10 Die Vielfalt der Empfehlung

In diesem Kapitel untersuche ich, wie verschiedene Typen von Empfehlungssystemen eine Brücke zwischen Dingen und Nutzerinnen schlagen und damit den Kreislauf der Momente der Datafizierung zugleich vollenden und neu starten. An den Vergleichsbegriff von Bettina Heintz (2010; 2016) angelehnt, bezeichne ich diesen Brückenschlag zwischen Nachfrage und Angebot, zwischen Nutzerinnen und Dingen als »Relationierung«. Damit meine ich die technische Operation, in der Nutzerinnen und Dinge zueinander in Beziehung gesetzt werden, um im Sinne von »good matches« die Beziehung zwischen Nutzerinnen und Unternehmen aufrechtzuerhalten und weitere Anschlusshandlungen zu motivieren.

In Auseinandersetzung mit Joseph Konstans und Michael Ekstrands Kurs »Introduction to Recommender Systems« identifiziere ich vier Idealtypen von Empfehlungssystemen, die sich in ihren Quantifizierungs-, Kategorisierungs-, Bewertungs- und Vergleichsweisen unterscheiden.[1] Während ich in den Kapiteln 6 bis 8 die drei Momente der Datafizierung analytisch isoliert und in ihrer sequenziellen Logik dargestellt habe, fokussiere ich mich nun auf das dritte Moment: Wie relationieren die vier unterschiedlichen Typen von Empfehlungssystemen Nutzerinnen und Dinge, d.h. wie verdichten sie die Beziehungen zwischen den verschiedenen Einheiten, um Empfehlungen abzuleiten?

Popularitätsmetriken bringen Dinge in besser/schlechter Relationen. Solche Ranglisten sind in dem Sinn unpersonalisiert und »objektiv«, da allen Nutzerinnen die gleichen Dinge (z.B. die besten Filme aller Zeiten) angezeigt werden. Sie unterstellen, dass das Populäre für alle gleichermaßen relevant ist, da sie keine Differenzen zwischen Nutzerinnen beobachten (10.1).

1 Sich im Betrieb befindende Empfehlungssysteme sind um einiges komplexer als die von mir beschriebenen Varianten (siehe Seaver 2019 zur Komplexitätsfrage von Algorithmen). Laufende Systeme kombinieren verschiedene dieser Logiken und basieren auch auf neueren Ansätzen, deren Beschreibung ich hier nicht vornehmen kann.

Die *stereotypisierende Empfehlung* beruht auf askriptiven, demografischen Kategorien wie Alter oder Geschlecht, die über Konventionen oder gruppenspezifische Popularitätsmetriken Relationen zwischen Gruppen und Dingen herstellen. Solche Verfahren beruhen auf einer Form der kategorialen Relationierung, einer Art Deduktion: Männer mögen Actionfilme, Frauen mögen Chick-Flicks. Nutzerin X ist eine Frau. Ergo: Empfehle Chick-Flicks (10.2).

Inhaltsbasierte Empfehlungen relationieren Nutzerin und Dinge über Matchings, d.h. über den Vergleich von paarweisen Nutzerin/Ding-Relationen (anstatt nur Dinge oder nur Nutzerinnen miteinander zu vergleichen). Die ontologische Differenz zwischen Nutzerin und Ding wird dabei irrelevant: In Suchapplikationen ist die Nutzerin als Liste von (Such-)Begriffen repräsentiert, wie auch Dinge als Kombination von Begriffen repräsentiert sind (10.3).

Collaborative Filtering relationiert Nutzerinnen und Dinge über ein doppeltes Matching. Das Matching von ähnlichen Nutzerpaaren führt zu fluiden, quasi-kategorialen »Nachbarschaften« der ähnlichsten Nutzerinnen, welche die Grundlage für das zweite Matching bieten: Den Vergleich von Nutzerin/Ding-Relationen (10.4).

Obwohl der Fokus auf der Art und Weise der Relationierung liegt, ist es nicht möglich, das erste und zweite Moment der Datafizierung vollständig auszuklammern. Insbesondere die Frage der Herstellung von Vergleichbarkeit erweist sich in den Relationierungsverfahren immer wieder als problematisch, da sich idiosynkratische Vergleichseinheiten (seien es Kommentare wie in 10.1 oder Nutzerinnen wie in 10.4) trotz weitgehendem Encoding immer noch der Vergleichbarkeit entziehen. Die Formalisierung von Verhaltensweisen wie zum Beispiel die encodierten Bewertungsmodi »upvote« und »downvote« auf Reddit oder die Bewertung von Filmen auf einer Skala von 1 bis 5 reichen nicht aus, um Kommensurabilität von Kommentaren oder Nutzerinnen herzustellen, da sie in den Augen der Expertinnen zu verzerrten Resultaten führen. Die Herstellung von Vergleichbarkeit (das zweite Moment der Datafizierung) überschneidet sich also mit der Ableitung prospektiver Relationen (dem dritten Moment der Datafizierung) und ist in die algorithmischen Verfahren der Ableitung neuer Relationen eingefaltet.

Meine Untersuchung von Collaborative-Filtering-Empfehlungssystemen macht zudem deutlich, dass für die Datenerzeugung nicht nur Reziprozität zwischen Unternehmen und Nutzerinnen eine Rolle spielt (siehe Kapitel 6), sondern auch jene zwischen Nutzerinnen. Insbesondere auf Social-Media-Plattformen sind die datengenerierenden Verhaltensweisen kommunikative

Handlungen, die sich in erster Linie an andere Nutzerinnen richten und so den Aufenthalt auf der Plattform für andere zu einem interessanten Zeitvertreib macht (siehe auch Kapitel 7.1). Würde niemand etwas teilen, wäre der Besuch einer solchen Plattform schnell langweilig.

> platform-based companies structure reciprocity in such a way that it appears to be directed at the collectivity of users rather than themselves: people ›share‹ their news with their ›friends‹ and acquaintances, rather than with Facebook; they expose their professional information for potential LinkedIn connections to peruse, rather than for LinkedIn itself; and their photos are directed at other Instagrammers, rather than Instagram. (Fourcade & Kluttz, 2020, S. 5)

Vor allem frühe Collaborative-Filtering-Systeme machen sich Formen der technisch vermittelten Sozialität zu Nutze, um ihre Empfehlungen – beziehungsweise die Art und Weise, wie sie Nutzerinnen und Dinge relationieren – darauf zu basieren. In diesem Sinne wird (algorithmische) Sozialität zu einer Ressource für »good matches« zwischen Unternehmen und Nutzerinnen (siehe Kapitel 10.4).

Wie ich zeigen werde, beruht die Personalisierung von Empfehlungssystemen ironischerweise gerade auf einer weitgehenden »Depersonalisierung«: Die formalisierten Verhaltensweisen und algorithmischen Relationierungen betrachten nicht das Individuum an sich: Nutzerinnen (und Dinge) erscheinen immer nur in Relationen zu Nutzerinnen und Dingen (Seaver, 2012). Algorithmische Sozialität ist nicht dasselbe wie Sozialität, da sie auf encodierten beziehungsweise standardisierten Verhaltensweisen beruht, die zwar auf subjektive Sinninvestitionen verweisen, diese aber gleichzeitig auch ignorieren müssen.

Die Analysen in diesem Kapitel wurden bereits in Form eines Buchkapitels und in Form eines Zeitschriftenaufsatzes veröffentlicht: »*›user item item user‹ – Zur Vergleichspraxis von Recommendersystemen*« ist im Band »Global beobachten und vergleichen: Soziologische Analysen zur Weltgesellschaft« von Bennani et al. (2020) erschienen (Unternährer, 2020b). Die Typologie der Empfehlungssysteme entwickle ich in einem Aufsatz für den Sonderband 73 »*Soziale Praktiken des Beobachtens: Vergleichen, Bewerten, Kategorisieren und Quantifizieren*« der Kölner Zeitschrift für Soziologie und Sozialpsychologie (herausgegeben von Bettina Heintz und Theresa Wobbe) (Unternährer, 2021). Für dieses Kapitel habe ich diese beiden Aufsätze neu sortiert und zusammen-

gestellt. Ein Großteil des Texts ist aus diesen Artikeln direkt übernommen und formal nicht als Zitat ausgewiesen.

10.1 Unpersonalisierte Popularitätsmetriken

Leserinnen von Onlinemedien können zwischen meistgelesenen, meistkommentierten oder meistgeteilten Artikeln wählen (Die Zeit). Wer sich überlegt, welcher Film die geeignete Feierabendsunterhaltung sein könnte, kann aus einer Liste populärer oder aktuell aufstrebender (»trending«) Filme auswählen (Netflix: »Popular on Netflix«, »Trending Now«). Musikstreamingdienste wie Spotify haben zahlreiche Listen im Angebot, die der Logik von Musikcharts folgen, zum Beispiel Today's Top Hits, Deutsche Charts oder die beliebtesten Songs einer Künstlerin. Onlineshops zeigen ihren potenziellen Kundinnen Bestseller, angesagte Neuheiten oder Top-10-Kategorien. Bei der Auswahl von Restaurants oder Hotels helfen mir Seiten wie Yelp.com oder Booking.com, mit durchschnittlichen User-Ratings.

Unpersonalisierte Empfehlungen scheinen im Gegensatz zur vieldiskutierten Personalisierung unspektakulär zu sein, doch sind sie allgegenwärtig. Für Konstan und Ekstrand heißt »unpersonalisiert«, dass allen Nutzerinnen dieselben Dinge angezeigt werden. Alle erhalten die gleichen Empfehlungen: Die »aggregierten Präferenzen« der klickenden oder explizit bewertenden Nutzerinnen dienen als »prediction« für die Präferenzen aller Nutzerinnen. Diese Art der Empfehlung fußt auf der Grundannahme (oder für Konstan und Ekstrand der Not geschuldeten Arbeitshypothese, weil zu wenig Daten über die Nutzerinnen verfügbar sind), dass die besten oder populärsten Dinge für alle gleichermaßen relevant sind. Gemäß Tarleton Gillespie identifizieren Trending-Algorithmen nicht, »was spezifisch für ›dich‹ relevant sein könnte«, sondern »was bei ›uns‹ allgemein populär ist« (Gillespie, 2017, S. 78).

Auf Coursera beschreiben Konstan und Ekstrand verschiedene Varianten, wie unpersonalisierte Empfehlungen hergestellt und dargestellt werden können. Je nach »domain« – d.h. je nach Anwendungsbereich wie News-Webseiten, E-Commerce oder Empfehlungsportale – eignen sich unterschiedliche Arten der Berechnung von Popularität. Die vermeintliche Logik unpersonalisierter Empfehlungen ist so verlockend einfach wie irreführend: Es handelt sich mitnichten darum, einfach zählen zu können, wie viele Nutzerinnen etwas angeklickt oder mit einem Like versehen haben.

Ein einfaches Beispiel vermag das zu verdeutlichen: Würde die Webseite einer Tageszeitung die meistgelesenen Artikel *aller Zeiten* auf der Startseite anordnen, hätte sie wohl bald keine Leserinnen mehr, da sich auf der Startseite kaum je etwas verändern würde. »Meistgelesen« oder ähnliche Ranking-Kriterien sind meistens mit einem Zeitindex versehen: z.B. am meisten gelesen *in den letzten 24 Stunden*.[2] Da in den meisten Fällen nicht absolute Popularität, sondern Informationen über zeitlich, räumlich und sozial begrenzte Popularität erwünscht ist, müssen Popularitätsmetriken festlegen, welches Zeitfenster, welche Nutzerinnen und welche Beiträge wie gezählt werden (Gillespie 2017 zu Twitter-Trends).

Ekstrand bespricht das Beispiel, wie die Webseite Hacker News[3] ihre Beiträge auf der Webseite ordnet. Die Nutzerinnen der Webseite konnten Beiträge mit Up- und Downvotes bewerten. Die naheliegende Variante, wie Beiträge sortiert werden könnten, besteht darin, die Differenz von Up- und Downvotes als Score beziehungsweise Vergleichskriterium zu verwenden. Da das Ziel aber ist, »to highlight hot or new stuff« (Salihefendic, 2015), ist die Zeit in der Berechnung des Scores zentral. In die Berechnungsformel ist ein »decay factor« eingebaut, der dafür sorgt, dass der Score mit zunehmendem Alter zuerst sehr schnell und dann immer langsamer abnimmt. Das bedeutet auch, so Ekstrand, dass die ersten Up- und Downvotes einen höheren Einfluss haben als spätere. Damit Artikel auf Hackernews in Bezug auf die Kriterien »Hotness« und »Neuheit« vergleichbar sind, reicht die Aggregation nominaler Nutzerbewertungen (i.e. +1, -1) oder die reine Sortierung nach dem Alter des Artikels nicht aus. Nutzerbewertungen und Alter müssen in einer mathematischen Formel zur Berechnung des »Hotness/Newness«-Scores zusammengezogen werden. Erst dann lassen sie sich in ein ordinales Ranking bringen, das den Ansprüchen von Hackernews genügt.

Die Formel von Hackernews beinhaltet zusätzlich einen »penalty term«, »to influence the kind of community that the site owner wants to develop […] to shape a little bit the kinds of things that come to the top to exert some influence over the flavor that the site has« (Coursera, 2017). So könnten beispielsweise der Score von Beiträgen der Kategorie »Umfragen« negativ beeinflusst werden, da die Eigentümer der Seite nicht zu viele davon auf der Seite wünschen. Auf Webseiten, die in einer anderen »Domain« operieren,

2 Auch analoge Rankings und Bestenlisten sind mit einem Zeitindex versehen. Sie berücksichtigen normalerweise festgelegte Perioden.

3 news.ycombinator.com

ist es womöglich sinnvoll, die Stimmen der Nutzerinnen unterschiedlich zu gewichten. Man denke an Seiten wie TripAdvisor, auf der Hoteleigentümerinnen versuchen, den eigenen Score zu erhöhen, indem sie mehrere Profile erzeugen, um sich selbst gute Bewertungen abzugeben. Der Einfluss solcher Praktiken ließe sich abschwächen, indem beispielsweise die Reputation oder das Alter des Profils als Gewichtungsfaktoren eingesetzt werden.[4]

Während im Beispiel von Hackernews Up- und Downvotes als Grundlage des Vergleichs von Artikeln dienen und ein Klick als explizite und diskrete Handlungs- und Zähleinheit relativ eindeutig bestimmbar ist, ist bei anderen Dingen weniger klar, was gezählt werden soll. Durch Entscheidung und Konvention muss festgelegt werden, was zählt: Ein Song auf Spotify zählt als gehört, wenn die Nutzerin mindestens 30 Sekunden des Songs gestreamt hat (Snickars, 2016); Netflix unterscheidet gemäß The Verge zwischen »starters«,»watchers« und »completers«: »Households that watch 70 percent of a movie or one episode in a series are considered ›watchers‹« (Alexander, 2019).

Bei manchen unpersonalisierten Empfehlungen fällt auf, dass der angegebene Score nicht mit der Positionierung im Ranking übereinstimmt. Auf Hackernews stehen Items mit niedrigeren Punktzahlen vor höheren; auf Spotify steht nicht unbedingt derjenige Song mit den meisten »listens« an erster Stelle der populären Songs einer Künstlerin.[5] Die Ordnung des Rankings unterscheidet sich von der Ordnung der aggregierten Ratings (i.e. welchen Score die Dinge ausweisen). Verschiedene Ordnungslogiken können sich überlagern, wie die Differenz zwischen »prediction« und »recommendation« zeigt:

> we typically focus on two tasks. The first is the predict task: given a user and an item, what is the user's likely preference for the item? [...] The second task is the recommend task: given a user, produce the best ranked list of n items for the user's need. An n-item recommendation list is not guaranteed to contain the n items with the highest predicted preferences, as predicted

4 Siehe zum Beispiel Pinch (2012) für die Anordnung von Nutzerreviews nach dem Kriterium der Nützlichkeit – bewertet durch andere Nutzerinnen. Orlikowski & Scott (2014) für das Beispiel TripAdvisor.

5 Die populärsten Songs einer jeweiligen Künstlerin sind nicht einfach nach totalen Streams geordnet. Sowohl zeitliche Faktoren als auch implizite Bewertungen (beispielsweise ob Nutzerinnen einen Song in ihre individuellen Playlisten aufgenommen haben) dürften eine Rolle spielen (Spotify Community, 2013).

> preference may not be the only criteria used to produce the recommendation list. (Ekstrand et al., 2011, S. 86f.)

Die Unschärfe in der angezeigten Ordnung der Dinge verweist auf ein epistemisches Problem. Popularität, Qualität oder Trends sind nicht Dinge, die irgendwo da draußen existieren und sich mit den richtigen Methoden beobachten und messen lassen.

> Trends sind keine unabhängigen Phänomene: Anders als beispielsweise Abonnentenzahlen oder die Anzahl an Likes, beanspruchen sie noch nicht einmal, verifizierbare Fakten darzustellen. ›Trends‹ können hunderte Bedeutungen annehmen oder auch gar nichts bedeuten. (Gillespie, 2017, S. 85)

Popularitätsmetriken werden hergestellt, indem kontingente, möglicherweise aber institutionalisierte Entscheidungen getroffen werden, was als Indikator für Popularität oder Qualität wie gezählt und verrechnet und wie für die Nutzerinnen aufbereitet wird. Sobald dies einmal festgelegt ist, nehmen die Scores als vermeintlich objektive Repräsentationen der Realität ein Eigenleben an.

Weil Twitters Bestenliste eine Aura der Objektivität und Relevanz anhaftet, ist höchst umstritten, was darin auftaucht oder nicht (Gillespie, 2012): Die Präsenz oder Absenz auf der Liste ist ein Politikum sondergleichen, so dass sich Twitter dem Vorwurf der Zensur aussetzt. Obwohl relativ opak ist (aber: Lotan 2011), wie Twitter die Viralität von Themen bewertet, haben soziale Bewegungen beziehungsweise Aktivistinnen Strategien entwickelt, wie sie ihre Anliegen zu Trends machen können (Tufekci, 2017). Zeynep Tufekci (2014, S. 10) beschreibt, wie türkische Aktivistinnen Hashtags im Vorfeld von Aktionen planen und in Stellung bringen, um sie von allen Followern zum gleichen Zeitpunkt posten zu lassen. Solche konzertierten Kampagnen führen »spikes« von Twitteraktivitäten herbei, die von Twitters Trending-Algorithmus registriert werden sollen. Hinter der Strategie der Spikes steckt die Annahme, dass nicht kontinuierliche, sondern plötzliche Aktivität einer bestimmten Anzahl von Accounts zu einer Aufnahme in der Bestenliste führt.[6]

Wer oder was »zählt«, ist immer auch eine Frage der Macht. Vermeintlich selbstverständliche Zählweisen werden von neuen Intermediären (Morris,

6 Für ein Update seit Musks Übernahme von Twitter siehe Narayanan (2023).

2015) wie Spotify irritiert, wie das Beispiel des französischen Musikpreises »Victoires de la musique« zeigt: In einem Jahresbericht beklagt die Organisation, die den Preis ausrichtet, eine »surexposition du répertoire rap, hip hop & R&B«, die sich korrigieren werde, wenn auch ältere Musikhörerinnen zu streamen beginnen. In der Zwischenzeit wird die Popularität von Hip-Hop nach unten korrigiert, indem nur bezahlte Musikstreams gezählt werden (Oliver, 2020).

Popularitätsmetriken stellen einerseits kategoriale Relationen von vergleichbaren Einheiten und andererseits ordinale Relationen zwischen den verglichenen Einheiten her, die das Populärere vom weniger Populären unterscheiden. Als Form der Empfehlung beziehen sich Popularitätsmetriken auf ein diffuses Wir, in dessen Namen und für dessen Erwägung sie Relevanz, Präferenz, Qualität oder Interessantheit suggerieren. Obwohl sich die Popularität beziehungsweise Qualität unpersonalisierter Empfehlungssysteme auf eingegrenzte Nutzergemeinschaften bezieht, wird sie oftmals als Ausdruck einer allgemeinen Popularität oder einer den Dingen inhärenten Qualität gelesen.

> [S]ie beanspruchen, etwas über öffentliche Aufmerksamkeit jenseits der von der Nutzerin ausgewählten Community aus Freunden oder Followers auszusagen. Sie sagen etwas über kulturelle Relevanz aus – vielleicht implizit, vielleicht auch fehlerhaft – in jedem Fall sind wir angehalten, sie als solche zu interpretieren. Sie kristallisieren populäre Aktivitäten zu etwas Lesbarem und leiten es dann an uns zurück, oftmals im gleichen Moment, in dem weitere Aktivitäten stattfinden. (Gillespie, 2017, S. 85)

Dominique Cardon (2016) identifiziert die Logik der Popularität als eine Bestärkung von Konformismus und Mainstream. Seine Kritik verweist auf die Grundannahme der unpersonalisierten Empfehlung, dass es sinnvoll sein kann, die Inhalte einer Webseite nach Popularität zu ordnen, da Popularität (zumindest in Abwesenheit besserer, personalisierter Alternativen) ein guter Prädiktor der Präferenzen aller Nutzerinnen oder einer durchschnittlichen Nutzerin ist. Cardons Kritik des Mainstreamings ist die Kehrseite des Arguments, dass Personalisierung zur Fragmentierung der Öffentlichkeit führt, die geteilte Erfahrungen zunehmend unwahrscheinlich machen. Vor dem Hintergrund der Fragmentierungsdiagnose erscheinen an allgemeiner Popularität und inhärenter Qualität orientierte Ordnungslogiken in einem positiveren Licht: »Anstatt sie zu zerreißen, rufen Trending-Algorithmen Öffentlichkeiten eher zusammen« (Gillespie, 2017, S. 85). Unpersonalisier-

te Empfehlungen generieren keine formalen Relationen zwischen Nutzerinnen, deren Aktivitäten aggregiert werden, und Nutzerinnen, denen Popularitätsrankings oder aggregierten Ratings angezeigt werden. Auf der formalen Ebene erzeugen unpersonalisierte Empfehlungen lediglich »ordinale« Relationen (Fourcade, 2016b) zwischen Dingen. Die Logik der Popularität erzeugt zwar die Suggestion eines gemeinsamen Wirs, das aber diffus im Sinn einer »imagined community« bleibt.

Fallanalyse Trending: »A smarter Reddit«

Angèle (Christin, 2018) beschreibt in ihrem Artikel »Counting Clicks: Quantification and Variation in Web Journalism in the United States and France« wie Besucherzahlen (»traffic numbers«) von Nachrichtenseiten die journalistische und redaktionelle Arbeit verändern und in den beiden nationalen Kontexten unterschiedlich interpretiert werden. »American and French journalists see different things when looking at web analytics« (Christin, 2018, S. 1410), je nach lokal vorherrschenden Rollenbildern, Professionalisierung und relativer Abschirmung gegenüber Marktzwängen.

Es werden aber nicht nur Journalistinnen und Redakteurinnen mit Besucherzahlen konfrontiert. Onlinemedien offerieren ihren Leserinnen unpersonalisierte Empfehlungen, die auf Klicks basieren: »Top Stories« (New York Times), »most popular« (The Guardian), »meistgelesen«, »meistkommentiert«, »meistgeteilt« (Die Zeit) oder »was heute wichtig ist« (NZZ). Diese Bezeichnungen suggerieren, dass »counting clicks« ausreichend wäre, um Artikel nach Rang zu sortieren. Die Ordnung dieser Listen ist zwar intuitiv verständlich, aber nur scheinbar unproblematisch.

Ich demonstriere im Folgenden, wie voraussetzungsreich die Anordnung von Kommentaren auf der Internetseite Reddit ist. Reddit ist die selbsternannte »front page of the internet«. Registrierte Benutzerinnen können auf der Seite sogenannte Posts, d.h. Geschichten, Nachrichten, Bilder, Witze, etc. erstellen, die von anderen Nutzerinnen bewertet und kommentiert werden. Bei jedem Post ist Benutzername, Zeitpunkt des Postings und Pfad des Unterforums (»subreddit«) vermerkt. Zusätzlich wird gezählt, wie viele Kommentare ein Post erhalten hat. Jeder Post erhält einen Score, der über seine Beliebtheit informiert. 2015 gab es gemäß Reddit 88 700 aktive Subreddits, in denen 8.7 Millionen User insgesamt 736 Millionen Mal kommentierten (Reddit, 2015).

Reddit ist ein lebhaftes und oftmals kontroverses Diskussionsforum: Leserinnen kommentieren und bewerten Artikel und Kommentare anderer Leserinnen. Hier konzentriere ich mich auf die Bewertung und Sortierung von Kommentaren. Diese einzelnen Kommentare erscheinen immer mit einem Score, der die Anzahl der »points« eines Beitrags ausweist. Die Benutzerinnen können Kommentare mit einem Up- oder Downvote bewerten, indem sie beim Beitrag auf den Pfeil nach oben (i.e Upvote) beziehungsweise unten (i.e. Downvote) klicken. Die Bewertungen der Leserinnen werden verrechnet und je nach Resultat wird die Position des Kommentars angepasst.

Nicht alle Klicks sind gleich: In Anlehnung an Christins Erkenntnis, dass Zahlen für amerikanische und französische Journalistinnen eine unterschiedliche Bedeutung haben, ist für die analytische Maschinerie des Vergleichs eine 5 nicht immer eine 5. Am Beispiel von Reddits System zur Anordnung von Kommentaren zeigt sich, wie Vergleichbarmachung, Bewertung und Quantifizierung miteinander verwoben sind.

2009 führte Reddit ein neues Verfahren zur Sortierung von Kommentaren ein. Das alte Top-Verfahren ordnete die Posts nach Punkten: Der Kommentar mit den meisten Punkten stand an erster Stelle. Diese Art der Anordnung von Kommentaren stieß auf Einwände, da Kritiker in Frage stellten, ob jene Kommentare mit den meisten Punkten tatsächlich die »besten« sind. Als Reaktion darauf wurde das neue Best-Verfahren entwickelt: Es sortiert die Kommentare nach »Qualität«, so dass der »beste« Kommentar an erster Stelle erscheint. Das muss nicht zwangsläufig der Kommentar mit den meisten Punkten sein. Das alte Verfahren zählte Klicks (i.e. Up- und Downvotes) und produzierte eine für die Kritiker unbefriedigende Anordnung der Kommentare. Das neue System misst die Qualität von Kommentaren und kann nun tatsächlich die in den Augen der Kritiker besten Kommentare priorisieren.

Wie diese Berechnungen durchgeführt und wie Kommentare verglichen und angeordnet werden, ist Gegenstand der folgenden Ausführungen. Die empirische Analyse beruht auf Erklärungen im Coursera Kurs, auf Blogbeiträgen sowie auf Reddit-Diskussionen, die den Wechsel des Rankingverfahrens für Kommentare begleiteten.

Ein Artikel auf der Frontseite von Reddit hat typischerweise einige hundert bis mehrere tausend Kommentaren.[7] Ein unpersonalisiertes[8] Empfeh-

7 2010 zählte der wohl meistkommentierte Artikel 358 000 Kommentare (Reddit, 2010).

8 Unpersonalisiert bedeutet hier, dass diese Anordnung für alle Leserinnen gleich aussieht.

lungssystem soll den Leserinnen helfen, mit diesem Überfluss an Kommentaren umzugehen, indem die Kommentare in einer (nicht nummerierten) Rangliste angeordnet werden. Die Sortierung der Kommentare ist allerdings keineswegs trivial. Im Gegenteil: Evan Miller[9] definiert in seinem Blog die Sortierung als zentrales Problem:

> PROBLEM: You are a web programmer. You have users. Your users rate stuff on your site. You want to put the highest-rated stuff at the top and lowest-rated at the bottom. You need some sort of ›score‹ to sort by. (Miller, 2009)

Reddits einfache Lösung – das Top-Verfahren – besteht darin, die aufaddierten Downvotes von den aufaddierten Upvotes abzuziehen. Die Differenz von Up- und Downvotes ergibt die Punkte, die für jeden Kommentar ausgewiesen werden. Top sortiert Kommentare absteigend nach Punkten. Miller (2009) kritisierte diese simple Art der Sortierung mit deutlichen Worten:

> Suppose one item has 600 positive ratings and 400 negative ratings: 60 % positive. Suppose item two has 5,500 positive ratings and 4,500 negative ratings: 55 % positive. This algorithm puts item two (score = 1000, but only 55 % positive) above item one (score = 200, and 60 % positive). WRONG. (Miller, 2009)

Da sich die Anzahl Bewertungen in ganz unterschiedlichen Größenordnungen bewegen kann (eine Handvoll, mehrere Hundert, mehrere Tausend), würden sowohl einfache Werte wie die absolute Differenz als auch der relative Anteil positiver Bewertungen falsche Resultate und eine falsche Ordnung produzieren. Doch weshalb ist das eine »falsche« Ordnung?

Auch für Randall Munroe, Ingenieur, Autor des Webcomics xkcd und einflussreicher Redditor, ist das Top-Verfahren problematisch. Er formuliert das Problem spezifisch für Reddit: »When a mediocre joke gets posted in the first hour a story is up, it will become the top comment if it's even slightly funny«, und er folgert: »The ›top‹ list was simply a list of the best jokes from within the first hour« (Munroe, 2009). Top identifiziere nicht die besten Kommentare, sondern die frühesten. Munroe untersucht als Beispiel

9 Miller entwickelt Statistiksoftware für Unternehmen und Wissenschaft. Der hier zitierte Blogeintrag gehört zu Courseras Lernmaterialien.

die Kommentare eines beliebigen, acht Stunden alten, Posts: Von den 10 Top-Kommentaren waren alle sieben oder acht Stunden alt, d.h. sie wurden in der ersten Stunde seit Erscheinen des Posts abgegeben. Beim Top-Verfahren kämen also oftmals nicht die besten, sondern nur mittelmäßige Kommentare an die Spitze des Rankings.

Munroe ist der Auffassung, dass es einen besten Kommentar gibt. Das Problem besteht nur darin, ihn zu erkennen und richtig einzuordnen. Zur Verdeutlichung gibt er ein Beispiel. In einem Post frage jemand: »Which Redditor took a pic of me while I was blackout drunk at the bar this weekend?« (Munroe, 2009). Gemäß Munroe sollte jener Kommentar am besten abschneiden, der tatsächlich die Frage des Posts beantwortet. Da Reddit seinen Nutzerinnen beide Ordnungsverfahren anbietet, lässt sich direkt vergleichen, wie die beiden Verfahren Top und Best abschneiden: Sortiert nach Top – dem alten Verfahren – erscheint ein zusammenhangsloser Kommentar an erster Stelle. Sortiert nach Best erscheint der richtige Kommentar, nämlich die Antwort des Fotografen – beziehungsweise das von ihm aufgenommene Bild des betrunkenen Redditors.

Das Problem besteht für Munroe darin, dass beim Top-Verfahren frühe gegenüber späteren Kommentaren einen Vorteil haben. Um dieses Problem zu beheben, schlagen Miller (2009) und Munroe (2009) vor, den unteren (konservativeren) Grenzwert des Wilson-Konfidenzintervalls anzuwenden. Die Vergleichbarkeit von Kommentaren mit wenigen Bewertungen und jenen mit vielen Bewertungen wird hergestellt, indem die momentane Anzahl Bewertungen als Stichprobe behandelt wird: Der Kommentar bekommt einen provisorischen Score, den er mit einer Sicherheit von 95 % erreichen wird. Kommen weitere Bewertungen hinzu, wird der Score laufend angepasst (Munroe, 2009; Salihefendic, 2015). Die Vergleichbarkeit von Kommentaren wird hergestellt, indem neue Kommentare mit wenigen Bewertungen mathematisch so behandelt werden, als hätten sie schon viele Bewertungen erhalten (Salihefendic, 2015). Es handelt sich um eine Vorhersage des Scores aufgrund des Verhältnisses von Up- und Downvotes zum Berechnungszeitpunkt.

Der vorliegende empirische Fall ist instruktiv, um die Differenz zwischen Zählen und Messen zu beleuchten, wie sie sich in den Verfahren Top und Best widerspiegelt. Die Ausgangslage für beide Verfahren ist die Formalisierung der auf Reddit möglichen Verhaltensweisen (Kommentieren sowie Bewerten mit Up- oder Downvote) und die Gleichheitsunterstellung der Vergleichseinheiten (i.e. Kommentare).

Auf Reddit sind verschiedene (Un-)Gleichheitsunterstellungen zu beobachten. Kommentare und Posts werden als unvergleichbar behandelt: Auf der Startseite sind nur die beliebtesten und bestbewerteten Posts zu finden, obwohl auch denkbar wäre, die besten Kommentare als lesenswerte Einheiten zu präsentieren.

Kommentare werden nur auf jeweils einer Ebene miteinander verglichen: Ein Hauptkommentar zu einem Post wird nur mit anderen Hauptkommentaren zum gleichen Post in Beziehung gesetzt. Ein Kommentar zu Kommentar X nur mit anderen Kommentaren zu Kommentar X – nicht aber mit Kommentaren zu Kommentar Y.[10] Die Handlungsmöglichkeiten auf Reddit sind durch die Benutzeroberfläche dermaßen formalisiert, dass keine Unsicherheit darüber bestehen kann, ob es sich nun um einen Hauptkommentar, einen Unterkommentar oder einen Post handelt.

Um Differenzen zwischen Kommentaren einer Ebene feststellen zu können, lässt Reddit die Kommentare von seinen Leserinnen mit einem einfachen Ratingsystem bewerten: Leserinnen können Kommentare entweder mit einem Up- oder einem Downvote versehen. Sowohl Top als auch Best geben vor, auf der Basis dieser Bewertungen Differenzen in der Qualität von Kommentaren feststellen zu können. Sie produzieren diese Differenzen aber auf unterschiedliche Weise.

Es lassen sich drei unterschiedliche Arten der Bewertung beobachten: nominale, kardinale und ordinale »judgements« (Fourcade, 2016b).[11] Die Nutzerinnen bewerten Kommentare nominal, indem sie Up- oder Downvotes *vergeben*. Das Top-Verfahren produziert kardinale Bewertungen, indem es Up- und Downvotes *zählt*. Best *misst* die »Qualität« von Kommentaren und erzeugt so eine ordinale Rangordnung.

Up- und Downvotes sind zufällige Symbole einer binären Bewertung. Es spielt im Grunde keine Rolle, ob es Pfeile, Daumen oder Einsen und Nullen

10 Reddit ignoriert diese Vergleichssperre, wenn Posts verglichen werden: Reddit zählt für jeden Post die Anzahl Kommentare, ohne zwischen Hauptkommentaren und Unterkommentaren zu unterscheiden.

11 Nominale Bewertungen sind an Essenzen orientiert: Sie definieren, was etwas oder jemand ist. Bei kardinalen Bewertungen wird gezählt: Die Anzahl beziehungsweise die Akkumulation von Dingen ist ausschlaggebend. Ordinale Bewertungen erzeugen relative Positionierungen: Nicht Größenordnungen oder absolute Differenzen sind von Interesse, sondern die Einordnung von Dingen als besser oder schlechter als andere Dinge (Fourcade, 2016b, S. 176 ff.).

sind (Heintz, 2018). Nutzerbewertungen sind »nominal judgements« (Fourcade, 2016b, S. 176f.), die in der Unterscheidung von Up- und Downvotes als diskrete und zählbare Ereignisse formalisiert sichtbar gemacht und regelmäßig produziert werden.[12]

Das Top-Verfahren ist ein »cardinal judgment«, das als Vergleichskriterium direkt Beobachtbares und Zählbares verwendet (Fourcade, 2016b, S. 177): Es zählt die Up- und Downvotes eines Kommentars und weist die Differenz aus, so dass jener Kommentar mit der höchsten positiven Differenz an erster Stelle steht. Das Vergleichskriterium, mit dem Differenz konstatiert wird, ist im Top-Verfahren alleine die Größenordnung von Up- und Downvotes, wenn auch die Benennung suggeriert, dass es sich um ein Qualitätskriterium handelt. Wie aus der Diskussion um das am besten geeignete Ranking-Verfahren hervorgeht, scheint das simple Zählen von Punkten aber nicht die erwünschten Resultate zu produzieren, da sich darin nicht die »Qualität« der Kommentare ausdrücke. Top als »cardinal judgment« funktioniert nicht als Indikator für Qualität.

Im Gegensatz zum Punktestand ist »Qualität« nicht direkt (beziehungsweise durch Zählen) beobachtbar.

> Wenn Erkenntnisgegenstände nicht physisch gegeben und direkt beobachtbar sind, wachsen die kognitiven Probleme ihrer Erfassung schlagartig. Wo man nicht mehr die Einheiten einer Gesamtheit zählen und einfache statistische Maßzahlen wie Durchschnitte, Verteilungen und Raten (zum Beispiel Anteil der Bevölkerung unterhalb der definierten Armutsgrenze) bilden kann, beginnt das Reich des Messens. (Mayntz, 2017, S. 7)

In der Argumentation für das Best-Verfahren wird in Frage gestellt, ob Top als Zählverfahren die gewünschte Ordnung der Kommentare produziert: Es kann nicht Qualität wiedergeben, sondern nur Kommentare mit den höchsten Punkteständen. Wie Munroe (2009) argumentiert, eruiere Top nicht die besten Kommentare, sondern die frühesten:

---

12 Leserinnen können sich zwar der binären Logik von Up- und Downvote verweigern, indem sie beispielsweise einen schlechten Kommentar mit einer vernichtenden Antwort kommentieren. Diese Art der Bewertung »zählt« aber nicht, wenn sie auch in der inhaltlichen Diskussion wichtiger sein mag.

> [O]nce a comment gets a few early upvotes, it's moved to the top. The higher something is listed, the more likely it is to be read (and voted on), and the more votes the comment gets. It's a feedback loop that cements the comment's position, and a comment posted an hour later has little chance of overtaking it. (Munroe, 2009)

Wie gesehen, plädiert er dafür, das simple Zählen des Punktestandes durch ein Messverfahren zu ersetzen, das diesen »time bias« und den damit verbundenen Rückkopplungseffekt in Betracht zieht. An dieser Diskussion um das »richtige« Verfahren lässt sich beobachten, dass Konventionen der Quantifizierung (Espeland & Stevens, 1998, 2008) und Gleichheitsunterstellungen kontingent und wandelbar sind. Zu Beginn galt Top – das Zählen der Punkte – als zuverlässiger Indikator, um den besten Kommentar zu ermitteln. Um Qualität verlässlich messen zu können, führte Reddit das Messverfahren Best ein, um die von Munroe kritisierte quasi-kategoriale Ungleichheit von früheren und späteren Kommentaren mathematisch einzuebnen. Das Best-Verfahren stellt so im Prozess der Differenzbeobachtung selbst kategoriale Gleichheit her: Es wird nicht mehr die Differenz von Up- und Downvotes gezählt, sondern die »Qualität« von Kommentaren gemessen. Während sich Top als kardinales Verfahren auf absolute Größen fokussiert, ist Best als ordinales Verfahren nicht an der Größe der Differenz interessiert, sondern an der relativen Rangordnung (Fourcade, 2016b, S. 178). Entsprechend werden weder ein Best-Score noch Rangziffern angezeigt: Entscheidend ist bloß die Anordnung.

Das Reddit-Beispiel verweist auf die Relevanz unterschiedlicher Arten von Zahlenförmigkeit, die für die Operationsweise des Vergleichs maßgebend sind. Es macht deutlich, wie im Falle des untersuchten Best-Verfahrens nicht Zahlenförmigkeit an sich einen für die Teilnehmenden sinnvollen Vergleich garantiert. Die Vergleichbarkeit von Kommentaren basiert auf der Formalisierung von Kommentar- und Bewertungsfunktion, welche die kategoriale Gleichheit der Kommentare und das Material für die Differenzbeobachtung (beider Verfahren) verlässlich und regelmäßig erzeugen (siehe Kapitel 7).

Diese Formalisierungsleistung alleine genügt aber nicht (mehr): »counting clicks« ist in den Augen der Entwickler nicht ausreichend. Am Beispiel der Entwicklung des Best-Algorithmus zeigt sich, wie Kritiker wie Munroe eine zu Beginn unproblematische Gleichheit von Kommentaren in Frage stellen, so dass die Vergleichbarkeit durch einen Messprozess neu hergestellt

werden musste. Das Problem der quasi-kategorialen Ungleichheit von frühen und späten Kommentaren löste Reddit durch ein mathematisch elaboriertes Messverfahren, das im Prozess selbst eine Gleichheit von frühen und späten Kommentaren herstellt.

10.2 Stereotypisierung: »I'm not a 15-year-old girl«

Die Stärke von Popularitätsmetriken liegt darin, dass sie auch dann Empfehlungen machen können, wenn über die Nutzerin nichts bekannt ist.[13] Was populär ist, trifft aber womöglich nicht den Geschmack der Nutzerin, der Empfehlungen präsentiert werden sollen:

> Popularity as I said can be useful. If I'm looking for popular new songs, I might look at the top 10 chart or the top 100 chart. On the other hand that chart might be dominated by songs popular among high schools girls. And I don't know how well you can see me in these videos but I'm not a 15-year-old girl. (Coursera, 2017)

Eine Lösung könne darin bestehen, »to loosely personalize if not to the individual then to some attributes of the individual« (Coursera, 2017). Etablierte Kategorien wie Geschlecht, Alter, Wohnort, Einkommen, etc. aber auch Kombinationen daraus wie Lebensstile könnten als Stellvertreter für Geschmack, Interessen und Konsumentscheide dienen. Solche Informationen werden bei der Registrierung erhoben oder können von Drittanbietern erworben werden (Bivens & Haimson, 2016; Crain, 2018). Gemäß Konstan sind demografische Daten für Empfehlungssysteme eine wertvolle Ressource, weil sie relativ einfach zu erheben sind und für neue Nutzerinnen – über die sonst wenig bekannt ist – stereotypisierende Empfehlungen generieren können. Ich bezeichne solche Empfehlungen als »stereotypisierend« (statt »stereotypisch«), da sie auf Konvention und vermeintlichen Selbstverständlichkeiten beruhende Relationen zwischen Gruppen und Dingen herstellen.

Die Herstellung einer stereotypisierenden Empfehlung schließt in ihrer Logik an Popularitätsmetriken an. Anstatt Popularität über die Gesamtpopulation von Nutzerinnen zu erheben, lässt sich mithilfe der demografischen Kategorien »desaggregieren«. Damit wird eine weitere Vergleichsebene eingeführt. Ein einfaches Beispiel: Eine Population von Nutzerinnen und Nut-

13 Das ist das sogenannte Kaltstart-Problem (siehe Kapitel 6).

zern bewertet Filme auf einer Skala von 1 bis 5. Aus diesen Bewertungen lassen sich Durchschnittswerte für alle bewerteten Filme berechnen, um sie dann absteigend nach bester Bewertung zu sortieren: Es resultiert ein Vergleich oder konkret eine Bestenliste der von der Community bewerteten Filme. Da die Nutzerinnen und Nutzer bei der Registrierung ihr Geschlecht angegeben haben, lassen sich die »global averages« zudem nach den Kategorien »männlich« und »weiblich« aufschlüsseln, wodurch zwei weitere Bestenlisten entstehen: »Beliebt bei Frauen« und »Beliebt bei Männern«.

> You're selling things, and it turns out that the most popular products for women are different from the most popular products for men. Break down those summary statistics, and display the relevant products on your page when a man or woman arrives. (Coursera, 2017)

Die Liste »Populäre Filme bei Frauen« behauptet keine allgemeine Relevanz, sondern bringt spezifische Nutzer-Kategorien mit spezifischen, für diese Gruppe relevanten, Dingen in eine Relation. Stereotypisierung beruht auf kategorialen Relationen zwischen Nutzerinnen und nutzt diese Relation, um Affinitäten zwischen diesen kategorialen Gruppen und Dingen festzustellen.

Kategoriale Zugehörigkeiten dienen als Indikatoren für bestimmte Interessen und Geschmackspräferenzen der Webseitenbesucher. Diese Intuition schließt sowohl an sozialwissenschaftliche Erkenntnisse und Methoden als auch an Methoden des »mass marketing« des 20. Jahrhunderts an (Fisher & Mehozay, 2019). Riedl & Konstan (2002) verweisen auf historische Vorläufer und Abgrenzungsfolien, von denen sich Technologien der Personalisierung rhetorisch abgrenzen – obwohl sie nach wie vor damit verquickt sind (Pridmore & Hämäläinen, 2017; Bolin & Andersson Schwarz, 2015). Joseph Turow (2003) erzählt die Geschichte der Segmentationspraktiken in der Werbebranche als eine der langen Präsenz, aber zögerlichen Anwendung: Demografisches »targetting« – »Personalisierung« avant-la-lettre – war zwar als Technik bereits bekannt, schien jedoch in der ersten Hälfte des 20. Jahrhunderts ineffizient zu sein, da Werber mit möglichst wenig Aufwand ein möglichst großes Publikum erreichen wollten. Targetting beginnt in den USA erst mit einer Ausdifferenzierung von Konsumprodukten und einer Ausdifferenzierung der Medienlandschaft Fuß zu fassen: Produzenten materieller Güter versuchen mit leicht unterschiedlichen Varianten derselben Produkte unterschiedliche Marktsegmente zu erreichen (Turow, 2000, S. 240f.) beziehungsweise ihre Produkte von generischen zu Produkten mit Wiedererkennungswert – d.h. Markenprodukten – zu machen (Beniger, 1986, K.

8). Radiosender, Magazine und später Kabelsender richten sich an spezifische demografische Segmente, die mit spezifischen Lebensstilen und Produktpräferenzen in Verbindung gebracht werden (Turow 2003, K. 2, Fisher & Mehozay 2019). Demografisches und später Lifestyle-Targetting setzt sich erst durch, als sich (unter Werbeleuten) Theorien einer zunehmend fragmentierten Gesellschaft verbreiten. Die kulturelle Plausibilisierung, dass unterschiedliche Gruppen unterschiedlich konsumieren sowie die technischen Entwicklungen von Computerisierung und statistischer Methode der Konsumentenforschung waren verbunden mit einer faktischen Ausdifferenzierung von Öffentlichkeiten (Turow, 2003, K. 3).

Sowohl die Publikumsforschung des »mass marketing« als auch die stereotypisierenden Empfehlungssysteme stellen Interessen und Präferenzen nicht nur beschreibend dar. Turows historische Aufarbeitung von Segmentationspraktiken, aber auch Konstan und Ekstrand selbst, weisen auf das ko-konstitutive Verhältnis von Konsumentenkategorien und Produkten hin. Unternehmen stellen nicht nur Produkte her, sondern, in Zusammenarbeit mit Marketingabteilungen und Datenbanken, auch die zugehörigen Konsumentinnen (Zwick & Denegri Knott, 2009). Konstan argumentiert anhand einer Anekdote, dass stereotypisierende Empfehlungen in vielen Fällen funktionieren, weil schon in der Produktion an bestimmte Zielgruppen gedacht wurde.

> [The recommender] worked way too well to believe that the machine learning could simply learn your preferences in many cases. With six or seven preferences, it nailed people's television watching behaviors. And as they explored this, part of what they explained was, this is not an accident. Television programs [...] are targeted towards particular demographics. If you watched back then 60 Minutes and Murder She Wrote, you were probably a senior citizen. If you watch certain sports programs, you have a 70 % or 80 % chance of being male. If you watch certain soap operas, you had a 70 % or 80 % chance of being female. [...] the reason that content was created to reach audiences was because advertisers were trying to reach audiences. And the history of mass media was advertisers driving content so that, that content could be used to pitch advertisements for products and generate sales. (Coursera, 2017)

Produktkategorien erzeugen nicht nur Differenz gegen außen und Homogenität gegen innen. Sie beinhalten oftmals auch relationale Komponenten wie die von Cynthia Cockburn und Susan Ormrod (1993, K. 4) beschriebe-

nen Kategorien der »white goods« – langweilige, simple Haushaltsgeräte für (Haus)Frauen – sowie »technische«, zur Unterhaltung gedachte, interessante »brown goods« für Männer. Marianne Lien schreibt in ihrer Studie zu den Marketingpraktiken eines norwegischen Nahrungsmittelherstellers: »there is a marked tendency to describe the product in terms of its consumers (and vice versa) thus blurring the conceptual distinction between product and consumer« (Lien, 2000, S. 162).

Konstan und Ekstrand beschreiben eine Art stereotypisierender Empfehlung, die dem Paradigma der Demografie- und Marktforschung verhaftet ist. Demografische Kategorien werden explizit abgefragt und über Häufigkeitsauswertungen oder Regressionsanalyse mit spezifischen Verhaltensweisen, Interessen und Werten in Verbindung gebracht. Fisher und Mehozay sprechen von einer »ascriptive conception of the individual: each individual could be assigned to a category, which could then be sociologically and culturally characterized« (2019, S. 1181). Soziale Kategorien sind weiterhin relevant: Werbekundinnen von Plattformen fragen weiterhin binäre Genderkategorien nach. Rena Bivens und Oliver Haimson (2016) konnten aufzeigen, dass auf der Nutzeroberfläche von Facebook mittlerweile zwar aus über 50 Genderkategorien ausgewählt werden kann, diese im Hintergrund für Werbekunden aber wieder binarisiert werden. Sie unterscheiden drei Arten, wie Genderkategorien auf Social-Media-Plattformen zugeschrieben werden. Erstens: Nutzerinnen sortieren sich selbst bei der erstmaligen Anmeldung auf der Plattform in ihre präferierte Genderkategorie ein. Zweitens: Gender wird indirekt über Daten von Drittanbietern zugeschrieben, die beispielsweise über Emailadressen oder über die angegebenen präferierten Pronomen zugeordnet werden können. Drittens: Aus datafizierten Verhaltensweisen wird (beispielsweise) auf Genderkategorien rückgeschlossen, d.h. die Relation zwischen Nutzerin und Ding wird nicht mehr über stereotype Kategorisierung hergestellt, sondern über den Vergleich von Verhaltensweisen mit einem etablierten Muster:

> While Twitter and LinkedIn maintain genderless sign-up pages and profile pages, both of these platforms use user data and actions to algorithmically infer a binary gender category to satisfy their advertising and marketing clients«. (Bivens & Haimson, 2016, S. 6)

Zwischen diesen beiden Arten der Zuschreibung von Genderkategorien und der dritten besteht eine qualitative Differenz, die mit Fisher & Mehozay als algorithmisches Paradigma bezeichnet werden kann: »the rubrics in the al-

gorithmic episteme […] represent not social categories, but patterns of data« (Fisher & Mehozay, 2019, S. 1185f.).

Was versteht Google unter »verheiratet sein«?

Wenn Google von einer kleinen Subpopulation weiß, dass sie verheiratet sind, können sie daraus ein Modell erstellen, um meine Verhaltensweisen mit diesem Modell zu vergleichen. Sind meine Verhaltensweisen dem Modell genügend ähnlich, stecken sie mich ebenfalls in die Gruppe der Verheirateten und präsentieren mir entsprechende Werbung.

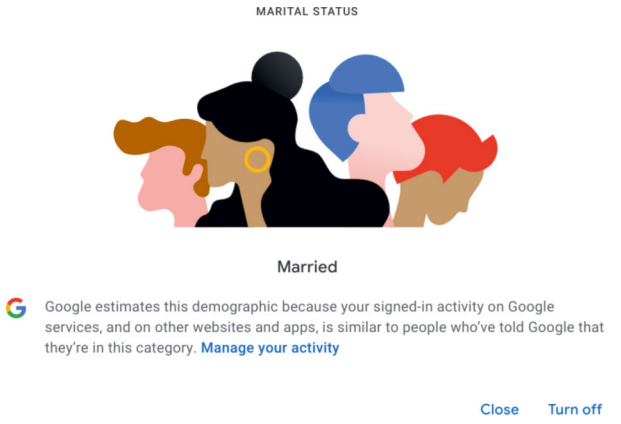

Abb. 7: Quelle: https://adssettings.google.com (nicht öffentlich verfügbar, Screenshot vom 30. Oktober 2020

Während demografische Identitätskategorien oftmals essenzialistisch gedacht sind, sind Identitätskategorien als »patterns of data« formbarer und sozusagen agnostisch (Cheney-Lippold, 2011). Cheney-Lippold beschreibt solche Protokategorien als »measurable types«:

> A measurable type is a data template, a nexus of different datafied elements that construct a new, transcoded interpretation of the world. These templates are most often used to assign users an identity, an algorithmic identification that compares streams of new data to existing datafied models. Categorical membership is assigned based on algorithmic fit: if one's data is spoken for ›as if‹ it was produced by a ›terrorist‹, for example, one is seen to be a terrorist. And fit is subsequently based only on what data is available to

be measured, so membership – and identity at large – is based exclusively on data. (Cheney-Lippold, 2017, S. 47)

Was versteht Facebook unter »hispanic«?

Zwischen Facebooks und Googles Kategorien der »profitable convenience« (Cheney-Lippold, 2017, S. 7) und den gebräuchlichen Alltagskategorien besteht eine qualitative Differenz. Alltagskategorien sind oftmals essenzialistisch oder biologistisch gedacht. Das Beispiel von Facebook macht deutlich, dass das bei algorithmischen Kategorien gerade nicht so ist. Ihre Kategorie »hispanic« ist eine »als-ob«-Kategorie: Es spielt für Facebook keine Rolle, ob jemand »tatsächlich« lateinamerikanischer Herkunft ist – was immer das auch heißen soll –, solange diese Nutzerin sich so verhält als ob, d.h. solange das Verhalten mit dem »Modell Hispanic« übereinstimmt und solange sie auf Werbung und Empfehlungen reagiert, die sich an entsprechende Personen richten, kategorisiert Facebook die Nutzerin als »hispanic«.

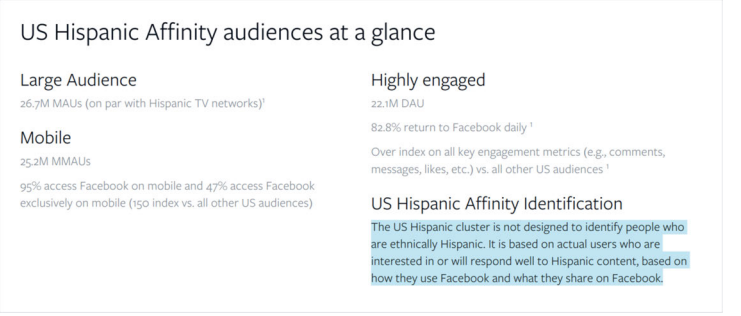

Abb. 8: Quelle: https://www.facebook.com/business/a/us-hispanic-affinity-audience (mittlerweile nicht mehr abrufbar; Screenshot vom 18. Januar 2019

Techniken des Maschinellen Lernens verändern, wie traditionelle Identitätskategorien verstanden werden: Plattformen wie Google und Facebook sind nicht mehr darauf angewiesen, dass Nutzerinnen ihnen diese Kategorien selbst mitteilen.

Der Referenzpunkt von »measurable types« ist aber nach wie vor die soziale Kategorie, zumindest wenn das Identitätsmodell mit Verfahren des »supervised learning« erstellt wurde. Demografische Kategorien werden also nicht gänzlich überflüssig. Sie müssen aber nicht mehr explizit nachgefragt

werden, sondern lassen sich aus Mustern in den Daten ableiten. Solange Kunden von Werbeplattformen mit stereotypisierenden Segmenten arbeiten, werden demografische Kategorien wohl relevant bleiben (Bivens, 2017; Bolin & Andersson Schwarz, 2015).

10.3 Inhaltsbasierte Empfehlungen: User = Item?

Stereotypische Empfehlungen verwenden demografische Variablen als Stellvertreter für Interessen und Präferenzen. Die Zugehörigkeit zu verschiedenen kategorialen Gruppen wie Frauen, Bisexuellen oder über 65-Jährigen kann ein Prädiktor für Risikoverhalten, Konsumentscheidungen oder Einstellungen sein – wie die Versicherungsstatistik, Marktforschung oder empirische Sozialforschung herausgefunden haben.

Im Gegensatz dazu basieren inhaltsbasierte Empfehlungen nicht auf kategorialen Zugehörigkeiten, um daraus Rückschlüsse auf etwas anderes zu ziehen. Bei inhaltsbasierten Empfehlungen spielt die Ähnlichkeit von Nutzerinnen keine Rolle: Nutzerinnen solcher Systeme werden weder kategorialen Kollektiven zugeordnet noch zu einer Nachbarschaft ähnlicher Nutzerinnen in Relation gesetzt. Was für den einzelnen als relevant oder interessant eruiert wird, basiert nicht auf einer Relationierung von Nutzern – sei es pauschal wie bei Popularitätsmetriken, kategorial wie bei Segmentation oder über Matching wie bei User-User-Collaborative-Filtering (UUCF) –, sondern auf einem Vergleich von Dingen (Englisch: »items«). Dementsprechend können sogar dann Empfehlungen berechnet werden, wenn nur eine einzige Nutzerin im System ist (Jannach et al., 2011, S. 51).

Steht bei der Segmentation die Kategorisierung der Nutzerinnen im Vordergrund, geht es in inhaltsbasierten Empfehlungssystemen primär um die Kategorisierung beziehungsweise Modellierung von Dingen. Nutzerinnen werden analog zu den Dingen modelliert, um die Vergleichbarkeit von Dingen und Nutzerinnen herzustellen: Nutzerinnen müssen wie Dinge zu »Vektoren« in einem gemeinsamen Vektorraum werden: Der zuletzt gelesene Artikel, angehörte Songs oder gesehene Film wird zum Stellvertreter für die Nutzerin. In einem dritten Schritt wird ein Matching von neuen, aber bereits kategorisierten Dingen und Nutzermodellen vorgenommen, um sie nach Item-User-Ähnlichkeit zu sortieren. Ich will nun am Beispiel eines fiktiven Recommenders von Zeitungsartikeln, den Konstan und Ekstrand in ihrem Kurs entwerfen, diese drei Schritte demonstrieren.

Die Kategorisierung der Dinge: Wie Konstan darlegt, ist die zentrale Frage, welche Attribute von Artikeln für die Leserinnen relevant sind: Welche »terms« sind »descriptive of our data set […] what we're looking for is a set of keywords that's descriptive of the items, that we can map to the items, and that seem related to people's preferences« (Coursera, 2017). Bei der Empfehlung von Artikeln geht es darum, Schlagwörter zu finden, die beschreiben, um was es in einem Artikel geht.[14] Diese Arbeit der »Indexierung« beziehungsweise der Akt des Coding (Stäheli, 2016, S. 15) kann auf mehrere Arten vorgenommen werden. Indexierung wird traditionellerweise von Expertinnen vorgenommen, wie es beispielsweise in Bibliotheken getan wird. Gerade wenn die Arbeit der Indexierung über eine Aufzählung technischer Eigenschaften hinausgeht und qualitative, subjektive Eigenschaften von Dingen benennen soll, werden auch in der digitalen Ökonomie Expertinnen hinzugezogen (Jannach et al., 2011, S. 52).[15] Typischerweise sind es aber nicht Expertinnen, sondern die Nutzerinnen selbst, welche die Dinge mit inhaltlichen Attributen versehen – wenn auch tendenziell auf chaotische Weise (siehe zum Beispiel Lamere (2008) zu »music information retrieval«). Die Möglichkeiten des »tagging« von Items durch normale Nutzerinnen gilt als eine der Errungenschaften des Web 2.0 (Beer & Burrows 2007; Keshet 2011 zu Folksonomies) durch welche die digitale Ökonomie die Arbeit von »prosumers« (Ritzer & Jurgenson 2010, technisch: Peters & Stock 2008, Lops et al. 2011, S. 94f.) zu Nutze machen kann (kritisch: Terranova 2000). Inhaltsbasierte Empfehlungssysteme können Empfehlungen generieren, selbst wenn sie nur eine einzige Nutzerin im System haben. Bei der Frage, wie die Artikel (oder andere kulturelle Produkte) zu ihren Attributen kommen, kann eine große Nutzerbasis aber von Vorteil sein: »Content-based techniques work without a large set of users, but they need that set of item data.

14 Auch hier bestehen Unterschiede in der Domain: Sollen Filme empfohlen werden, erscheinen Schauspielerinnen womöglich als relevante Attribute von Filmen.

15 Pandoras »music genome project« beschäftigt ein »team of trained musicologists«, die Musik hören und jeden Song mit bis zu 450 Attributen versehen. Netflix ließ Filmexperten jeden Film mit Attributen taggen: »Using large teams of people specially trained to watch movies, Netflix deconstructed Hollywood. They paid people to watch films and tag them with all kinds of metadata. This process is so sophisticated and precise that taggers receive a 36-page training document that teaches them how to rate movies on their sexually suggestive content, goriness, romance levels, and even narrative elements like plot conclusiveness. They capture dozens of different movie attributes. They even rate the moral status of characters« (Madrigal, 2014, k.S.).

And sometimes a large set of users helps you get that item data« (Coursera, 2017).

Die Repräsentation von Dingen durch eine »explicit list of features for each item« (Jannach et al 2011: 52) kann aber auch automatisiert werden. Gerade bei Nachrichtenartikeln, die schnell wieder verschwinden und vor allem aus Text bestehen, erscheint die automatische Indexierung von Vorteil. Die zentrale Differenz zu den oben beschriebenen Kategorisierungsmodi durch Expertinnen oder die Crowd besteht darin, dass die Kategorien (i.e. »features«, »attributes«) nicht mehr von außen kommen, sondern von »innen«. Im ersten Fall entscheiden Dritte, welche Kategorien ein Item beschreiben. Im zweiten Fall wird der Inhalt selbst durch ein automatisiertes »pre-processing« in beschreibende Features transformiert.[16] Die Unterscheidung von Daten und Metadaten – Daten über Daten bzw. die Repräsentation von Daten in einfacherer Form (Pomerantz, 2015) – wird damit unscharf: »everything is metadata« (Weinberger, 2007).[17]

Eine wichtige Methode zur Repräsentation von Textdokumenten ist das Vektorraummodell (vorgeschlagen von Salton et al. 1975; Jannach et al. 2011; Ricci et al. 2011, siehe Kapitel 7). Dazu werden Artikel von einer analogen, kontinuierlichen Einheit in Form eines Narrativs oder Arguments in eine »digitale«, diskontinuierliche Liste von Attributen transformiert (siehe Stäheli 2016 zur Transformationsarbeit der Indexierung). Die Einheit des Artikels wird in eine Liste von Wörtern aufgebrochen, so dass ein Artikel in der Datenbank durch eine Liste von Attributen repräsentiert wird. Die Attribute sind alle Wörter, die im Korpus vorkommen (d.h. das Vokabular des Gesamtkorpus) – repräsentiert durch eine 1 oder eine 0 (oder eine Zahl, die über die relative Relevanz des Worts im Text Auskunft gibt), je nach dem, ob sie in einem bestimmten Text vorkommen oder nicht. Dieser »bag of words«-Ansatz behandelt Wörter als Einheiten, deren semantische Beziehung zu vorhergegangenen und nachfolgenden Wörtern gekappt wurde (vgl. Burrell 2016,

16 Nicht nur Texte, sondern auch Audioinhalte können maschinell in Features transformiert werden, wie an der Differenz zwischen Pandora und Spotify exemplarisch verdeutlicht werden kann: Pandora lässt Musikstücke von »musicologists« kategorisieren; Spotify nutzt die Software von EchoNest, die formale Charakteristiken von Musikstücken in Features übersetzt.

17 Das zeigt sich beispielsweise daran, dass die Suchrichtung umgekehrt werden kann: Die klassische Suche nach einem Musikstück verfährt über Titel, Künstlerin, Album – klassische Metadaten. Die Mobile-App Shazam (o.Ä.) »hört« (d.h. transformiert) einige Takte des Stückes, um daraufhin Künstlerin und Titel des Stückes auszugeben.

S. 7–8, 9): Ein Dokument mit dem Inhalt »Franz liebt Fritz« wird dadurch ununterscheidbar von »Fritz liebt Franz«.

Nicht alle Wörter kommen als Attribute in Frage: Wörter sind nicht gleich Wörter. Bestimmte Wörter werden transformiert oder entfernt: Im Bereich des »natural language processing« spricht man von Lemmatisierung. Wörter werden auf ihre Grundform reduziert (aus Häuser, Hauses, hausen, etc. wird die Grundform Haus) und »stopwords« – d.h. Wörter wie »der«, »eine«, »und«, etc. – werden entfernt, da sie in allen Texten sehr oft vorkommen, aber nicht bedeutungstragend seien.[18]

Sind diese Reduktionsmaßnahmen durchgeführt, ist ein Artikel durch die in ihm vorkommenden Lemmata definiert. Sie lassen sich nun auf verschiedene Arten relationieren, um herauszufinden, wie wichtig sie für den vorliegenden Artikel sind. Zwei »naive« Möglichkeiten bestünden darin, einfach zu zählen, wie oft ein Wort vorkommt, oder binär mit 1 oder 0 zu codieren, ob ein Wort vorkommt oder nicht (Coursera, 2017). Je nach Domain mag es ausreichen, die binäre Option zu wählen. Bei Texten gilt die Annahme: Je öfter ein Wort in einem Text vorkommt, umso besser charakterisiert es diesen Text. Andererseits scheint ein Wort, das im Gesamtkorpus verhältnismäßig selten vorkommt, für einen Text, in dem es vorkommt, umso aussagekräftiger zu sein. Diese beiden Annahmen stecken in der Formel zur Berechnung des TF-IDF-Scores (»term frequency–inverse document frequency«). »Term frequency« *zählt*, wie oft ein Wort in einem Text vorkommt. »Inverse document frequency« ist ein *Maß* dafür, wie selten ein Wort im Gesamtkorpus ist.

> What TF-IDF does as a concept is create a profile of a document that says: Hey, here's an indicator for each keyword, tag or term in the document of how important this term is as a descriptive term for this document. This book you just said that you like, well it's very much about the Civil War, and a little bit about romance, and a lot about General Grant, and a little bit about Robert E. Lee. (Coursera, 2017)

TF-IDF ist ein algorithmischer Score zur automatischen Indexierung von Texten. Er bewertet, welche Wörter einen Text besonders auszeichnen. Ein

18 Verschiedene Probleme, die dabei entstehen, werden hier nicht behandelt: Wörter oder Phrasen, die in Kombination mit anderen auftauchen oder die höhere Relevanz von Titeln und Überschriften.

Wort, das nur in einem Text vorkommt und dort gleich mehrmals, gilt dementsprechend als für diesen Text charakteristisch – möglicherweise werden die Wörter auch ordinalisiert und zum Beispiel nur die wichtigsten Wörter weiter verwendet (Jannach et al., 2011, S. 57).

TF-IDF macht die Relevanz von Wörtern in Texten im Verhältnis zum Gesamtkorpus vergleichbar. Das heißt aber noch nicht, dass die Texte untereinander vergleichbar wären. Interessanterweise verhindert nicht ein qualitatives Merkmal wie beispielsweise Stil, die Vergleichbarkeit, sondern ein quantitatives: Textlänge. Das selbe Wort taucht in langen Texten tendenziell öfter auf als in kurzen. Und: Ein langer Text enthält tendenziell mehr verschiedene Wörter als ein kurzer. Längere Texte haben deshalb eine höhere Wahrscheinlichkeit, der Leserin empfohlen zu werden (Singhal et al., 2017). Um »Äquivalenz« von kürzeren und längeren Texten herzustellen, wird »normalisiert«: Wörter in einem längeren Text (ein Text mit mehr Attributen) erhalten niedrigere Gewichtungen, die im Verhältnis zur Anzahl Attribute stehen: Es wird so getan, als ob jeder Text (oder jeder Vektor) gleich lang wäre.

Aus der Sicht eines inhaltsbasierten Empfehlungssystems ist jedes Item durch einen Punkt in einem multidimensionalen Raum repräsentiert. Dieser Raum hat so viele Dimensionen, wie unterschiedliche Item-Attribute (hier: lemmatisierte Wörter) über alle Items hinweg aufsummiert vorkommen.

Das Nutzerprofil: Inhaltsbasierte Empfehlungssysteme sind eng mit Systemen zur Informationssuche – wie Bibliothekskataloge sowie Datenbank- und Websuche – verwandt. Solche Systeme müssen auch dann funktionieren, wenn sie nur eine Nutzerin haben und (fast) nichts über sie wissen.[19] Bei der Suche in einem Bibliothekskatalog oder in einer Datenbank bilden die eingegebenen Suchbegriffe sozusagen ein kurzfristiges Nutzerprofil. Sucht die Nutzerin nach den Begriffen »Pferde in der Landwirtschaft« wird sie als Vektor repräsentiert, der ihr Suchinteresse analog zu den formalisierten Items im Katalog darstellt. Die Nutzerin beziehungsweise ihr kurzfristiges Profil wird zu einem Item, das mit anderen Items verglichen werden kann. Die direkte Eingabe von Suchbegriffen nach dem »information retrieval«-Modell ist aber nur eine Variante, wie Nutzerprofile generiert werden. In-

19 Es besteht hier in mehreren Hinsichten auch eine Verwandtschaft zu Produktassoziationen (Apriori-Algorithmen bei Mackenzie 2018), die ausgehend von einem ausgewählten Produkt weitere Produkte empfehlen, die »oft zusammen gekauft« wurden oder andere Gemeinsamkeiten aufweisen.

Tab. 1: Suchanfrage als Nutzerprofil

	wort-1…	landwirtschaft	pferd	…wort-n
Suchanfrage1	0	1	1	0

haltsbasierte Empfehlungssysteme können auch längerfristige Interessensprofile generieren, die als Filter funktionieren, um aus allen möglichen Items jene herauszupicken, die den breiten Interessen einer Nutzerin entsprechen. Die Idee eines solchen personalisierten inhaltsbasierten Empfehlungssystems popularisierte Nicholas Negroponte, Direktor am MIT, am fiktiven Beispiel des »Daily Me«, einer Zeitung mit der Auflage von 1.

> What if a newspaper company were willing to put its entire staff at your beck and call for one edition? It would mix headline news with ›less important‹ stories relating to acquaintances, people you will see tomorrow, and places you are about to go to or have just come from. It would report on companies you know. In fact, under these conditions, you might be willing to pay the Boston Globe a lot more for ten pages than for a hundred pages, if you could be confident that it was delivering you the right subset of information. You would consume every bit (so to speak). Call it The Daily Me. (Negroponte, 1996, S. 153)

Die Idee koinzidierte mit der am MIT entwickelten Software Fishwrap, die es den Studierenden und Angestellten des MIT ermöglichte, News online zu lesen und ihren Interessen anzupassen (Chesnais et al. 1995; aber auch andere inhaltsbasierte Recommender zu dieser Zeit, z.B. Krakatoa Chronicle).[20] Erstmalige Nutzerinnen von Fishwrap mussten verschiedene Fragen beantworten, aus denen ihr Nutzerprofil generiert wurde (Chesnais et al. 1995, S. 275, Harper 2009).

Statt die Nutzerinnen direkt nach ihren Interessen zu befragen, können komplementär dazu auch explizite oder implizite Bewertungen (z.B. Verweildauer) von Artikeln Teil des Profils werden. Ein Beispiel: Eine Nutzerin gibt einem Artikel zu Pferden in der Landwirtschaft und einem zur Zucht

20 Die Computerwissenschaftliche Erforschung und wirtschaftliche Anwendung von Empfehlungssystemen gehen Hand in Hand (vgl. Cohn 2019 zu Netperceptions, mitbegründet durch Joseph Konstan und John Riedl), wie ich auch in meiner ethnografischen Feldforschung in der Schweiz beobachten konnte.

von Pferden je einen Upvote und einem Artikel zu Traktoren in der Landwirtschaft einen Downvote: »[w]e're going to model that as a vector and that vector will be folded into the user's profile« (Coursera, 2017).

Tab. 2: Item-Vektoren (ungewichtet)

	wort-1…	zucht	land-wirtschaft	pferd	traktor	..wort-n
item-1	0	0	1	1	0	0
item-2	0	1	0	1	0	0
item-3	0	0	1	0	1	0

Im Referenzbeispiel verwendet Konstan »binary accumulation«, um das Nutzerprofil anzupassen: Items mit positiven Bewertungen werden addiert, solche mit negativen subtrahiert. Grundsätzlich lassen sich also drei Modi un-

Tab. 3: Nutzerprofil

	wort-1…	zucht	land-wirtschaft	pferd	traktor	..wort-n
user-1	0	1	0	2	-1	0

terscheiden, wie Interessen ins Nutzerprofil kommen: Per Eingabe im Suchfeld, per Fragebogen oder per expliziter oder impliziter Bewertung. In allen drei Fällen resultiert jeweils ein spezifischer Vektor oder ein »bag of words«, der ein formales Äquivalent der Item-Vektoren darstellt.

Matching: Sobald Items in ihre Bestandteile zerlegt und ein kurz- oder langfristiges Nutzerprofil erstellt wurde, können Items und Interessen gematcht werden. Oftmals wird dazu die »cosine similarity« verwendet. Dieses Ähnlichkeitsmaß beinhaltet bereits eine Normalisierung der Textlänge. Zwei Vektoren – hier: Item-Vektor und Profil-Vektor – werden zueinander ins Verhältnis gesetzt, indem der Winkel zwischen ihnen gemessen wird. Je kleiner der Winkel zwischen den Vektoren, desto höher die Ähnlichkeit der beiden Vektoren und, so die Annahme, umso höher die Ähnlichkeit zwischen Text und Interesse der Nutzerin. Dieses Ähnlichkeitsmaß lässt sich für jede Kombination von Nutzerprofil und Item berechnen, um die Items hinsichtlich ihrer Ähnlichkeit zum aktiven Nutzerprofil vergleichen zu können und

sie in der Empfehlungsliste entsprechend zu positionieren.[21] Statt einer Relationierung von neuen Items und Nutzerprofil könnte auch nach den fünf ähnlichsten, von der Nutzerin bereits bewerteten, Items gesucht werden. Wurden vier von fünf Items positiv bewertet, wird das entsprechende Item präsentiert (Jannach et al., 2011, S. 58f.).

Inhaltsbasierte Empfehlungen gelten als »schwach personalisiert«. Matchings werden über paarweise Ähnlichkeitsberechnungen hergestellt, die dann nach den höchsten Ähnlichkeitswerten geordnet werden können. Demografische Zugehörigkeiten oder User-User-Affinitäten spielen keine Rolle (auch wenn bei Fishwrap davon ausgegangen wird, dass sich die Nutzer für News ihres Heimatortes interessieren). In dieser Hinsicht ließen sich inhaltsbasierte Empfehlungen auch als personalisierteste aller Empfehlungen beschreiben, da sie die Nutzerin »in splendid isolation« (Vgl. Prey 2018, hier: S. 1090) betrachten.

Bei inhaltsbasierten Empfehlungssystemen geht es aber nur am Rand um Nutzerinnen. Im Zentrum steht die Katalogisierung von Dingen. Nehmen wir an, wir haben einen Katalog von drei Filmen: »Crazy Rich Asians«, »Forrest Gump« und »Matrix«. Expertinnen haben die Filme nach Romantiklevel und Schlüssigkeit des Plots bewertet. »Crazy Rich Asians« ist sehr romantisch und sehr schlüssig; »Matrix« und »Forrest Gump« haben beide einen relativ hohen Romantikwert, »Forrest Gump« ist aber einiges schlüssiger. Wir haben nun eine Nutzerin, die sich den Film »Crazy Rich Asians« angesehen hat. Wir wollen ihr eine Empfehlung machen, welchen Film sie sich als nächstes anschauen soll. Dieser Film soll »Crazy Rich Asians« – das temporäre Nutzerprofil – möglichst ähnlich sein.

Um der Nutzerin eine Empfehlung zu machen, matcht das System das Profil der Nutzerin – i.e. »Crazy Rich Asians« – mit allen zur Verfügung stehenden Filmen. Konkret berechnen wir, wie ähnlich sich »Crazy Rich Asians« und »Matrix« und wie ähnlich sich »Crazy Rich Asians« und »Forrest Gump« sind. Die Filme lassen sich dazu in einen Vektorraum übertragen. Damit das Beispiel anschaulich bleibt, haben wir uns auf die zwei Dimensionen der Romantik und der Schlüssigkeit beschränkt. Bei Netflix hätten

21 Die Anordnung der Resultate muss nicht unbedingt nach dem Kriterium der Ähnlichkeit geschehen. Auch hier: Die Ordnung der Prediction muss nicht der Ordnung der Recommendation entsprechen. Beispielsweise könnte zusätzlich nach einem Autoritätsmaß (Cardon, 2016) wie dem plutokratischen PageRank (Introna, 2007, S. 19), einem Popularitätsmaß oder einem Aktualitätsmaß gerankt werden.

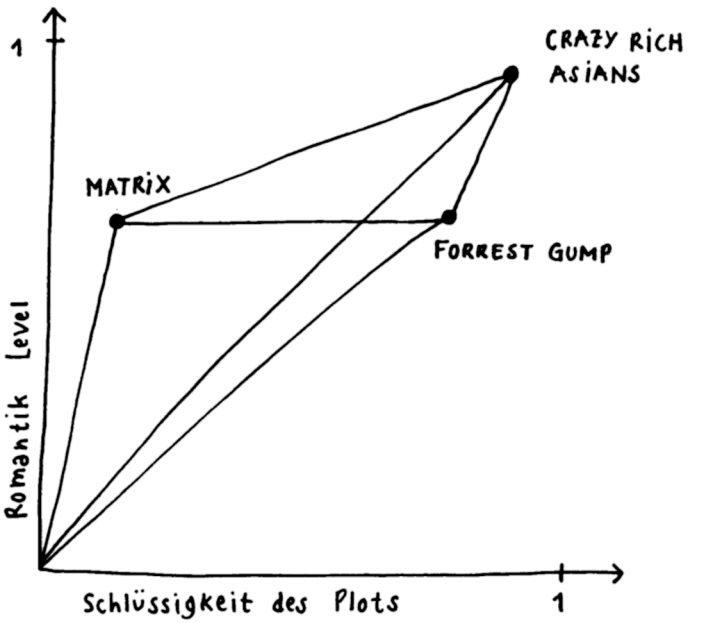

Abb. 9: Hypothetischer Vektorraum eines inhaltsbasierten Empfehlungssystemes

wir einen Raum mit mehreren Dutzend Dimensionen. Wenn wir die Filme als Punkte in einem zweidimensionalen Raum betrachten, geben wir ihnen eine quantitative Form, so dass sich Ähnlichkeiten berechnen lassen. Beispielsweise indem wir den Winkel zwischen den Vektoren messen oder die Distanz zwischen den Punkten. Dann kommen wir zum Beispiel auf Ähnlichkeitsscores von 0.2 für »Crazy Rich Asians« und »Forrest Gump« und 0.8 für »Crazy Rich Asians« und »Matrix«. Wir empfehlen also »Forrest Gump« als nächsten Film, weil dieser näher liegt.

Inhaltsbasierte Empfehlungssysteme scheinen einzelne Nutzerinnen in Isolation zu betrachten, tatsächlich behandeln sie Nutzerinnen aber als äquivalent zu Dingen.

10.4 Collaborative Filtering

Algorithmische Sozialität

Stereotypisierende Empfehlungen brauchen kategoriale Informationen über die Nutzerinnen. Inhaltsbasierte Empfehlungen brauchen Informationen über die »Substanz«, i.e. den Inhalt von Dingen. Collaborative Filtering braucht weder das eine noch das andere.

> In Collaborative Filtering, we ignore the user and item attributes. We don't care what's in the item. We don't care who the user is. We only look at the interactions between users and items. And we mine patterns from these, such as looking at what people like you also bought. (Coursera, 2017)

Mit »interactions« sind hier Bewertungen durch die Nutzerinnen gemeint. Das klassische Collaborative Filtering benötigt solche Bewertungen als Grundlage aller Berechnungen. Die Nutzerinnen werden entweder dazu aufgefordert, explizite Bewertungen abzugeben, oder aus Verhaltensdaten werden implizite Bewertungen abgeleitet (Seaver, 2018). Damit radikalisiert Collaborative Filtering das schon in der Stereotypisierung angelegte dualistische Verständnis von Individuen: Eine Nutzerin ist durch ihre Bewertungen von Dingen definiert; ein Ding durch seine Bewertungen der Nutzerinnen (Seaver, 2012). John Riedl und Joseph Konstan sehen in diesem Verzicht auf Identitätskategorien eine Chance, Empfehlungen nicht mehr auf diskriminierende Stereotypisierungen basieren zu müssen und grenzen dies explizit von Segmentationstechniken des Marketings ab:

> The problem is, simple demographics don't begin to tell the story of individuals. People who like chess can also like football and chili contests. People cross taste lines all the time if they're permitted to. […] The vestiges of mass marketing still hound us, however. Businesses insist on feeding the same products and ads to people who share the same age, race, gender, and class. […] If it's drilled into us that because we're young, white, middle-class girls we should drink Pepsi and like Britney Spears, we probably will. But it doesn't mean, in a world where we weren't pigeonholed, that we wouldn't buy something else. […] Racial profiling and profiling your customers both spring from the same lazy, prejudiced philosophy. (Riedl & Konstan, 2002, S. 112f.)

Stereotypisierung beruht auf einer Relationierung von Nutzerinnen und Dingen über kategoriale Zugehörigkeit. Wie aber wird eine Relation hergestellt, ohne sich auf Kategorien zu berufen? Collaborative Filtering eruiert »patterns of data« über zweifache Matchings – eine spezifische Variante des Vergleichs, die die paarweise Relationierung von Einheiten (Nutzerinnen, Dinge) als Vergleichskriterium verwendet: Erstens durch ein Matching der Nutzerinnen untereinander (user-user), um »Nachbarschaften« festzulegen; zweitens durch ein Matching von Nutzerinnen und Items, um Empfehlungen zu generieren (user-item).

Im ersten User-User-Matching wird eine Nutzerin Alice mit allen anderen Nutzerinnen Barbara, Carla und Nora paarweise in Relation gesetzt, um ihre Ähnlichkeit mit anderen Nutzerinnen festzustellen. Es wird berechnet, wie ähnlich sich Alice und Barbara, Alice und Carla und Alice und Nora, usw. sind. Ähnlichkeit mit Alice wird zum relationierenden Vergleichskriterium, durch das Barbara, Carla und Nora verglichen und als temporäre Nachbarinnen (»people like you«) von Alice »kategorisiert« und von »people unlike you« unterschieden werden können. Die Nachbarinnen bilden die Berechnungsgrundlage für den zweiten Schritt.

Im zweiten User-Item-Matching wird ein Item ausgewählt, das Alice noch nicht bewertet hat, ihre Nachbarinnen aber schon. Aufgrund der Bewertungen ihrer Nachbarinnen wird eine Vorhersage errechnet, die darüber Auskunft gibt, wie gut Alice dieses Item bewerten wird. Diese Berechnung wird für alle Items, die Alice noch nicht bewertet hat, durchgeführt. Noch nicht bewertete Items können jetzt anhand des Kriteriums der höchsten, vorhergesagten Bewertung durch Alice verglichen werden. Jenes Item mit der höchsten vorhergesagten Bewertung wird Alice auf der Benutzeroberfläche präsentiert.

Steht Stereotypisierung exemplarisch für ein askriptives Verständnis von Individuen, so drückt sich im hier präsentierten Empfehlungsmodus des User-User-Collaborative Filtering ein algorithmisches Verständnis von Individuen aus (vgl. Fisher & Mehozay 2019; Cheney-Lippold 2011). Im Gegensatz zu den relativ fixen und benennbaren Identitätskategorien, transformieren sich algorithmische Identitäten »from one click to the next« (Cohn, 2019, S. 46) und tauchen gar nicht erst als benennbare Kategorien auf. Sobald Alice und die anderen Nutzerinnen zusätzliche Bewertungen abgeben, lassen sich Nachbarschaften und vorhergesagte Bewertungen neu berechnen. Nachbar-

schaften mögen sich durch eine relative Persistenz auszeichnen.[22] Die »patterns of data« sind aber prinzipiell fluid und nicht durch Grenzen sozialer Kategorien fixiert. Alice wird bei der nächsten Empfehlungsberechnung möglicherweise neue Nachbarinnen dazugewinnen und andere verlieren.

Collaborative Filtering beinhaltet nicht nur eine algorithmische Identität, sondern vor allem auch eine algorithmische Sozialität (Vgl. Bucher 2013, S. 490, Alaimo & Kallinikos 2017, zu »computed sociality«). Genügt inhaltsbasierten Empfehlungen eine einzige Nutzerin in »splendid isolation«, ist hier die Relationierung zu allen anderen der Ausgangspunkt einer Empfehlung. Personalisierung im Sinne einer singularisierenden Adressierung auf der Nutzungsoberfläche beruht ironischerweise auf algorithmischer, technisch hergestellter Sozialität. Einer Sozialität, die in digitalen Infrastrukturen ohne das Wissen der Nutzerinnen hergestellt wird. Nutzerinnen, die nie miteinander in direkten Kontakt getreten sind oder jemals treten werden, »kooperieren« insofern, als ihre Bewertungen von Dingen zur Ressource für Empfehlungen für andere werden. Collaborative Filtering ermöglicht Kollaboration ohne Gemeinschaft. Die Empfehlungen für Alice kommen nur zustande, weil andere Nutzerinnen ebenfalls Dinge bewerten.

Matchings als spezifische Form des Vergleichs verdeutlichen, dass Nutzerinnen auf der Hinterbühne des User-User-Collaborative-Filtering (UUCF) nur in Relationen vorkommen. Um zu sehen, was mit algorithmischer Sozialität gemeint ist – was das Kollaborative an Collaborative Filtering ist –, ist es erhellend, die ersten Collaborative-Filtering-Empfehlungssysteme anzuschauen, die als Online-Interessensgemeinschaften konzipiert wurden. Eine frühe, nicht-automatisierte Variante von Collaborative Filtering war »Tapestry«, ein Empfehlungssystem für Dokumente.

> Collaborative filtering simply means that people collaborate to help one another perform filtering by recording their reactions to documents they read. Such reactions may be that a document was particularly interesting (or particularly uninteresting). These reactions, more generally called annotations, can be accessed by others' filters. (Goldberg et al., 1992, S. 61)

22 Die tatsächliche, positive Bewertung einer Empfehlung macht Alice möglicherweise ihren Nachbarinnen ähnlicher, d. h. sie erhält weiterhin Empfehlungen, die auf ungefähr derselben Nachbarschaft beruhen. Das ist das Filterblasen-Argument (Pariser, 2011).

Diese frühe Form des Collaborative Filtering hebt dessen soziale Komponente hervor (Cohn, 2019, S. 72 ff.): Der Filter ist ein kollektives Produkt, das ermöglicht wird, indem unterschiedliche Leute mit den Dokumenten »interagieren« (i.e. Kommentare schreiben, antworten, weiterleiten, etc.). Vorausgesetzt ist, dass Nutzerinnen für andere Nutzerinnen identifizierbar und untereinander bereits bekannt sind:

> you know that Smith, Jones and O'Brien read all of comp.unix-wizards newsgroup material, and reply to the more interesting documents. Tapestry allows you to filter on ›documents replied to by Smith, Jones, or O'Brien‹. (Goldberg et al., 1992, S. 62)

Tapestry verlangte von den Nutzerinnen, explizite Filter zu definieren, indem sie sich auf bekannte und vertrauenswürde Beziehungen berufen: Es brauchte »human effort to establish the relationship between the people making and the people receiving the recommendation« (Riedl & Konstan, 2002, S. 5f.).

Das Musikempfehlungssystem RINGO und seine Weiterentwicklung Firefly automatisierten »word of mouth«-Empfehlungen: »instead of having to ask a couple friends about a few items, a social information filtering system can consider thousands of other people« (Shardanand & Maes, 1995, S. 211). Pattie Maes entwickelte RINGO zu Firefly weiter, das zu einem der ersten sozialen Netzwerke werden sollte. Collaborative Filtering, so Maes, habe großes Potenzial in »fostering community« (zitiert in: Cohn 2019, S. 74). Um das Potenzial auszuschöpfen, ermöglichte es Firefly über eine Chat-Funktion mit ähnlichen Nutzerinnen Kontakt aufzunehmen (Cohn, 2019, S. 74). In diesem Sinne empfahl Firefly nicht nur Musik, sondern auch ähnliche Nutzerinnen. Der Begriff Collaborative Filtering kommt daher, dass einige dieser frühen Systeme tatsächlich als Interessensgemeinschaften konzipiert waren: Nutzer kollaborierten miteinander, um Empfehlungen für Musik und andere Dinge auszutauschen – vermittelt über das Empfehlungssystem.

Im Unterschied zu Bestenlisten, von denen ich auch dann profitieren kann, wenn ich selbst keine Bewertungen abgebe, ist das bei Collaborative Filtering gerade nicht möglich. Und im Unterschied zu inhaltsbasierten Empfehlungen reicht es nicht aus, wenn das System über mein Profil verfügt, ohne Informationen über andere zu besitzen. Die Relationierungsmethode der Stereotypisierung, die anhand kategorialer Zugehörigkeiten auf Präferenzen schließt, erscheint gegenüber Collaborative Filtering als diskriminierend.

User-User-Matching ist eine algorithmische Form der Relationierung, die mich (verstanden als »interactions« mit Dingen) und andere (verstanden als deren »interactions« mit Dingen) in Bezug auf »Ähnlichkeit« vergleicht. Relationen zu ähnlichen Nutzerinnen dienen als Ressource für die Berechnung von Empfehlungen. Am Beispiel von Tapestry – das erste Empfehlungssystem, das unter der Bezeichnung Collaborative Filtering figurierte – lässt sich erkennen, dass diese Art der Empfehlung auf sozialen Beziehungen beruht. Collaborative Filtering automatisiert die Erzeugung dieser Beziehungen unter dem Motto: Wir wissen nicht, ob zwischen euch beiden eine Beziehung besteht. Basierenden auf euren Gewohnheiten, eurem Geschmack oder Ähnlichem tun wir aber so, als ob es eine Beziehung gäbe, um daraus eine Empfehlung zu generieren. Algorithmische Relationen treten an die Stelle sozialer Beziehungen. Wie Bucher für Facebook argumentiert, haben unsere digitalen Freundschaften Konsequenzen dafür, was uns angezeigt wird:

> the [digital] connections we forge with other people may have real consequences as the conditions of the intelligible and sensible is increasingly calculated on the basis of who our friends are, what they have done, and how many of them there are. (Bucher, 2013, S. 490)

Collaborative Filtering nimmt uns sogar die Aufgabe ab, digitale »Freundschaften« selbst zu artikulieren.

Herstellung von Vergleichbarkeit

Die personalisierte Empfehlung basiert nicht auf einem abstrakten und dekontextualisierten Verständnis von Qualität wie in Kapitel 10.1 gesehen sondern auf der Annahme, dass es sinnvoll ist, unterschiedlichen Nutzerinnen unterschiedliche Dinge zu empfehlen. Das System sucht auf der Basis bereits aufgezeichneter Verhaltensweisen wie Bewertungen – »interactions between users and items« – nach ähnlichen Nutzerinnen und empfiehlt mir Dinge, die solche ähnlichen Nutzerinnen positiv bewertet haben. Um mir ein Produkt zu empfehlen, ist es nicht entscheidend, dass möglichst viele Nutzerinnen etwas positiv bewertet haben, sondern was mir ähnliche Nutzerinnen mögen. Anstatt mir die Nummer eins der aktuellen Hitparade vorzuschlagen, würde mir ein UUCF ein Lied vorschlagen, das Nutzerinnen positiv bewertet haben, die mir in Bezug auf ihre Bewertungen und ihr Hörverhalten ähnlich sind (zu Empfehlungssystemen im Bereich der Musik siehe Passoth et al. 2014; Prey 2016).

Ich will nun aufzeigen, wie Formalisierung, Bewertung und Vergleich im Prozess eines UUCF ineinander verschachtelt werden. Dabei möchte ich sichtbar machen, auf welche Weise verschiedene Arten der Kategorisierung und des Vergleichs aufeinander aufbauen (Fourcade & Healy, 2017a). Zweitens zeige ich, analog zum Beispiel von Reddit, dass und wie Vergleichbarmachung im Prozess der Quantifizierung stattfindet. Eine 5 ist eben nicht immer eine 5. Und drittens werde ich darstellen, wie UUCF von Relationen zwischen Nutzerinnen ausgeht, diese Relationen aber neue sortiert und angeordnet werden, so dass am Ende Empfehlungen als Relationen zwischen Dingen präsentiert werden können.

Joseph Konstan, Computerwissenschaftler an der Universität Minnesota, erklärt die Funktionsweise eines UUCF für Filme folgendermaßen:

> The story is very simple. Let's say we want to build a recommender system for movies. So we have a *bunch of people*. They're all sitting around here, sometimes those people are in a role where they're *rating movies*. They're inputting into our system their opinion of some movie. And when we get that rating, we're gonna *store it in a database* and we're also going to *compute the pairwise correlations between all pairs of people* in our system. That correlation is a measure of agreement between people and it's something that's very easy to maintain and incrementally update as you go. Now once we have that pairwise correlation table and we have the ratings we wait for somebody to ask for something. So somebody comes in and makes a request. They might just say find me a good movie, or they might say what am I gonna think about this particular movie. When that request comes in, the first thing we're gonna do is we *look at those correlations to find a good neighborhood*. Who are the people who are most like – in their past opinions – this person who is asking the question? Once we have that neighborhood, we can look up those people's ratings, and our collaborative filtering engine really does just three things: *it takes these ratings and it normalizes them*, it weights them by the correlations and then it combines them and denormalizes them back to the user's scale«. (Konstan, 2012, 31′–34′)

Anhand von Konstans Erklärung greife ich nun drei Aspekte auf, die ich in diesem Zitat kursiv hervorgehoben habe: Wie »Leute« zu Usern werden (I), wie Ähnlichkeit zwischen Usern konstruiert wird (II) und wie unterschiedliche Bewertungsweisen mathematisch gleich gemacht werden (III).

I) »have a bunch of people […] rating movies. […] store ratings in a database«

Die Grundlage von Empfehlungssystemen sind Datenbanken, in denen Relationen zwischen Dingen und Nutzerinnen, hier: zwischen Filmen und Zuschauern, erfasst werden. Es dürfte zwar jede Kinogängerin Meinungen und Lieblingsfilme haben: Für das Recommendersystem einer Internetplattform sind diese Relationen aber weder verfügbar noch beobachtbar und auch nicht in der Sprache der Datenbank verfasst. Nur Nutzer der Plattform lassen sich in den Vergleichsraum inkludieren. Damit aus einem »Haufen Leute« Nutzerinnen (fortan: User) werden, müssen Personen Profile anlegen. Durch das Anlegen eines Benutzerprofils und das Login werden Leute zu Usern und damit für die Datenbank sichtbar – wenn auch ein UUCF damit alleine noch keine Vergleiche anstellen kann.

Im Alltag ist eine Vielzahl verschiedenster Relationen zwischen Filmen und ihren Zuschauern möglich: Jemand ist ein Fan und schaut sich denselben Film jeden Tag an. Jemand anderes hasst den Film und schreibt vernichtende Blogposts darüber. Ein anderer empfiehlt den Film einer Freundin weiter. Im Gegensatz dazu ist für das UUCF-Empfehlungssystem nur eine dieser Relationen sichtbar und relevant: die formalisierte Bewertung im System. Beispielsweise können User auf der Webseite von Movielens.org Filme mit 1 bis 5 Sternen bewerten. Das Verhältnis von User und Film muss über die Benutzeroberfläche und die dahinterliegende Datenbank auf formale Weise erfasst werden: »Their opinion of some movie« ist nur als Rating von 1 bis 5 Sternen relevant, alles andere bleibt unsichtbar.[23]

Auch wenn Personen durch den Akt des Logins zu »Usern« werden: Um personalisierte Empfehlungen zu bekommen, müssen sie erst selbst Filme bewerten. Ein neuer »User« – also jemand, der sich ein Benutzerkonto zugelegt hat – ist erst dann ein User (ohne Anführungszeichen) im Sinn des Recommendersystems, wenn »interactions between users and items« vorliegen: In der Fachsprache wird dies das Kaltstart-Problem genannt (siehe: Schein et al. 2002, Kapitel 6). Das Problem kann durch verschiedene Strategien gelöst werden: Beispielsweise zeigt Movielens.org Popularitätsmetriken an, die dem »User« nicht-personalisierte Empfehlungen anzeigen (zum Beispiel Filme mit der besten durchschnittlichen Bewertung). Sobald ein User 15 Filme bewertet hat, sind personalisierte Empfehlungen möglich. Netflix zeigt neuen »Usern« beim ersten Login Listen von Filmen an, die sie eventuell

23 Zumindest in einem basalen UUCF, das nur Ratings berücksichtigt.

schon gesehen haben und bewerten können. Eine weitere Möglichkeit wäre die Verwendung traditioneller Kategorien – wie zum Beispiel auf Geschlecht oder Alter basierende Empfehlungen.

Um Differenzen konstatieren zu können, müssen User insofern gleich sein, als dass sie eine minimale Anzahl von Items bewertet haben. Erst das macht sie zu Usern, die mit anderen Usern vergleichbar sind. Vorher sind sie nur eine leere Zeile in einer User-Item-Kreuztabelle oder ein leerer Datenbankeintrag. User_a hat in der folgenden Tabelle beispielsweise keine Bewertung abgegeben und bekommt deshalb keine personalisierten Empfehlungen. User_b hat hingehen bereits einige Filme bewertet und kann zum Beispiel für das noch nicht bewertete item_1 eine Empfehlung erhalten (falls dafür genügend Bewertungen vorhanden sind).

Tab. 4: User-Item-Ratings

	item-1	item-2	item-3	item-n
user-a	?	?	?	
user-b	?	3	5	
user-c	1	4	5	
user-m				

Was Konstan als relativ unproblematisch voraussetzt – »have a bunch of people […] rating movies« –, erweist sich als voraussetzungsreicher Formalisierungsprozess: Bevor überhaupt eine Empfehlung berechnet werden kann, müssen also User *definiert*, Verhaltensweisen *formalisiert* und in ausreichender Menge *registriert* werden.

II) »compute pairwise correlations [and] find a good neighborhood«

Die Berechnung einer Vorhersage, wie ein user_a den Film »Wonder Woman« bewerten wird, basiert nicht auf den Bewertungen aller anderen User, die »Wonder Woman« bereits gesehen und bewertet haben. Nur User, die user_a genügend ähnlich sind – »people like you« bzw. »Nachbarn« – und »Wonder Woman« bereits bewertet haben, bilden die Grundlage der Berechnung der Empfehlung. Bevor die Empfehlung für »Wonder Woman« für user_a berechnet werden kann, wird festgelegt, auf der Basis welcher anderen User in der »Nachbarschaft« von user_a diese Berechnung durchgeführt werden soll.

Aber wie werden »ähnliche« User identifiziert und von »nicht-ähnlichen« unterschieden?

Die User wurden durch die Formalisierung ihres Bewertungsverhaltens bereits vereinheitlicht: Es lassen sich nur User vergleichen, deren Verhaltensweisen als Bewertungen von 1 bis 5 registriert werden und die eine Mindestmenge an Bewertungen vorgenommen haben. Das ist die Voraussetzung für die folgenden Kategorisierungsprozesse, welche die Ähnlichkeit für jedes User-Paar messen. Darauf basierend lassen sich »ähnliche« und »nicht-ähnliche« User kategorial unterscheiden: Nachbarinnen, oder eben nicht. Im Falle eines UUCF entsteht »Ähnlichkeit« erst durch eine Kombination unterschiedlicher »classificatory judgements« (Fourcade, 2016b).

Bevor eine konkrete Empfehlung berechnet werden kann, muss in einem vorhergehenden Schritt der Pearson-Korrelationskoeffizient für alle Userpaare bestimmt werden, um die hinreichend ähnlichen Nachbarn für die Berechnung der konkreten Empfehlung für den Film »Wonder Woman« zu selektieren. Dieses Ähnlichkeitsmaß soll darüber Auskunft geben, wie ähnlich sich zwei User sind. Um den Pearson-Korrelationskoeffizient überhaupt berechnen zu können, müssen zwei User in einem weitergehenden Sinn hinreichend »gleich« sein: Sie müssen nicht nur für sich eine Mindestmenge an Bewertungen abgegeben haben. Es braucht auch Überschneidungen der bewerteten Filme. Mindestens zwei gleiche Filme müssen sowohl von user_a als auch von user_b bewertet sein (auch wenn die Filme nicht gleich bewertet sein müssen). Haben sie keine Filme gemeinsam, kann nichts berechnet werden. Bei nur einer Überschneidung resultiert eine 0 beziehungsweise keine Ähnlichkeit von user_a und user_b; bei zwei gemeinsamen Filmen resultiert eine 1, d.h. eine sehr hohe Ähnlichkeit von user_a und user_b. Der hohe Ähnlichkeitswert bei nur zwei von beiden bewerteten Filmen kann aber trügerisch sein, wenn beide User noch viele weitere, aber je unterschiedliche Filme bewertet haben (Ekstrand, 2013).[24]

Für jedes Userpaar wird nun periodisch – zum Beispiel jede Woche – der Pearson-Korrelationskoeffizient berechnet. Es resultiert eine Kreuztabelle aller User-User-Kombinationen, die für jede User-User-Kombination den Ähnlichkeitswert ausweist. Ein hoher Wert steht für eine große Übereinstim-

24 Wenn zwei User viele individuelle Bewertungen abgegeben haben, aber nur wenige Filme von beiden bewertet wurden, kann eine Signifikanzgewichtung angewendet werden, welche die Ähnlichkeit abschwächt, bis eine Mindestmenge an paarweisen Bewertungen erreicht ist (Herlocker et al., 1999).

Tab. 5: User-User-Ähnlichkeit

	user-a	user-b	user-c	user-m
user-a	1			
user-b	0.7	1		
user-c	0.2	0.1	1	
user-m				1

mung der Bewertungsweisen bzw. Präferenzen der zwei in Relation gesetzten User.

Nachbarn zu selektieren heißt, aufgrund der für alle User-Kombinationen errechneten Ähnlichkeitswerte ein »nominal judgement« (Fourcade, 2016b) darüber vorzunehmen, wer die »people like you« (im Unterschied zu »people unlike you«) sind. Das heißt, der Ähnlichkeitswert wird als Vergleichskriterium verwendet, um genügend ähnliche Nachbarinnen auszuwählen. Diese ausgewählte Nachbarschaft dient als Basis für die Berechnung einer konkreten Empfehlung für einen von user_a noch nicht bewerteten Film. Dabei bestehen drei grundsätzliche Möglichkeiten, Nachbarschaften zu definieren: Alle User als Nachbarn definieren, einen Grenzwert der Ähnlichkeit festlegen oder eine begrenzte Zahl der ähnlichsten Nachbarn auswählen.

Wenn sehr viel Rechenkapazität zur Verfügung steht – oder die Datenbank nicht zu viele User oder Items beinhaltet – liegt die naheliegende Möglichkeit darin, alle User als Nachbarschaft zur Berechnung einer Filmempfehlung zu verwenden. Angesichts der Menge an Usern und Items ist diese Möglichkeit aber in vielen Fällen schon technisch ausgeschlossen: »Commercial collaborative filtering systems are beginning to handle millions of users, making considerations of every neighbor infeasible. The system must select the best neighbors, discarding the remaining users« (Herlocker et al., 1999, S. 234). Die schiere Menge möglicher Berechnungen ohne eingeschränkte Nachbarschaften ist ein Problem der Rechenkapazität, wie Konstan verdeutlicht: »All the pairwise correlations is m squared times n. You think about that and m is 100 million users, and n is a million items. Ew. That's 10'000 trillion or 10 quadrillion. Yuck. That's a lot of computation« (Coursera, 2017). Zudem könne die Verwendung aller User als Nachbarschaft dazu führen, dass die höher korrelierten Nachbarinnen im »noise« der tiefer korrelierten untergehen würden.

Es gibt zwei weitere Möglichkeiten, um Nachbarschaften auszuwählen, die »people like you« von »people unlike you« unterscheiden: Grenzwerte der Ähnlichkeit sowie Festlegen der Nachbarschaftsgröße (Herlocker et al., 1999). Wird ein Grenzwert von beispielsweise 0.5 festgelegt, sind für user_a jene User Nachbarn, deren Ähnlichkeitswerte höher liegen. Beim sogenannten Top-N-Verfahren entscheiden sich die Entwickler im Voraus für eine optimale Nachbarschaftsgröße: Entscheiden sie sich beispielsweise für Top-30, werden für user_a alle anderen User nach ihren Ähnlichkeitswerten mit user_a geordnet und die dreissig mit den höchsten Werten als Nachbarn ausgewählt, um daraus Empfehlungen zu berechnen. Die Wahl zwischen einem Grenzwert und Top-N ist ein Zielkonflikt: »If you have a minimum similarity, you may not get very many neighbors. If you limit the size, you may not have very good similarity« (C2W1-02A).

User_a unterscheidet sich von anderen Usern durch die spezifische Kombination der ihr ähnlichsten anderen User. Die Nachbarn von user_a sind mit aller Wahrscheinlichkeit nicht dieselben wie jene von user_b, obschon user_a möglicherweise Teil der Nachbarschaft von user_b ist und umgekehrt. Während im ersten Schritt User als »user-item-interactions« erfasst wurden (I.), sind sie nun über ihre User-User-Relationen, d.h. über ihre Nachbarschaften definiert.

III) »normalize ratings«

Im dritten Schritt wird das Problem adressiert, dass nicht alle User die Bewertungsskala auf die gleiche Weise verwenden. User_a ist schnell begeistert und vergibt meist gute Bewertungen. User_b ist sehr kritisch und vergibt nur selten gute Bewertungen. Um nun Ähnlichkeiten festzustellen, werden die Bewertungsweisen »normalisiert«:

> When I talk about normalization and denormalization, it's a compensation for the fact that people have different levels of positivity or enthusiasm. What we learn when we run these systems is that [for] some people [...] everything is a four or five, you know. A four means it was okay and a five means I loved it. There are other people who've never actually gotten to the number five. Most things are twos and there are a few threes and a couple of fours, and a two means it's okay, and a three means, okay I guess I might spend money on it, and a four means this is wonderful but I've never seen anything better than that. And if you don't make your adjustment it's really hard to match somebody who thinks everything is wonderful with somebo-

> dy who ehm is the other side: I think everything is pretty miserable. But if you normalize you can usually match those people if they at least agree that the most wonderful things are about the same. (Konstan, 2012, 33′–34′)

Die »interactions between [user_a] and items« und die »interactions between [user_b] and items« liegen zwar in einer gemeinsamen Metrik vor, dennoch ist eine 5 von user_a nicht mit einer 5 von user_b gleichzusetzen, wie das konstruierte Beispiel in der Tabelle 3 zeigt. Beide haben item_4 mit einer 5 bewertet. Das heißt aber nicht, dass sie sich in ihrem Urteil einig sind. Um die spezifischen Bewertungsweisen miteinander vergleichen zu können, wird für jeden User der Mittelwert all seiner Bewertungen berechnet und von den ursprünglichen Bewertungen abgezogen. Erst dann wird die »richtige« Differenz von user_a und user_b sichtbar. Obwohl beide auf den ersten Blick die gleiche Bewertung vergeben haben, ist die normalisierte Bewertung unterschiedlich: Zwischen den normalisierten Bewertungen von user_a und user_b für item_4, das sie beide mit einer 5 bewertet haben, besteht nun eine Differenz von 1,5.

Tab. 6: Nicht normalisierte Bewertungen

	user-a	user-b
item-1	5	3
item-2	4	2
item-3	4	2
item-4	**5**	**5**
Mittelwert	4.5	3

Die Bemessung der Ähnlichkeit zweier User beruht darauf, die vorliegenden Bewertungen eines Items zum Durchschnitt aller Bewertungen eines Users in Beziehung zu setzen. Die Vergleichbarkeit der User gilt erst dann als gegeben, wenn einkalkuliert wird, wie User die Ratingskalen nutzen. Auch wenn numerische Daten vorliegen, werden diese Daten mathematisch behandelt, um die Vergleichbarkeit von Usern zu gewährleisten. »Commensuration« (Espeland & Stevens, 1998, 2008) ist nicht ein Prozess, der abgeschlossen ist, sobald ein Sachverhalt in numerischer Form vorliegt. Im vorliegenden Fall heißt das, dass ich erst einmal nur meine eigenen Filmbewertungen vergleichen kann – und auch nur dann, wenn man davon aus-

Tab. 7: Normalisierte Bewertungen

	user-a	user-b
item-1	0.5	0
item-2	-0.5	-1
item-3	-0.5	-1
item-4	**0.5**	**2**
Mittelwert	0	0

geht, dass mein Bewertungsverhalten und Geschmack zeitlich stabil sind. Um die Ratings eines Users mit den Ratings anderer User vergleichen zu können, muss die Differenz der Bewertungsweisen miteinbezogen und mathematisch auf eine gemeinsame Ebene gebracht werden. Erst dann gelten Ratings verschiedener User als kommensurabel.

Das Ziel des Empfehlungssystems besteht darin, einem user_c eine oder mehrere Empfehlungen von Filmen zu präsentieren. Die Erzeugung von Empfehlungen umfasst mehrere Schritte der Vergleichbarkeitsherstellung und des Vergleichs. Wir haben mit einer voraussetzungsreichen Tabelle begonnen, die User und Items über Bewertungen zueinander in Beziehung setzt. Um überhaupt in potenziell bedeutungsvoller Weise in dieser Tabelle aufzutauchen, müssen »User« zu Usern werden, indem sie Items bewerten. Erst dann können User miteinander verglichen werden.[25] User_c muss also eine Mindestmenge an Filmen bewerten, die auch andere User bewertet haben.

Besteht eine genügend große Überschneidung von user_c und anderen Usern, kann die Ähnlichkeit dieser User berechnet werden. Es resultiert eine Kreuztabelle aller möglichen User-User-Relationen. Für jeden User können nun genügend ähnliche User ausgewählt werden, die als Basis zur Berechnung von Empfehlungen dienen sollen. Für user_c wird nun eine bestimmte Menge an Nachbarn ausgewählt, die entweder einen bestimmten Schwellenwert der Ähnlichkeit überschreiten oder, zum Beispiel, die dreissig höchsten Ähnlichkeitswerte aufweisen.

25 Das gilt auch für die umgekehrte Perspektive: Ein »Item« ist erst dann ein mit anderen vergleichbares Item, wenn es von Usern bewertet wurde.

Obwohl die User des Systems alle die gleiche Bewertungsskala verwenden müssen, sind ihre Bewertungen noch nicht mit den Bewertungen anderer User kommensurabel. Verschiedene User gelten erst dann als vergleichbar, wenn ihre Bewertungsweisen mathematisch normalisiert, das heißt ihre Differenzen – »different levels of positivity and enthusiasm« (Konstan, 2012) – eingeebnet sind.

Um nun vorherzusagen, wie user_c noch nicht bewertete Items bewerten wird, werden die Bewertungen aller Nachbarn von user_c normalisiert und miteinander verrechnet. Dabei erhalten die einzelnen Nachbarn einen nach ihrer Korrelation gewichteten Einfluss. Für jedes noch nicht bewertete Item wird diese Berechnung durchgeführt, woraus eine Liste vorhergesagter Bewertungen resultiert, die sich nach der Höhe des Scores sortieren lässt. Das Empfehlungssystem präsentiert den Film mit dem höchsten Score.

Alle hier beschriebenen Prozesse werden in Sekundenschnelle im Verborgenen abgewickelt, sobald sie in laufende Systeme implementiert sind. Was die Endanwenderin sieht, das heißt, wie ihre Welt am Bildschirm geordnet ist, ist das Endprodukt einer verdichtenden Relationierung von Nutzerinnen und Dingen. Was sichtbar wird, ist kaum als Vergleich zu erkennen, insbesondere wenn es sich bloß um die Präsentation einzelner Items handelt. Der mehrdimensionale Vergleichsprozess, der User und Items, User und User sowie Bewertungsverhältnisse von Usern und Items zueinander in Relation setzt, wird im letzten Schritt zu einem Vergleich von Items. Dabei werden die Relationen zwischen verschiedenen, immer auch wechselnden Benutzerkollektiven unsichtbar gemacht: Die Beziehungen zwischen Usern werden zu Beziehungen zwischen Dingen.

Präsentiert wird dieser Vergleich aber als Ausdruck meiner Persönlichkeit: In der Sprache der Anbieter heißt es dann zum Beispiel »Playlists extra für dich« (Spotify) oder »Top Picks for Markus« (Netflix). Der Prozess der Empfehlung führt raffinierte Gleichheitskonstruktionen durch, damit ich als User mit anderen Usern verglichen werden kann, verdeckt aber gleichzeitig seine Vergleichsoperationen und die daraus entstehenden User-User-Relationen, um meine Einzigartigkeit als Person hervorzuheben.[26]

Diese Umkehrung ist bezeichnend für ein spezifisches Verständnis von Publikum (Fisher & Mehozay, 2019; Passoth et al., 2014), das man

26 Andere Beschreibungen, die auf User-User-Relationen aufmerksam machen, sind möglich, wie das Beispiel von Amazon deutlich macht: »Kunden, die diesen Artikel gekauft haben, kauften auch«.

sich nicht mehr notwendigerweise in demografischen Termini von Alter, Wohnort oder Geschlecht vorstellen muss. Entscheidend für die Gruppierung von Usern sind Gemeinsamkeiten und Differenzen der beobachteten und quantifizierten Verhaltensweisen. User beziehungsweise Usergruppen können durch die (kulturellen) Produkte definiert werden, die sie konsumieren. Produkte können durch die User definiert werden, von denen sie konsumiert werden. Seaver bringt es auf den Punkt: »users are known as a collection of relations to items and items are known as a collection of relations to users«, oder zugespitzter: »pizzas in terms of people and people in terms of pizza« (Seaver, 2012, k.S.). Dies trifft zwar zu, doch ist die Relation User-Item (»people«-»pizza«) nur der Ausgangspunkt eines Collaborative-Filtering-Recommendersystems. Wie ich zeigen konnte, durchläuft ein UUCF-Empfehlungsdienst für Filme im Zuge seiner eingebetteten (mathematischen) Vergleichbarkeitskonstruktionen und Differenzbeobachtungen eine Reihe anderer Relationierungen: User-Item, User-User, User/Item-User/Item, Item-Item.

Inhaltsbasierte Empfehlungssysteme und Collaborative Filtering bewerten und vergleichen nicht Nutzerinnen oder Dinge, sondern Relationen zwischen Nutzerinnen und Nutzerinnen und Relationen zwischen Nutzerinnen und Dingen. Solche »matchings« – eine spezifische Variante der Relationierung – relationieren die Nutzerinnen paarweise zu allen anderen Nutzerinnen. Es sind diese Relationen zwischen den Nutzerinnen (oder zwischen Nutzerinnen und Dingen), die im Zentrum von Empfehlungssystemen stehen.

Die Analyse verschiedener Typen von Empfehlungssystemen zeigt eines deutlich: Empfehlungssysteme bewerten keine Nutzerinnen. Im Gegensatz zu Loyalty-Marketing (Turow, 2008) oder Credit-Scoring (Fourcade & Healy, 2017b) geht es nicht darum, jene Kundinnen mit dem höchsten Wert für das Unternehmen ausfindig zu machen, um sie zu belohnen – oder gar Kundinnen mit niedrigem Wert von Leistungen auszuschließen. Netflix, Amazon, Facebook, etc. zeigen allen ihren Nutzerinnen Empfehlungen. Wer die Plattform stärker nutzt, mag »passendere« Empfehlungen erhalten, doch werden keine Nutzerinnen davon ausgeschlossen. Neue Nutzerinnen erhalten möglicherweise nur unpersonalisierte oder stereotypisierende Empfehlungen. Es ist aber kaum vorstellbar, einer Nutzerin gar nichts anzuzeigen, wenn sie auf der Seite erscheint.

Die Bedeutung der Personalisierung liegt nur vordergründig in einer Berücksichtigung des Individuums. Personalisiert wird mit formalen, für alle

gleichen Mitteln der Relationierung. Nutzerinnen sind in diesen Empfehlungssystemen relational gedacht und haben keine »ontologische Dignität« (Heintz, 2021). Sie treten in Relation zu Dingen in Erscheinung – was sie anklicken, anschauen, kaufen (durch Encoding) – und in Relation zu anderen Nutzerinnen (durch algorithmische Relationierung). Digitale Personalisierung operiert über De-Personalisierung, d.h. einer mit den formalen Mitteln des Encoding und des algorithmischen Vergleichs erzeugten Verdichtung von Relationen von Nutzerin, anderen Nutzerinnen und Dingen.[27]

27 Das heißt auch, dass die kategorialen Grenzen des Vergleichs tendenziell wegfallen. Besser gesagt: Sie werden durch proto-kategoriale Grenzen ersetzt, d.h. die in die Software eingelassenen prä-kategorialen Formalisierungsprozesse entscheiden darüber, welche Nutzerinnen zueinander in Relation gesetzt werden können – zum Beispiel Nutzerinnen die eine Mindestmenge an Items bewertet haben und nicht: Nutzerinnen, die einer gemeinsamen Kategorie wie Geschlecht angehören.