grundlagen der praktischen Information und Dokumentation.* Berlin: De Gruyter Saur.

Wu, X., Yan, J., Liu, N., Yan, S., Chen, Ying, & Chen, Z. (2009). Probabilistic Latent Semantic User Segmentation for Behavioral Targeted Advertising*. *ADKDD '09 Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising.* Paris: ACM.

Wurman, R. (2000). *Information Anxiety 2.* Hayden/Que.

Yeung, K. Y., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics, 17*(10), 977-987.

Zaiane, O. R., Srivastava, J., Spiliopoulou, M., & Masand, B. (2003). *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles ...* Springer.

Zhang, X., Liu, J., Du, Y., & Lv, T. (2011). A novel clustering method on time series data. *Expert Systems with Applications*, S. 11891-11900.

7 Anhang

7.1 Relative Clustervalidierung DHC, PLSA, K-Means

7.1.1 Einleitung

In diesem Kapitel wird das Vorgehen für die relative Clustervalidierung der drei Algorithmen DHC, PLSA und K-Means vorgestellt. Zuerst wird aufgezeigt, welche Clickstream Modellierung genutzt worden ist und danach die Ergebnisse der 10 Clusterdurchläufe pro Algorithmus in Bezug auf Reproduzierbarkeit und Stabilität aufgezeigt.

7.1.2 Code

- DHC: https://github.com/svenlenz/clickstream-survey/tree/master/clustering
- K-MEANS: https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/kmeans/KMeansClickstream.java
- PLSA: https://github.com/svenlenz/clickstream-survey/tree/master/PLSA
- Clickstream Converter: https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/utils/ClickstreamConverter.java
- Cluster Comparsion: https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/utils/CompareClusters.java

7.1.3 Clickstream Modellierung

Die Tabelle 26 illustriert jeweils die Modellierung der Klickstream Sequenzen für das DHC sowie K-MEANS Verfahren. Gemäss Wang et al. (2013) wird bei der Repräsentation der Events ein besseres Ergebnis erzielt, wenn high-level Kategorien (bzw. Konzepte) anstelle der rohen Klickdaten verwendet werden. Eine mögliche Erklärung ist, dass dadurch eine bessere Toleranz zu Noise erreicht werden kann. Auch Fu et al. (2000) und Bandari et al. (2017) nutzen eine Generalisierung bzw. eine Hauptkomponentenanalyse der Click-Events zur Reduktion der Vektorgrösse vor dem Clustering.

Um die Tests möglichst ohne Einfluss von Noise oder Anfälligkeiten der Algorithmen auf grosse oder zerstreute Vektoren zu haben, werden die Click-Events möglichst hoch generalisiert. Verwendet werden nur die grundlegenden Event-Typen wie open / video / home / back etc. sowie fix definierte diskrete Zeitintervalle beim DHC-Verfahren. Dazu wurden die folgenden diskreten Zeitintervalle genutzt: kleinste Zeiteinheit = 1; < 1s = 2; < 2s = 3; <3s = 4; < 5s = 5; <8s = 6; < 13s = 7 < 20s = 8; < 30s = 9; < 60s = 10; > 60s = 11

DHC	open(1)open(1)open(1)video(1)video(7)video(8)video(9)video(8)video(2)video(2)home(5)open(4)open(1)open(1)back(1)open(1)back(1)open(8)back(1)open(5)open(1)back(1)open(10)home(1)open(5)open(1)open(1)back(1)open(6)open(1)back(1)open(5)open(1)back(1)open(3)home(1)open(3)open(1)open(1)back(2)open(10)home(1)open(5)open(1)home(1)open(3)open(1)open(1)home(1)open(9)
PLSA/K-Means	open open open video video video video video video video home open open open backopen back open back open open back open home open open open back open open back open open back open home open open open back open home open open home open open open home open

Tabelle 26: Clickstream Modellierung

7.1.4 Divisive Hierarchical Clustering

	Cluster 1	Cluster 2	Cluster 3
1	1, 3, 7, 9, 14, 17, 22, 23, 27, 32, 34, 35, 38, 39, 40, 43, 46, 47, 49	2, 4, 5, 10, 11, 12, 13, 15, 20, 21, 24, 25, 28, 31, 33, 41, 42, 48	6, 8, 16, 18, 19, 26, 29, 30, 36, 37, 44, 45, 50
2	...	...	...
N	...	...	...
	100%	**100%**	**100%**

Tabelle 27: DHC mit 50er Testset

100% Reproduzierbarkeit

Durch das Feature-Pruning werden bei jedem Durchgang dieselben Bäume gebildet.

Halbieren des Testset (1-25) und Vergleich mit 50

Cluster 1	Cluster 2	Cluster 3
1, 3, 7, 9, 14, 17, 22, 23	2, 4, 5, 12, 13, 15, 21, 24, 25	6, 8, 10, 11, 16, 18, 19, 20
100%	**100%**	**62.5%**

Tabelle 28: DHC mit 25er Testset

Clusterstabilität nach Halbierung des Testset gegenüber dem ganzen Testset: 87.5%

7.1.5 K-MEANS

	Cluster 1	Cluster 2	Cluster 3
1	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 23, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 17, 18, 19, 21, 26, 29, 30, 40, 41, 48, 49, 50	12, 37
2	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50	37
3	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50	37
4	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 23, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 17, 18, 19, 21, 26, 29, 30, 40, 41, 48, 49, 50	12, 37
5	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 23, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 17, 18, 19, 21, 26, 29, 30, 40, 41, 48, 49, 50	12, 37
6	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	37
7	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45,	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50	37

	46, 47		
8	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50	37
9	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50	37
10	2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50	37
	79%	**75%**	**77%**

Tabelle 29: K-Means mit 50er Testset

77% Reproduzierbarkeit

Halbieren des Testset (1-25) und Vergleich mit 50

Cluster 1	Cluster 2	Cluster 3
2, 3, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25	1, 4, 6, 12, 19, 21, 23	7, 17, 18
45%	**27%**	**0%**

Tabelle 30: K-Means mit 25er Testset

Clusterstabilität nach Halbierung des Testset gegenüber dem 50er Testset: 24%

Verdoppelung des Testset (1-100) und Vergleich mit 50

Cluster 1	Cluster 2	Cluster 3
2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 15, 16, 20, 22, 24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 38, 39, 42, 43, 44, 45, 46, 47, 51, 52,	1, 6, 7, 12, 17, 18, 19, 21, 23, 26, 29, 30, 40, 41, 48, 49, 50, 53, 54, 55, 58, 60, 68, 75, 77, 81, 83, 84, 92	37, 64

56, 57, 59, 61, 62, 63, 65, 66, 67, 69, 70, 71, 72, 73, 74, 76, 78, 79, 80, 82, 85, 86, 87, 88, 89, 90, 91, 93, 94, 95, 96, 97, 98, 99, 100		
96%	**100%**	**50%**

Tabelle 31: K-Means mit 100er Testset

Clusterstabilität nach Verdoppelung des Testset gegenüber dem 50er Testset: 82%

7.1.6 PLSA

Beim PLSA wird jeweils eine Wahrscheinlichkeit berechnet, zu welchem Cluster ein Dokument zugeordnet werden kann. Für den Vergleich wurde jeweils das Cluster mit der höchsten Wahrscheinlichkeit ausgewählt.

	Cluster 1	Cluster 2	Cluster 3
1	1,2,3,4,5,6,7,8,11,12,13,14,17,18,19,20,21,22,24,28,30,32,37,41,49,50	9,10,16,25,29,31,35,38,39,42,44,45,46	15,23,26,27,33,34,36,40,43,47,48
2	1,2,3,4,5,6,7,8,11,12,13,17,18,19,20,21,22,24,28,30,32,37,49,50	10,14,16,25,29,31,38,39,41,44,45,46	9,15,23,26,27,33,34,35,36,40,42,43,47,48
3	1,2,3,4,5,6,7,8,11,12,13,17,18,19,20,21,22,24,28,30,32,37,41,49,50	9,10,16,25,29,31,35,38,39,42,44,45,46	14,15,23,26,27,33,34,36,40,43,47,48
4	1,2,3,4,5,6,7,8,11,12,13,14,17,18,19,20,21,22,24,28,30,32,37,49,50	25,36,38,45	9,10,15,16,23,26,27,29,31,33,34,35,39,40,41,42,43,44,46,47,48
5	2,4,5,6,12,13,14,17,21,24,25,38,41,50	1,3,7,8,9,18,19,20,22,28,30,32,35,37,49	10,11,15,16,23,26,27,29,31,33,34,36,39,40,42,43,44,45,46,47,48
6	1,2,3,4,5,6,7,8,11,12,13,17,18,19,20,21,22,24,28,30,32,37,41,49,50	9,10,16,25,29,31,35,38,39,42,44,45,46	14,15,23,26,27,33,34,36,40,43,47,48
7	2,4,5,6,8,12,13,17,21,24,25,38,41,50	1,3,7,9,18,19,20,22,28,30,32,35,37,49	10,11,14,15,16,23,26,27,29,31,33,34,36,39,40,42,43

			,44,45,46,47,48
8	1,2,3,4,5,6,7,8,11,12,13,17,18,19,20,21,22,24,28,30,32,37,41,49,50	9,10,16,25,29,31,35,38,39,42,44,45,46	14,15,23,26,27,33,34,36,40,43,47,48
9	1,2,3,4,5,6,7,8,11,12,13,17,18,19,20,21,22,24,28,30,32,37,41,49,50	9,10,16,25,29,31,35,38,39,42,44,45,46	14,15,23,26,27,33,34,36,40,43,47,48
10	1,2,3,4,5,6,7,8,11,12,13,14,17,18,19,20,21,22,24,28,30,32,37,41,49,50	9,10,16,25,29,31,35,38,39,42,44,45,46	15,23,26,27,33,34,36,40,43,47,48
	78%	**53%**	**72%**

Tabelle 32: PLSA mit 50er Testset

68% Reproduzierbarkeit

Halbieren des Testset (1-25) und Vergleich mit 50

Cluster 1	Cluster 2	Cluster 3
1,2,3,4,5,6,7,8,11,12,13,17,18,19,20,21,22,24,25	9,10,14	15,16,23
95%	**67%**	**67%**

Tabelle 33: PLSA mit 25er Testset

Clusterstabilität nach Halbierung des Testset gegenüber dem ganzen Testset: 76%

Verdoppelung des Testset (1-100) und Vergleich mit 50

Cluster 1	Cluster 2	Cluster 3
1,2,3,4,5,6,7,12,14,15,19,21,22,23,24,25,26,27,28,31,33,35,36,37,39,40,41,43,44,45,46,47,48,50,54,55,58,59,61,62,63,64,65,66,67,69,70,72,73,74,75,80,81,82,85,87,90,91,93,94,97,99	10,29,30,32,38,53,60,68,71,76,77,84,88,92,96,100	8,9,11,13,16,17,18,20,34,42,49,51,52,56,57,78,79,83,86,89,95,98
65%	**23%**	**9%**

Tabelle 34: PLSA mit 100er Testset

Clusterstabilität nach Verdoppelung des Testset gegenüber dem 50er Testset: 33%

7.1.7 Vergleich

Vergleich der Übereinstimmung zwischen den Zuteilungen durch die einzelnen Cluster

	Cluster 1	Cluster 2	Cluster 3	Total
PLSA/DHC	26%	18%	18%	21%
K-MEANS/DHC	30%	26%	15%	24%
PLSA/K-MEANS	62%	36%	27%	42%

Tabelle 35: Übereinstimmung der Clusterings

Die beste Übereinstimmung haben PLSA und K-MEANS mit durchschnittlich 42%.

7.2 Referenz Prototypen

Für die Erstellung der Referenz Prototypen wurde das Programm CreatePrototypes WithKMeans[18] verwendet, welches in etwa dem Pseudocode der Tabelle 36 entspricht.

```
LOOP 1...10000
    LOOP b5results
        convertToVector
        addToB5Vectors
    KMEANS euclidian distance, max 200 iterations -> b5vectors
    orderClusterCentroidsBySize
    collectCentroidsOfEachClusterInArray
    calculateAveragePerClusterArray
```

Tabelle 36: Pseudocode CreatePrototypesWithKMeans

Für drei Cluster ergibt sich das Profil:

6 6 5 4 4
5 5 4 4 3
5 5 3 4 3

[18] https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/kmeans/CreatePrototypesWithKMeans.java

Für fünf Cluster:

6 5 4 4 4
5 5 5 4 4
5 5 4 4 4
5 5 4 4 4
5 5 4 4 3

Zuteilungen der Sessions mittels dem Programm StatDump[19] zu den Profilen

Profil	Zuteilungen	Anzahl	%
6 6 5 4 4	1, 2, 5, 6, 7, 8, 10, 13, 15, 16, 19, 24, 26, 29, 30, 32, 36, 43, 44, 46, 49, 50, 54, 57, 58, 60, 61, 62, 63, 64, 66, 67, 68, 69, 71, 77, 78, 81, 82, 88, 91, 92, 93, 94, 95, 96, 97, 100, 101, 102, 103, 106, 107, 110, 111, 113, 114, 115, 116, 117, 118, 120, 124	63	50
5 5 4 4 3	12, 20, 21, 22, 23, 25, 27, 35, 39, 41, 45, 47, 48, 55, 59, 72, 73, 79, 83, 84, 85, 89, 105, 108, 109, 112, 122	27	21
5 5 3 4 3	3, 4, 9, 11, 14, 17, 18, 28, 31, 33, 34, 37, 38, 40, 42, 51, 52, 53, 56, 65, 70, 74, 75, 76, 80, 86, 87, 90, 98, 99, 104, 119, 121, 123, 125, 126	36	29

Tabelle 37: 3er Cluster Testdatenzuteilung

Profil	Zuteilungen	Anzahl	%
6 6 5 4 4	2, 3, 5, 6, 8, 15, 18, 19, 22, 24, 26, 29, 30, 31, 32, 35, 36, 37, 39, 40, 43, 44, 49, 57, 58, 60, 61, 67, 68, 71, 77, 78, 84, 88, 90, 94, 95, 97, 99, 100, 101, 103, 106, 107, 113, 115, 116, 121, 124, 125	50	40
5 5 5 4 4	1, 10, 12, 16, 20, 21, 23, 41, 45, 46, 47, 50, 54, 62, 63, 66, 72, 81, 92, 93, 96, 102, 110, 111, 114, 117, 118, 120	28	22
5 5 4 4 4	7, 9, 13, 53, 64, 65, 69, 75, 76, 82, 87, 89, 91, 104, 108, 109, 123, 126	18	14

[19] https://github.com/svenlenz/clickstream-survey/blob/master/processing/src/main/java/processing/utils/StatDump.java

5 5 4 4 3	4, 11, 14, 17, 25, 27, 28, 33, 34, 38, 42, 48, 51, 52, 55, 56, 59, 70, 73, 74, 79, 80, 83, 85, 86, 98, 105, 112, 119, 122	30	24

Tabelle 38: 5er Cluster Testdatenzuteilung

7.3 Statistiken Clusterergebnisse

Dieses Kapitel enthält Visualisierungen der Clusterings mittels dem LLN1 und MLN2 Modell. Die detaillierten Rohdaten finden sich unter:

- https://github.com/svenlenz/clickstream-survey/blob/master/results/clickstreams/LLN1.xlsx
- https://github.com/svenlenz/clickstream-survey/blob/master/results/clickstreams/MLN1.xlsx

7.3.1 LLN1

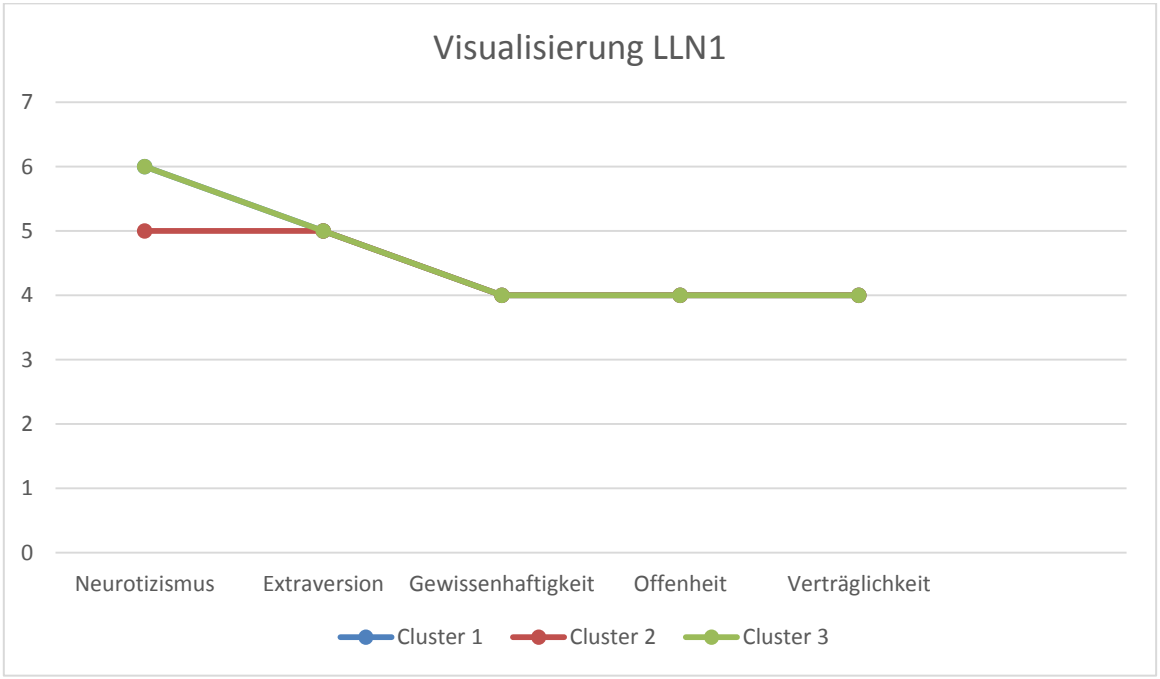

Abbildung 14: Visualisierung LLN1 (eigene Darstellung)

Cluster	Technik	Männer	Frauen	Alter	Klicks	Duration
Cluster 1	3	25	19	3	33	136331
Cluster 2	3	19	23	3	9	63317
Cluster 3	3	26	13	3	41	244961

Tabelle 39: Statistik LLN1

Cluster	Produkt 1	Produkt 2	Produkt 3	Produkt 4
Cluster 1	336	195	256	481
Cluster 2	121	62	97	40
Cluster 3	452	359	296	301

Tabelle 40: Events pro Produkt (LLN1)

7.3.2 MLN1

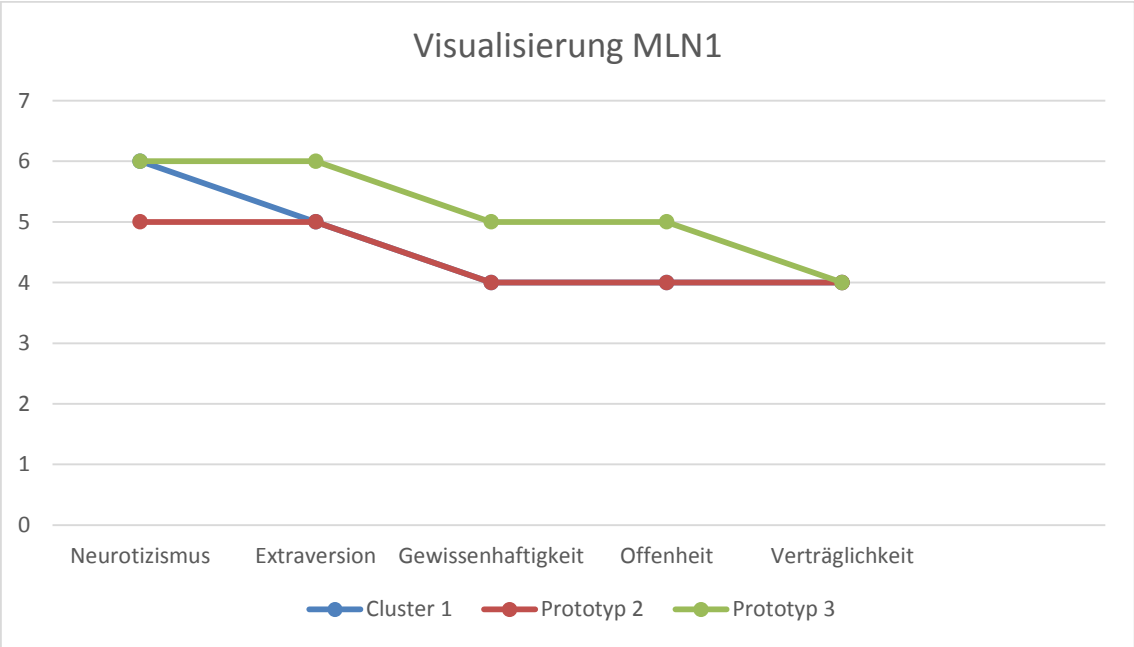

Abbildung 15: Visualisierung MLN2 (eigene Darstellung)

Cluster	Technik	Männer	Frauen	Alter	Klicks	Duration
Cluster 1	3	44	19	3	36	204165
Cluster 2	3	21	26	3	15	94453
Cluster 3	3	5	9	3	29	43895

Tabelle 41: Statistik MLN1

Cluster	Produkt 1	Produkt 2	Produkt 3	Produkt 4
Cluster 1	583	391	385	615
Cluster 2	207	138	138	97
Cluster 3	93	73	103	92

Tabelle 42: Events pro Produkt (MLN2)

Bisher erschienene Schriften

Ergebnisse von Forschungsprojekten erscheinen jeweils in Form von Arbeitsberichten in Reihen. Sonstige Publikationen erscheinen in Form von alleinstehenden Schriften.

Derzeit gibt es in den Churer Schriften zur Informationswissenschaft folgende Reihen:
Reihe Berufsmarktforschung

Churer Schriften zur Informationswissenschaft – Schrift 1
Herausgegeben von Josef Herget und Sonja Hierl
Reihe Berufsmarktforschung – Arbeitsbericht 1:
Josef Herget
Thomas Seeger
Zum Stand der Berufsmarktforschung in der Informationswissenschaft in deutschsprachigen Ländern
Chur, 2007 (im Druck)
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 2
Herausgegeben von Josef Herget und Sonja Hierl
Reihe Berufsmarktforschung – Arbeitsbericht 2:
Josef Herget
Norbert Lang
Berufsmarktforschung in Archiv, Bibliothek, Dokumentation und in der Informationswirtschaft:
Methodisches Konzept
Chur, 2007 (im Druck)
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 3
Herausgegeben von Josef Herget und Sonja Hierl
Reihe Berufsmarktforschung – Arbeitsbericht 3:
Josef Herget
Norbert Lang
Gegenwärtige und zukünftige Arbeitsfelder für Informationsspezialisten in privatwirtschaftlichen Unternehmen und öffentlich-rechtlichen Institutionen
Chur, 2004
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 4
Herausgegeben von Josef Herget und Sonja Hierl
Sonja Hierl
Die Eignung des Einsatzes von Topic Maps für e-Learning
Vorgehensmodell und Konzeption einer e-Learning-Einheit unter Verwendung von Topic Maps
Chur, 2005
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 5
Herausgegeben von Josef Herget und Sonja Hierl
Nina Braschler
Realisierungsmöglichkeiten einer Zertifizierungsstelle für digitale Zertifikate in der Schweiz
Chur, 2005
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 6
Herausgegeben von Josef Herget und Sonja Hierl
Reihe Berufsmarktforschung – Arbeitsbericht 4:
Ivo Macek
Urs Naegeli
Postgraduiertenausbildung in der Informationswissenschaft in der Schweiz:
Konzept – Evaluation – Perspektiven
Chur, 2005
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 7
Herausgegeben von Josef Herget und Sonja Hierl
Caroline Ruosch
Die Fraktale Bibliothek:
Diskussion und Umsetzung des Konzepts in der deutschsprachigen Schweiz.
Chur, 2005
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 8
Herausgegeben von Josef Herget und Sonja Hierl
Esther Bättig
Information Literacy an Hochschulen
Entwicklungen in den USA, in Deutschland und der Schweiz
Chur, 2005
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 9
Herausgegeben von Josef Herget und Sonja Hierl
Franziska Höfliger
Konzept zur Schaffung einer Integrationsbibliothek in der Pestalozzi-Bibliothek Zürich
Chur, 2005
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 10
Herausgegeben von Josef Herget und Sonja Hierl
Myriam Kamphues
Geoinformationen der Schweiz im Internet:
Beurteilung von Benutzeroberflächen und Abfrageoptionen für Endnutzer
Chur, 2006
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 11
Herausgegeben von Josef Herget und Sonja Hierl
Luigi Ciullo
Stand von Records Management in der chemisch-pharmazeutischen Branche
Chur, 2006
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 12
Herausgegeben von Josef Herget und Sonja Hierl
Martin Braschler, Josef Herget, Joachim Pfister, Peter Schäuble, Markus Steinbach, Jürg Stuker
Evaluation der Suchfunktion von Schweizer Unternehmens-Websites
Chur, 2006
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 13
Herausgegeben von Josef Herget und Sonja Hierl
Adina Lieske
Bibliotheksspezifische Marketingstrategien zur Gewinnung von Nutzergruppen:
Die Winterthurer Bibliotheken
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 14
Herausgegeben von Josef Herget und Sonja Hierl
Christina Bieber, Josef Herget
Stand der Digitalisierung im Museumsbereich in der Schweiz
Internationale Referenzprojekte und Handlungsempfehlungen
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 15
Herausgegeben von Josef Herget und Sonja Hierl
Sabina Löhrer
Kataloganreicherung in Hochschulbibliotheken
State of the Art Überblick und Aussichten für die Schweiz
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 16
Herausgegeben von Josef Herget und Sonja Hierl
Heidi Stieger
Fachblogs von und für BibliothekarInnen – Nutzen, Tendenzen
Mit Fokus auf den deutschsprachigen Raum
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 17
Herausgegeben von Josef Herget und Sonja Hierl
Nadja Kehl
Aggregation und visuelle Aufbereitung von Unternehmensstrategien mithilfe von Recherche-Codes
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 18
Herausgegeben von Josef Herget und Sonja Hierl
Rafaela Pichler
Annäherung an die Bildsprache – Ontologien als Hilfsmittel für Bilderschliessung und Bildrecherche in Kunstbilddatenbanken
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 19
Herausgegeben von Josef Herget und Sonja Hierl
Jürgen Büchel
Identifikation von Marktnischen – Die Eignung verschiedener Informationsquellen zur Auffindung von Marktnischen
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 20
Herausgegeben von Josef Herget und Sonja Hierl
Andreas Eisenring
Trends im Bereich der Bibliothekssoftware
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 21
Herausgegeben von Josef Herget und Sonja Hierl
Lilian Brändli
Gesucht – gefunden? Optimierung der Informationssuche von Studierenden in wissenschaftlichen Bibliotheken
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 22
Herausgegeben von Josef Herget und Sonja Hierl
Beatrice Bürgi
Open Access an Schweizer Hochschulen – Ein praxisorientierter Massnahmenkatalog für Hochschulbibliotheken zur Planung und Errichtung von Institutional Repositories
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 23
Herausgegeben von Josef Herget und Sonja Hierl
Darja Dimitrijewitsch, Cécile Schneeberger
Optimierung der Usability des Webauftritts der Stadt- und Universitätsbibliothek Bern
Chur, 2007
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 24
Herausgegeben von Nadja Böller, Josef Herget und Sonja Hierl
Brigitte Brüderlin
Stakeholder-Beziehungen als Basis einer Angebotsoptimierung
Chur, 2008
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 25
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Jonas Rebmann
Web 2.0 im Tourismus, Soziale Webanwendungen im Bereich der Destinationen
Chur, 2008
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 26
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Isabelle Walther
Idea Stores, ein erfolgreiches Bibliothekskonzept aus England – auf für die Schweiz?
Chur, 2008
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 27
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Scherer Auberson Kirsten
Evaluation von Informationskompetenz: Lässt sich ein Informationskompetenzzuwachs messen?
Eine systematische Evaluation von Messverfahren
Chur, 2009 (im Druck)
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 28
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Nadine Wallaschek
Datensicherung in Bibliotheksverbünden.
Empfehlungen für die Entwicklung von Sicherheits- und Datensicherungskonzepten in Bibliotheksverbünden
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 29
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Laura Tobler
Recherchestrategien im Internet
Systematische Vorgehensweisen bei der Suche im Internet, dargestellt anhand ausgewählter Fallstudien
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 30
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Bibliotheken und Dokumentationszentren als Unternehmen:
Antworten von Bibliotheken und Dokumentationszentren auf die Herausforderungen der digitalen Gesellschaft
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 31
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Karin Garbely, Marita Kieser
Mystery Shopping als Bewertungsmethode der Dienstleistungsqualität von wissenschaftlichen Bibliotheken
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 32
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Tristan Triponez
E-Mail Records Management
Die Aufbewahrung von E-Mails in Schweizer Organisationen als technische, rechtliche und organisatorische Herausforderung
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 33
Herausgegeben von Robert Barth, Nadja Böller, Urs Dahinden, Sonja Hierl
und Hans-Dieter Zimmermann
Die Lernende Bibliothek 2009
Aktuelle Herausforderungen für die Bibliothek und ihre Partner im Prozess des
wissenschaftlichen Arbeitens
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 34
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Rene Frei
Die Informationswissenschaft aus Sicht des Radikalen Konstruktivismus
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 35
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Hans-Dieter Zimmermann
Lydia Bauer, Nadja Böller, Sonja Hierl
DIAMOND Didactical Approach for Multiple Competence Development
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 36
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Wolfgang Semar
Michaela Spiess
Einsatz von Competitive Intelligence in Schweizer Spitäler
Chur, 2009
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 37
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Wolfgang Semar
Jasmine Milz
Informationskompetenz-Vermittlung an Deutschschweizer Fachhochschulen:
eine quantitative Inhaltsanalyse der Curricula
Chur, 2010
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 38
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Wolfgang Semar
Corinne Keller
RFID in Schweizer Bibliotheken – eine Übersicht
Chur, 2010
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 39
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Wolfgang Semar
Bibliotheksbau in der Schweiz 1985 – 2010
Planung – Nutzung – Ästhetik
Herausgegeben von Robert Barth und Iris Kuppelwieser
Chur, 2010
ISSN1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 40
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Wolfgang Semar
Stephan Becker
Klassifikationsraster zur Relevanzanalyse aktueller Themenanfragen an einer
Mediendokumentationsstelle in der Schweiz
Chur, 2010
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 41
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Wolfgang Semar
Reihe Berufsmarktforschung – Arbeitsbericht 5:
Iris Capatt, Urs Dahinden
Absolventenbefragung 2010
Bachelorstudiengang Informationswissenschaft und Diplomstudiengang Information und Dokumentation der HTW Chur
Chur, 2010
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 42
Herausgegeben von Robert Barth, Nadja Böller, Sonja Hierl und Wolfgang Semar
Saro Adamo Pepe Fischer
Bestandserhaltung im Film-/Videoarchiv des Schweizer Fernsehens
Chur, 2010
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 43
Herausgegeben von Robert Barth, Iris Capatt, Sonja Hierl und Wolfgang Semar
Patricia Dürring
Ökonomischer Mehrwert von Bibliotheken, aufgezeigt anhand ausgewählter Dienste der Zentral- und Hochschulbibliothek Luzern
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 44
Herausgegeben von Robert Barth, Iris Capatt, Sonja Hierl und Wolfgang Semar
Pia Baier Benninger
Model Requirements for the Management of Electronic Records (MoReq2).
Anleitung zur Umsetzung
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 45
Herausgegeben von Robert Barth, Iris Capatt, Sonja Hierl und Wolfgang Semar
Martina Thomi
Überblick und Bewertung von Musiksuchmaschinen
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 46
Herausgegeben von Robert Barth, Iris Capatt und Wolfgang Semar
Regula Trachsler
Angebote für Senioren in Deutschschweizer Bibliotheken
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 47
Herausgegeben von Robert Barth, Iris Capatt und Wolfgang Semar
Wolfgang Semar (Hrsg.)
Arge Alp Tagung 23.-24. September 2010, Chur
Informationsgesellschaft und Infrastrukturpolitik im Alpenraum
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 48
Herausgegeben von Robert Barth, Lydia Bauer, Iris Capatt und Wolfgang Semar
Heinz Mathys
Jungs lesen weniger als Mädchen.
Was können Bibliotheken gemeinsam mit den Schulen tun, um dies zu ändern?
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 49
Herausgegeben von Robert Barth, Lydia Bauer, Iris Capatt und Wolfgang Semar
Anina Baumann
Stärken und Schwächen von Discovery Diensten am Beispiel des EBSCO Discovery Service
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 50
Herausgegeben von Robert Barth, Lydia Bauer, Iris Capatt und Wolfgang Semar
Reihe Berufsmarktforschung – Arbeitsbericht 6:
Iris Capatt, Urs Dahinden
Absolventenbefragung 2011
Hochschule für Technik und Wirtschaft HTW Chur Weiterbildungsstudiengänge Informationswissenschaft.
Externer Bericht.
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 51
Herausgegeben von Robert Barth, Lydia Bauer, Iris Capatt und Wolfgang Semar
Reihe Berufsmarktforschung – Arbeitsbericht 7:
Iris Capatt, Urs Dahinden
Absolventenbefragung 2011
Hochschule für Technik und Wirtschaft HTW Chur Weiterbildungsstudiengänge Management.
Externer Bericht.
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 52
Herausgegeben von Robert Barth, Lydia Bauer, Iris Capatt und Wolfgang Semar
Salome Arnold
Auf den Spuren der Barrieren für ein barrierefreies Webdesign
Chur, 2011
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 53
Herausgegeben von Robert Barth, Lydia Bauer, Iris Capatt und Wolfgang Semar
Laura Stadler
Die Gläserne Decke in Schweizer Bibliotheken
Chur, 2012
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 54
Herausgegeben von Robert Barth, Lydia Bauer, Brigitte Lutz und Wolfgang Semar
Ruth Süess
Evaluation von Web Monitoring Tools zur softwaregestützten Informationsbeschaffung am Beispiel ausgewählter Open Source Web Monitoring Tools
Chur, 2012
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 55
Herausgegeben von Robert Barth, Lydia Bauer, Brigitte Lutz und Wolfgang Semar
Michael Hunziker
Approval Plans und andere Outsourcing-Formen im Bestandesaufbau an den Wissenschaftlichen Bibliotheken der Deutschschweiz
Chur, 2012
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 56
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Urs Dahinden, Michael Aschwanden und Lydia Bauer
Verpasste Chancen? Altersspezifische digitale Ungleichheiten bei der Nutzung von Mobilkommunikation und Internet
Chur, 2012
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 57
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Grégoire Savary
Eine Konservierungsstrategie für das Archiv der Siedlungsgenossenschaft Freidorf bei Muttenz. Eine Hilfestellung für kleine Archive mit gemischten Beständen
Chur, 2013
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 58
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Patrick Wermelinger
Die Georeferenzierung von Katalogdaten mit Hilfe von Linked Open Data
Chur, 2013
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 59
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Carla Biasini
E-Books in öffentlichen Bibliotheken der Schweiz – Determinanten der Akzeptanz bei Kunden
Chur, 2013
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 60
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Nadja Böller
Modell zur strategischen Analyse von Konzepten zur Förderung der Informationskompetenz durch Hochschulbibliotheken – MOSAIK-PRO
Chur, 2013
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 61
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Nina Santner
Von der Mediothek zum Recherchezentrum
Chur, 2013
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 62
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Daniela Denzer
Gründe für die Nichtnutzung von Bibliotheken bei Pensionierten in der Deutschschweiz
Chur, 2013
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 63
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Verena Gerber-Menz
Übernahme von born-digital Fotobeständen und Fotografennachlässen ins Archiv
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 64
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Vanessa Kellenberger
E-Shop Analytics und Erfolgsoptimierung – Die wichtigsten Kennzahlen
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 65
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Matthias Dudli
Open Innovation in Bibliotheken – Eine Konzeptstudie der ETH-Bibliothek Zürich
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 66
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Sarah Carbis
Welche Verbandszeitschrift wünschen sich die Mitglieder des BIS?
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 67
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Yvonne Lingg
Patientenverfügung als Informations- und Kommunikationsinstrument
Analyse der Vielfalt sowie Dokumentation der Inhalte und Standardisierungsmöglichkeiten
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 68
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Mara Sophie Hellstern
Förderung von Engagement in GLAM (Galleries, Libraries, Archives and Museums) durch Wikipedians in Residence (WiR)
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 69
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Philipp Trottmann
Die epochale Trendwende: Der Benutzerrückgang an öffentlichen Bibliotheken der Deutschschweiz
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 70
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Ursula Huber
10 Jahre Open Access Initiative – Eine Zwischenbilanz für die Schweiz
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 71
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Beat Mattmann
Die Möglichkeiten von RDA bei der Erschliessung historischer Sondermaterialien
Chur, 2014
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 72
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Diane Golay
User-center redesign of the Biotechgate portal: a remote usability testing case study
Chur, 2015
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 73
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Felicitas Isler
Inklusion von Mitarbeitenden mit einer Beeinträchtigung in Bibliotheken
Chur, 2015
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 74
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Tamara Müller
Die Schwierigkeiten bei der Recherche im Archiv(-katalog): Ursachenforschung und Vorschläge zur Problembehebung
Chur, 2015
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 75
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Benjamin Fischer
Potential von automatischen Videoanalysen im Fussball am Beispiel der Schweizer Super League
Chur, 2015
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 76
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Simon Schultze
Videospielturniere in öffentlichen Schweizer Bibliotheken
Ein Pilotprojekt der St. Galler Stadtbibliothek Katharinen
Chur, 2015
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 77
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Charlotte Frauchiger
Barrierefreie E-Books
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 78
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Stefanie Dietiker
Cognitive Map einer Bibliothek
Eine Überprüfung der Methodentauglichkeit im Bereich Bibliothekswissenschaft –
am Beispiel der Kantonsbibliothek Graubünden
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 79
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Sharon Alt
Konzeption und Evaluation eines Online-Tutorial zur Förderung der
E-Health-Literacy von Männern im Alter von 50 bis 80 Jahren
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 80
Herausgegeben von Wolfgang Semar und Brigitte Lutz
Bettina Wille
Automatisierung und Digitalisierung in den wissenschaftlichen Bibliotheken der Schweiz
Ein Oral History Projekt
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 81
Herausgegeben von Wolfgang Semar
Michael Mente
Ansichtskarten sind Ansichtssache – Bilder, Grüsse und Metadaten
Über den Wert topografischer Ansichtskarten in Archivbeständen und
Einsichten in Fragen ihrer archivischen Erschliessung
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 82
Herausgegeben von Wolfgang Semar
Fabian Muster
Datenstrategiemodell: Ein Referenzmodell zur Entwicklung von Datenstrategien
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 83
Herausgegeben von Wolfgang Semar
Sandro Lorenzo
Bibliotheken und Integration
Aspekte der interkulturellen Bibliotheksarbeit und deren Einfluss auf die Integration von
Migranten und Migrantinnen sowie Menschen mit Migrationshintergrund in der Deutschschweiz
mit einem Fokus auf den deutschsprachigen Teil des Kantons Bern
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 84
Herausgegeben von Wolfgang Semar
Johannes Reitze
Was öffentliche Bibliotheken meinen, wenn sie vom Dritten Ort sprechen
Chur, 2016
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 85
Herausgegeben von Wolfgang Semar
Simone Beeler
Sonntagsöffnungszeiten in öffentlichen Bibliotheken in der Schweiz
Chur, 2017
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 86
Herausgegeben von Wolfgang Semar
Marco Humbel
Die Umsetzung von Open Data an Wissenschaftlichen Bibliotheken der Schweiz:
Eine qualitative Untersuchung
Chur, 2017
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 87
Herausgegeben von Wolfgang Semar
Flurina Huonder
Medieninhaltsanalyse Big Data:
Big Data, Datenschutz und Privatsphäre in Schweizer und US-amerikanischen Zeitungen
Chur, 2017
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 88
Herausgegeben von Wolfgang Semar
Marcel Hanselmann
Makerspaces in öffentlichen Bibliotheken:
Eine Untersuchung der didaktischen Ziele und eine Evaluation der Technologie littleBits
Chur, 2017
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 89
Herausgegeben von Wolfgang Semar
Franziska Brunner
Überlieferungsbildung 2.0:
Eine Untersuchung zum Mehrwert von Partizipation Dritter in staatlichen Archiven
Chur, 2017
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 90
Herausgegeben von Wolfgang Semar
Marcella Haab-Grothof
„Kleider machen BibliothekarInnen“:
Der Einfluss von Kleidung des Bibliothekspersonals auf die Kontaktaufnahme von Benutzenden
Chur, 2017
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 91
Herausgegeben von Wolfgang Semar
Sven Lenz
Customer Engagement Analytics: Clustering User Navigation Behaviour
Chur, 2017
ISSN 1660-945X

Über die Informationswissenschaft der HTW Chur

Die Informationswissenschaft ist in der Schweiz noch ein relativ junger Lehr- und Forschungsbereich. International weist diese Disziplin aber vor allem im anglo-amerikanischen Bereich eine jahrzehntelange Tradition auf. Die klassischen Bezeichnungen dort sind Information Science, Library Science oder Information Studies. Die Grundfragestellung der Informationswissenschaft liegt in der Betrachtung der Rolle und des Umgangs mit Information in allen ihren Ausprägungen und Medien sowohl in Wirtschaft und Gesellschaft. Die Informationswissenschaft wird in Chur integriert betrachtet.

Diese Sicht umfasst nicht nur die Teildisziplinen Bibliothekswissenschaft, Archivwissenschaft und Dokumentationswissenschaft. Auch neue Entwicklungen im Bereich Medienwirtschaft, Informations- und Wissensmanagement und Big Data werden gezielt aufgegriffen und im Lehr- und Forschungsprogramm berücksichtigt.

Der Studiengang Informationswissenschaft wird seit 1998 als Vollzeitstudiengang in Chur angeboten und seit 2002 als Teilzeit-Studiengang in Zürich. Seit 2010 rundet der Master of Science in Business Administration das Lehrangebot ab.

Der Arbeitsbereich Informationswissenschaft vereinigt Cluster von Forschungs-, Entwicklungs- und Dienstleistungspotenzialen in unterschiedlichen Kompetenzzentren:

- Information Management & Competitive Intelligence
- Collaborative Knowledge Management
- Information and Data Management
- Records Management
- Library Consulting
- Information Laboratory

Diese Kompetenzzentren werden im **Swiss Institute for Information Research** zusammengefasst.

IMPRESSUM

Verlag & Anschrift

Arbeitsbereich Informationswissenschaft

HTW - Hochschule für Technik und Wirtschaft
University of Applied Sciences
Ringstrasse 37
CH-7000 Chur

www.informationswissenschaft.ch

www.htwchur.ch

ISSN 1660-945X

Institutsleitung

Prof. Dr. Niklaus Stettler
Telefon: +41 81 286 24 61
Email: niklaus.stettler@htwchur.ch

Sekretariat

Telefon : +41 81 286 24 24
Fax : +41 81 286 24 00
Email: clarita.decurtins@htwchur.ch