HAW HAMBURG

BACHELORTHESIS
Ali Soltani

Konzept und Implementierung eines Nachrichten Recommender Systems

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Ali Soltani

Konzept und Implementierung eines Nachrichten Recommender Systems

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang Bachelor of Science Wirtschaftsinformatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Jan Sudeikat

Eingereicht am: 17. Januar 2020

Ali Soltani

Thema der Arbeit

Konzept und Implementierung eines Nachrichten Recommender Systems

Stichworte

Recommender System, Collaborative Filtering, Nachrichten, Personalisierung, Data Mining

Kurzzusammenfassung

Diese Arbeit beschäftigt sich mit Recommender Systeme für Online-Nachrichtenplattformen. Hierfür werden Data-Mining-Techniken untersucht und eine prototypische Implementierung eines Recommender Systems, für eine deutschsprachige Online-Nachrichtenplattform entwickelt und evaluiert. Für die prototypische Entwicklung wird das weit verbreitete Collaborative-Filtering Verfahren verwendet. Als Datenquelle dienen hier reale Cookies der Plattform.

Ali Soltani

Title of Thesis

Concept and implementation of a news recommendation system

Keywords

Recommender system, collaborative filtering, news, personalization, data minig

Abstract

This thesis deals with recommender systems for online news platforms. Data mining techniques are analyzed and a prototypical implementation of a recommender system for a German online news platform is developed and evaluated. For the prototypical development is the widely used collaborative filtering method used. Real cookies of the platform serve as data source.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

In dem Film *The Social Network* [7] sagt Marck Zuckerberg:

> „Einst lebten wir auf dem Land, dann in Städten und von jetzt an im Netz."

Zurzeit befindet sich die Menschheit im Zuge der Digitalisierung in einem Informationszeitalter. Das Internet ist ein wichtiger und omnipräsenter Teil der Gesellschaft geworden. Sie bietet die Möglichkeit Informationen fast zu jeder Zeit abzurufen. So nutzen 66,5 Mio. Personen ab 10 Jahren in Deutschland das Internet [1] und 77% davon kaufen Waren und Dienstleistungen über das Internet ein [4]. Suchmaschinen helfen hier Nutzern die Information aus dem Internet zu filtern bzw. einzugrenzen, welche tatsächlich oder ungefähr ihrer Suche entsprechen. So ist Google als Information Retrieval System die meistbesuchte Webseite der Welt [27]. Unternehmen stehen hier vor der Herausforderung durch ihre große Sparte an Produkten (z.B. Artikel und Filme), das Produkt schnellstmöglich zu liefern, welches das Bedürfnis des Kunden befriedigt bzw. das Interesse oder Bedürfnis danach weckt. Somit versuchen Unternehmen die Produkte zu präsentieren, welche mit einer großen Wahrscheinlichkeit gekauft werden. Empfehlungssysteme helfen Unternehmen anhand der Analyse großer Datenmengen, welches z.B. ein Kunde auf einer Webseite hinterlässt, personalisierte und individuell angepasste Produkte zu empfehlen. Sie werden somit immer mehr ein wichtiger Bestandteil der Marketingstrategie. Ab Mitte der 90er haben sich Empfehlungssysteme mit der Zeit zu einem unabhängigen Forschungsgebiet entwickelt. In den letzten Jahren hat sich das Interesse an Empfehlungssystemen weiter verstärkt. Zudem gibt es spezielle Konferenzen und Workshops zu diesem Thema, wie die 2007 gegründete ACM Recommender Systems (RecSys), welches jetzt die wichtigste jährliche Veranstaltung zur Förderung von Technologieforschung ist. Wissenschaftlichen Zeitschriften, wie AI Communications, IEEE Intelligent Systems, Journal of Electronic Commerce, greifen das Thema durch Sonderausgaben mehr auf [42]. Aus der Wirtschaft hat Amazon mit ihrer E-Commerce Plattform 35% der Umsätze durch Produktempfehlung erzielt [6]. Netflix stieß mit einem Wettbewerb [10] in Empfehlungsalgorithmen im Jahr 2006 durch einen Gewinnerpreis von 1 Mio. Dollar die Forschung nochmal an. Somit

ist eine E-Commercestrategie basierend auf Empfehlungen unabdingbar für Verlage und Nachrichtenseiten, deren Hauptverkaufszweig bislang offline war.

1.1 Ziel dieser Arbeit

Diese Arbeit beschäftigt sich mit einem prototypischen Konzept der Implementierung eines Empfehlungssystems für eine Nachrichtenplattform. Hierfür werden wissenschaftliche Beiträge zu diesem Thema vorgestellt und verglichen. Ziel dieser Arbeit ist es zu untersuchen, inwieweit eine Empfehlung basierend auf dem Collaborative-Filtering Ansatz mit Cookies möglich ist. Als Datengrundlage dienen extrahierte Daten aus realen Cookies einer Online Zeitung aus dem deutschsprachigen Raum. Die Daten bestehen aus 80 Spalten und 1 Million Zeilen.

1.2 Thematische Abgrenzung

Diese Arbeit liefert eine prototypische Herangehensweise für eine Produktivplattform. Zudem beschäftigt sich diese Arbeit hauptsächlich mit Data-Mining Verfahren in Bezug auf Empfehlungssysteme. Hierfür werden die Verfahren verglichen und ein Modell implementiert. Zudem wird das vorliegende System evaluiert. Diese Arbeit beschäftigt sich nicht mit einer produktiven Anwendung eines Empfehlungssystems auf einer Webseite.

1.3 Inhaltlicher Aufbau

Der erste Abschnitt dient der Vorstellung, Einleitung und der thematischen Einführung. Anschließend folgt ein Abschnitt zur gesellschaftlichen Nachrichtennutzung in Deutschland, welches ein Überblick über die mögliche Zielgruppe und das Potenzial der Branche darstellt. Im nächsten Kapitel werden Konzepte der Empfehlungssysteme mit ihren Vor und– Nachteilen und ihren Herausforderungen dargestellt. Im vierten Abschnitt fließen die Erkenntnisse des letzten Kapitels mit ein. Hier wird die Vorgehensweise, die Datenvorbereitung, technische Umgebung und ein Modell erstellt. Anschließend werden die Ergebnisse evaluiert und im letzten Kapitel wird ein Ausblick auf mögliche Optimierungen und relevante Themen für diese Arbeit aufgeführt.

2 Verhalten und Entwicklung der Nachrichtennutzung in Deutschland

In diesem Kapitel geht es um den Umschwung der Verlage und das Nachrichtennutzungsverhalten innerhalb der deutschen Gesellschaft. Im Folgenden soll ein kurzer Überblick zur historischen Entwicklung der Verlage wiedergegeben werden. Dabei werden auch die Unterschiede von Zeitungen aufgezeigt. Anschließend wird die Digitalisierung der Verlage thematisiert.

2.1 Geschichte der Presse

Der Begriff Medien ist ursprünglich Latein und stammt vom *Medius* was Mitte, einen Mittelpunkt oder etwas Vermittelndes bezeichnet. Ein Medium ist ein Hilfsmittel, womit sich Informationen verbreiten lassen [8]. Hingegen stammt der Begriff Nachricht aus dem 17. Jahrhundert und bedeutete *das, wonach man sich zu richten hat.* In der Gegenwart wird Nachricht als eine Mitteilung über einen Sachverhalt beschrieben. Nachrichten werden durch unterschiedliche Medien wie Fernseher, Zeitungen oder Funk verbreitet [9].

Typologie der Kommunikation Schon in der Antike nutzten Menschen unterschiedlichste Medien zur Kommunikation. Anfangs waren es noch Beschriftung der Wände oder auf Steintafeln. Mit der Einführung des Papyrus wandelte sich das meist genutzte Medium. So ergab sich 500 vor Christus die Möglichkeit erstmals Informationen zu transportieren.

Mit der Erfindung der Drucktechnik im 15. Jahrhundert fand in der Mediennutzung eine Revolution statt. Das Monopol lag bei der Kirche, welche durch den Buchdruck den Zugang zu Büchern für die gesamte Bevölkerung ermöglichte [23]. Mit der neuen Technologie verbreiteten sich auch die ersten Nachrichten zügiger. Mit der *Neuen Zeytung* wurden

die ersten, gedruckten Nachrichtenblätter bezeichnet. Neben den Zeitungen wurden auch besonders Flugblätter verteilt [45].

Die moderne Presse hat ihren Ursprung im 17. Jahrhundert. Sie entstand ohne große Aufmerksamkeit zu genießen. Die ältesten Ausgaben der Wochenzeitungen von *Aviso* und *Relation* führen auf 1609 zurück, welches als die Geburtsstunde der modernen Presse bezeichnet wird. Die Wochenzeitungen grenzen sich in ihren Eigenschaften zu anderen zuvor erschienen Medien ab, wie zum Beispiel der *Novellanten*. Die Wochenzeitungen vereinen nämlich die typischen Eigenschaften der modernen Presse. Diese sind:

- Periodizität
- Aktualität
- Universalität
- Publizität

Periodizität steht dabei für einen zeitlichen Rhythmus, bei dem Zeitungen erscheinen. Die Aktualität spielt eine entscheidende Rolle bei der Unterscheidung von Tageszeitungen und Wochenzeitungen. Während bei Tageszeitungen die Spanne zwischen dem Geschehen und der Herausgabe gering ist, liegt sie bei Wochenzeitungen um einiges höher. Schließlich liegt der Fokus bei Wochenzeitungen auf Hintergrundinformationen. Die Universalität impliziert die Vielfältigkeit von Zeitungen bei der Aufbereitung von Themen. Der allgemeine Zugang zur Öffentlichkeit wird als Publizität bezeichnet. [45]

2.2 Wandel der Zeitschriften

Die Schreibmaschine und der Computer Weitere Erfindungen, die die Verbreitung von Medien revolutionierten, war die Schreibmaschine und der Computer. Mit dem Aufkommen der Schreibmaschine war nun das Schreiben um einiges erleichtert worden. Mit dem Aufkommen des Computers und des Internets fand die Digitalisierung der Medien Einzug.

Zu keinem Zeitpunkt der Mediengeschichte war die Dynamik des Wandels so hoch, wie sie heute im Soge des Internets zu beobachten ist. Zum Ende der Weimarer Republik gab es in Deutschland so viele Zeitungen wie dies in der früheren und der späteren Geschichte der Zeitung nie mehr der Fall war. Mit der Verbreitung des Radios, des Fernsehens und

auch des Internets sind neue Medien entstanden, die Zugang zu Nachrichten ermöglichen. Dieser Entwicklung steuern die Zeitungen jedoch durch beispielsweise Online-Plattformen entgegen.

Zahlen und Fakten über Zeitschriften Deutschland ist der größte Zeitungsmarkt Europas und gehört mit USA, Japan, China und Indien zu den Top 5 weltweit. Pro Erscheinungstag werden 14,7 Millionen Wochen- und Sonntagszeitungen verkauft. Die Umsätze der Zeitschriftenverlage in Deutschland lagen 2018 bei 14,6 Milliarden Euro [47]. Am allgemeinen Werbemarkt liegen Zeitungen nach dem Fernsehen auf Platz 2 [49]. Die Werbeeinnahmen für Januar bis November 2019 lagen bei 4,4 Milliarden Euro [21].

Digitalisierung der Verlage Die erste deutsche Online-Tageszeitung, war die *Schweriner Volkszeitung*, die seit dem 05. Mai 1995 online ist und der mittlerweile viele andere Tages- und Wochenzeitungen nachgezogen sind [3]. So ist auch die Zeitschriftenbranche vom Wandel betroffen. Im Jahr 1991 hatten die Tageszeitungen eine tägliche Auflage von rund 27,3 Millionen Exemplaren. 27 Jahre später lag die verkaufte Auflage bei rund 14,1 Millionen Exemplaren. Einer der Faktoren für den Rückgang der Exemplare ist vor allem auf das Internet zurückzuführen. Im Jahr 1995 gab es laut dem Bundesverband Deutscher Zeitungsleger e. V. fünf Online-Angebote. Im Wandel der Digitalisierung gab es 2017 knapp 700 Online-Angebote [5]. Die verkaufte Auflage der Tageszeitungen ist in Deutschland hingegen seit Jahren rückläufig, deshalb machen die erhöhten Auflagen von E-Papers und Paid-Content-Modelle Hoffnung. So stieg Zahl der Paint-Content Angebote von 40 im Jahr 2014 auf 212 in 2018. Der Paid-Content Umsatz stieg von 2013 auf 2018 um insgesamt 260 % [37]. Angebote der Zeitungen im Internet rufen mittlerweile 38,7 Millionen Unique User über 14 Jahren pro Monat auf [50]. Die Reichweite der online Angebote für Nutzer ab 16 Jahren lag 2018 bei 68 % [21]. Vor allem ist sie durch ihren Umsatz von 14,6 Milliarden Euro eine nicht zu unterschätzende Branche. Wie in der Abbildung 2.1 zusehen ist die Anzahl der Online Angebote der Zeitungen erheblich gestiegen. Somit ist eine E-Commerce Strategie und die damit verbundene Personalisierung ein wesentlicher Bestandteil der Online Zeitschriften.

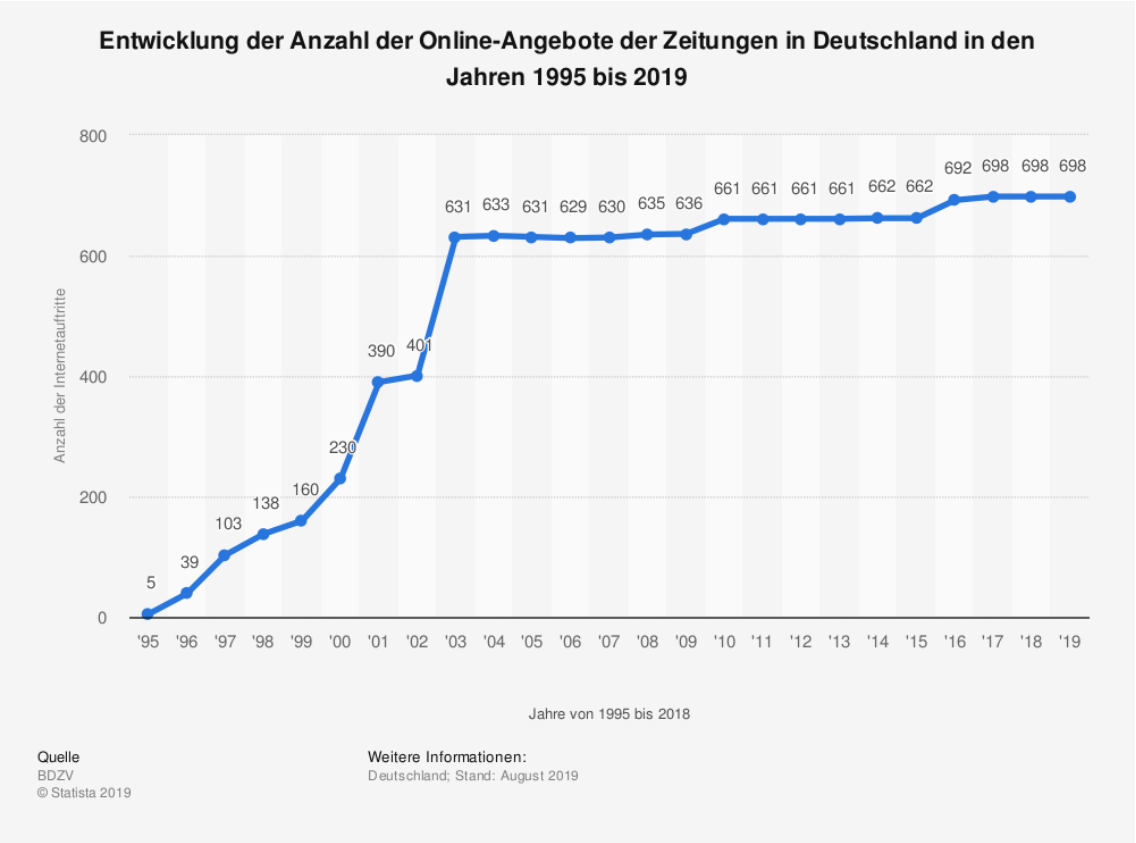


Abbildung 2.1: Entwicklung der Online Angebote der Zeitungen in Deutschland bis 2019 [48]

3 Recommender Systems

Das folgende Kapitel gibt eine allgemeine Übersicht von Recommender Systemen, sowie die Definition und den Zweck von Recommender Systemen wieder. Anschließend wird der Aufgabenbereich der jeweiligen Filter-Methoden und ihre Vor- und Nachteile erläutert.

3.1 Einführung in Recommender Systems

Recommender Systeme sind nicht auf ein bestimmtes Forschungsgebiet zurückzuführen. Ihre Wurzeln sind in den Forschungsbereichen des Information Retrievals, Expertensysteme, Kognitionswissenschaften und Konsumentenentscheidungsmodelierung im Marketing zu finden [12].

Im Zuge der Digitalisierung ab den 1990er wächst das World Wide Web stetig an. Parallel dazu erlangte der E-Commerce seine Entfaltung. Prognosen zu Folge steigt der Umsatz des E-Commerce immer weiter an. Die Produktpaletten der Anbieter vergrößert sich täglich. Es ist möglich fast alles aus dem Internet zu beziehen, von Medikamenten, Spielzeuge, Filme bis hin zu Autos. Mit dem Anstieg der Produkte und der großen Vielfalt steigen ebenfalls auch die Entscheidungsmöglichkeiten. Diese Realität erfordert ein System, bei der individuellen Bedürfnisse der Kunden im E-Commerce berücksichtigt werden. Diese Entscheidungshilfe kann durch Personalisierung erfolgen.

> „[Die] Personalisierung [...] dient dem Hersteller eines Produktes als Argument und Instrument der Vermarktung. Dem Verbraucher eines personalisierbaren Produktes hilft sie hingegen, Arbeitsabläufe effektiver zu gestalten. Beide Sichten zeigen, dass Personalisierung in der modernen Zeit ein wichtiges Werkzeug ist, um den Erfolg eines Produktes positiv zu beeinflussen, indem letztlich die Wünsche seines Verbrauchers beachtet werden.“ [13]

Recommender Systeme sollen in dem Prozess der Entscheidungshilfe ihren Beitrag leisten [30]. Bereits erste E-Commerce-Unternehmen, wie Amazon, hatten Recommender Systeme als festen Bestandteil ihrer E-Commerce Strategie [16, 44].

Schafer et al. [44] stellen drei Wege auf, wie der Umsatz der Unternehmen im E-Commerce gesteigert werden kann.

Surfer in Käufer umwandeln Besucher von Webseiten surfen herum, ohne etwas zu kaufen. Es kann verschiedene Gründe haben, warum es nicht zu einem Kauf kommt, z. B. kann ein gewünschtes Produkt bei der riesigen Auswahl nicht gefunden werden. Recommender Systeme können den Nutzer helfen Produkte zu finden, die sie suchen und ihren Interessen entsprechen könnten.

Steigerung von Cross-Selling Beim Cross-Selling werden dem Käufer weitere passende Produkte empfohlen. Sind die Empfehlungen *gut*, sollte die Bestellmenge und somit der Umsatz steigern. Die Empfehlung kann im Bestellprozess angezeigt werden, z. B. auf Grundlage von gesuchten oder den im Warenkorb gewählten Produkten.

Loyalität bilden Im E-Commerce sind die Konkurrenten nur einen oder zwei Klicks entfernt. Deshalb gehört das Aufbauen einer Kundenbindung zu den wesentlichen Bestandteilen einer E-Commerce Strategie. Recommender Systeme sorgen für eine wertschöpfende Beziehung zwischen Plattform und dem Kunden. Dadurch entsteht eine engere Kundenbindung. Je mehr die Bedürfnisse und Interessen eines Kunden befriedigt werden, desto loyaler wird er der Plattform gegenüber. Durch die Loyalität werden die E-Commerce Plattformen wiederkehrend benutzt. Dadurch können mehr Daten zu den Nutzern gesammelt werden. Dies führt zur Verbesserung der Recommender Systeme [40, 41].

Die Welt der Recommender Systeme ist groß. Sie bedienen sich verschiedenster Verfahren aus den Forschungsgebieten des Data-Minings, welche Subfields von Künstliche Intelligenz sind [42]. Die Ansätze können verschieden sein. Dazu gehören algorithmische, wie Collaborative-Filtering, nicht-algorithmische, personalisierte, wie nicht personalisierte und weitere Ansätze, die im Laufe dieses Kapitels vorgestellt werden. Hierbei werden Verfahren angewendet, um das Interesse der Nutzer an Informationen und Produkten vorherzusagen. Die Systeme verwenden für die Empfehlung Produktdaten und nützliche Daten des Nutzers, u. a. die bisherigen Präferenzen und Interessen [38]. Die Daten können aus verschiedenen Quellen gesammelt und aufbereitet werden. Hier können die Daten aus expliziten oder impliziten Informationen bestehen.

Explizite Informationen Bei der expliziten Informationssammlung wird der Anwender mit Formularen oder direkten Fragen konfrontiert. Er wird beispielsweise nach seinem Geburtsdatum, Geschlecht, Interessengebiete, seinen Hobbys, seinem Familienstatus, Herkunft und einigen anderen Informationen gefragt. Das System erstellt daraufhin ein Benutzerprofil und speichert die gesammelten Daten darin ab. Eine andere und bekannte Art der Sammlung ist das Auslesen von Bewertungen von Objekten, z. B. Artikel, die ein Nutzer getätigt hat. Die Bewertungen des Nutzers zu einem Objekt können ebenfalls in einem Benutzerprofil gespeichert werden.

Implizite Informationen Im Gegensatz zu explizit gesammelten Informationen, bestehen bei der impliziten Informationssammlung die Daten aus Mustererkennung und Analyse der Kundeninteraktion. Im E-Commerce werden z. B. jegliche Kundeninteraktionen auf der Plattform gespeichert. Dazu zählen z. B. das Klickverhalten, die Verweildauer auf einer Seite, Kaufgewohnheiten und Suchmuster der Nutzer.

Im Fokus eines Recommender Systems steht letztendlich die Kunden zugeschnittene Leistung bzw. die Produktempfehlung [33].

3.2 Definition

In der Literatur finden sich zwei verschiedene Schreibarten *Recommendation System* und *Recommender System*, zu Deutsch Vorschlagssystem oder Empfehlungssystem. Letzterer wird in der aktuellen Literatur verwendet. Für die vorliegende Arbeit wird durchgängig die Bezeichnung *Recommender System* genutzt. In der Forschung gibt es keine einheitliche Definition eines Recommender Systems. Sie nehmen je nach Anforderung und Anwendung eine spezifische Bedeutung an. Der Begriff darf nicht gleichgesetzt werden mit *Personalisierung* oder mit *Information-Retrieval*.

Personalisierung Der Begriff der Personalisierung umfasst mehr als Recommender Systeme. Unter Personalisierung wird die Anpassung der Informationen, Diensten oder Produkten verstanden. Diese Anpassung kann auf Grundlage von bestimmten Informationen geschehen oder durch *aktive* Personalisierung z. B. durch das Anpassen eines Interfaces einer Webseite durch den Nutzer. So kann Personalisierung als Oberkategorie für Recommender Systeme bezeichnet werden [32].

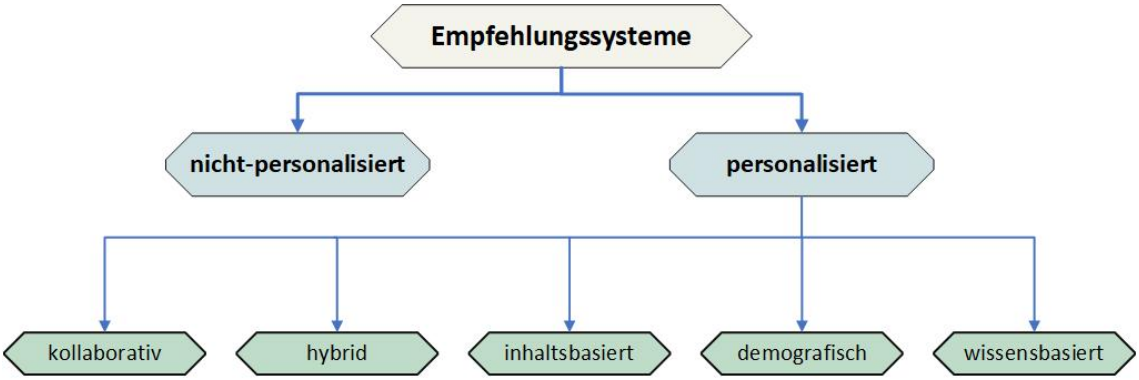


Abbildung 3.1: Klassifizierung von Recommender System (angelehnt an Burke et. al[16])

Information Retrieval Unter dem Begriff Information Retrieval, wird das aktive Formulieren einer Anfrage an einem System verstanden. Das Verfahren ist bekannt z. B. von Desktop-Suchen oder Suchmaschinen. Das Differenzierungsmerkmal zu Empfehlungssystemen besteht darin, dass der Benutzer die Empfehlung nicht automatisch erhält, sondern erst nach Formulierung einer Anfrage. In der Praxis fließen beide Verfahren ineinander. Werden Elemente aus Recommender System in der Suche eingebunden, wie z. B. Interessen des Benutzers, dann handelt es sich hierbei um einen Mischverfahren [32].

Der Begriff Recommender System wird auch häufig mit dem Begriff *Collaborative-Filtering* gleichgesetzt. Laut Resnick und Varian ist dies nicht ganz korrekt. Empfehlungen lassen sich kategorisieren - wie im Organigramm zusehen - indem sie personalisiert oder nicht personalisiert sind. [30]. Die jeweiligen Techniken unterscheiden sich anhand ihrer Datenquellen für die Empfehlung [16]. Techniken wie Collaborative, User und Item -based Filtering ordnen sich somit den personalisierten Empfehlungen unter. Nicht personalisierte Empfehlungen sind typische Suchanfragen wie in Information Retrieval Systemen, die jedem Nutzer gleiche Empfehlungen geben [30]. Die einzelnen Methoden werden ab Abschnitt 3.3 näher erläutert.

Einige zentrale Aussagen zu Recommender Systeme dienen der Annäherung der Definition:

- „Recommendersystems are personalized information agents that provide recommendations: suggestions for items likely to be of use to a user.“ [16]
- „Recommender Systems [...] are software tools and techniques providing suggestions for items to be of use to a user [...]. The suggestions relate

to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read.“ [42]

- „Recommendation systems apply data mining techniques and prediction algorithms to predict users interest on information and products among the large amount of available items. [...] Recommendation system is a software of predict the useful information regarding product using user's past preference and interest.“ [38]

Das Problem lässt sich formal wie folgt durch eine Nützlichkeitsfunktion beschreiben: Sei C der Raum der Nutzer (User) im System und S der Raum der Objekte (Items), dann wird eine Funktion c gesucht, welche für jeden Nutzer eine Nützlichkeit r für alle Objekte angibt. Hierbei steht jeder Nutzer c zu einer Beziehung zu jedem Objekt s. Somit ergibt sich die Abbildung [12]:

$$f_c : C \times S \longrightarrow R \tag{3.1}$$

Hier stellt R die geordnete Menge aller Nützlichkeitswerte r dar und besteht aus nicht negativen, reellen oder natürlichen Zahlen:

$$R = r \mid r \in \mathbb{R}_+, r \in \mathbb{N}_0 \tag{3.2}$$

Zusammengefasst ist ein Empfehlungssystem oder auch Recommender System:

> „[...] ein System, das einem Benutzer in einem gegebenen Kontext aus einer gegebenen Entitätsmenge aktiv eine Teilmenge *nützlicher* Elemente empfiehlt.“ [32]

3.3 Filter Methoden für ein Recommender System

Recommender Systeme lassen sich auf Grundlage ihrer Filtermethode unterscheiden. Die Filtermethoden bedienen sich verschiedener Ansätze, welche in den nächsten Abschnitten thematisiert werden. Es gibt nicht die eine Methode, welches für jedes Szenario passt. Je nach Anforderung kommt eine bestimmte Methode infrage, weil jede Methode ihre eigenen Vor- und Nachteile hat und bestimmte Parameter voraussetzt. In der Forschung wurden bereits verschiedene Verfahren aufgestellt. Angelehnt an Ricci et al. [46] und Burke et al. [16] können folgende Kategorien aufgestellt werden:

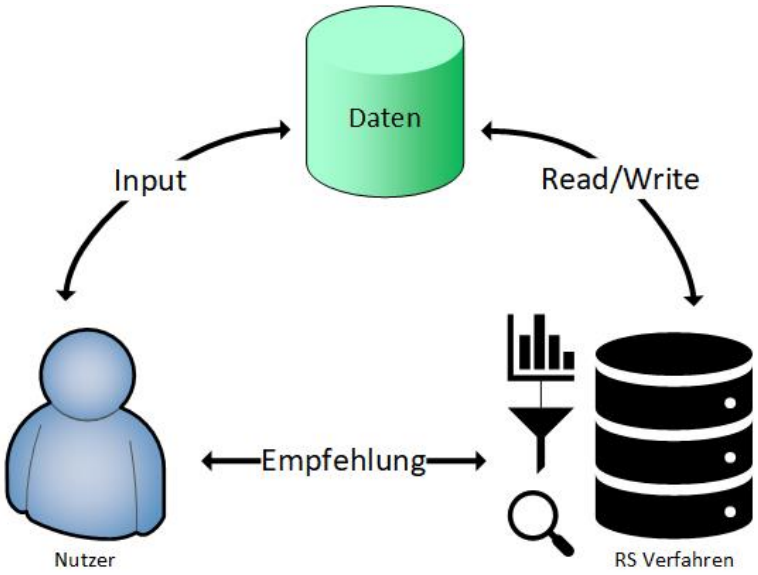


Abbildung 3.2: Vereinfachte Darstellung der Funktionsweise eines Empfehlungssystems (angelehnt an Hohfeld et al. [30])

Nicht-personalisiertes Filtern Die Empfehlungen sind für alle Nutzer gleich, wie z. B. eine einfache Desktop-Suchanfrage oder die Empfehlung des neusten Produktes.

Collaborativ-Filtering Hier basieren die Empfehlungen auf historisch gesammelte Bewertungen und Daten. Einfach ausgedrückt werden ähnlich *gut* bewertete Produkte von Kunden mit ähnlichem Profil vorgeschlagen. Im Detail wird noch zwischen $User-based$ und $Item-based$ Methode unterschieden.

Content-based Filtering Die Empfehlungen basieren auf gesammelte Daten und Bewertungen von Produkten. Hierbei werden Produkte mit ähnlichen Eigenschaften von bereits *gut* bewerteten Produkten empfohlen.

Demografisches Filtern In dieser Filter-Methode werden nur Informationen über den Nutzer verwendet, wie z. B. Geschlecht, Alter und Beschäftigungsstatus. Das System generiert somit Empfehlungen auf Grundlage von demografischen Ähnlichkeiten.

Wissensbasiertes Filtern Empfehlungen werden auf Grundlage von explizitem Wissen, wie z. B. über bestimmte Interessen des Nutzers gegeben.

Hybrides Filtern Empfehlungen bestehen aus Kombinationen von mehreren Filter-Methoden, wie z. B. einer Kombination aus Collaborative-Filtering und Content-based-Filtering.

3.4 Nicht-personalisiertes Filtern

Wie dem Namen schon entnommen werden kann, berücksichtigt die Methode des Nicht-personalisierten Filterns keine Informationen des Benutzers. Daraus folgt, dass alle Empfehlungen identisch sind. Beispiele für diese einfachen Systemen, die in E-Commerce Webseiten angewendet werden, sind einfache Empfehlungen von $Top - N$ neu erschienene Produkte, die dem Nutzer auf der Startseite eingeblendet werden. Ein konkretes Beispiel lässt sich bei Amazon.com finden. Anonyme Benutzer bekommen auf der Webseite Empfehlungen von Produkten, die zurzeit von anderen Mitgliedern angesehen werden. Somit sind die Empfehlungen einfache Vorschläge, die dem Nutzer gefallen könnten, unabhängig von konkreten Benutzerinformationen [39].

Vorteil Nicht-personalisiertes Filtern ist einfach in der Implementierungen, da die Empfehlungen aus beliebten oder einfach hoch bewerteten Produkten bestehen. Somit sind auch die erforderlichen Daten für diese Empfehlungen leicht zu erfassen [24].

Nachteil Durch die Methode des Nicht-personalisiertes Filtern sind für jeden Benutzer die Empfehlung gleich. Somit sprechen möglicherweise die Empfehlungen nicht alle an [24].

3.5 Demografisches Filtern

Die demografische Filter Methode benutzt demographische Daten des Nutzers, wie z. B. Alter, Geschlecht, Herkunft und Beruf, um bestimmte *Klassen* von Nutzer zu identifizieren, die ein bestimmtes Produkt mögen könnten. Einer der ersten Recommender Systeme wurde von Grundy basierend aus dieser Methode erstellt [43]. Hier werden Bücher empfohlen, die auf persönlichen Informationen basieren. Diese Daten wurden durch einen interaktiven Dialog gesammelt. Die Antworten der Nutzer wurden anschließend mit verschiedenen Benutzerstereotypen verglichen. Dadurch werden $Mensch - zu - Mensch$

Korrelationen abgebildet [16]. So beruhen die Empfehlungen hier auf eine Übereinstimmung zwischen den demographischen Daten und den Eigenschaften und Attributen der Produkte [30].

Vorteil Der Vorteil dieser Methode besteht wahrscheinlich darin, dass es keine historischen Daten benötigt, wie es bei anderen Filter-Methoden wie es bei Collaborative- und Content-based-Filtering der Fall ist [16].

Nachteil Bei der Erfassung der gesamten demographischen Informationen kann es zu Konflikten mit dem Datenschutz kommen, da es sich um sensible und persönliche Daten handelt [11]. Zudem gehört es zu den traditionellen Techniken Profile von Nutzern zu erstellen und es fließen keine erweiterte Data-Mining Techniken hinein. Dadurch entsteht kein Wettbewerbsvorteil und die aktuelle Forschung wird nicht berücksichtigt [12].

3.6 Wissensbasiertes Filtern

Der Nutzer hat bestimmte Interessen, die in Korrelation mit den Eigenschaften bestimmter Objekte stehen. Hier wird dann funktionales Wissen durch Schlussfolgerungen abgeleitet. Dieses Wissen in Kombination mit einem Nutzerprofil in der Datenbank hinterlegt und mit anderen Produkteigenschaften verglichen [30]. Wissensbasiertes Filtern wird in den Rubriken der am wenigsten gekauften Produkte eingesetzt z. B. Immobilien und Autos. Sie nutzen Regel- und Ähnlichkeitsinformationen. Nutzer haben die Möglichkeit Eigenschaften der Produkte auszuwählen, um die Empfehlungen gemäß Interessen einzuschränken. Auf der Plattform mobile.de, haben die Nutzer die Möglichkeit eine Vielzahl von Eigenschaften eines Autos festzulegen, wie z. B. Baujahr, Modell, Marke, Kilometerstand, Zustand und Preis. Durch das Wissen der Präferenzen kann dann das Recommender System eine Empfehlung generieren. Wissensgenerierung stellt hier eine Herausforderung dar. Es muss hier identifiziert werden welche der Produkteigenschaften relevant sind, z. B. Farbe oder Marke des Autos [15].

Vorteil Benutzer haben hier den Vorteil Produkte durch die Einschränkungen zu finden, mit dem sie vertraut sind oder Produkte, die ähnlich sind und welche, die den aufgesetzten Kriterien entsprechen [15].

Nachteil Das System muss einen Zugang zu einer Wissensdatenbank haben, wo die Informationen leicht erkennbar oder ableitbar sind. Zudem besteht die Herausforderung für eine gute Empfehlung, zu untersuchen welche Produkteigenschaften genau ausschlaggebend sind für den Benutzer [15].

3.7 Collaborative-Filtering

Collaborativ-Filtering ist eine weit verbreitete Filter Methode bei Recommender Systeme. In diesen Systeme werden auf Grundlage von gesammelten Bewertungen anderen Nutzern zu Objekten (Texte und Produkte) Empfehlungen für einen Nutzer generiert. Die Bewertungen können boolesch *ja / nein*, *interessant / nicht interessant* oder metrisch skaliert sein.

Im Modell m existieren Nutzer $U = \{u_1, u_2, ..., u_n\}$ und eine Liste von Objekten I. Jeder Nutzer u_i besitzt eine Liste von Objekten, über die das Profil des Nutzers errechnet wird. Somit ist I_{ui} die Bewertung eines Objektes I_i von dem Nutzer U_i und ist Teilmenge von I und kann auch die Nullmenge enthalten [32]. Daher ist die Empfehlung eines Objekts I_i für den Nutzer u_i oder die Empfehlung einer $Top-N$ Liste, bestehend aus sehr wahrscheinlich interessanten Objekten aus der Menge I, auf den Bewertungen des Nutzers und den Bewertungen anderer Nutzer mit ähnlichen Profilen. Ziel ist es ähnliche Nutzerprofile bzw. dem Nutzer am nächsten Nachbarprofil zu identifizieren, um daraus eine Empfehlung zu generieren. Hierfür wird eine gewichtete Kombination der Beurteilung der Nutzergruppe, deren Mitglieder als Nachbarn infrage kommen berechnet und auf Basis dessen eine Empfehlung für den aktiven Nutzer generiert [44]. Gemeinsamkeiten in Nutzerprofilen können mittels u. a. des Vektorraummodells realisiert werden. Der konkrete Vektor des Nutzers N ergibt sich durch die angesehenen, gekauften oder als positiv bewerteten Objekte. Die Ähnlichkeit zu anderen Nutzern ergibt sich durch die Berechnung des Cosinus der jeweiligen Vektoren. Über einen Schwellenwert der Ähnlichkeit (Cosinus) erkennt das System diejenigen Nutzer, die dem Ausgangsnutzer am ähnlichsten sind. Aus dem Bewertungsverhalten der am ähnlichsten Nutzer werden letztendlich die Empfehlungen abgeleitet [30]. Weitere Methoden zu Berechnung von Ähnlichkeiten werden in Abschnitt 4.5 behandelt.

Es gibt zwei verschieden Herangehensweisen, eine Memory- und Model-based Variante.

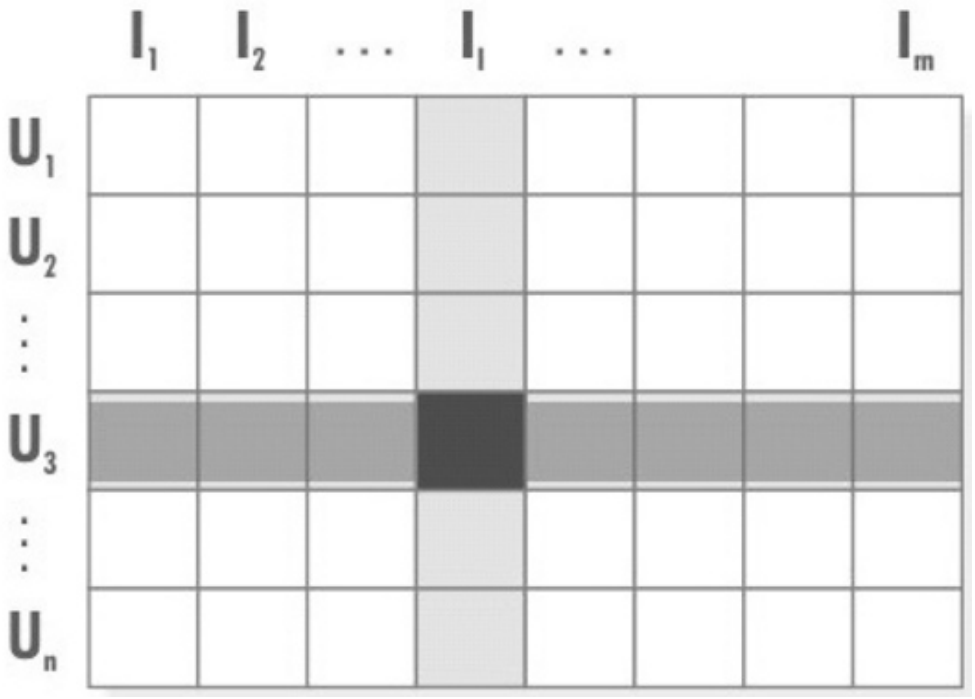


Abbildung 3.3: Die CF-Matrix der Benutzerempfehlungselement Beziehungen [32]

Memory-based Hier basieren alle Berechnungen, also Empfehlungen, auf Cosinus- oder Korrelationsbasiertem Ähnlichkeitsmaß. Die Berechnungen werden direkt auf der Datenmatrix ausgeführt. Vorteil dieser Methode ist die Beachtung der gesamten Datenbasis. Daraus ergibt sich auch ein Nachteil, weil dadurch die Berechnungszeiten höher sind [30].

Model-based Mithilfe stochastischer Verfahren, andere Techniken u. a. Clusteranalyse und Neuronalen Netzen und eines Trainingsdatensatzes wird $offline$ ein Modell gelernt. Das gelernte Modell wird dann *online* für die Berechnung benutzt. Ein Vorteil dieser Methode ist, dass die Berechnungszeiten kürzer sind, weil nicht die komplette Datenbasis aufgerufen wird. Nachteil kann der Informationsverlust bei der Reduktion auf ein Modell sein [30].

Das Collaborative-Filtering Konzept kann mit zwei verschiedenen Verfahren implementiert werden.

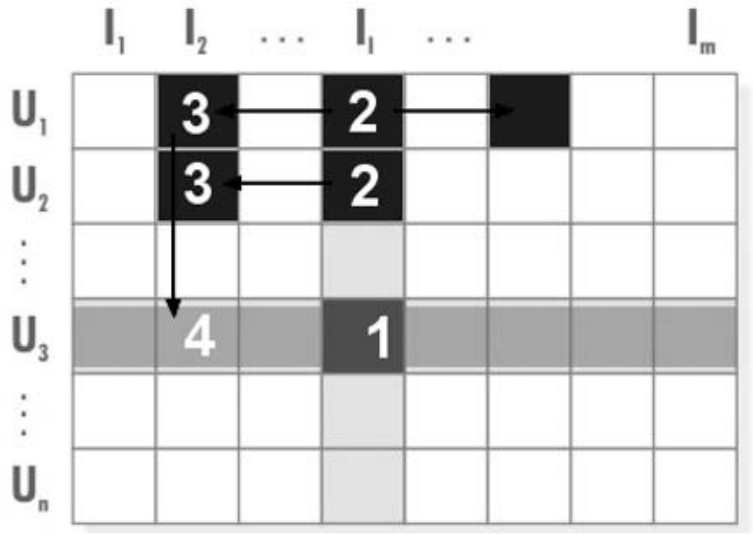


Abbildung 3.4: Elementbasiertes Collaborative-Filtering Konzept [32]

Item-based Verfahren Das Item-based Verfahren basiert auf Ähnlichkeiten von Objekten und dem Bewertungsmuster der Benutzer, um Empfehlungen zu generieren. Mit der Annahme, dass die Benutzer dieselben Präferenzen für die gleichen Objekte haben, werden die Objekte als gleich angesehen, die dieselbe Bewertung haben. Auf der Basis von n Benutzer und m Objekte wird eine Matrix $R = (r_ij)$ mit $i = 1..n$ und $j = 1..m$ aufgestellt. Ziel ist es eine Bewertung bzw. Relevanz R (I_y, U_x) für ein Objekt I_y für den Nutzer U_x zu finden. Für die Berechnung der Ähnlichkeit wird ein Vektor gebildet. Der Vektor besteht aus allen Paaren der Empfehlungsobjekten $(I_y, I1), ..., (I_y, Iu)$ anderer Benutzer, die das Zielobjekt I_y und Iu bewertet haben. Dieser Vektor wird dann für die Ähnlichkeitsberechnung verwendet.

User-based Algorithmus Im User-based Verfahren werden Ähnlichkeiten auf Grundlage von ähnliche Nutzer definiert. Die Ähnlichkeiten bzw. Nachbarschaften zwischen den Nutzern werden durch das Bewertungsverhalten ermittelt. Auf der Basis von n Benutzer und m Empfehlungsobjekten wird die Matrix $R = (r_ij)$ mit $i = 1..n$ und $j = 1..m$ erzeugt. Hierbei stellt der Wert r_ij die Bewertung des Objektes j durch den Nutzer i. Die Bewertung kann explizit $gut/schlecht$ oder implizit $gelesen/nicht - gelesen$ sein. Ziel ist es somit für ein Nutzer U_x ein Empfehlungsobjekt I_y zu empfehlen durch die Benutzer $U = \{U_1, .., U_n\}$, die U_x, die am Ähnlichsten sind und Objekte gut bewertet haben, die der Benutzer U_x, noch nicht bewertet hat. Für dieses Verfahren wird häufig die

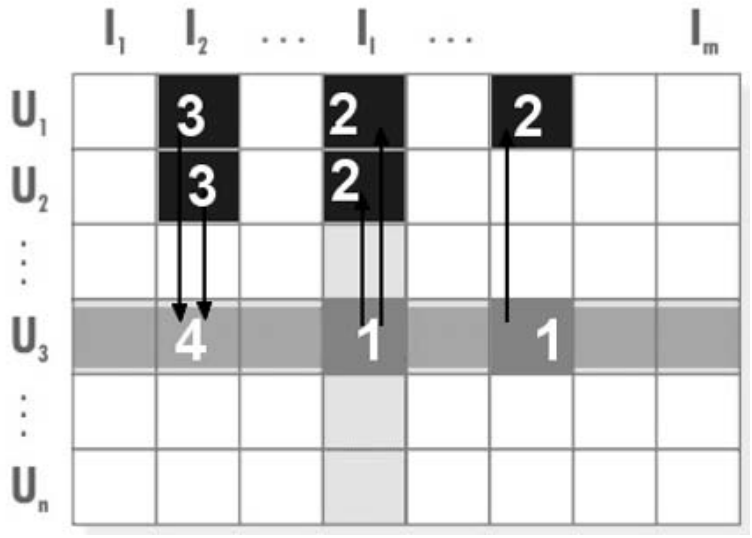


Abbildung 3.5: User-based Collaborative-Filtering Konzept [32]

Pearson-Korrelation verwendet, um die Ähnlichkeiten zwischen zwei Nutzer U_1 und U_x zu berechnen. Sind die k nächsten Nachbarn gefunden wurden, wird die Bewertung durch die Summe des gewichteten Durchschnitts der Bewertung über die k nächsten Nachbarn berechnet.

Herausforderungen des Collaborative-Filtering Die wesentlichen Nachteile des Collaborative-Filtering ist das Coldstart-Problem, das Problem der Spärlichkeit (sparsity) und des Lemming-Effekts. Diese werden im Folgenden beschrieben.

Coldstart-Problem Die Empfehlungen beim Collaborative-Filtering Verfahren basieren auf Bewertungsmuster anderer Benutzer, daher kann eine Empfehlung nur erfolgen, wenn auch eine bestimmte Anzahl an Bewertung anderer Nutzer auch vorhanden ist. Das Problem gilt ebenfalls für neue Nutzer $New - User - Problem$. Sind keine implizite oder explizite Informationen vorhanden, kann das Collaborative-Filtering Verfahren keine Empfehlungen generieren. Diese Herausforderung wird auch als $Coldstart - Problem$ bezeichnet.

Spärlichkeit (sparsity) Die Plattformen haben oftmals eine große Anzahl von Produkten. Nutzer haben meistens wenige Bewertungen abgegeben im Verhältnis zur gesam-

ten Produktpalette. So erhalten Benutzer mit einem sehr $spezifischen$ Interessenprofil Empfehlungen, die nicht relevant sind, weil es wenige ähnliche Benutzer gibt.

Lemming-Effekt Der Lemming-Effekt zeichnet sich dadurch aus, dass ein Objekt sehr begehrt ist und oft gut bewertet wird. So kommt es dazu, dass das dieses Objekt oft empfohlen wird. Dadurch werden diese Objekte immer wieder empfohlen und dies kann dazu führen, dass die neuen Objekte keine Chance haben in die $Top-N$ Liste zu gelangen und werden somit nicht empfohlen. Ein bekanntes Beispiel hierfür wurde bei Amazon.com beobachtet. Beim Büchershop von Amazon.com wurden 2006 sehr oft die Bücher von Dan Brown empfohlen, obwohl diese Bücher keinen konkreten Zusammenhang mit dem Profil des Nutzers haben. Dieses Produkt wurde einfach zu oft gekauft und positiv bewertet [32].

3.8 Content-based Filtering

Das Content-based-Filtering kann als $Objekt-zu-Objekt-Korrelation$ verstanden werden. Hier basieren die Empfehlungen im Wesentlichen auf die Eigenschaften der Objekte. Dem Nutzer werden Objekte empfohlen, die zu seinem eigenen Nutzerprofil passen. Im Gegensatz zu Collaborativ-Filtering wird das Wissen vom Profil des Nutzers abgeleitet und bezieht sich nicht auf die Profile und Bewertungen anderer Nutzer. Hierfür werden die Bewertungen des Nutzers zu den Objekten genutzt, die der Nutzer in der Vergangenheit gemacht hat. Dafür werden mit verschiedenen Techniken die Eigenschaften der Objekte analysiert $Feature-Selection$ und anschließend für die Empfehlung Funktionen benutzt, die die Ähnlichkeit zwischen Objekten berechnen. Für die Laufzeitkomplexität der Eigenschaftsanalyse ist es wichtig ein gutes Verhältnis zwischen der Menge der Eigenschaften und dem Ausschließen anderer Objekte zu erzielen. Das System lernt Präferenzen von Nutzern zu analysieren, indem diese ihm ein Feedback abgeben, welches *implizit* oder *explizit* sein kann. Anhand der gewonnenen Daten wird ein Profil erstellt, das durch Nutzer-Feedback weiter modifiziert werden kann. Wird dem Nutzer eine Webseite vorgeschlagen, so können die Wörter jenes Dokuments mit ihren Gewichtungen – sofern er die Empfehlung als positiv erachtet – in sein Nutzerprofil aufgenommen werden [30, 32]. Entstanden ist diese Form der Empfehlungssysteme aus der Technik der Informationsfilterung. Die Lernmethoden entscheiden über die Art des Nutzerprofils. Mögliche Anwendungen sind Entscheidungsbäume, Neuronale Netze oder vektorbasierte

Repräsentationen [16]. Content-based Empfehlungsansätze eignen sich i. d. R. eher für Texte als für Musik oder Bilder, weil der Inhalt von Texten einfacher analysiert und dargestellt werden kann. Bestimmte Worte innerhalb der Texte werden gewichtet, um die Relevanz eines Dokuments für einen bestimmten Nutzer feststellen zu können, indem die gewichteten Wörter mit seinen Präferenzen abgeglichen werden. Sie eignen sich auch bedingt für die Empfehlung von Produkten. Dies ist aber nur dann der Fall, wenn der Benutzer das Angebot eines Online-Shops regelmäßig nutzt und seine Präferenzen in Form von Produktbewertungen abgibt. Der News-Alert-Service von Google ist ein Beispiel für Content-based-Filtering. Nach Eingabe verschiedener Suchwörter durch den Nutzer wird dieser informiert, wenn es im Nachrichtenbereich Meldungen gibt, die die angegebenen Suchwörter enthalten. [32]

3.9 Hybrides Filtern

Hybride Filter Methode verwenden eine Kombination aus den zuvor erwähnten Methoden. Wie bereits erwähnt haben die jeweiligen Methoden ihre Vor- und Nachteile. Durch die Kombination der Methoden versucht das Hybride-Verfahren die jeweiligen Stärken zu nutzen, um die Schwächen letztendlich zu minimieren. Das $Coldstart - Problem$, wie in Abschnitt 3.7 beschrieben, ist eine Herausforderung für Collaborative-Filtering, weil es keine Objekte empfehlen kann, wenn ein Nutzer noch keine Bewertungen abgegeben hat. Durch das Kombinieren mit *Wissensbasierte Filtern*, in dem vorher einige Interessen abgefragt werden, kann diese Schwäche des Collaborativ-Filtering abgefangen werden. Burke et. al [16] zählen folgende Kombinationen der Ansätze auf:

Weighted Bei einem gewichteten Ansatz erstellen die Verfahren unabhängig voneinander eine $Top - N$ Liste aus Empfehlungen. Anschließend werden die Ergebnisse zusammengetragen anhand einer bestimmten Gewichtung. Die Herausforderung hier ist eine möglichst zutreffende Gewichtung zu wählen.

Switching In diesem Ansatz werden unter einer bestimmten Bedingung oder Ereignis zwischen verschiedenen Filter Verfahren gewechselt.

Mixed Dieser Ansatz ist besonders geeignet, wenn die $Top-N$ Liste aus vielen Objekten bestehen kann. Denn hier erstellen verschiedene Verfahren unabhängig voneinander die $Top-N$ Liste und am Ende werden die Ergebnisse zusammengeführt.

Feature Combination Bei der *Feature Combination* werden Informationen und Erkenntnisse aus einem Verfahren abgeleitet und fließen in dem anderen Verfahren mit ein. Ein Beispiel wäre durch das Collaborative-Filtering Benutzerklassen zu erstellen und diese Merkmale anschließend in einem Content-based-Filtering zu überführen.

Feature Augmentation Hier gibt es Ähnlichkeiten zu dem *Feature Combination* Ansatz. Anstatt Features aus dem einem Verfahren direkt zu übernehmen, werden für jedes Element ein neues Feature auf Basis des Verfahrens generiert. Diese Informationen fließen anschließend in das nächste Verfahren hinein.

Cascade Der *cascade* Ansatz ist hierarchisch, wobei ein Verfahren dem anderen untergeordnet ist. Das *schwächere* Verfahren kann die Empfehlungen des *stärkeren* nicht ganz überstimmen, aber dient dazu eine Verfeinerung in den Empfehlungen vorzunehmen.

Meta Level In diesem Ansatz werden die abgleitet Information eines Verfahrens als *Input* für das nächste Verfahren verwendet. Der Unterschied zum *Feature Augmentation* ist, dass in diesem Ansatz das erste Verfahren vollständig die Wissensquelle durch ein gelerntes Modell ersetzt. Dieses Modell wird als Berechnungsgrundlage anschließend verwendet.

Durch hybride Systeme können einige der Nachteile der einzelnen Verfahren abgeschwächt werden. Im Idealfall werden nur die positiven Eigenschaften der Grundsysteme übernommen und die Nachteile vollständig eliminiert. Häufig wird als Grundlage ein Collaborative-System verwendet und mit anderen Verfahren kombiniert, um das $New-Item$ oder $New-User$ Problem zu lösen. Zum Beispiel kann ein weiteres Content-based-System eingesetzt werden, um neue Objekte zu empfehlen, die noch nicht von Benutzern bewertet wurden [16].

4 Konzept und Implementierung eines Recommender Systems für Nachrichten

In diesem Kapitel wird die Vorgehensweise für das Entwickeln des Konzepts, die Implementierung und die Entwicklungsumgebung einschließlich der Ressourcen beschrieben. Zudem wird die Datenquelle beschrieben und anschließend wird der Datenverarbeitungsprozess für diesen Fall erläutert. Zum Schluss wird das Modell mittels der Software KNIME implementiert und die Ergebnisse präsentiert.

4.1 Zielsetzung

Wie in den vorherigen Kapiteln erläutert, sind Recommender Systeme essenziell für Nachrichtenplattformen. Die Online Nachrichtenplattform, aus dem die Cookies entnommen sind, bietet zurzeit eine Filterfunktion an, in dem der Nutzer Artikeln nach Kategorien filtern kann. Im Kontext dieser Plattform soll untersucht werden, ob aus den gesammelten Cookies Empfehlungen generiert werden können. Es soll überprüft werden, ob eine Empfehlung mittels dem Collaborative-Filtering Ansatz möglich ist. Um eine Empfehlung generieren zu können muss aus den Daten Wissen generiert werden. Für den Collaborative-Filtering Ansatz ist es wichtig aus den Daten Profile von Nutzer zu erstellen, die ähnlich in ihren Interessen und Nutzerverhalten sind. Dafür müssen die Daten aus den Cookies verarbeitet und transformiert werden. Für die Berechnung der Ähnlichkeiten von Profilen werden Distanzfunktionen genutzt. Anschließend sollen die Profile geclustert werden. Hierfür soll überprüft werden, welche Algorithmen aus dem Data-Mining Umfeld geeignet sind. Anschließend sollen für einen Nutzer Artikeln empfohlen werden, die er noch nicht gelesen hat. Die Empfehlung basiert somit auf der Grundlage des ähnlichen Nutzerverhaltens und Interessen des Nutzers zu anderen Nutzern aus seinem Cluster.

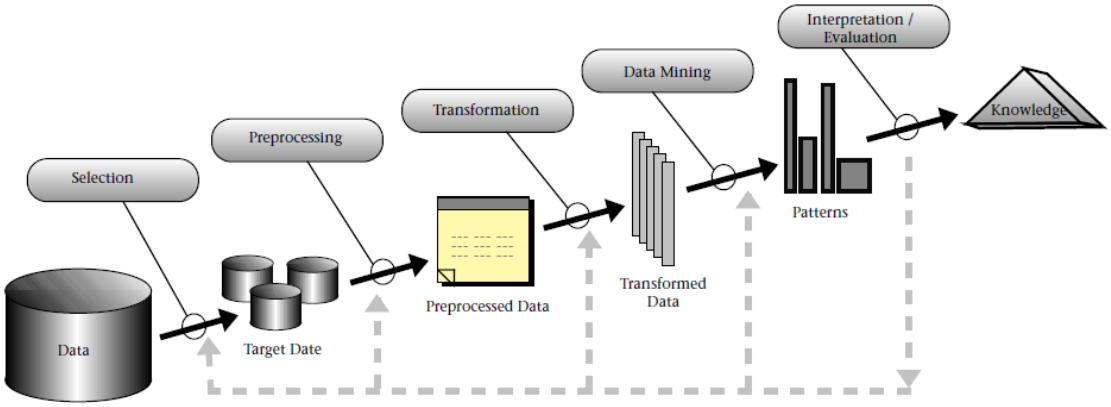


Abbildung 4.1: KDD-Prozess nach Fayyad et al. [25]

4.2 Vorgehensweise CrispDM/Fayyad

Die Vorgehensweise wird angelehnt an dem KDD - Knowledge Discovery in Databases - Prozess von Fayyad et al. [25]. Recommender Systeme nutzen wie in Kapitel 3 ausgeführt Data-Mining Techniken. Fayyad definiert KDD wie folgt:

> „The nontrivial process of identifying valid, novel, potentiallyuseful, and ultimately understandable patterns in data.“ [25]

Dieser Ansatz stellt einen nichttrivialen Prozess dar, dessen Ziel es ist, Muster aus den Datenquellen zu erstellen. Diese Muster haben die Eigenschaft, dass sie gültig sind, unbekannt, potenziell nützlich und leicht verständlich. Das CRISP-DM Modell, *Crossindustry standard process for data mining*, hat im Mittelpunkt die betriebswirtschaftliche Problemstellung und stellt den zyklischen Charakter des Prozesses dar. Das Modell wurde durch ein Konsortium von Firmen wie NCR Corporation, Daimler AG, SPSS, Teradata und OHRA entwickelt [19].

KDD nach Fayyad In dem KDD-Prozess nach Fayyad wird im ersten Schritt die vom Datenbestand, die Daten selektiert, die relevant sind für das Ziel. Anschließend werden die Daten bereinigt und ggf. Daten angereichert. In der Transformationsphase werden die Daten in eine Form gebracht, die für das Analyseverfahren benötigt werden. Die transformierten Daten werden unter bestimmten Data-Mining Verfahren auf Mustererkennung

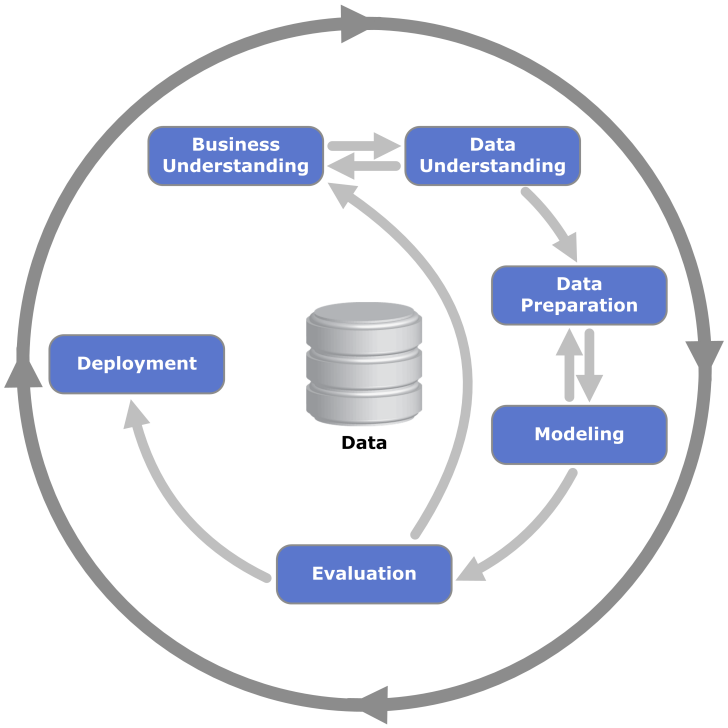


Abbildung 4.2: CRISP-DM Prozess [2]

untersucht. Mithilfe eines Experten werden in der Interpretationsphase die entdeckten Muster interpretiert und Wissen abgeleitet. Zum Schluss wird dann das ganze evaluiert. Ist das Ergebnis nicht zufriedenstellend können Rücksprünge in vorherige Phasen gemacht werden [25].

CRISP-DM Im ersten Schritt des Modells, *Business Understanding*, werden mit den Fachabteilungen und Beteiligten die Problemstellung und Zielsetzung des Projektes festgelegt. Anschließend wird die Datenbasis im Hinblick der definierten Zielsetzung und Problemdarstellung, *Data-Understanding*, untersucht. Hier gibt es die Möglichkeit einen Schritt zurückzuspringen wie in der Abbildung 4.2 dargestellt. Im *Data-Preparation* Schritt wird die Datenbasis bereinigt und für das Data-Mining Verfahren transformiert. Im *Modeling* Schritt werden Data-Mining Verfahren angewendet, Muster innerhalb der Datenbasis gesucht und durch Modelle beschrieben. Anschließend werden im Evaluationsschritt die Ergebnisse unter Berücksichtigung der anfangs definierten Anforderungen ausgewertet. Im *Deployment* Schritt wird das gewonnen Wissen eingesetzt. Ist die Evaluation erfolgreich ist der Prozess abgeschlossen, ansonsten gibt es wieder die Möglichkeit

zum *Business Understanding* Schritt zurückzuspringen [17].

Das CRISP-Modell unterscheidet sich vom KDD nach Fayyad. Die Prozessschritte 1-2 und 6 sind im Fayyad-Modell nicht explizit aufgeführt. Zudem widerspiegelt die 3. Phase des CRISP-Modells die ersten 3 Phasen des Fayyad-Modells. Das CRISP-Modell findet häufig Anwendung in der Wirtschaft und vor allem in der Industrie, während das Fayyad-Modell in der Forschung Anwendung findet [19].

Diese Arbeit beschäftigt sich zwar in einem wirtschaftlichen Kontext, doch im Vordergrund steht die Überprüfung der Zielsetzung. Hier werden die Methoden in der Forschung untersucht und auf diesen Fall projiziert. Der Autor dieser Arbeit hat sich für die Vorgehensweise des Fayyad-Modells entschieden.

4.3 Überblick Data-Mining

Data-Mining bedeutet das Graben von Daten mit dem Ziel aus den Daten Wissen zu generieren. Daten bzw. Wissen entspricht dem heutigen Gold für Unternehmen, da daraus Umsätze und Gewinne generiert werden können. Durch die Informationsüberflutung und der Speicherung großer Datenmenge könne interessante Beziehungen zwischen den Daten versteckt sein. Data-Mining ist eingebettet in analytische Informationssysteme und kann alleine integriert in Business Intelligence oder als Baustein in einem Data-Warehouse angewendet werden. Fayyad et al. definiert Data-Mining wie folgt:

> „Data mining is the application of specific algorithms for extracting patterns from data.“ [25]

Die Analyse der Datenbeziehung mit verschiedenen Verfahren, um Muster in den Datenbeständen zu erkennen, wird also als Data-Mining verstanden. Der gesamte Prozess von Datenselektion bis zur Wissensgenerierung wird als Knowledge Discovery in Databases bezeichnet. Somit ist Data-Mining eine Phase aus dem KDD Prozess. Die Methoden des Data-Mining werden in der Literatur in verschiedenen Kategorien eingeteilt. P. Alpar unterteilt Data-Mining in zwei Ebenen. Die erste Ebene wird als Aufgabe bezeichnet. Die Aufgaben ergeben sich aus dem konkreten Anlass für Data-Mining. Hierzu gehören:

- Klassifikation
- Segmentierung

- Prognose
- Abhängigkeits- und Abweichungsanalyse

J. Cleve et al. ergänzen Text-Mining und Web-Mining als Aufgabenbereich des Data-Minings [19].

Überwachtes und unüberwachtes Lernen Data-Mining arbeitet mit einer gegebenen Menge an Daten. Das Training des Modells wird in zwei Lernstrategien zusammengefasst: überwachtes und unüberwachtes Lernen. Beim unüberwachtes Lernen sind die zu entdeckenden Muster unbekannt. Ein Beispiel hierfür ist eine Cluster-Analyse, dazu mehr in 4.3.1. Das überwachte Lernen beschreibt die Lernstrategie, wenn Beispiele vorgegeben sind, die bereits Resultate enthalten. Beispiel hierfür ist die Zuordnung von Personen anhand von bestimmten Merkmalen. Ein typisches Beispiel ist die Zuordnung von Gehaltslevel in Kreditwürdigkeitsklassen *gut* oder *schlecht*.

Die zweite Ebene sind die Methoden, Verfahren und Algorithmen, die den verschiedenen Aufgaben zugeordnet werden.

4.3.1 Aufgabenbereiche

Klassifikation Ziel der Klassifikation ist es die Objekte in einer vorher bestimmten Klasse zuzuordnen. Anhand der Objektmerkmale und Klasseneigenschaften findet die Zuordnung statt. Eine Klassifikation kann die Einteilung der Kunden in normaler Kreditwürdigkeit oder in sehr guter Kreditwürdigkeit sein. Durch Trainingsdaten, bei denen die Kreditwürdigkeit der Kunden bekannt ist, kann ein Modell entwickelt werden. Da die Klassen und Daten zu einer Klasse vorher bekannt sind, wird von einem überwachten Lernen gesprochen. Die Nachfrage nach Vorhersagen eines Kundenverhaltens auf Basis von Daten ist sehr groß. Dementsprechend gehört die Klassifikation zu den am meisten benutzten Aufgabenbereiche von Data-Mining. Einige Verfahren für diesen Aufgabenbereich sind K-Nearest-Neighbour Verfahren, Entscheidungsbäume, Naives-Bayes-Algorithmus und Support Vector Machines [14, 19].

Prognose (Numerische Vorhersage) Klassifikation und Prognose können als eine Kategorie betrachtet werden. Der Unterschied liegt darin, dass eine Klassifikation diskrete Werte vorhersagt, während eine Prognose Zahlen vorhersagt. Für die numerische Vorhersage wird üblicherweise eine Funktion approximiert. Dafür werden auf Basis von Trainingsdaten die Werte zukünftiger Datensätze berechnet, welche dafür genutzt werden, eine Funktion zu berechnen. In der Praxis findet dieser Aufgabenbereiche eine starke Anwendung, bekannt durch Vorhersage von Aktienkursen und Verkaufszahlen. Einige Verfahren für diesen Aufgabenbereich sind Lineare Regression, Regressionsbäume und k-Nearest Neighbour [14, 19].

Segmentierung Bei der Segmentierung werden Datenmengen in Teilmengen zerlegt. Objekte der Menge werden in Teilmengen, also Gruppen bzw. Cluster zusammengefasst. Die Objekte innerhalb des Clusters sollen so ähnlich wie möglich und so unähnlich wie möglich zu Objekten anderer Cluster sein. Für die Berechnung ist eine Distanz- bzw. Abstandsfunktion erforderlich, um die Ähnlichkeit der Objekte zueinander berechnen zu können [14, 19]. Durch eine Cluster-Analyse können bisher unbekannte Klassen gebildet werden, um z. B. gezielt für spezifische Kundenprofile, Angebote zu erstellen. Einige Verfahren zu diesem Aufgabenbereich sind k-Means-Algorithmus und k-Medoid-Verfahren.

Abbhängigkeitsanalyse und Abweichungsanalyse Bei der Abhängigkeitsanalyse oder Assoziationsanalyse werden nach Beziehungen zwischen Objekten gesucht. Diese Beziehung kann in einem bestimmten Kontext stehen, wie z. B. bei einem Warenkorb. Bei dem klassischen Anwendungsfall einer Warenkorbanalyse steht die Frage im Mittelpunkt, welches Produkt wird mit welchem Produkt am häufigsten gekauft. Somit wird nach der Aussage gesucht:

> Wer Produkt Y kauf, kauft auch Produkt Z.

Durch das gewonnene Wissen, kann ein Unternehmen dem Kunden bei einem Kauf von Produkt Y dem Kunden den Kauf von Produkt Z empfehlen. In einem Supermarkt können z. B. meist zusammengekaufte Produkte nebeneinander aufgestellt werden. Diese Analyse eignet sich gut, um Zusammenhänge zwischen verschiedene Waren zu erkennen und Kundenverhalten zu analysieren. Es analysiert die Daten, um Muster zu identifizieren und das Verhalten neuer Datensätze vorherzusagen. Für die Analyse werden Verfahren wie A-Priori-Verfahren und Frequent Pattern Growth verwendet.

Text-Mining Ein wesentlicher Unterschied zu Data-Mining ist, dass Texte meistens im Gegensatz zu Datenbanken und Web-Seiten unstrukturiert sind. Somit ist ein Kerngebiet des Text-Minings die Analyse von Textdokumenten. Die Nachfrage nach der Klassifizierung eines Dokuments ist ein bekanntes Beispiel. Hierfür werden relevante Begriffe aus dem Dokument extrahiert, um daran das Dokument einem Thema, also einer Klasse, zuzuordnen. Eine andere Möglichkeit wäre die Berechnung der Ähnlichkeit zu anderen Dokumenten. Das Vorgehensmodell zu Text-Mining ähnelt dem Data-Mining Vorgehensmodell. Hinzu kommt das Extrahieren relevanter Informationen aus den Textdokumenten. Für vertiefte Informationen wird auf die Literatur *The Text Mining Handbook* verwiesen [26].

Web-Mining Data-Mining Prozesse, die das Internet als Datenquelle für die Modelle nutzen werden als Web-Mining bezeichnet. Wie in Abbildung 4.3 zusehen, werden je nach inhalts- oder nutzungsorientierter Analyse das Themengebiet in Web-Content Mining und Web-Usage Mining aufgeteilt.

Web-Content Mining befasst sich mit der Analyse des Inhaltes im Internet, z. B. Texte und multimediale Informationen.

Web-Usage Mining befasst sich mit der Analyse des Verhaltens von Nutzern im Internet. Hier werden Data-Mining Methoden auf gesammelte Daten des Web-Servers angewendet. Web-Log Mining bezeichnet die ausschließliche Nutzung von Protokolldateien des Web-Servers. Werden weitere Datenquelle für die Data-Mining Prozesse herangezogen, wird diese Form als Web-Usage Mining bezeichnet [19].

Das Internet hat sich zu einer bedeutenden Plattform für geschäftliche Prozesse entwickelt 2. Web-Log Mining bekommt dadurch einen hohen Stellenwert. Unternehmen nutzen automatisch erzeugte Nutzungsdaten in Logdateien über die Nutzer der Plattform. Zu den Web-Usage Daten gehören Daten aus dem Web-Server, Proxy-Server Logs, Browser Logs, Benutzer Profile, Benutzer Sessions, Transaktionen, Cookies, Mausklicks und mehr [22]. Anschließend können die gesammelten Daten in Data-Mining Prozesse in Ergebnisse und Wissen umgewandelt werden. Diese Erkenntnisse fließen in die Gesamtmarketingstrategie ein, vor allem für die Verbesserung des Webauftritts und Angebote.

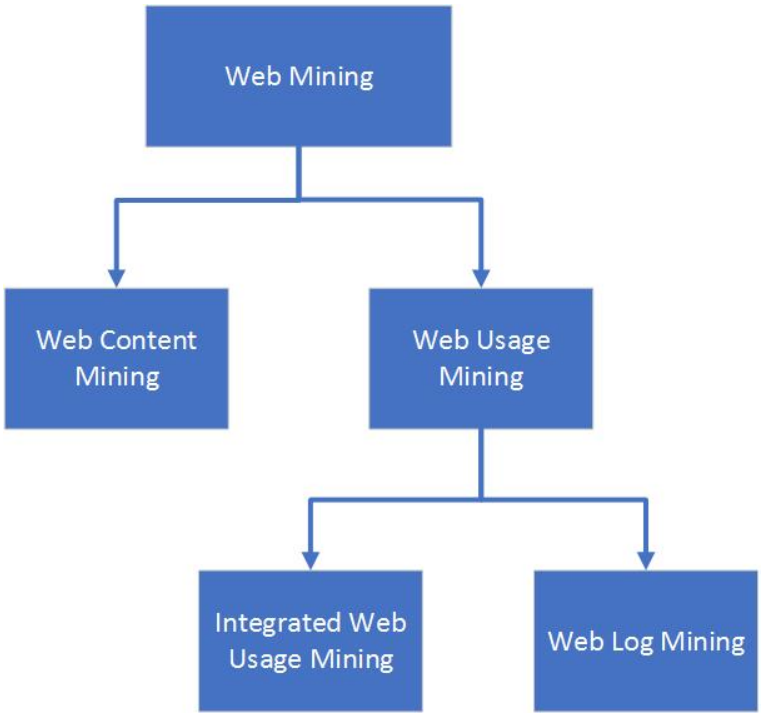


Abbildung 4.3: Web Mining Übersicht (angelehnt an Cleve et al. [19])

4.3.2 Verfahren

In Kapitel 3 werden verschiedene Konzepte für Recommender Systeme präsentiert. Diese Konzepte bedienen sich wie erläutert verschiedener Verfahren aus dem Data-Mining. Das ausgewählte Konzept, Collaborative-Filtering, ist dem Aufgabenbereich der Klassifikation zugeordnet. Beim Collaborative-Filtering spielen Ähnlichkeiten eine große Rolle. Das Item-based Verfahren ermittelt Empfehlungen auf Basis von Ähnlichkeiten von Objekten und dem Bewertungsmuster der Nutzer. Das User-based Verfahren ermittelt Empfehlungen auf Basis von ähnliche Nutzer. Für beide Verfahren werden also Ähnlichkeiten gesucht. Da es keine bereits vorhandenen Klassen gibt und eine Cluster-Analyse notwendig ist, gehört dieses Szenario zu dem Aufgabenbereich der Segmentierung. Für die Cluster-Analyse kommen unter anderem Algorithmen wie k-Means und k-Medoid infrage. Bei der Cluster-Analyse ist die Grundannahme, dass ähnliche Objekte sich durch einen geringeren Abstand als unähnlich auszeichnen [19]. Cluster-Analysen lassen sich in 6 Unterkategorien einteilen:

- Partitionierende Clusterbildung

- Hierarchische Clusterbildung
- Dichtebasierte Clusterbildung
- Clusterbildung mit Neuronalen Netzen
- Fuzzy und Graph-basierte Clusterbildung

Für den Gegenstand dieser Arbeit wird die partitionierende Clusterbildung näher analysiert und dargestellt. Weitere Informationen zu diesem Thema in dem Werk von J. Han et al. [29].

Partitionierende Clusterbildung Clusterbildung Das Ziel der Partitionierende Clusterbildung ist es eine Menge von Objekten in k Cluster zu zerteilen. Diese Zerteilung erfolgt durch die Auswahl k für die Anfangspartionierung. Repräsentiert kann es durch den Medoid oder Centroid werden.

Centroid Ein Cluster lässt sich als Vektor des Mittelwerts der Attributwerte seiner Objekte darstellen. Der Centroid kann eine künstlich geschaffene Instanz sein und muss nicht in der Datenmenge vorkommen [19].

Medoid Der Medoid ist ein Element aus der Datenmenge. Dafür kann z. B. ein Datensatz, welches am nächsten zum Centroid liegt, gewählt werden [19].

k-Means-Algorithmus In der Grundvariante legt das k-Means Verfahren die Zentren der Cluster fest und ändert es im Verlauf des Verfahrens iterativ. Die Anzahl der gesuchten Cluster werden vorgegeben, wobei ein Cluster durch die Centroide repräsentiert wird. Der k-Means-Algorithmus hat mehrere Schritte. Am Anfang werden initiale Cluster generiert. Anschließend werden die Centroide berechnet. In Schritt 3 erfolgt eine neue Zuordnung zu den Clustern. Die Zuordnung wird anhand der Berechnung der Abstände zu den Centroiden berechnet, wobei jedes Centroid ein eigenes Cluster repräsentiert. Hierbei kann es dann zu einer Verschiebung der Centroide kommen. In jedem Durchlauf findet eine Neuzuordnung der Knoten zu den Clustern und eine Neuberechnung der Centroide statt. Dieses Vorgehen wird so lange iterativ fortgesetzt, bis kein Punkt mehr seine Cluster wechselt [19].

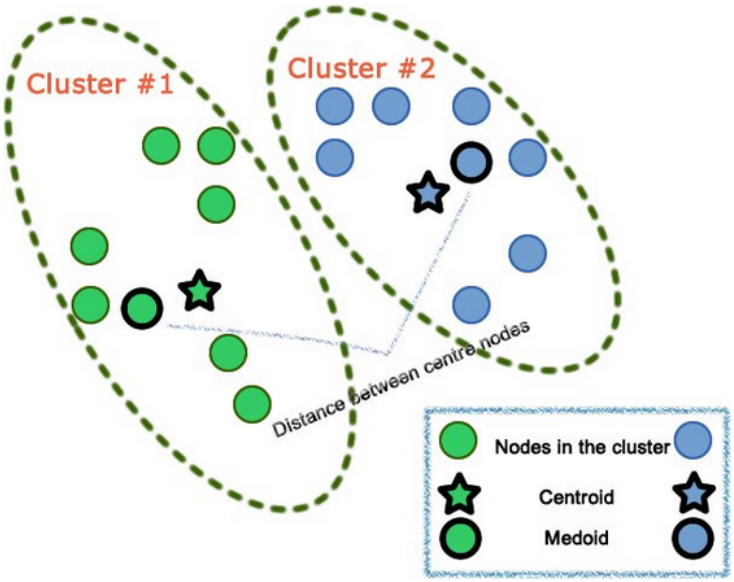


Abbildung 4.4: Centroid und Medoid [18]

Vorteile

- Erfahrungen belegen, dass Iterationen vergleichsweise klein sind
- Sehr anschaulich
- Einfach zu implementieren, Verfahren besteht aus Abstandsberechnungen und Neuzuordnung

Nachteile

- Qualität der initialen Zerlegung hat einen großen Einfluss auf die Ergebnisse
- Alle Objekte fließen in die Berechnung der Centroiden ein, Ausreißer kann Ergebnis beeinflussen
- Hoher Aufwand, da in jeder Iteration Neuberechnung der Centroiden erfolgt
- Ordinale und nominale Daten müssen in nummerische Werte umgewandelt werden für den Algorithmus

k-Medoid-Algorithmus Das k-Medoid-Verfahren ist dem k-Means-Verfahren sehr ähnlich. Der wesentliche Unterschied liegt in der Repräsentation eines Clusters. Während im k-Means der Centroid ein Cluster repräsentiert, ist es beim k-Medoid der Medoid, welches ein Element der Eingabedatenmenge ist. Als Auswahlstrategie für den Medoid kann das Element genommen werden, welches als Nächstes zum Centroid liegt. Zu Problemen kommt es, wenn die Datensätze nominal oder ordinal sind. So ist eine Berechnung des Centroids nicht mehr möglich. Dafür werden bessere Medoide durch Tauschen gesucht. Eines der ersten Algorithmen war Partioning Around Medoids kurz PAM. Im ersten Schritt werden k Objekte als Clusterrepräsentanten ausgewählt, nämlich die Medoide. Im zweiten Schritt werden die Objekte dem nächsten Medoid zugeordnet. Anschließend werden für jeden Medoid und Nicht-Medoid die Rollen vertauscht. In Schritt vier werden für jede Vertauschung die Distanzen ermittelt. In Schritt 5 werden die neue Medoide aus der Berechnung zuvor mit der kleinsten Distanz bzw. Unähnlichkeit ausgewählt. Zum Schluss werden die Schritte 2-5 solange wiederholt, bis die Medoide sich nicht mehr verändern [31].

4.4 Datenvorbereitung

Datensatz und Selektion Im Folgenden werden die Daten dargestellt, welche für den KDD Prozess benutzt werden. Es handelt sich hierbei um Daten aus einem Zeitraum von drei Monaten. Die Informationen um welche Monate es sich handelt, sind nicht gegeben. Die Daten basieren auf Cookies aus einer Online Nachrichtenplattform.

Cookies Cookies können als Ergänzung zu HTTP verstanden werden. Sie sind eine Textinformation, die erzeugt wird, wenn ein Client auf eine Webseite zugreift. Cookies können auf dem Browser des Clienten abgespeichert werden. Für eine erneute Benutzung einer Webseite können die Cookies vom Webserver direkt ausgelesen werden oder über ein Skript an dem Server übermittelt werden. Die Aufgaben bestehen unter anderem darin den Client zu identifizieren, Warenkörbe und Anmeldungen abzuspeichern und mehr. Das Thema kann in dem Artikel Kristol et al. vertieft werden [34].

Die Annahme liegt sehr nahe, dass es sich um Session Cookies handelt. Session Cookies oder In-Memory-Cookies, existieren nur temporär, während der Benutzer sich auf der Webseite befindet. Meistens löschen die Browser diese Cookies, wenn der Browser geschlossen wird.

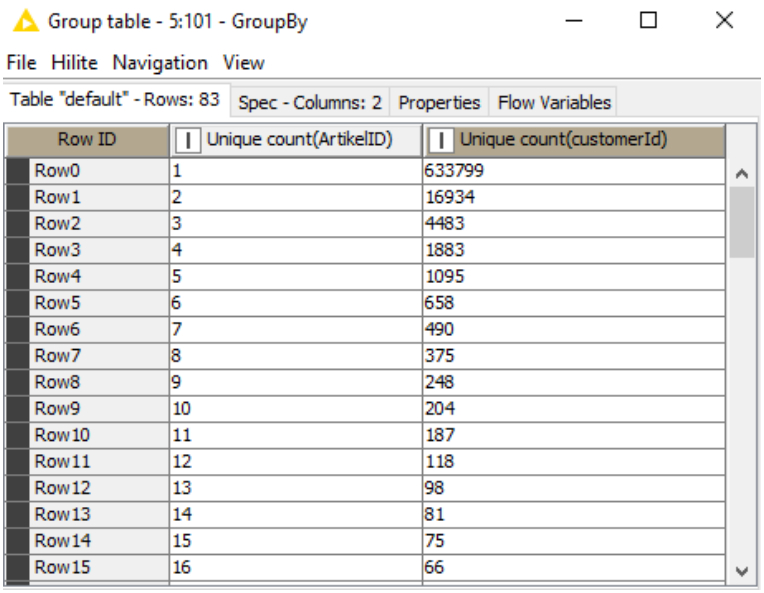
Politik	Unternehmen	Technologien	Finanzen
1. Deutschland	1. Industrie	1. IT + Telekommunikation	1. Börsenkurse
2. Konjunktur	2. Energie	2. Gadgets	2. Märkte (hat U)

Abbildung 4.5: Auszug Kategorien der Nachrichtenplattform

Datenbeschreibung Die Daten wurden von einer IT-Dienstleistungsfirma in dem Dateityp *table* zur Verfügung gestellt. Die Rohdatei ist 111 MB groß. Sie enthält 80 Spalten und 1 Mio. Zeilen. Alle Daten sind in dem Datentyp *String*. Jede Zeile repräsentiert eine Session Cookie, in dem ein Nutzer auf die Online Nachrichtenplattform einen Artikel aufgerufen hat. Jeder Artikelaufruf ist in einer Zeile repräsentiert. Die Cookies für diese Arbeit sind in einer Festplatte zur Verfügung gestellt wurden. Der vollständige Rohdatensatz und alle weiteren darauffolgenden bearbeiteten Datensätze befinden sich im Anhang. Aus den Cookies sind diese relevanten Informationen extrahiert worden und auf diese Dimensionen reduziert.

- FirstLevelCategory
- SecondLevelCategory
- ArtikelID
- CustomerID
- Device

Das sind die Informationen, die interpretierbar sind und für den weiteren Verlauf der Arbeit in Betracht gezogen werden. Die restlichen Spalten waren Duplikate der Informationen in einer anderen Form und nicht identifizierbare Informationen. In Abbildung 4.5 werden die Kategorien der Nachrichtenplattform dargestellt. Im Laufe dieser Arbeit wurden Änderungen von der Plattform vorgenommen, wie die Kategorie $Meine-News$. Diese Änderungen werden in dieser Arbeit nicht mehr berücksichtigt. Der Content der Seite ist in Kategorien aufgeteilt. Es gibt Hauptkategorien mit mindestens eine Unterkategorie. Unterkategorien haben 0 oder mehr Unterkategorien (in der Abbildung 4.5 mit *hat U* gek.). Teilweise ist der Content mehreren Unterkategorien zugeordnet, weil dieser wahrscheinlich nicht klar differenzierbar ist. Für diese Arbeit werden die Kategorien bis zur ersten Ebene berücksichtigt, weil es ab der zweiten Ebene wenig Content in Relation zum Gesamtkorpus gibt.



Row ID	Unique count(ArtikelID)	Unique count(customerId)
Row0	1	633799
Row1	2	16934
Row2	3	4483
Row3	4	1883
Row4	5	1095
Row5	6	658
Row6	7	490
Row7	8	375
Row8	9	248
Row9	10	204
Row10	11	187
Row11	12	118
Row12	13	98
Row13	14	81
Row14	15	75
Row15	16	66

Abbildung 4.6: Anzahl wie oft die Anzahl der gelesenen Artikel vorkommt

Tools für den KDD Prozess Für den KDD Prozess wird größtenteils KNIME benutzt. KNIME ist eine Opensource Data-Mining Software und wurde von der Universität Konstanz entwickelt. Es ist kompatibel mit allen Betriebssystemen [19]. Für einige Schritte wird auch Excel und Java benutzt.

Datensäuberung Die Daten wurden auf fehlende, verrauschte, falsche und inkonsistente Daten untersucht. Es wurden 133.094 (13,09 %) Nullwerte in der $FirstLevelCategory$ gefunden. Diese Zeilen wurden entfernt, da die Artikel nicht einer Kategorie zugeordnet werden konnten. Die Zeilen, die Nullwerte in der $SecondLevelCategory$ aufweisen, aber in $FirstLevelCategory$ einen Wert haben, werden behalten, da eine Zuweisung stattfinden kann. Es gibt keine verrauschten, inkonsistenten und falschen Daten und Ausreißer. Für eine aussagekräftige Empfehlung sind genügend Daten notwendig. Es gibt viele verschiedene Herausforderungen für Collaborative-Filtering, wie in Kapitel 3 beschrieben u. a. das Cold-Start Problem. Wie in Abbildung 4.6 zu sehen ist, haben 633.799 (63,37 %) nur 1 gelesen und 16.934 (1,69 %) haben 2 Artikel gelesen. Diese Nutzer werden entfernt, um einen Anhaltspunkt für die Empfehlung zu haben. Somit wurden alle Benutzer mit mehr als 2 gelesene Artikel behalten. Nach diesem Datensäuberungsprozess sind 73.778 (7,37 %) Zeilen übrig. Die Nullwerte in der Spalte *Device* werden durch den String *Unknown* ersetzt und bleiben in dem Datensatz bestehen.

Filtered - 4:95 - Row Filter

File Hilite Navigation View

Table "default" - Rows: 73778 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	customerId	ArtikelID	device	FirstLevel	Second...	Unique count(ArtikelID)
Row33_Row1...	39ef81dd-5a07-4714-b836-e026d3bdf460	22976952	desktop	politik	deutschland	15
Row105_Row...	2ef3662f-7e17-4717-bdb8-42a2eaec22a3	22975972	mobile	politik	deutschland	25
Row534_Row...	559deded-a105-4b8b-b3f0-4c3dcc86d8d1	22977246	desktop	video	unternehmen	15

Abbildung 4.7: Daten nach der Säuberung

Group table - 5:101 - GroupBy

File Hilite Navigation View

Table "default" - Rows: 83 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Unique count(ArtikelID)	Unique count(customerId)
Row0	1	633799
Row1	2	16934
Row2	3	4483
Row3	4	1883
Row4	5	1095
Row5	6	658
Row6	7	490
Row7	8	375
Row8	9	248
Row9	10	204
Row10	11	187
Row11	12	118
Row12	13	98
Row13	14	81
Row14	15	75
Row15	16	66

Abbildung 4.8: Daten nach dem Transformationsprozess

Datentransformation Jede Zeile des jetzigen Datensatzes stellt einen Artikelaufruf durch einen Nutzer dar. Der Nutzer kann durch die $CustomerID$ identifiziert werden. Um zu erkennen, welche Artikel ein Nutzer gelesen hat, wird die bisherige Tabelle, wie in Abbildung 4.8 zu sehen, transformiert. Die neuen Spalten der neuen Tabelle enthalten zusätzlich alle Kategorien. Die Werte stellen die Anzahl der gelesenen Artikel in der jeweiligen Kategorie dar. Somit gibt es insgesamt 10.467 Benutzer.

In der Boxplotanalyse wird ersichtlich, dass die Anzahl der gelesenen Artikel stark variiert. Hier werden der aktuellen Tabelle neue Features hinzugefügt. Alle Benutzer, die weniger als 4 Artikel gelesen haben werden als $WenigLeser$ dargestellt durch die Zahl 1, $NormalLeser$ sind alle zwischen 4 und 17 und $VielLeser$ alle über 17. So wurde eine neue Spalte *LeseAktivit"ät* zu dem Profil der Leser hinzugefügt. Für neue Featureextraction aus den Daten wird der Gesamtkorpus analysiert.

Robust Statistics - 4:60 - Box Plot (local)

File Hilite Navigation View

Table "default" - Rows: 7 | Spec - Column: 1 | Prop

Row ID	D Unique ...
Minimum	3
Smallest	3
Lower Quartile	4
Median	7
Upper Quartile	17
Largest	36
Maximum	516

Abbildung 4.9: Boxplot Anzahl gelesener Artikel pro User

Occurrences Table - 4:3 - Statistics (2)

File Hilite Navigation View

Table "default" - Rows: 90 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	S FirstLevel	I Count...	D Relativ...	S Second...	I Count ...	D Relativ...
Row0	politik	25467	0.345	international	12691	0.172
Row1	unternehmen	18105	0.245	deutschland	11787	0.16
Row2	finanzen	15758	0.214	industrie	5427	0.074

Abbildung 4.10: Statistiken Artikelkorpus

In dem gesamten Korpus gibt es 6.184 Artikel. Wie in Abbildung 4.10 zu erkennen ist, sind die Top 3 der *FirstLevelCategory* Politik, Unternehmen und Finanzen. Die Top 3 der *SecondLevelCategory* sind International, Deutschland und Märkte. Als zusätzliches Feature wurde zu den Profilen der Benutzer die relative Häufigkeit der gelesenen Artikel in Bezug auf die Anzahl der Artikel in der jeweiligen Kategorie ergänzt. Die transformierte Tabelle besteht somit aus 111 Spalten und 10.467 Zeilen. Die Werte aller Spalten sind numerisch. Jede Zeile repräsentiert einen Benutzer und die Spalten stellen die Attribute der Benutzer dar.

4.5 Data-Mining und Modell

Distanzfunktionen Bevor die Algorithmen angewendet werden können, müssen die Quelldaten in eine bestimmte Form vorliegen. Bei der Bildung von Cluster, werden die Ähnlichkeiten von zwei Datensätze quantifiziert. Dies wird meistens durch Distanzfunktionen realisiert [19]. Im Folgenden werden einige bekannte Distanzfunktionen genannt.

Euklidische Distanz Die euklidische Distanz ist die räumliche Distanz zwischen zwei Punkte im n-dimensionalen Raum. Zwei Vektoren werden als unähnlich betrachtet, je kleiner deren euklidische Distanz ist. Sie ist für metrische Daten geeignet [19].

Cosinus Ähnlichkeitsmaß Die Ähnlichkeit zwischen zwei Vektoren kann als cos(a,b) ausgedrückt werden. Hierbei wird der Cosinus des Winkels zwischen beiden Vektoren bestimmt. Dadurch kann bestimmt werden, ob zwei Vektoren in gleiche Richtung zeigen.

Hamming-Distanz Die Hamming-Distanz zählt, an wie vielen Positionen die Datensätze sich unterscheiden. Als Beispiel ist der Abstand der Datensätze {Gustav, 1993, Hamburg} und {Ali, 1993, Hamburg} gleich 1, da sich die Datensätze an einer Stelle unterscheiden. Die Hamming-Distanz ist somit auf alle Datentypen anwendbar.

Modellerstellung Die Werte der vorliegenden Daten sind numerisch, daher eignet sich wie zuvor beschrieben für den Ansatz des Collaborative-Filtering der k-means-Algorithmus. Zudem ist der k-means-Algorithmus für partitionierte Clusterverfahren weit verbreitet. Die gefundenen Cluster stellen die Grundlage für das Collaborative-Filtering

Labeled input - 7:100 - k-Means

File Hilite Navigation View

Table "default" - Rows: 9943 | Spec - Columns: 113 | Properties | Flow Variables

Row ID	customerId	Finanze...	Technik...	LeseAk...	Cluster
Row0	000648f1-55e7-43ee-b247-7efe0b60fcca	0	0	1	cluster_6
Row1	000893bf-4edf-4741-b996-3ff05872acb0	0	0	1	cluster_6
Row2	0009ffb2-0b13-4aad-8a7c-c8ca703dc745	0.001	0	2	cluster_8
Row3	0014239f-999b-46e2-9c85-cf885c9701a7	0.001	0	2	cluster_6
Row4	00146052-0b29-4547-ac64-13499724190c	0.004	0	3	cluster_0
Row5	0016e461-6812-4060-8732-0c75c855ef9d	0.001	0	2	cluster_0
Row6	00170ef4-4643-45cc-a09a-6d89b2f71833	0.001	0	2	cluster_6
Row7	0017597a-a0fe-4051-8c75-669ae2b5f483	0	0	1	cluster_6

Abbildung 4.11: k-means-Algorithmus Ergebnisse

dar. Eine bekannte und etablierte Distanzfunktion als Vorbereitung für den k-means-Algorithmus ist die euklidische Distanzfunktion. Es wird die euklidische Distanz errechnet und anschließend der k-means-Algorithmus angewendet mit der Auswahl für k = 9, angelehnt an die 9 First-Level Kategorien. Die Iterationen worden variiert. Ab 30 Iterationen haben sich die Ergebnisse nicht mehr verändert. Jedes Cluster repräsentiert ein bestimmtes Verhaltensmuster der Nutzer. Als Ergebnis wird eine neue Spalte *Cluster* zu jedem Profil hinzugefügt, die aussagt, zu welchem Cluster dieses Profil gehört (siehe Abbildung 4.11).

Empfehlung Die Ergebnisse des k-means-Algorithmus zeigen die Zugehörigkeit eines Profils zum jeweiligen Cluster. Im Korpus der Cluster sind viele verschiedene Artikel zu finden, auch mit Überschneidungen zu anderen Clustern. Die Frage stellt sich hier welche Artikel genau zu empfehlen sind. Dafür wird eine Reduktion der jeweiligen Cluster vorgenommen. In diesen Cluster werden, die am häufigsten *Top*10 gelesenen Artikel selektiert. Anschließend wird überprüft, welche Artikel die Nutzer gelesen haben und welche nicht. Die Artikel, die nicht gelesen wurden sind, werden dem Nutzer empfohlen. Somit basiert die Empfehlung auf der Basis der jeweiligen Profile und ihrer Zugehörigkeit zu einem bestimmten Clusterprofil, welches aus ihrem Leseverhalten generiert wurde.

5 Evaluation und Fazit

In diesem Kapitel werden am Anfang die Datenqualität und das Modell insgesamt ausgewertet. Anschließend wird ein Gesamtfazit gezogen.

5.1 Evaluation

Datenqualität Die Daten für diese Arbeit basieren hauptsächlich aus Cookies. Aus den Cookies konnten einige Informationen entnommen werden, jedoch basiert hier die Annahme, dass eine CustomerID auch einem Profil zugeordnet ist. Da die Cookies meistens automatisch beim Schließen des Browsers gelöscht werden, werden dem gleichen Nutzer beim erneuten Aufruf der Seite eine neue CustomerID zugeordnet. Für den prototypischen Ansatz hat diese Annahme zwar ausgereicht, aber für ein produktives System ist es besser zwischen Nutzer, die einen Account angelegt haben und die ohne einen Account, zu unterscheiden. Die vorhandenen Cookies weisen im Ganzen 133.094 (13,30 %) Nullwerte auf. Dies hängt damit zusammen, dass aus dem jeweiligen Cookie die Kategorien nicht ermittelt werden konnten. Der Datenverarbeitungsprozess kann erheblich vereinfacht werden, wenn ein Zugriff auf die Datenbank der Nachrichtenplattform besteht und so sich die Möglichkeit ergibt mit strukturierten Daten zu arbeiten.

Bewertung Modell und Ergebnis Die Bewertung von Cluster-Verfahren lässt sich als allgemein schwierig darstellen. Einfacher ist es, wenn Beispiele schon gegeben sind und die Untersuchung darin besteht, wie ähnlich die Cluster zu den bereits gegebenen Clustern sind. Bei einem Clusterverfahren sind jedoch meistens die Zielcluster nicht bekannt. Das Modell kann bewertet werden, wenn das erwartete Ergebnis zum Vergleich vorliegt [19]. Im Rahmen dieser Arbeit wurde durch den KDD Prozess und Nutzung der jeweiligen Methoden Empfehlungen generiert. Das Modell kann z. B. durch die *Genauigkeit* ausgewertet werden. Diese Berechnung wird in vielen verschiedenen KDD Prozesse verwendet u. a. auch bei Burke et al. [16, 36]. Hier wird dann der relative Anteil der falsch

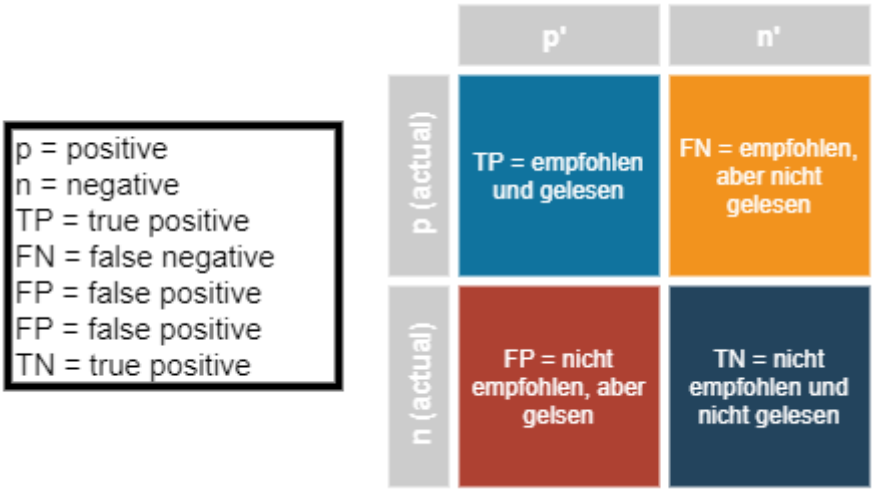


Abbildung 5.1: F1 Score Matrix Artikel Recommendation

klassifizierten Daten zu der gesamten Zuordnung berechnet. Voraussetzung für diese Methode ist jedoch, dass das zu erwartete Ergebnis zum Vergleich vorliegt. Im Falle einer Empfehlung kann beobachtet werden, ob ein Nutzer das Produkt kauft. In diesem Fall ist zu beobachten, ob ein Nutzer den empfohlenen Artikel auf der Webseite liest bzw. anklickt.

$$accuracy = \frac{\text{number of successful recommendations}}{\text{number of recommendations}} \tag{5.1}$$

Für die Bewertung kann auch der $F1$ Score ermittelt werden.

$$F1 = 2x\frac{\text{precesion x recall}}{\text{precesion} + \text{recall}} \tag{5.2}$$

Hier kann es dann eine Gegenüberstellung geben, wie in Abbildung 5.1 zu sehen ist. Da diese Daten in diesem System nicht gegeben sind, kann die Genauigkeit nicht berechnet werden. Hierfür müsste das Modell auf die Produktivplattform implementiert werden, um herauszufinden inwieweit, die generierten Empfehlung Akzeptanz bei den Nutzern finden. Im Rahmen dieser Arbeit wurden die Timestamps nicht berücksichtigt. Aus den Timestamps kann die Aktualität eines Artikels erkannt werden. Schließlich sind aktuelle Nachrichten meistens relevant. Zudem kann auch die Empfehlungsstruktur der Top 10 Artikel verbessert werden. Unter den Top 10 könnte es ein Ranking geben, das zeigt, welcher Artikel innerhalb der Top 10 am häufigsten gelesen wurde.

5.2 Fazit

In dieser Arbeit wurde ein Einblick auf die bekanntesten Konzepte und Verfahren aus den Themengebieten der Recommender Systeme und dem Data-Mining gegeben. Zudem wurde ein prototypischer Ansatz eines Recommender Systems mit dem Collaborative-Filtering Verfahren auf Grundlage von Cookies für eine digitale Nachrichtenplattform entwickelt. Mit den Cookies, die für diese Arbeit benutzt wurden, konnten durch den Collaborative-Filtering Ansatz Empfehlungen generiert werden. Hierbei wurden die ähnlichen Nutzer durch Clusterverfahren identifiziert. Die Profile der Nutzer wurden anhand der Informationen über Artikel, die sie gelesen haben, erstellt. Anschließend wurden den Nutzern Empfehlungen aus ihrem Cluster empfohlen. Hier bleibt die Auswertung aus, da das entwickelte Modell in dem produktiven Umfeld beobachtet werden muss. Zudem kann gesagt werden, dass die Datenquelle durch die Nullwerte wenig Information hergibt, die für die Profilerstellung der Nutzer relevant ist. An dem Modell und der Datenquelle können noch viele Optimierungen vorgenommen werden, wie zuvor beschrieben. Aus der Untersuchung geht hervor, dass ein Collaborative-Filtering Verfahren für Online Nachrichtenplattformen geeignet ist und schon mit wenig Nutzerinformationen Empfehlungen generiert werden können. Web-Mining Data-Mining in Bezug zu E-Commerce bleiben weiterhin spannende Themen, welche ihre Potenziale noch nicht vollends ausgeschöpft haben. Von der Auswahl der Ansätze, bis hin zum Analysieren der Nutzerverhalten und Algorithmen werden immer wieder neue Ansätze veröffentlicht. Zudem investieren Unternehmen weiterhin in diese Themengebiete.

5.3 Generalisierung

Fast jede Zeitschrift in Deutschland hat mittlerweile einen Online Nachrichtenportal, hinzu kommen reine online Nachrichtenportale wie www.ka-news.de. Das Thema der Personalisierung spielt in solchen Portalen eine immer wichtigere Rolle. Nachrichten können immer Kategorien zugeschrieben werden und diese Kategorien lassen sich auf Newsportale auch finden. Die Cookies können relativ einfach implementiert werden, um das erstellte Modell auch auf die Plattformen anzuwenden. Jedoch bedarf es, wie in den Abschnitten zuvor beschrieben, einiger Optimierungsprozesse. Empfehlungen nur auf Grundlage von Cookies mit den Informationen, die in dieser Arbeit aus den Cookies extrahiert wurden, sind möglich, aber es ist zu überprüfen, wie die Güte dieser Empfehlungen ist.

6 Ausblick

Optimierung der Datenquelle In dieser Arbeit wurde auf der Grundlage von Cookies gearbeitet. Ziel war es auf der Grundlage dieser Daten eine Empfehlung zu generieren. Da es keinen Ratingsystem für Artikel auf der Plattform gibt, ist die Annahme, dass ein Artikel interessant ist oder nicht, durch einen Klick auf den Artikel festzustellen. Es wäre daher nützlich zu messen, wie lange die Verweildauer auf einen Artikel ist, um der Wahrheit näher zu sein, dass ein aufgerufener Artikel auch gelesen wurde. Das hätte sehr wahrscheinlich eine Auswirkung auf die Profilerstellung. Natürlich kann auch ein Ratingsystem eingeführt werden. Um dem Cold-Start Problem entgegenzuwirken, können auf der Webseite Interessen hinterlegt werden. Diese Daten schärfen nochmal die Profilerstellung für das Collaborative-Filtering. Zudem kann überprüft werden, ob der Device eine Auswirkung darauf hat, ob kurze oder lange Artikel gelesen werden, um dementsprechend Maßnahmen zu treffen z. B. beim Nutzen des Smartphones, kurze Artikel wie Ticker zu empfehlen. Als Kategorien über das sich das Interessenprofil der Nutzer definiert, wurden hier die vorgegebenen Kategorien der Webseite benutzt. Es könnte interessant werden mittels natural language processing Methoden, den Content der Seite zu analysieren und statt z. B. Kategorie Politik und Deutschland, speziell Angela Merkel als Kategorie bzw. Interesse zu einem Nutzerprofil zuzuordnen. Diese und ähnliche Optimierungen können in weiteren Arbeiten untersucht werden, inwiefern sie ein Profil eines Nutzers schärfen und ob sich das positiv auf die Empfehlungen auswirkt.

Optimierung des Modells In dem Kontext der Nachrichten spielt die Aktualität eine große Rolle. Artikel bzw. Nachrichten, haben die Eigenschaft, dass sie eine aktuelle neue Nachricht vermitteln. Daher sollte das Verfahren für die Empfehlung nicht alle Artikel gleich behandeln, sondern den aktuellen Nachrichten aus den Top 10 eine höhere Gewichtung beimessen. Anders wäre es bei Büchern oder ähnliche Produkte, wo die Aktualität nicht das Hauptkriterium ist und sie somit eher zeitlos sind. So ist es auch mit Breaking News, welche meistens viele Nutzer interessiert, auch wenn es nicht unbedingt dem Interessenprofil entspricht. Zudem sind saisonabhängige Ereignisse, wie die Wahlen

in Deutschland und Fußball WM, Faktoren, die in der Modellerstellung berücksichtigt werden können. Diese Ansätze können in weiteren Arbeiten untersucht werden. Dabei kann analysiert werden, inwiefern diese Gewichtungen in den Algorithmen eine Rolle spielen.

Nutzerverhalten Es ist weiterhin zu untersuchen, inwiefern sich Nutzerverhalten im Laufe der Zeit ändern. Es kann sein, dass Nutzer im Laufe der Zeit sich für andere Themenbereiche interessieren. Hier kann eine Idee sein, dass aktuelle Interessensbereiche mehr auf die Profilerstellung durch höhere Gewichtung einwirken. Zudem kann es zu einem schizophrenen Nutzerverhalten kommen. Deshalb ist es weiterhin zu analysieren, ob mehrere Nutzer gleichzeitig ein Account benutzen, oder der Nutzer sich gerade im Urlaub befindet. Es gibt weitere Trends in Richtung Web-Log Analyse, die speziell die Log Files auswerten und Analysieren, um Erkenntnisse zu gewinnen, von welcher Seite ein Nutzer den Aufruf gestartet hat und mittels der IP-Adresse Nutzer zu identifizieren. Dazu zählen u. a. die Analyse der Verweildauer und das Klickverhalten, die mit in dem Nutzerverhalten berücksichtigt werden können. Solch ein Verhalten kann die Empfehlung stark beeinflussen.

Wie in Kapitel 7 beschrieben bekommt das Web-Log Mining immer mehr einen hohen Stellenwert. Aus den Logfiles können unter anderem folgende Information extrahiert werden:

- User-name: Identifikation des Nutzers durch IP-Adresse oder die Anmeldung durch einen Nutzerprofil.
- Visiting-path: Durch welchen Pfad die Webseite aufgerufen wurde, direkt durch einen Hyperlink oder einer Suchmaschine.
- Path-raversed: Pfade, die innerhalb einer Webseite benutzt wurden, um einen Link aufzurufen.
- Timestamp: Die Verweildauer, die ein Nutzer auf einer Webseite einnimmt.

  Aus diesen Daten können verschiedene Informationen durch Mustererkennung in der Pfadanalyse z. B. durch Assoziationsanalyse und weitere Data-Mining Verfahren, generiert werden. Zudem liefert Web-Usage Mining noch weitere Erkenntnisse über Nutzerverhalten die bei LK Grace et al. [28] vertieft werden können.

Es ist weiterhin für eine weiterführende Arbeit interessant zu überprüfen, welchen Mehrwert genau solche Verfahren einer Nachrichtenplattform bringen.

Optimierung des Verfahrens Yao D. et al. [20] haben im Rahmen ihrer Forschung untersucht, welches Verfahren bessere Ergebnisse erzielt bei News Recommendation und sind zu dem Entschluss gekommen, dass ein Hybrides Collaborative-Filtering Verfahren aus Item-based und User-based, deutlich zur Optimierung beiträgt. Sie haben auch einige Optimierung an den Distanzfunktionen vorgenommen. Zudem haben J. Liu et al. [35] für einen der erfolgreichsten News Recommendation Plattformen Google News einen Hybriden Ansatz aus Wissensbasiertes Filtern und Collaborative-Filtering entwickelt. Diese Ansätze können in weiteren Arbeiten mitberücksichtigt werden.

Datenschutz als Herausforderung Für die Erstellung der Nutzerprofile und dem Clustern ist wichtig aus den gesammelten Daten Interessen abzuleiten. Hierfür werden von vielen Plattformen, viele Daten über Nutzer gesammelt. Manchmal bekommt das der Nutzer nicht mit. Es gibt allgemeine Richtlinien, die diese Sammlungen regulieren sollen, welche auch in der neuen DGSVO, Datenschutz Grundverordnung, ausgeführt werden. Es ist weiterhin zu überprüfen welche Daten mittels Cookies überhaupt und inwieweit gesammelt werden dürfen.

Literaturverzeichnis

[1] *90 Prozent der Bevölkerung in Deutschland sind online.* https://www.destatis.de/DE/Presse/Pressemitteilungen/2018/09/PD18_330_634.html;jsessionid=5AD112E620963E8438E913DB45D39606.InternetLive2. – Eingesehen am 14.01.2020

[2] *CRISP-DM Process Diagram.* https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png. – Accessed: 2019-11-20

[3] *Die Geschichte der Tageszeitungen.* https://www.deutsche-tageszeitungen.de/pressefachartikel/die-geschichte-der-tageszeitungen/. – Eingesehen am 11.09.2019

[4] *Einkauf Online.* https://www.destatis.de/DE/Presse/Pressemitteilungen/2018/11/PD18_427_634.html;jsessionid=5AD112E620963E8438E913DB45D39606.InternetLive2. – Eingesehen am 14.01.2020

[5] *Entwicklung der Anzahl der Online-Angebote der Zeitungen in Deutschland in den Jahren 1995 bis 2018.* https://de.statista.com/statistik/daten/studie/4191/umfrage/anzahl-der-online-angebote-von-zeitungen-seit-1995/. – Eingesehen am 11.09.2019

[6] *How retailers can keep up with consumers.* https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers. – Eingesehen am 11.09.2019

[7] *Marck Zuckerberg Zitat aus dem Film The Social Network.* https://www.faz.net/aktuell/wirtschaft/digitec/sean-parker-ueber-facebooks-nutzer-manipulation-15286051.html. – Eingesehen am 15.01.2020

[8] *Medium*. https://www.duden.de/rechtschreibung/Medium_Vermittler_Traeger. – Eingesehen am 30.09.2019

[9] *Nachricht*. https://www.duden.de/rechtschreibung/Nachricht. – Eingesehen am 20.09.2019

[10] *Netflixpreis*. https://www.netflixprize.com/. – Eingesehen am 11.09.2019

[11] Aamir, Mohammad ; Bhusry, Mamta: Recommendation system: state of the art approach. In: *International Journal of Computer Applications* 120 (2015), Nr. 12

[12] Adomavicius, Gediminas ; Tuzhilin, Alexander: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. In: *IEEE Transactions on Knowledge & Data Engineering* (2005), Nr. 6, S. 734–749

[13] Aehnelt, Mario: Personalisierung als Schlüssel zum Erfolg. In: *Multimedia & Bildung: Beiträge zu den* 4 (2003), S. 129–140

[14] Alpar, Paul ; Niedereichholz, Joachim: Data mining im praktischen Einsatz. In: *Braunschweig/Wiesbaden, Vieweg Verlagsgesellschaft* (2000)

[15] Burke, Robin: Integrating knowledge-based and collaborative-filtering recommender systems. In: *Proceedings of the Workshop on AI and Electronic Commerce*, 1999, S. 69–72

[16] Burke, Robin: Hybrid recommender systems: Survey and experiments. In: *User modeling and user-adapted interaction* 12 (2002), Nr. 4, S. 331–370

[17] Chapman, Peter ; Clinton, Janet ; Kerber, Randy ; Khabaza, Tom ; Reinartz, Thomas ; Shearer, C. Russell H. ; Wirth, Robert: CRISP-DM 1.0: Step-by-step data mining guide, 2000

[18] Cheong, Se-Hang ; Si, Yain-Whar: CWBound: boundary node detection algorithm for complex non-convex mobile ad hoc networks. In: *The Journal of Supercomputing* 74 (2018), Nr. 10, S. 5558–5577

[19] Cleve, Jürgen ; Lämmel, Uwe: *Data Mining. vol. 2.* Berlin: Walter de Gruyter GmbH, 2016

[20] Dong, Yao ; Liu, Shan ; Chai, Jianping: Research of hybrid collaborative filtering algorithm based on news recommendation. In: *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* IEEE (Veranst.), 2016, S. 898–902

[21] *Zeitungen in Deutschland.* https://de.statista.com/statistik/studie/id/6551/dokument/zeitungen-in-deutschland/. – URL https://de.statista.com/statistik/studie/id/6551/dokument/zeitungen-in-deutschland/. – Eingesehen am 13.01.2020

[22] Eirinaki, Magdalini ; Vazirgiannis, Michalis: Web Mining for Web Personalization. In: *ACM Trans. Internet Technol.* 3 (2003), Februar, Nr. 1, S. 1–27. – URL http://doi.acm.org/10.1145/643477.643478. – ISSN 1533-5399

[23] Fahlenbrach, Kathrin: *Medien, Geschichte und Wahrnehmung: Eine Einführung in die Mediengeschichte.* Springer-Verlag, 2018

[24] Fan, Yongjian ; Shen, Yanguang ; Mai, Jianying: Study of the Model of E-commerce Personalized Recommendation System Based on Data Mining. In: *2008 International Symposium on Electronic Commerce and Security* (2008), S. 647–651

[25] Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: The KDD process for extracting useful knowledge from volumes of data. In: *Communications of the ACM* 39 (1996), Nr. 11, S. 27–34

[26] Feldman, Ronen ; Ronen ; Sanger ; James: *The text mining handbook: Advanced approaches in analyzing unstructured data.* 01 2007

[27] *Meistbesuchte Seite der Welt.* https://www.alexa.com/topsites/. – URL https://www.alexa.com/topsites/. – Eingesehen am 14.01.2020

[28] Grace, LK ; Maheswari, V ; Nagamalai, Dhinaharan: Analysis of web logs and web user in web mining. In: *arXiv preprint arXiv:1101.5668* (2011)

[29] Han, Jiawei ; Pei, Jian ; Kamber, Micheline: *Data mining: concepts and techniques.* Elsevier, 2011

[30] Hohfeld, Stefanie ; Kwiatkowski, Melanie: Empfehlungssysteme aus informationswissenschaftlicher Sicht-State of the Art. In: *Information Wissenschaft und Praxis* 58 (2007), Nr. 5, S. 265

[31] Kaufman, Leonard ; Rousseeuw, Peter J.: *Finding groups in data: an introduction to cluster analysis.* Bd. 344. John Wiley & Sons, 2009

[32] Klahold, André: Empfehlungssysteme. In: *Vieweg+ Teubner, Wiesbaden* (2009)

[33] Konstan, Joseph A. ; Riedl, John: Recommender systems: from algorithms to user experience. In: *User modeling and user-adapted interaction* 22 (2012), Nr. 1-2, S. 101–123

[34] Kristol, David M.: HTTP Cookies: Standards, privacy, and politics. In: *ACM Transactions on Internet Technology (TOIT)* 1 (2001), Nr. 2, S. 151–198

[35] Liu, Jiahui ; Dolan, Peter ; Pedersen, Elin R.: Personalized news recommendation based on click behavior. In: *Proceedings of the 15th international conference on Intelligent user interfaces* ACM (Veranst.), 2010, S. 31–40

[36] Olmo, Félix Hernández del ; Gaudioso, Elena: Evaluation of recommender systems: A new approach. In: *Expert Syst. Appl.* 35 (2008), 10, S. 790–804

[37] *Paid Conten Umsätze der Publikumspresse.* https://www.editorial.media/2019/01/21/paid-content-boomt-400-millionen/. – URL https://www.editorial.media/2019/01/21/paid-content-boomt-400-millionen/. – Eingesehen am 13.01.2020

[38] Patel, Yagnesh G. ; Patel, Vishal P.: A survey on various techniques of recommendation system in web mining. In: *International Journal of Engineering Development and Research* 3 (2015), Nr. 4

[39] Poriya, Anil ; Bhagat, Tanvi ; Patel, Neev ; Sharma, Rekha: Non-personalized recommender systems and user-based collaborative recommender systems. In: *Int. J. Appl. Inf. Syst* 6 (2014), Nr. 9, S. 22–27

[40] Reichheld, Frederick F.: Loyalty-based management. In: *Harvard business review* 71 (1993), Nr. 2, S. 64–73

[41] Reichheld, Frederick F. ; Sasser, W E.: Zero defeofions: Quoliiy comes to services. In: *Harvard business review* 68 (1990), Nr. 5, S. 105–111

[42] Ricci, Francesco ; Rokach, Lior ; Shapira, Bracha: Introduction to recommender systems handbook. In: *Recommender systems handbook*. Springer, 2011, S. 1–35

[43] Rich, Elaine: User modeling via stereotypes. In: *Cognitive science* 3 (1979), Nr. 4, S. 329–354

[44] Schafer, J B. ; Konstan, Joseph A. ; Riedl, John: E-commerce recommendation applications. In: *Data mining and knowledge discovery* 5 (2001), Nr. 1-2, S. 115–153

[45] Schröder, Thomas: *Die ersten Zeitungen: Textgestaltung und Nachrichtenauswahl*. Gunter Narr Verlag, 1995

[46] Shapira, Bracha: *Recommender systems handbook*. Springer-verlag New York Incorporated, 2015

[47] *Umsätze der Zeitschriften*. https://de.statista.com/statistik/daten/studie/205874/umfrage/prognose-zum-umsatz-der-zeitschriftenverlage-in-deutschland/. – URL https://de.statista.com/statistik/daten/studie/205874/umfrage/prognose-zum-umsatz-der-zeitschriftenverlage-in-deutschland/. – Eingesehen am 13.01.2020

[48] *Umsätze der Zeitschriften*. https://de.statista.com/statistik/daten/studie/12551/umfrage/umsatzentwicklung-der-zeitschriften-seit-2003/. – URL https://de.statista.com/statistik/daten/studie/12551/umfrage/umsatzentwicklung-der-zeitschriften-seit-2003/. – Eingesehen am 30.09.2019

[49] *Werbemarkt in Deutschland Januar bis April 2018*. https://de.statista.com/statistik/daten/studie/189855/umfrage/marktanteile-der-mediengattungen-im-werbemarkt/. – URL https://de.statista.com/statistik/daten/studie/189855/umfrage/marktanteile-der-mediengattungen-im-werbemarkt/. – Eingesehen am 13.01.2020

[50] Zeitungsverleger, Bundesverband D.: Die deutschen Zeitungen in Zahlen und Daten 2018. In: *Berlin: Bund deutscher Zeitungsverleger* (2018)

A Anhang

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Gemäß der Allgemeinen Prüfungs- und Studienordnung ist zusammen mit der Abschlussarbeit eine schriftliche Erklärung abzugeben, in der der Studierende bestätigt, dass die Abschlussarbeit „– bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit [(§ 18 Abs. 1 APSO-TI-BM bzw. § 21 Abs. 1 APSO-INGI)] – ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt wurden. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich zu machen."

Quelle: § 16 Abs. 5 APSO-TI-BM bzw. § 15 Abs. 6 APSO-INGI

Erklärung zur selbstständigen Bearbeitung der Arbeit

Hiermit versichere ich,

Name: _____

Vorname: _____

dass ich die vorliegende Bachelorarbeit – bzw. bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit – mit dem Thema:

Konzept und Implementierung eines Nachrichten Recommender Systems

ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

_____	_____	_____
Ort	Datum	Unterschrift im Original