technische universität
dortmund

Bachelorarbeit

Eine Fallstudie zur Wirkungsanalyse verschiedener Empfehlungsverfahren im E-Commerce

Aaron Larisch
Oktober 2015

Gutachter:
Prof. Dr. Dietmar Jannach
Michael Jugovac, M. Sc.

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl für Dienstleistungsinformatik (LS13)
http://ls13-www.cs.tu-dortmund.de

In Kooperation mit:
TriMeXa GmbH
Friedrich-Ebert-Platz 5B
51373 Leverkusen

Inhaltsverzeichnis

Kapitel 1

Einleitung

Bei einem Empfehlungssystem handelt es sich um eine Software, die verschiedene Algorithmen einsetzt, um beispielsweise den Besucher eines Online-Shops auf bestimmte Produkte aufmerksam zu machen. Diese Produkte werden dabei automatisiert als besonders interessant für ihn eingestuft. Empfehlungssysteme dienen dazu, den Besucher vor einer Überflutung von zu vielen Informationen zu bewahren und ihm möglichst nur die für ihn zu diesem Zeitpunkt relevanten Artikel zu präsentieren. Sie tragen außerdem dazu bei, ihn zum Kauf von möglichst vielen Produkten anzuregen und somit den Umsatz des Online-Shops zu steigern. Typischerweise werten die eingesetzten Verfahren dabei das Kauf- und Aufrufverhalten der Besucher aus und nutzen zudem vorliegende Artikelinformationen wie ihre Beschreibungen.

> „Ein Empfehlungssystem (oft auch ‚Recommender System' genannt) ist ein System, das einem Benutzer in einem gegebenen Kontext aus einer gegebenen Entitätsmenge aktiv eine Teilmenge ‚nützlicher' Elemente empfiehlt." [17]

Empfehlungssysteme haben dabei Wurzeln in verschiedenen Themengebieten der Informatik wie dem *information filtering* und dem *data mining*.

1.1 Motivation und Hintergrund

Da die Wünsche eines Besuchers in einem Online-Shop im Gegensatz zu einem realen Geschäft nicht durch einen Fachverkäufer sofort abgefragt und erkannt werden können, bietet es sich an, dafür automatisierte Verfahren einzusetzen. Diese bieten ihm beispielsweise aufgrund seines Aufruf- und Kaufverhaltens oder des typischen Verhaltens anderer Benutzer Produkte an, die ihn vermutlich interessieren werden. Dazu wurden in der Forschung bereits verschiedene Verfahren entwickelt. Zur richtigen Auswahl des Verfahrens ist es jedoch erforderlich, diese zunächst miteinander zu vergleichen und herauszufinden, welche sich für den jeweiligen Einsatzzweck am besten eignen. Dazu können z. B. verschiedene Verfahren

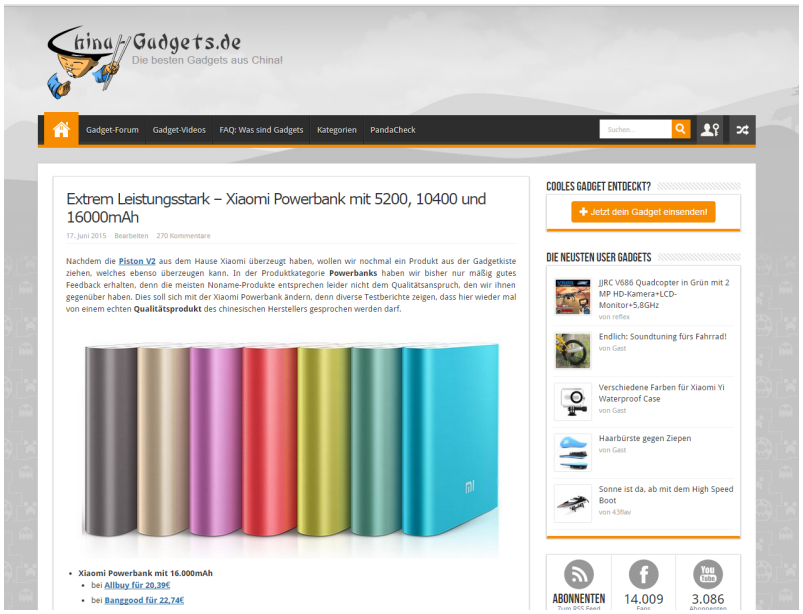

Abbildung 1.1: Startseite von China-Gadgets.de

in Form eines A/B-Tests evaluiert werden. In der Wissenschaft wird jedoch insbesondere die Evaluierung von Empfehlungsverfahren auf Live-Systemen oftmals vernachlässigt, da es im akademischen Umfeld häufig schwierig ist, dafür geeignete Systeme zu finden. So gibt es zwar viele Untersuchungen, die auf bestehenden historischen Datensätzen arbeiten und bewerten, welche Verfahren den größten Erfolg versprechen, allerdings finden diese Tests offline statt und berücksichtigen zum Beispiel keine realen psychologischen Auswirkungen auf den Nutzer [19].

Für die Untersuchungen in dieser Arbeit wurden verschiedene Empfehlungsverfahren implementiert und auf einer Plattform im realen Betrieb eingesetzt. Der Einsatzort war dabei die Seite China-Gadgets.de, ein sogenannter „Schnäppchenblog“ (Abbildung 1.1), der sich auf die Vermarktung von chinesischen Produkten im deutschsprachigen Raum spezialisiert hat. In diesem Blog werden täglich mehrere Beiträge zu verschiedenen neuen Artikeln veröffentlicht, die zum Teil einfach nur deren Verwendung erläutern oder aber auch ganze Testberichte darstellen und die Besucher zum Kauf anregen sollen. Aus dem jeweiligen Beitrag heraus kann über einen Link direkt auf den Online-Shop des Händlers zugegriffen werden, bei dem das jeweilige Produkt zum möglichst günstigen Preis zum Verkauf angeboten wird. Diese geben dabei einen Teil des Verkaufserlöses als Provision ab und beteiligen somit China-Gadgets.de an ihren Einnahmen, über die sich der Blog schließlich finanziert.

1.2 Aufbau der Arbeit

Die Arbeit stellt in einem theoretischen Teil zuerst bestehende Ergebnisse im Praxiseinsatz von Empfehlungssystemen vor und zeigt mögliche Unterschiede zu Offline-Evaluationen auf. Außerdem wird auf die grundsätzliche Funktionsweise einiger Empfehlungsverfahren eingegangen und es wird gezeigt, wie diese evaluiert werden können. Anschließend werden die konkreten technischen Einzel- und Besonderheiten bei der Realisierung des Systems auf China-Gadgets.de näher betrachtet. Im letzten Teil der Arbeit werden die aus der Fallstudie gewonnenen Ergebnisse schließlich vorgestellt und aus diesen ein Fazit gezogen.

Kapitel 2

Empfehlungssysteme und ihre Evaluation

Dieses Kapitel geht auf die Funktionsweise von Empfehlungssystemen ein und stellt die verschiedenen Möglichkeiten, die zur Erzeugung von Vorschlägen zur Verfügung stehen, vor. Außerdem nennt es geeignete Evaluierungsverfahren und Metriken, die zur Beurteilung von Empfehlungen verwendet werden können. Diese sind im Anschluss anhand bestehender Studien auf ihre Aussagekraft im Vergleich zu Ergebnissen von Online-Evaluationen zu untersuchen.

2.1 Grundlagen

Die **Produkte** oder **Artikel** bilden die Entitätsmenge eines Empfehlungssystems und können verschiedener Art sein. Denkbar sind z. B. Kleidungsstücke, elektronische Geräte, Musiktitel, Filme oder Zeitungsartikel, aber auch völlig andere Entitäten, die aufgrund ihrer höheren Anzahl zur Unübersichtlichkeit führen können. Die Schwierigkeit liegt darin, die Wahl des richtigen Empfehlungsverfahrens möglichst gut auf die jeweilige Zielgruppe und die Art der Artikel abzustimmen [13, 16, 5, 9]. Um die eingesetzten Algorithmen die Auswahl von relevanten Artikeln bestmöglich treffen zu lassen, bietet es sich an, möglichst viele Informationen über den Benutzer und die Produkte auszuwerten. Viele Empfehlungsverfahren arbeiten daher personalisiert, generieren also auf die jeweiligen Interessen der Benutzer angepasste Empfehlungen und benötigen dazu Präferenzinformationen. Dazu ist es notwendig, Benutzerprofile zu erzeugen, die diese Präferenzen möglichst gut widerspiegeln. Diese Profile können auf zwei verschiedene Arten erzeugt werden [15, S. 22]:

Bei der **expliziten Bewertung** wird der Benutzer direkt nach seiner persönlichen Meinung zu einem bestimmten Produkt gefragt, welches er sich möglicherweise gerade ansieht. Dazu kann er je nach Implementierung entweder wie in Abbildung 2.1 auf einer vorgegebenen Skala auswählen, wie sehr ihm der Artikel gefällt oder auch wie in Abbildung

Abbildung 2.1: Explizite Bewertung bei Amazon

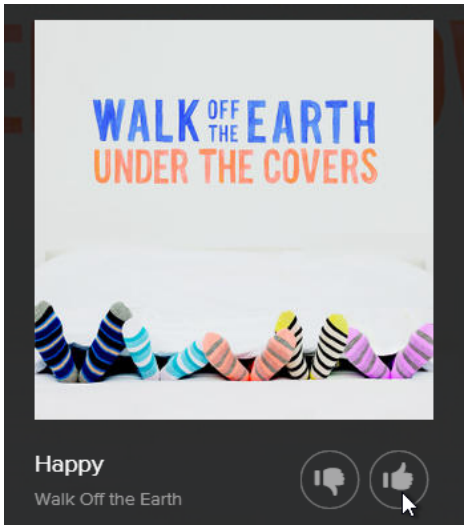

Abbildung 2.2: Explizite Bewertung bei Spotify

2.2 nur „dafür“ oder „dagegen“ stimmen. Die Auswahl wird dann in sein persönliches Benutzerprofil übernommen und gespeichert.

Hingegen wird das Profil des Benutzers bei der **impliziten Bewertung** indirekt über sein Verhalten erzeugt. So wird dabei ausschließlich protokolliert, welche Artikel er sich ansieht, in den Warenkorb legt oder kauft und anhand dessen anschließend abgeleitet, welche Artikel er besonders interessant fand. Das führt allerdings zu dem Problem, dass nicht sichergestellt werden kann, ob den Benutzer das Produkt auch tatsächlich interessiert hat. So kann es sein, dass er sich es zwar angesehen oder sogar gekauft hat, aber schlussendlich nicht davon überzeugt war oder es für jemanden anders bestellt hat.

Je nach Einsatzort können zudem weitere Informationen über den **Kontext** mit einfließen, der jedoch in der Literatur nicht einheitlich definiert wird [15, S. 291]. Nach [22] enthält dieser z. B. Informationen darüber, an welchem Ort sich der Benutzer zu der Zeit befindet oder welche weiteren Personen und Begebenheiten in seiner Nähe sind. Diese Informationen können dazu verwendet werden, um einem Benutzer, der gerade nach einem Kinofilm sucht, beispielsweise direkt die Kinos aus seiner unmittelbaren Umgebung anzuzeigen, in denen dieser Film gezeigt wird.

Außerdem gibt es zwei wesentliche Arten von Empfehlungsverfahren. Diese unterscheiden sich darin, ob sie primär die Benutzerprofile verschiedene Benutzer miteinander vergleichen (*kollaborativ*) oder ob sie die einzelnen Produkte inhaltlich z. B. über ihre Produktbeschreibungen vergleichen und jeweils darüber weitere Artikel heraussuchen (*inhaltsbasiert*). Ebenfalls lassen sich hybride Verfahren realisieren, die beide Verfahren miteinander kombinieren.

Zudem muss zwischen personalisierten und unpersonalisierten Empfehlungsverfahren unterschieden werden. So wird bei **personalisierten Verfahren** das aktuelle Profil des

aktiven Benutzers mit berücksichtigt, sodass die Vorschläge diesbezüglich optimiert werden. Das setzt allerdings auch voraus, dass der Benutzer bereits einige Zeit auf der Seite verbracht hat, um ausreichend Profildaten erzeugt zu haben. Ein typisches Verfahren wäre es hier, dem Benutzer die Artikel aus seinem vorherigen Besuch als Erinnerung noch einmal anzuzeigen. Hingegen sind **unpersonalisierte Verfahren** nicht von bereits vorliegenden Profildaten des Benutzers abhängig und geben ausschließlich allgemeingültige Empfehlungen aus, die auf alle Besucher gleichermaßen zutreffen können. Diese sind somit selbst bei neuen Besuchern sofort einsetzbar.

Die Verfahren können dabei entweder nach interessanten und uninteressanten Artikeln klassifizieren und somit anschließend eine Rangfolge von möglicherweise interessanten Artikeln ausgeben oder für ein Objekt der Vorhersage prognostizieren, welchen Wert dieses für einen Benutzer hätte. Die Produkte können dabei anhand ihres prognostizierten Wertes durch Festsetzen eines sinnvollen Schwellenwertes oder Auswahl der Top-N ebenfalls leicht als interessant oder uninteressant klassifiziert werden.

2.2 Verfahren

Empfehlungsverfahren lassen sich in kollaborative Verfahren (*collaborative filtering*) und in inhaltsbasierte Verfahren (*content-based filtering*) unterscheiden, die in diesem Kapitel näher behandelt werden sollen.

2.2.1 Kollaborative Verfahren

Bei **kollaborativen Verfahren** werden die während des Betriebs der Seite gesammelten Benutzerdaten z. B. zu den aufgerufenen Artikeln und dessen Bewertungen ausgewertet. Dazu werden die Profile verschiedener Benutzer miteinander verglichen und über deren Gemeinsamkeiten und Unterschiede versucht, passende Empfehlungen zu generieren. Dabei gilt das Grundprinzip aus den historischen Daten Rückschlüsse auf die Interessen und Vorlieben eines Benutzer zu ziehen und diesen in Zukunft gezielt nachzukommen [15, S. 13]. Gibt es z. B. zwei Benutzer A und B, die beide dieselben bisherigen Artikel interessiert haben, wobei Benutzer A einen weiteren Artikel mehr interessiert hat, so ist es nach dem *Nächste-Nachbarn-Verfahren* (s. Kapitel 2.2.1) naheliegend, Benutzer B diesen Artikel ebenfalls vorzuschlagen.

Als Eingabe für einen kollaborativen Algorithmus dienen dabei typischerweise entweder die abgegebenen Artikelbewertungen der einzelnen Benutzer oder schlicht die aufgerufenen Artikel. Anschließend lässt sich für einen konkreten Benutzer und ein vorgegebenes Produkt, das noch nicht von ihm bewertet wurde, abschätzen, wie gut oder schlecht es ihm vermutlich gefallen würde. Aufgrund dieser Informationen lässt sich somit ebenfalls eine Liste erstellen, die die voraussichtlich bestbewertesten Artikel enthält (*Top-N*).

Abbildung 2.3: Verfahren mittels Co-occurrence Patterns bei Amazon in der Detailansicht eines Monitors

Tabelle 2.1: Co-occurrence Häufigkeiten bei zwei Benutzern

	Artikel 1	Artikel 2	Artikel 3	Artikel 4
Artikel 1	-	1	1	0
Artikel 2	-	-	2	1
Artikel 3	-	-	-	1
Artikel 4	-	-	-	-

Co-occurrence Patterns (frequent patterns) Ein beliebtes Verfahren, das in Online-Shops häufig anzutreffen ist (s. Abbildung 2.3), ist es die gegenseitigen Vorkommen der Produkte z. B. in Warenkörben auszuwerten, also mögliche Zusammenhänge im Aufruf- oder Kaufverhalten aufzudecken. Das Ziel ist es dabei, herauszufinden, ob bestimmte Produkte immer zusammen mit anderen gekauft oder angesehen werden.

Es gibt verschiedene Möglichkeiten das gewünschte Verhalten zu implementieren. Ein bekanntes Verfahren ist der Apriori-Algorithmus aus der Assoziationsanalyse [15, S. 31ff]. Bei diesem Verfahren werden Teilmengen von Artikeln gebildet und für diese jeweils bestimmt, wie wahrscheinlich es ist, dass Personen, die eine Menge gekauft haben, auch eine bestimmte andere gekauft haben. Daraus können schließlich Regeln abgeleitet werden, die angeben, wie viel Prozent der Besucher, die Produkt A und B gekauft haben, auch C gekauft haben.

Ein etwas einfacherer Ansatz ist es, lediglich die gemeinsamen Käufe oder Interaktionen zu zählen. Als beispielhafter Datensatz seien zwei Benutzer gegeben, von denen der eine Artikel 1, 2 und 3 gekauft hat und der andere Artikel 2, 3 und 4 gekauft hat. Daraus ergibt sich Tabelle 2.1, in der durch einfaches Mitzählen aller Artikel-Paare, die von einem Benutzer gemeinsam gekauft wurden, bestimmt werden kann, welche Produkte allgemein besonders häufig zusammen gekauft werden. So wäre es hier sinnvoll, einem dritten Besucher, der gerade Artikel 2 in seinen Warenkorb gelegt hat, ebenfalls Artikel 3 vorzuschlagen, da dieser von vielen Benutzern mitbestellt wurde.

Hierbei handelt es sich somit um ein unpersonalisiertes Verfahren, welches mit relativ trivialen Mitteln einfache Zusammenhänge aufdecken kann.

Tabelle 2.2: Beispiel-Bewertungen auf einer Skala von 1 bis 5 in Form einer Benutzer-Artikel-Matrix

Benutzer	Artikel 1	Artikel 2	Artikel 3	Artikel 4
A	4	2		4
B	4	3	5	5
C	3	2	4	3
D	2	3	2	3
E	1	5	3	2

Zu beachten ist dabei jedoch, dass je nach konkretem Einsatz unterschiedliche Arten von Produkten empfohlen werden können. Bei Betrachtung von gemeinsam *gekauften* Artikeln handelt es sich eher um Produkte mit ihrem jeweiligen Zubehör, währenddessen bei gemeinsam *betrachteten* Produkten eher weitere alternative Artikel empfohlen werden.

Nächste-Nachbarn-Verfahren Beim Nächste-Nachbarn-Verfahren handelt es sich hingegen um ein personalisiertes Verfahren, welches entweder benutzer- oder artikelbasiert eingesetzt werden kann. Dabei wird versucht, die Bewertung eines Artikels für einen Benutzer vorherzusagen, der ihn bislang noch nicht bewertet hat. Dies geschieht dabei durch Finden von Benutzern oder Artikeln, die aufgrund ihrer Bewertungen ähnlich zueinander sind. Der Grundgedanke dabei ist, dass ein Benutzer, der Produkte größtenteils ähnlich wie ein anderer bewertet hat, vermutlich auch die bislang noch nicht vom ihm selbst bewerteten Artikel ähnlich wie der andere einstufen wird [15, S. 13ff]. Dabei wird angenommen, dass sich der Geschmack der Benutzer über die Zeit nicht wesentlich ändert und gefundene Gemeinsamkeiten aus der Vergangenheit auch in der Zukunft noch bestehen werden.

Nachfolgend soll das benutzerbasierte kNN-Verfahren (k-Nächste-Nachbarn) beispielhaft als typisches kollaboratives Verfahren skizziert werden. Als Eingabe dafür wird eine Benutzer-Artikel-Matrix übergeben, die die bislang erfolgten Bewertungen der Benutzer enthält, wie sie beispielhaft in Tabelle 2.2 dargestellt ist. Zunächst wird eine k-große Nachbarschaft an weiteren Benutzern bestimmt, die alle ein ähnliches Bewertungsverhalten wie beim aktuellen Benutzer haben. Zur Bestimmung der Nachbarschaft wird häufig die Pearson-Korrelation [15, S. 14f] verwendet. Diese berechnet wie ähnlich die Bewertungen zweier Benutzer waren und respektiert ebenfalls unterschiedliche Auffassungen der Bewertungsskala, sodass ein Benutzer, der konsequent schlechter bewertet als ein anderer, dadurch nicht zwangsläufig weniger ähnlich ist. Die Pearson-Korrelation liegt dabei stets im Bereich zwischen -1 und 1 und berechnet sich bei der Menge Benutzer $U = u_1, ..., u_n$, der Menge $P = p_1, ..., p_m$ und der $n \times m$ Matrix R mit den Bewertungen $r_{i,j}$ $(i \in 1...n, j \in 1...m)$ für die beiden Benutzer a und b wie folgt:

$$sim(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \overline{r_a})(r_{b,p} - \overline{r_b})}{\sqrt{\sum_{p \in P}(r_{a,p} - \overline{r_a})^2}\sqrt{\sum_{p \in P}(r_{b,p} - \overline{r_b})^2}} \tag{2.1}$$

Wird nun der aktuelle Benutzer a mit allen anderen verglichen, kann daraus schließlich die Nachbarschaft des Benutzers berechnet werden, indem die Top-k ähnlichen Benutzer bestimmt werden. Die Bewertung eines Produktes p ergibt sich damit aus:

$$pred(a,p) = \overline{r_a} + \frac{\sum_{b \in N} sim(a,b) \cdot (r_{b,p} - \overline{r_b})}{\sum_{b \in N} sim(a,b)} \tag{2.2}$$

Beispielhaft ergibt Benutzer A verglichen mit den Benutzern B bis E die Pearson-Korrelationen von jeweils $0,83$; $0,82$; $-0,47$ und $-0,97$. Für $k = 2$ besteht die Nachbarschaft N damit aus den Benutzern B und C, sodass die fehlende Bewertung für Artikel 3 mit $3,21$ abgeschätzt werden kann.

Mit kNN existiert also ein typisches Verfahren, welches kollaborativ Bewertungen abschätzen kann, indem Zusammenhänge zu anderen Benutzern entsprechend ausgenutzt werden. Der Nachteil bei diesem Verfahren liegt allerdings im relativ hohen Rechenaufwand, der dazu benötigt wird. Um diesen zu reduzieren, existieren alternative, etwas einfachere Verfahren wie z. B. SlopeOne [15, S. 41ff].

2.2.2 Inhaltsbasierte Verfahren

Bei **inhaltsbasierten Verfahren** werden die Artikel als solche über die zusätzlichen Informationen, die über sie zur Verfügung stehen, miteinander verglichen. Dabei muss grundsätzlich zwischen zwei verschiedenen Arten von Informationen unterschieden werden:

1. Informationen in Form von **Artikeleigenschaften** d. h. strukturierten Angaben z. B. im Falle von Smartphones über die verfügbare Displayauflösung, der Taktrate des verbauten Prozessors oder dem integrierten Speicher

Tabelle 2.3: Strukturierte Daten am Beispiel von Smartphones

Modell	Auflösung	Prozessor	Speicher	Preis
A	1920x1080	2x1,8Ghz	64GB	459 Euro
B	2560x1440	4x2,1Ghz	32GB	509 Euro

2. oder Informationen in Form einer **Artikelbeschreibung**, die als Text den Artikel näher beschreibt.

Während im ersten Fall die Daten bereits, wie in Tabelle 2.3 aufgelistet, geordnet vorliegen und diese aufgrund einer häufig bereits bestehenden Rangordnung (z. B. je mehr Speicher, desto besser) leicht verglichen werden können, muss im zweiten Fall der Text zunächst auf seinen Inhalt hin untersucht werden.

Dazu findet in der Regel eine Analyse der Häufigkeiten der im Text vorkommenden Wörter statt. Die Häufigkeiten repräsentieren dann den Inhalt des Textes und ermöglichen es anschließend, besonders ähnliche oder konträre Artikel zu finden. Ebenfalls wird hier häufig das Benutzerprofil mit hinzugezogen, sodass Artikel bestimmt werden können, die einen inhaltlichen Zusammenhang zu den präferierten Artikeln des Benutzers haben.

TF-IDF Zur Bestimmung der Worthäufigkeiten werden in vielen Fällen *TF-IDF-Vektoren* erzeugt. **TF-IDF** steht für *term frequency-inverse document frequency* [15, S. 55f]. Bei diesem Verfahren wird die Häufigkeit eines jeden Wortes, das in einem Text vorkommt, bestimmt und mit der Häufigkeit dieses Wortes in allen Texten verglichen. Kommt beispielsweise ein Wort besonders häufig in einem Text vor und ist in allen Texten zusammen jedoch nur selten anzutreffen, so handelt es sich demnach vermutlich um ein Wort, welches besonders charakteristisch für diesen Text ist. Ist genau das Gegenteil der Fall, so ist dieses wahrscheinlich nicht sonderlich repräsentativ. Für jeden Text wird somit ein Vektor erzeugt, der für jedes im Text enthaltene Wort angibt, wie charakteristisch es für ihn ist. Um zu vermeiden, dass besonders lange Texte besonders hohe Wertungen erhalten, müssen diese noch über die Gesamtanzahl der Wörter normiert werden, sodass lediglich relative Häufigkeiten berücksichtigt werden. Zusätzlich sollten die Wörter vorher auf ihre Wortstämme reduziert werden (*stemming*), um eine unnötige Verteilung des sinngemäß selben Wortes z. B. auf verschiedene Konjugationen und Numeri zu reduzieren und zeitgleich die Vektoren selbst nicht unnötig zu vergrößern.

Kosinus-Ähnlichkeit Um die Ähnlichkeit zweier TF-IDF-Vektoren ($\vec{a}$, $\vec{b}$) und somit zweier Artikel zu bestimmen, wird in der Regel die Kosinus-Ähnlichkeit verwendet. Diese stammt aus der Vektorrechnung und ähnelt der Bestimmung des eingeschlossenen Winkels [15, S. 19]:

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|} \tag{2.3}$$

Dabei liegt $sim(\vec{a}, \vec{b})$ im Bereich von 0 bis 1, wobei 0 eine besonders niedrige und 1 eine besonders hohe Ähnlichkeit angibt.

Ähnliche Artikel In Situationen, in denen das derzeit betrachtete Produkt eines Benutzers ihn noch nicht vollständig zum Kauf überzeugt, kann es sinnvoll sein, ihm weitere ähnliche Artikel zu präsentieren, die seine Bedürfnisse möglicherweise besser erfüllen. Dazu kann der TF-IDF-Vektor des aktuell betrachteten Artikels verwendet werden. Das Ziel ist es, weitere TF-IDF-Vektoren zu finden, die möglichst ähnlich zum vorgegebenen sind, sodass über diese darauf geschlossen werden kann, dass auch die Inhalte der zugehörigen Artikelbeschreibungen ähnlich sind. Wird nun der vorgegebene Vektor mit allen anderen

Vektoren, die in der Produktdatenbank enthalten sind, verglichen und werden die zugehörigen Kosinus-Ähnlichkeiten berechnet, bilden die Produkte mit dem höchsten Wert schließlich die gewünschte Produktmenge.

Content-based Personalized Filtering Die voran beschriebenen Methoden können ebenfalls dazu eingesetzt werden, um einem Benutzer Artikel zu empfehlen, die inhaltlich besonders gut zu seinen Interessen passen. Dazu werden zunächst die TF-IDF-Vektoren aller Artikel, die ihm bislang gefallen haben oder die im Falle von implizitem Feedback von ihm aufgerufen wurden, zu einem Durchschnittsvektor zusammengerechnet. Seine persönlichen Präferenzen werden damit durch diesen Durchschnittsvektor repräsentiert. Indem anschließend die Kosinus-Ähnlichkeiten zu allen anderen vorhandenen Vektoren bestimmt werden, können wieder die für ihn besonders interessanten Produkte herausgefiltert werden [15, S. 58]. Dadurch ergibt sich eine Produktmenge, die inhaltlich zu seinen Interessen passen sollte.

2.2.3 Zusammenfassung

Die Auswahl der dargestellten Verfahren gibt einen Überblick darüber, welche vielfältigen Möglichkeiten zur Generierung von Empfehlungen zu Verfügung stehen und lässt erahnen, dass die Wahl der einzusetzenden Verfahren für einen Produktivbetrieb sorgfältig getroffen werden sollte. Es gibt jedoch ebenfalls zahlreiche weitere, teils deutlich komplexere Verfahren, die z.B. auf probabilistischen Ansätzen [12] oder Matrixfaktorisierung [18] basieren.

2.3 Offline-Evaluation

Um die richtige Auswahl von geeigneten Verfahren für einen bestimmten Einsatzzweck zu erleichtern, bietet es sich an, vorab *offline* einige Untersuchungen durchzuführen. Diese Untersuchungen können dabei auf zwei verschiedene Arten realisiert werden: Entweder durch Analyse eines historischen Datensatzes oder durch Probanden, die im Rahmen einer Laborstudie auf ihr Verhalten hin untersucht werden bzw. darum gebeten werden, die erzeugten Empfehlungen nach verschiedenen vorgegebenen Kriterien zu bewerten. Eine Laborstudie erfordert dabei jedoch im Vergleich zur Datensatzanalyse einen erheblich höheren Aufwand.

Bei Verwendung eines historischen Datensatzes wird dieser zunächst aus einem Live-System exportiert. Je nachdem, welche Daten erfasst werden, enthält dieser die expliziten oder impliziten Bewertungen, d. h. die Käufe, Aufrufe oder direkten Beurteilungen verschiedener Artikel jedes Benutzers [21, 3]. Anschließend wird eine Situation simuliert, in der für eine bestimmte Person Empfehlungen generiert werden sollen. Dazu wird für eine

Auswahl von Empfehlungsverfahren der Datensatz als Eingabe verwendet und das Ergebnis untersucht. Die Schwierigkeit liegt jedoch darin, zu beurteilen, ob die auswählten Artikel geeignet waren oder nicht. Gäbe es Menge, die die optimalen Empfehlungen für einen Benutzer enthalten würde, könnte bestimmt werden, welcher Anteil davon durch das Verfahren ausgewählt werden konnte. Da diese Menge jedoch nicht für jeden Benutzer mit seinen jeweiligen Interessen präzise bestimmt werden kann, ist eine zuverlässige objektive Bewertung schwierig [17, S. 37f].

Um dennoch eine Bewertungsmöglichkeit zu schaffen, wird in der Wissenschaft fast ausnahmslos immer nach demselben Verfahren vorgegangen. Dazu wird zunächst ein Teil des exportierten Datensatzes als Testdatenmenge versteckt und der Rest als Trainingsdatenmenge zur Eingabe für einen bestimmten Algorithmus zum Lernen verwendet. Das Ziel ist es anschließend mittels des Verfahrens die Daten aus der versteckten Testdatenmenge vorherzusagen. Die Qualität der Vorhersage wird schließlich danach beurteilt, wie präzise die Elemente aus der Testdatenmenge prognostiziert werden konnten.

Die Qualität kann jedoch im Detail auf verschiedene Weisen beurteilt werden, die zur Verwendung unterschiedlicher Metriken führen. Eine Metrik erfasst dabei numerisch eine bestimmte Eigenschaft der Empfehlungen, die als Vergleichswert zwischen verschiedenen Verfahren herangezogen werden kann. Durch Kombination verschiedener unterschiedlicher Metriken kann so entschieden werden, ob die generierten Empfehlungen für den jeweiligen Einsatzzweck sinnvoll waren oder nicht und welche am besten abgeschnitten haben.

Dieses Vorgehen wird dabei mit verschiedenen Aufteilungen von Trainings- und Testdatenmenge wiederholt und die bestimmten Metriken anschließend gemittelt (Kreuzvalidierung). Dadurch werden zufällige Ergebnisse aufgrund besonders günstiger oder ungünstiger Segmentierungen reduziert.

Im Anschluss daran können die eingesetzten Verfahren in eine Rangfolge gesetzt werden, die unterschiedliche betrachtete Aspekte gegeneinander abwägen. Um auszuschließen, dass die Ergebnisse nicht zu stark vom Zufall beeinflusst wurden oder zu nahe beieinander liegen, kann anschließend ebenfalls ein Signifikanztest durchgeführt werden. Damit lässt sich abschätzen, wie wahrscheinlich es ist, dass allfällig beobachtete Unterschiede nur durch Zufall entstanden sind.

Zu beachten ist jedoch, dass die Aussagekraft nur dann gewährleistet ist, wenn angenommen werden kann, dass sich der Benutzer beim Übergang in den Live-Betrieb möglichst genauso verhält wie in den Offline-Daten aufgezeichnet wurde [15, S. 254]. Damit wird nicht berücksichtigt, dass das Empfehlungssystem die Kaufentscheidung eines Benutzers möglicherweise sogar aktiv beeinflussen kann [2].

Die Offline-Evaluation stellt damit ein praktikables Werkzeug dar, um den Aufwand für eine Online-Evaluation zu reduzieren und bereits von Beginn an eine möglichst gute Vorauswahl zu treffen [15, S. 187f].

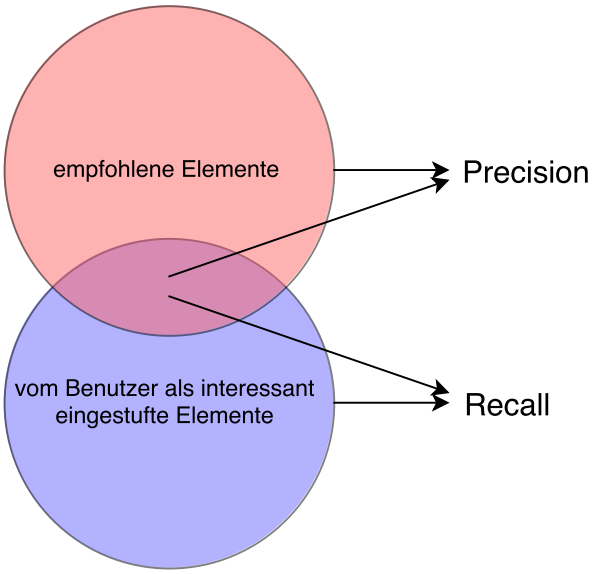

Abbildung 2.4: Precision und Recall veranschaulicht

2.3.1 Metriken

Eine Metrik dient dazu, die Qualität z. B. einer Lösung zu messen und bezieht sich stets auf einen bestimmten Teilaspekt. Für ein Verfahren, das in einem Straßennetz für ein Auto eine Route zwischen zwei Punkten bestimmen soll, können ebenfalls verschiedene Metriken bestimmt werden. Denkbar wären z. B. die Distanz, die Fahrtdauer oder der nötige Kraftstoffverbrauch. Je nachdem welche Lösung als *optimal* bezeichnet werden soll, können so verschiedene Metriken eine Rolle spielen, die individuell abgewogen werden müssen.

Im Fall eines Empfehlungssystems messen diese die Qualität der Klassifizierungen bzw. prognostizierten Bewertungen und helfen dabei verschiedene Verfahren gegeneinander abzuwägen und die richtige Wahl für den Live-Betrieb zu treffen.

Die populärsten Metriken zur Qualitätsmessung von Klassifizierungen sind *Precision* und *Recall* (s. Abbildung 2.4).

Die **Precision** [15, S. 180] gibt an, wie viele der Empfehlungen, die ausgegeben wurden, tatsächlich korrekt klassifiziert wurden. Sie entspricht somit dem Quotienten der Anzahl der korrekten Empfehlungen $|hits_u|$ und der Gesamtanzahl der ausgegebenen Empfehlungen $|recset_u|$.

$$P_u = \frac{|hits_u|}{|recset_u|} \tag{2.4}$$

Der **Recall** [15, S. 180] ist ein Maß für das Verhältnis zwischen den korrekt ausgegebenen Empfehlungen $|hits_u|$ und der Gesamtanzahl der Artikel $|testset_u|$, die den Benutzer interessiert haben.

$$R_u = \frac{|hits_u|}{|testset_u|} \tag{2.5}$$

Beide Metriken beziehen sich dabei häufig auf die Top-N empfohlenen Artikel eines Verfahrens und werden dementsprechend mit Precision@N und Recall@N bezeichnet.

Zur Bestimmung der Genauigkeit von Bewertungsvorhersagen wird hingegen häufig der **mittlere absolute Fehler (MAE)** [15, S. 179f] herangezogen, auf den jedoch nur kurz eingegangen werden soll, da sich der Großteil der Arbeit eher mit Verfahren auseinandersetzt, die klassifizieren und somit nur eine Rangfolge von Artikeln erzeugen. Der MAE gibt die durchschnittliche Abweichung der generierten Bewertungen $rec(u, i)$ von den tatsächlichen Bewertungen $r_{u,i}$ aller Benutzer $u \in U$ und Artikel $i \in I$ in ihren Testmengen $testset_u$ an:

$$MAE = \frac{\sum_{u \in U} \sum_{i \in testset_u} |rec(u, i) - r_{u,i}|}{\sum_{u \in U} |testset_u|} \tag{2.6}$$

Zur Auswertung von Verfahren, welche eine Rangfolge von Artikeln erzeugen, lässt sich z. B. mittels des **Mean Reciprocal Ranks (MRR)** [24] einschätzen, wie gut diese ist. Dabei wird angenommen, dass die Artikel in der Rangfolge absteigend nach ihrer vorausgesagten Relevanz sortiert wurden. $RR(u)$ entspricht für Benutzer $u \in U$ dabei dem Kehrwert des besten Platzes in der Rangfolge für alle Artikel aus seiner Testdatenmenge oder ist 0, wenn kein Artikel in der Rangfolge enthalten ist. MRR berechnet den Durchschnitt über alle Benutzer:

$$MRR = \frac{\sum_{u \in U} RR(u)}{|U|} \tag{2.7}$$

Diese Metrik beruht auf der Annahme, dass vordere Platzierungen von einem Besucher stets deutlich mehr Beachtung geschenkt bekommen, als hintere, sodass durch den verwendeten Kehrwert die Bestplatzierungen die höchste Wertung bekommen.

Es existieren außerdem weitere Metriken [11, 19, 2, 26], die mögliche persönliche Wünsche eines Benutzers abdecken. Dazu gehören

- die Ähnlichkeit der Produkte innerhalb einer Liste von Empfehlungen,
- die Gesamtanzahl der verschiedenen Produkte, die überhaupt ausgegeben werden können,
- die Durchschnittspopularität der empfohlenen Produkte oder auch
- die Neuheit der Empfehlungen.

Auf die Relevanz dieser Werte soll in Kapitel 2.4 etwas näher eingegangen werden.

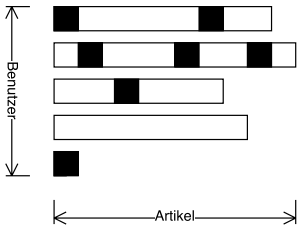

Abbildung 2.5: Zufällige Auswahl von 20% als Testdatensatz

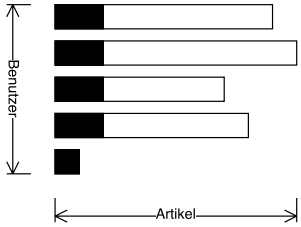

Abbildung 2.6: Auswahl von genau N Artikeln pro Benutzer als Testdatensatz (*all but N*)

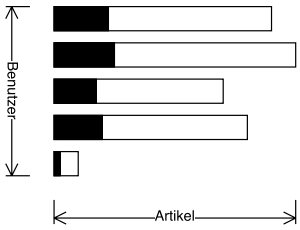

Abbildung 2.7: Auswahl von 25% pro Benutzer als Testdatensatz

2.3.2 Evaluationsmethodik

Wie eingangs erwähnt, ist für die Berechnung der Metriken in der Regel eine Aufteilung in Trainings- und Testdatenmenge notwendig, die üblicherweise relativ ähnlich geschieht:

Ein typisches Verfahren dafür ist nach [15, S. 177] die **N-fache Kreuzvalidierung**. Dazu wird ein Teil der Benutzerprofile als Trainingsdatensatz und ein anderer Teil als Testdatensatz verwendet. Mittels des Trainingsdatensatzes werden die Trainingsmodelle generiert, die für die Algorithmen benötigt werden, um anschließend ein effizientes Erzeugen der Empfehlungen zu ermöglichen. Der Testdatensatz dient der anschließenden Auswertung.

Dabei werden die Benutzerprofile in N gleich große Gruppen aufgeteilt und bei insgesamt N Durchläufen jeweils eine als Testdatensatz und die restlichen $N-1$ als Trainingsdaten ausgewählt. Die so erzeugten Empfehlungen werden schließlich mittels geeigneter Metriken bewertet und über die einzelnen Durchläufe hinweg gemittelt. Entspricht N der Gesamtanzahl aller Benutzer, so wird dies auch als **Leave-One-Out-Kreuzvalidierung** bezeichnet. Bei dieser Methode steht dem Algorithmus die größtmögliche Trainingsdatenmenge zur Verfügung, erfordert jedoch auch den größten Aufwand, da das Modell für jeden Benutzer einzeln neu berechnet werden muss.

Die Auswahl des Testdatensatzes kann völlig zufällig im gesamten Datensatz vorgenommen werden (Abbildung 2.5), pro Benutzer eine feste Anzahl von N Artikeln verwendet werden (Abbildung 2.6) oder pro Benutzer ein fester Prozentanteil seiner Artikel ausgewählt werden (Abbildung 2.7). Bei Auswahl von genau N Artikeln pro Benutzer als Testdatensatz spricht man auch von **all but** N, wohingegen die Auswahl von genau N Artikeln als Trainingsdatensatz als **given** N bezeichnet wird. Dabei hat *all but N* den Vorteil, dass für jeden Benutzer beim Testen bezogen auf die Klassifikationsmetriken die gleichen Bedingungen herrschen. Dahingegen hat *given N* den Vorteil, dass jedem Verfahren für jeden Benutzer eine gleich große Trainingsmenge zur Verfügung steht, sodass die Genauigkeit der Empfehlungen aufgrund gleicher Ausgangsbedingungen besser verglichen werden kann [15, S. 178].

Um den gesamten Datensatz bestmöglich ausnutzen, bietet es sich an, die Evaluierung pro Benutzer einzeln durchzuführen. Dies geschieht nach [21] wie folgt: Zunächst werden nahezu alle Daten aller Benutzer als Trainingsdaten verwendet und lediglich einige Artikel eines einzelnen Nutzers zum Testen davon ausgenommen. Diese Artikel können dabei entweder zufällig oder anhand ihrer Relevanz ausgewählt werden. Als relevant wird dabei z. B. eine feste Anzahl Artikel bezeichnet, die eine bestimmte Mindestbewertung überschreiten. Anschließend kann pro Benutzer genau gezeigt werden, welche Verfahren die gewünschten Artikel tatsächlich ausgewählt haben und es lassen sich diese über *Precision@N* und *Recall@N* miteinander vergleichen. Diese Relevanz-Auswahl hat den Vorteil, dass verhindert wird, dass selbst ein optimal arbeitendes Verfahren, das die Artikel unerlaubterweise aus dem Testdatensatz vorschlägt, die Chance hat, ein optimales Ergebnis von 1 zu erhalten. Nur dadurch kann sichergestellt werden, dass bei Auswertung der Top-N Empfehlungen überhaupt mindestens N relevante Artikel im Testdatensatz enthalten sind. Ebenfalls ist es möglich, die Daten zeitlich zu sortieren und nur die Interaktionen, die älter als ein vorgegebener Zeitpunkt sind, zum Training zu verwenden und auf die neueren Daten eines Benutzers zu testen. Dadurch wird eine möglichst realitätsnahe Situation aus der Vergangenheit nachgebildet.

Abschließend sollten die bestimmten Ergebnisse noch auf ihre Aussagekraft hin untersucht werden. Dazu gilt es die Nullhypothese zu widerlegen und zu zeigen, dass die Ergebnisse signifikant genug sind um weitere Aussagen darüber machen zu können (s. Kapitel 4.2). Dies kann dabei abgesehen vom in dieser Arbeit behandelten Chi-Quadrat-Test auch mittels einer paarweisen Analyse der ermittelten Varianzen (*ANOVA*) oder mittels T-Tests geschehen [4]. Zuletzt darf auch der praktische Nutzen dieser Untersuchungen nicht außer Acht gelassen werden. Es gilt somit ebenfalls herauszufinden, ob die möglicherweise vorgenommenen Optimierungen an einem Algorithmus dem Benutzer trotz ihrer statistischen Relevanz überhaupt einen Mehrwert bieten [15, S. 184]. Es gilt somit am Beispiel eines Online-Shops zu untersuchen, ob sich die Verbesserung einer Metrik auch an einer Umsatzsteigerung feststellen lässt.

2.4 Online-Evaluation

Bei der Online-Evaluation werden mehrere Empfehlungsalgorithmen in Form eines A/B-Tests miteinander verglichen. Dazu werden einige Verfahren implementiert und z. B. zur Bewerbung von Produkten im Rahmen eines Online-Shops eingesetzt. Das individuell angewandte Verfahren wechselt dabei von Benutzer zu Benutzer oder von Sitzung zu Sitzung und wird individuell ausgewertet.

Für diese Auswertung muss dabei eine sinnvolle Messung des Erfolges bestimmt werden. Als Standard gilt hier die **click-through rate** (CTR), die das Verhältnis von Klicks zu Impressionen einer Empfehlung (oder typischerweise einer Werbeanzeige) angibt. Eine

Impression bezeichnet dabei die einmalige Darstellung des Objektes auf dem Bildschirm des Benutzers. Allerdings gibt es auch hier Untersuchungen, die aufzeigen, dass dieser Wert nicht immer zwangsläufig die Relevanz des Produktes widerspiegelt, sondern diese z. B. auch durch die Reihenfolge, in der die Empfehlungen angezeigt werden, beeinflusst wird, obwohl alle gleichwertig gewesen wären [25].

In diesem Kapitel werden nachfolgend einige Arbeiten bzw. Studien behandelt und so auf mögliche Unterschiede im Vergleich zur Offline-Evaluation aufmerksam gemacht.

2.4.1 S. McNee - Being accurate is not enough: How accuracy metrics have hurt recommender systems [19]

Dieser Artikel stellt Alternativen zu den klassischen Genauigkeitsmetriken wie MAE vor, die in bestimmten Anwendungsfällen besser geeignet sind, um die gewünschten Ergebnisse zu erzielen und liefert einen kleinen Überblick über die Ergebnisse von bereits durchgeführten Studien und Untersuchungen.

In diesem Rahmen wird auf drei wesentliche Probleme eingegangen:

Ähnlichkeit Einige Algorithmen tendieren dazu, immer besonders ähnliche Artikel auszugeben. Insbesondere beim artikelbasierten *collaborative filtering* passiert dies sehr häufig, da die Genauigkeitsmetriken dieses Verhalten positiv honorieren. Das liegt daran, dass ähnliche Produkte vom Benutzer wahrscheinlich auch ähnlich gut bewertet werden, was ihm in Form einer Empfehlungsliste allerdings wenig Vorteil bietet, da er nichts grundlegend Neues entdecken kann. Als Gegenmaßnahme lässt sich hier die *Intra-List Similarity*[26] einführen, die angibt, wie ähnlich sich die darin enthaltenen Artikel sind. Dadurch ist es möglich, die Empfehlungsliste, die Benutzer zu Gesicht bekommt, als Ganzes zu betrachten und nicht lediglich auf die Einzelelemente einzugehen.

Unerwartetheit Die Unerwartetheit eines Artikels ist schwierig zu messen, da es sich dabei eher um ein persönliches Empfinden eines Benutzers handelt. Das Empfehlen von Artikeln, mit denen er nicht gerechnet hätte, kann jedoch durchaus von Vorteil sein, da sich bei ihm dadurch das Gefühl, eine neuen Entdeckung gemacht zu haben, einstellen kann. Das Gegenteil davon, die Vorhersagbarkeit, lässt sich jedoch relativ einfach messen und entspricht im Grunde den klassischen Evaluierungsverfahren. Es erscheint daher naheliegend, dass es manchmal nicht sinnvoll ist, ein Produkt zu empfehlen, welches zwar jeder Besucher mögen würde und besonders gut vorhersagbar ist, aber ebenfalls bereits jedem Benutzer bekannt ist und er es entweder bereits besitzt oder nicht besitzen möchte. Daher sollte auch das Empfehlen von Produkten, die besonders unvorhersagbar und somit unerwartet sind, in Betracht gezogen werden. Diese zu messen, ist allerdings fast ausschließlich in Labor-Studien durch konkrete Befragung möglich und kann höchstens

in Form der umgekehrten Vorhersagbarkeit abgeschätzt werden, wobei nicht sichergestellt werden kann, dass der gewünschte Überraschungseffekt dadurch erzielt wird.

Benutzererfahrungen Beim erstmaligen Besuch einer Webseite, auf der einem Benutzer Empfehlungen dargestellt werden, ist es durchaus wichtig, zuerst eine Vertrauensbasis aufzubauen. Um diesen daher vom Nutzen des System zu überzeugen, können anfangs besonders populäre und interessante Artikel präsentiert werden, die ihn möglichst fest daran binden sollen. Wird dieses Vertrauen jedoch direkt zu Anfang verspielt, ist es anschließend unter Umständen nicht mehr möglich, dieses zurückzuerlangen. Ebenfalls wird deutlich, dass es wichtig ist, einem Benutzer den Einstieg in die Empfehlungen so einfach wie möglich zu machen: So wurden demnach Empfehlungen, die in der Muttersprache des Benutzers ausgesprochen wurden, denen einer Fremdsprache vorgezogen, obwohl in beiden Fällen die Detailbeschreibungen der Artikel fremdsprachig waren. Die muttersprachliche Empfehlung z. B. in Form des übersetzten Titels führt somit bereits zu einer Herabsetzung der Hemmschwelle einen anderssprachigen Artikel zu lesen.

Zusammenfassung Insgesamt verdeutlicht diese Arbeit, dass auch andere Aspekte neben der Genauigkeit berücksichtigt werden sollten und sich nicht ausschließlich auf diese verlassen werden kann, um ein Empfehlungssystem zu bewerten.

2.4.2 D. Jannach - A case study on the effectiveness of recommendations in the mobile internet [13]

Bei dieser Fallstudie wurden verschiedene Empfehlungsverfahren auf einer Plattform zum Verkauf von Handyspielen miteinander verglichen. Das primäre Ziel dieser Untersuchung war es, herauszufinden, ob personalisierte Algorithmen zu mehr Aufrufen von Spielen führten als unpersonalisierte und ob dadurch höhere Verkaufszahlen erreicht werden konnten. Dazu wurde ein Feldversuch durchgeführt, bei dem jeder Besucher einem individuellen Verfahren zugelost wurde, welches über den gesamten Testzeitraum von 4 Wochen für ihn verwendet wurde. Dabei wurden ein Großteil aller Auflistungen auf der Startseite und innerhalb der einzelnen Spiele-Kategorien anhand dieses individuellen Verfahrens sortiert, sowie eine eigene Kategorie mit dem Namen „Meine Empfehlungen“ eingeführt. Verglichen wurden 7 verschiedene Verfahren inklusive einer Kontrollgruppe, die zum Vergleich auf dem Stand vor Beginn des Experiments beibehalten wurde. Zum Einsatz kamen dabei 6 verschiedene Algorithmen:

1. TopSeller: Sortierung nach der durchschnittlichen Benutzerbewertung
2. TopRating: Sortierung nach der Anzahl der Gesamtverkäufe
3. CF-Item: Kollaborative Artikel-Empfehlung basierend auf den Benutzerinteressen

4. CONTENT-BASED: Inhaltsbasierte Empfehlung mittels TF-IDF-Vektoren und Kosinus-Ähnlichkeiten

5. HYBRID: Kombination von Verfahren 3 und 4, die auf Verfahren 4 zurückgreift, wenn nicht ausreichend Bewertungen vorlagen

6. SLOPEONE: Kollaborative Vorhersage der Bewertungen eines Benutzer basierend auf denen anderer [15, S. 41ff]

TOPSELLER und TOPRATING waren unpersonalisierten Algorithmen, währenddessen die 4 restlichen Verfahren personalisiert waren. Als Datengrundlage dienten die direkt von den Benutzern abgegebenen Bewertungen sowie die jeweiligen Artikelaufrufe und -käufe.

Zu den Ergebnissen lässt sich zusammenfassend sagen, dass die personalisierten Verfahren in der Regel besser abgeschnitten haben als die unpersonalisierten.

Allerdings gab es einige Ausnahmen: So fiel z. B. auf, dass CONTENT-BASED innerhalb der Sektion „Meine Empfehlungen“ bezogen auf die Anzahl der Downloads schlechter als die anderen personalisierten Verfahren abgeschnitten hat. Bezogen auf die Anzahl der Klicks war es jedoch genauso gut. Das lässt sich dadurch erklären, dass sich zwar scheinbar ein großer Teil der Nutzer durch die Empfehlungen inspirieren lässt, allerdings schlussendlich nur eines der empfohlenen Spiele tatsächlich heruntergeladen wird. Dies beruht darauf, dass das inhaltsbasierte Verfahren dazu neigt, zueinander relativ ähnliche Empfehlungen innerhalb eines bestimmten Genres auszugeben. Der Nutzer sieht sich zwar die verschiedenen Spiele an, ist jedoch nicht gewillt, mehrere ähnliche Spiele auf einmal zu kaufen. Dahingegen decken die anderen Verfahren eine größere Bandbreite ab, sodass der Besucher möglicherweise sogar mehrere Spiele nacheinander kauft. Ebenfalls auffällig war die unterschiedliche Verteilung der heruntergeladenen Spiele zwischen Demo- und echten Vollversionen, die Geld kosten. So fällt hier auf, dass durch SLOPEONE und TOPRATING relativ viele Demoversionen heruntergeladen wurden, jedoch verhältnismäßig wenig Spiele tatsächlich verkauft wurden. Das liegt an einer besonderen Eigenschaft der Plattform, die es nur ermöglicht, eine Bewertung abzugeben, nachdem ein Spiel bereits heruntergeladen wurde. Das führt wiederum dazu, dass die kostenlosen Demoversionen deutlich mehr Bewertungen haben und somit Verfahren, die auf den direkten Benutzerbewertungen aufbauen, diese Produkte bevorzugt präsentieren.

Vor allem der letzte Punkt macht dabei deutlich, dass die eingesetzten Verfahren sinnvoll auf die besonderen Eigenschaften der Plattform abgestimmt werden müssen. Insgesamt konnte die Studie durch Einsatz der personalisierten Verfahren eine Verkaufssteigerung von 3,6% feststellen. Sie macht jedoch auch deutlich, dass durch weitere Abstimmung auf die Interessen und Erwartungen der Benutzer und das Ausnutzen weiterer Informationen, die durch den Benutzer zur Verfügung gestellt werden, der Wert noch weiter gesteigert werden kann.

Tabelle 2.4: Success- und Klickraten der verschiedenen Verfahren

Verfahren	Success offline	Success online	CTR
Kontextbaum	14%	19%	6%
beliebteste	16,5%	17,5%	4%
zufällige	1%	7%	6%

2.4.3 F. Garcin - Offline and online evaluation of news recommender systems at swissinfo.ch [9]

Dieser Artikel vergleicht offline und online evaluierte Empfehlungsverfahren am Beispiel einer Nachrichtenseite[1] und stellt dar, welche Unterschiede sich dabei erkennen lassen.

Bei diesem Vergleich wurde ein spezielles Verfahren verwendet, welches mit Kontextbäumen arbeitet [8] und besonders dazu geeignet ist, auf Nachrichtenseiten eingesetzt zu werden, da es neue Trends und Benutzerinteressen schnell berücksichtigt. Dabei wird stets die aufgerufene Sequenz an Beiträgen eines Besuchers berücksichtigt. So wird für jede Teilsequenz, die als Kontext bezeichnet wird, ein Modell berechnet, das einen wünschenswerten Aspekt wie Beliebtheit oder Neuheit berücksichtigt. All diese Kontexte werden schließlich unterschiedlich gewichtet und zu einer persönlichen Gesamtempfehlung verarbeitet, die somit die aktuellen Interessen immer mit berücksichtigt.

Zunächst wurde offline auf einem historischen Datensatz von 3 Wochen untersucht, wie gut das Verfahren verglichen mit zwei Standardverfahren abschneidet. Bei diesen beiden Standardverfahren handelte es sich um die Auswahl der beliebtesten Artikel und von zufälligen Artikeln. Dazu wurden dem jeweiligen Algorithmus stückweise die Daten eingegeben und anschließend gezählt, wie oft die derzeitigen 3 ausgegebenen Empfehlungen für einen Benutzer tatsächlich dem nächsten getätigten Aufruf entsprachen (*success@3*). Dabei stellte sich, wie in Tabelle 2.4 dargestellt ist, heraus, dass das Zufallsverfahren mit einer Trefferquote von ungefähr 1% am schlechtesten abgeschnitten hat, währenddessen sich die beliebtesten Artikel nach dem Eingeben der Daten von 7 Tagen bei einer Trefferquote im Bereich von 16,5% kurz vor dem kontextbaumbasierten Verfahren mit 14% stabilisierten.

Im Online-Versuch zeigte sich schließlich eine deutlich abweichende Verteilung. So schnitt hier das Zufallsverfahren mit 8% deutlich besser ab als in der Offline-Evaluation, währenddessen das kontextbaumbasierte Verfahren mit ca. 19% knapp vor dem Verfahren mit den beliebtesten Artikeln mit 17,5% liegt.

Anhand dieser Werte lässt sich mittels der *click-through rate* abschließend bestimmen, wie erfolgreich die einzelnen Verfahren tatsächlich waren und welche endgültig zu einer Verlängerung des Besuches beigetragen haben. So fällt auf, dass das Verfahren, das die

[1] `http://www.swissinfo.ch/`

beliebtesten Artikel empfiehlt, zwar bezogen auf die Success-Rate gute Ergebnisse erzielt, jedoch den offline bestimmten Wert nur um 1% im Vergleich zum Online-Wert erhöht. Bei einer CTR von 4% bedeutet das, dass 3/4 der Besucher den jeweiligen Artikel auch ohne Empfehlung angeklickt hätten und dessen Besuch dadurch also nicht bereichert wurde. Dahingegen schneidet das kontextbaumbasierte Verfahren deutlich besser ab, da nur 1/6 der Besucher auch selbstständig auf die empfohlenen Artikel gestoßen wären. Auffallend ist hier jedoch, dass nach dieser Messung das Zufallsverfahren am besten abgeschnitten haben muss, da alle Besucher dadurch ihre Artikelvielfalt vergrößert haben. Das liegt darin begründet, dass das kontextbaumbasierte Verfahren zuerst vom Benutzer lernen muss und aufgrund der hohen Anzahl an neuen Besuchern das zufällige Verfahren bei diesen besser abschneidet. Konzentriert man sich ausschließlich auf Besucher, die mindestens 3 Klicks getätigt haben, schneidet das Kontextbaum-basierte Verfahren jedoch besser ab.

In dieser Arbeit stellt sich heraus, dass zwischen Online- und Offline-Evaluationen durchaus große Unterschiede bestehen können. Außerdem ist erkennbar, dass es Situationen geben kann, in denen die CTR keinen guten Anhaltspunkt bietet, da sie durch populäre Artikel künstlich in die Höhe getrieben wird. So liefert sie für das beliebteste Artikel-Verfahren zwar einen relativ hohen Wert, der jedoch für den Benutzer keinen Vorteil bringt, da dieser ebenfalls alleine darauf gestoßen wäre, was ebenfalls in [25] näher behandelt wird.

2.4.4 M. Dias - The value of personalised recommender systems to e-business: A case study [5]

Im Rahmen dieses Artikels wurde für einen Online-Supermarkt[2], der insbesondere Lebensmittel und Dinge für den alltäglichen Gebrauch verkauft, ein Empfehlungssystem implementiert und über einen Testzeitraum von 20 Monaten ausgewertet. Das Ziel war es, die Verfahren auf das spezifische Einsatzgebiet anzupassen und die Einnahmen des Online-Shops zu steigern. Zu Einblendung der Empfehlungen wurden zwei verschiedene Orte ausgewählt.

Zum einen wurde bei der Ansicht des Warenkorbes eine Liste mit 6 Produkten angezeigt, dessen beide Teilhälften nach zwei verschiedenen Verfahren bestimmt wurden. Dabei handelte er sich einmal um Produkte, die sonst typischerweise regelmäßig vom Kunden gekauft werden, derzeit jedoch nicht im Einkaufswagen liegen und außerdem um Produkte, die statistisch gesehen zur aktuellen Produktauswahl am wahrscheinlichsten hinzugefügt werden könnten.

Zum anderen gab es später innerhalb der Kategorie-Ansichten eine neue Sektion, in der ebenfalls 8 Plätze für Empfehlungen zur Verfügung standen, wobei nur 2 durch personali-

[2] `https://www.leshop.ch/`

sierte Verfahren gefüllt wurden und die restlichen bis zu 6 Plätze vom Betreiber händisch ausgesucht wurden.

Im Rahmen dieser Testumgebung wurde untersucht, wie sehr das Empfehlungssystem überhaupt akzeptiert wurde, wie stark der Umsatz des Online-Supermarktes durch das Empfehlungssystem direkt gesteigert wurde und welche langfristigen Umsatzsteigerungen dadurch erzielt werden konnten. Da das Datenmodell während des Testzeitraums nur 5 Mal aktualisiert werden konnte, wurde außerdem untersucht, welche Auswirkungen eine Aktualisierung des Datenmodells mit sich bringt.

Um herauszufinden, wie gut das Empfehlungssystem angenommen wurde, wurde dazu die Anzahl der Leute, die einen empfohlenen Artikel „akzeptiert“ haben, ins Verhältnis zur Gesamtzahl der Besucher, die mindestens einen Artikel auf der Seite gekauft habe, gesetzt und monatlich aufgetragen. Dabei konnte beobachtet werden, dass die Benutzer sich mit dem neuen System zuerst vertraut machen müssen. So stieg dieses Verhältnis von anfangs 0,7% bis 1,8% schließlich am Ende auf bis zu 3,9% an. Außerdem war gut zu erkennen, dass jede Aktualisierung des Datenmodells zu einer Steigerung im Bereich von 0,1 bis 0,5 Prozentpunkten führte. Das Ausbleiben der Aktualisierung führte im Gegenzug wieder zu einer leichten Verschlechterung.

Bei Betrachtung der Umsatzsteigerung durch das Empfehlungssystem konnte gezeigt werden, dass diese bei einem ausreichend aktuellen Datenmodell um knapp 0,1% gesteigert werden konnte, jedoch nach 1-2 Monate alten Daten wieder auf fast 0% absank. Betrachtet man außerdem die indirekten Auswirkungen, d.h. erneut verkaufte Produkte, die ursprünglich aus einer Empfehlung stammten, oder Produkte, die aus einer Kategorie stammten, auf die der Kunde lediglich über ein empfohlenes Produkt aufmerksam gemacht wurde, so sind ebenfalls deutlich höhere Umsatzsteigerungen festzustellen. In diesem Fall können zusätzlich zu der direkten Umsatzsteigerung noch einmal 0,12 bis 0,15 Prozentpunkte hinzuaddiert werden, die aufgrund ihrer längerfristigen Auswirkungen auch weniger stark von veralteten Daten beeinflusst werden.

Allerdings lässt sich an dieser Stelle anmerken, dass insbesondere indirekte Umsatzsteigerung nur schwer messbar ist, da nicht sichergestellt werden kann, dass die Produkte nicht auch auf anderem Wege vom Benutzer hätten entdeckt werden können. So hätte es ohne Empfehlungssystem ebenfalls sein können, dass er durch selbstständiges Stöbern auch auf die Produkte gekommen wäre oder diese möglicherweise aufgrund der Empfehlung eines Freundes entdeckt hätte.

Insgesamt macht die Arbeit jedoch deutlich, welches Potential in Empfehlungssystemen steckt und legt dar, dass es wichtig ist ein stets aktuelles Datenmodell zu pflegen, um eine kontinuierlich gute Empfehlungsqualität zu gewährleisten. Außerdem ist zu beachten, dass eine Steigerung des Umsatzes um bereits 0,1% bei einem Unternehmen in dieser Größe durchaus erheblich sein kann.

2.4.5 Zusammenfassung

Beim realen Einsatz der zuvor offline evaluierten Empfehlungsverfahren fällt möglicherweise auf, dass andere Merkmale abseits der klassischen Metriken noch besser geeignet sind, um Aussagen über die Qualität zu treffen. Dafür gibt es eine Reihe von Ursachen, die bereits in verschiedenen wissenschaftlichen Untersuchungen näher betrachtet wurden.

In [19] wird angesprochen, dass eine zu große Ähnlichkeit der Empfehlungen problematisch sein kann. So werden die Algorithmen derzeit mittels der bestehenden Metriken für ihre gute Treffsicherheit belohnt. Das führt jedoch dazu, dass diese stets dem Geschmack des Benutzers folgen und ihn wenig auf unterschiedliche Produkte lenken. Stellt dieser dann fest, dass alle Empfehlungen gleich sind und ihn alle nicht interessieren, kann es passieren, dass er sie einfach ignoriert und sich nicht mehr darauf einlässt, da es für ihn keinen Mehrwert bringt. Zur Berücksichtigung dieses Problems ist es daher sinnvoll, sich nicht nur auf die klassischen Genauigkeitsmetriken zu verlassen, sondern ebenfalls die Ähnlichkeit eines Produktes zur Gesamtliste z.B. mittels der *Intra-List Similarity Metric*[26] zu berücksichtigen.

Ebenfalls kann es z.B. bei einem Reiseportal durchaus gewinnbringend sein, dem Nutzer völlig neue unbekannte Reiseziele vorzuschlagen, die ihn in eine neue Richtung lenken, auf die er selbst nicht gekommen wäre [19]. Diese Unerwartetheit kann somit einen Überraschungseffekt erzielen, der ihn schließlich zum Kauf anregt. Die Schwierigkeit liegt jedoch darin, diesen Faktor zu messen, da er stark vom persönlichen Befinden beeinflusst wird und schwierig vorherzusagen ist.

Weitere Unterschiede in der Live-Evaluation ergeben sich bei der Empfehlung von überdurchschnittlich populären Artikeln. Auch diese werden von den klassischen Metriken besonders honoriert, da sie naturgemäß von den meisten Benutzern für gut befunden werden. Hier hat aber die Untersuchung einer Nachrichtenseite gezeigt, dass dort genau das Gegenteil der Fall ist [9]. Dies ist dadurch zu erklären, dass der Leser die bekanntesten Artikel bereits vorab auf einer anderen Seite gelesen hat oder bereits ohne Empfehlung auf der Startseite darauf gestoßen ist. In dieser Untersuchung wurde jedoch deutlich, dass das dort eingesetzte personalisierte Verfahren, das auf Basis von Kontextbäumen arbeitet, insgesamt eine ungefähr doppelt so hohe *click-through rate* erzielt, wie das, das ausschließlich die populärsten Artikel empfiehlt. Allerdings wird auch aufgezeigt, dass die richtige Positionierung der Empfehlungen ebenfalls relevant ist und Seiteneffekte mit anderen z.B. manuell ausgewählten Artikeln auf derselben Seite zu berücksichtigen sind.

Ein weiterer wichtiger Gesichtspunkt ist das Vertrauen des Benutzers. So ist es durchaus wichtig, einem neuen Benutzer von Anfang an gute Empfehlungen zu liefern, die ihm den Nutzen des Systems möglichst gut verdeutlichen. Ist dies nicht der Fall, so kann es sein, dass dieser das Vertrauen sofort verliert und das Empfehlungssystem zu dem Zeitpunkt, als es genügend von ihm gelernt hat, bereits gar nicht mehr beachtet wird [19].

Insbesondere beim Empfehlen von unbekannteren Dingen ist es wichtig, das Vertrauen dadurch nicht zu zerstören [7].

Empfehlungssysteme lassen sich dafür einsetzen, um einen Benutzer von neuen Dingen zu überzeugen und sein Verhalten und seine Interessen damit aktiv zu beeinflussen. So wurde in [2] am Beispiel eines Video-on-Demand-Dienstes gezeigt, dass die Wahl eines Films dadurch gezielt gelenkt werden kann und sogar dazu beitragen kann, die Vielfältigkeit der gekauften Filme zu vergrößern.

Ansonsten gilt es zu berücksichtigen, dass die Modelle zur Generierung der Empfehlungen stets aktuell gehalten werden sollten, da sich sonst nach einigen Wochen die erzielte Umsatzsteigerung durch das Empfehlungssystem wieder nachweislich verringert [5].

Abschließend lässt sich zur Evaluierung also festhalten, dass sich die Ergebnisse in einem Live-Einsatz durchaus von den offline bestimmten unterscheiden können, allerdings durch sinnvolle und auf den Einsatzzweck abgestimmte Wahl der richtigen Metriken durchaus eine gute Annäherung erreicht werden kann [14, 3]. Dabei können die klassischen Metriken wie Precision und Recall ebenfalls gute Anhaltspunkte liefern, decken aber nicht das gesamte zu berücksichtigende Spektrum vollständig ab. Es gibt jedoch auch Untersuchungen, die im Online-Betrieb verglichen mit den offline berechneten Ergebnissen sogar eine gegensätzliche Rangfolge der Algorithmen zu Tage brachte [9].

Insgesamt geht bei Online-Evaluationen die Tendenz dahin, dass personalisierte Verfahren einen größeren Erfolg erzielen als unpersonalisierte [9, 13]. In einem Vergleich von 20 verschiedenen Empfehlungsverfahren auf der Webseite Forbes.com hat sich sogar herausgestellt, dass eine Mischung aus inhaltsbasierten und kollaborativen Verfahren mit einem Vorsprung von 23% gegenüber dem zweitbesten Algorithmus die höchste *click-through rate* erzielt hat.

Nicht zuletzt müssen verschiedene weitere praxisrelevante Gesichtspunkte mit berücksichtigt werden. Zum einen muss sichergestellt werden, dass die eingesetzten Algorithmen für den Produktivbetrieb geeignet sind und somit z.B. im E-Commerce die Empfehlungen während des Seitenaufbaus schnell genug generiert und eingeblendet werden können. Das macht es je nach Verfahren erforderlich, bereits vorab offline ein Modell zu berechnen, das die zu verarbeitende Datenmenge bei jedem Aufruf verringert und somit die nötige Generierungszeit sowie den erforderlichen Speicherverbrauch reduziert. Zum anderen dürfen auch Aspekte wie der Datenschutz oder der sinnvolle Darstellung auf mobilen Geräten nicht außer Acht gelassen werden [15, S. 26].

Kapitel 3

Technischer Aufbau

Dieses Kapitel beschäftigt sich mit dem technischen Aufbau des Empfehlungssystems, das auf dem Schnäppchenblog China-Gadgets.de zum Einsatz kam und soll aufzeigen, welche spezifischen Anforderungen bei der Implementierung zu berücksichtigen waren. Das Ziel des Aufbaus ist es, zu untersuchen, welche Auswirkungen Empfehlungssysteme auf das Klick- und Kaufverhalten eines Benutzers haben. Dafür wurde ein Empfehlungssystem entwickelt, welches dem Besucher weitere, für ihn interessante Produkte anzeigt und alle Interaktionen mit diesen Vorschlägen zur Auswertung protokolliert. Die Empfehlungen sollen dabei nach einem zufällig bestimmten, aber für einen einzelnen Besucher stets festen Verfahren generiert werden. Dabei kamen sowohl inhaltsbasierte als auch kollaborative Verfahren zum Einsatz.

Die ausgewählten Artikel werden unterhalb des ersten Beitrags auf der Startseite eingeblendet (s. Abbildung 3.1) und können sich je nach verwendetem Verfahren entweder auf das obige Produkt beziehen oder für den Besucher personalisiert sein.

Außerdem muss sichergestellt werden, dass die Webseite auch bei einem Ausfall des Empfehlungssystems weiterhin erreichbar bleibt und dieses die Seite nicht negativ beeinflussen kann. Damit geht auch einher, dass die Empfehlungen so schnell wie möglich generiert werden müssen, da der Besucher die Seite sonst bereits verlassen oder daran vorbei gescrollt hat, bevor diese angezeigt wurden.

3.1 Domänenspezifische Besonderheiten

Die Besonderheit bei einem Schnäppchenblog im Gegensatz zu einem klassischen Online-Shop liegt darin, dass auf der Seite selbst keine Produkte verkauft, sondern diese lediglich beworben werden. Das bedeutet, dass Redakteure dort täglich mehrere Beiträge veröffentlichen, die jeweils in unterschiedlich ausführlicher Form ein bestimmtes Produkt präsentieren. Ebenfalls gibt es keinen klassischen Warenkorb, der zum Generieren der Empfehlungen mit ausgewertet werden könnte. Daher ist es erforderlich, sich ausschließlich auf die Auf-

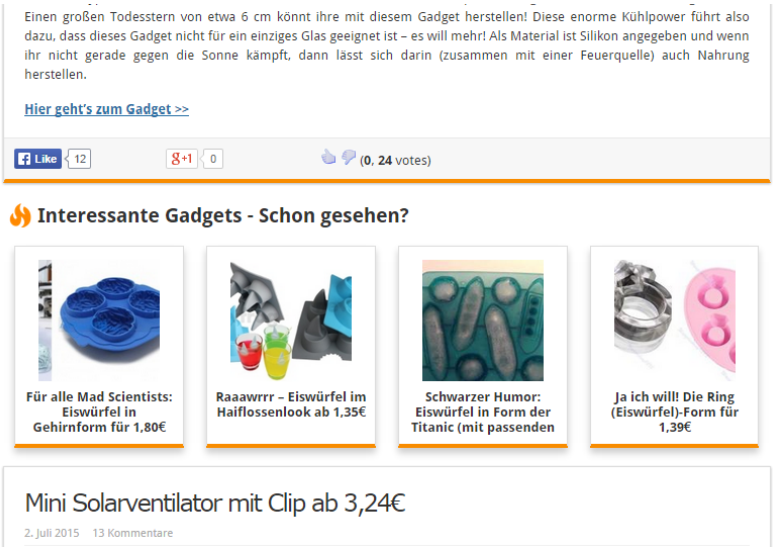

Abbildung 3.1: Eingeblendete Empfehlungen

rufe der Beiträge und der daraus resultierenden ausgehenden Verweise auf die Webseiten der jeweiligen Anbieter zu verlassen.

3.2 Realisierung

Als Web-Plattform wird auf dem Hauptserver China-Gadgets.de die freie Weblog-Software *WordPress*[1] eingesetzt, die es aufgrund ihrer offenen Struktur sehr einfach ermöglicht, das Aussehen und Verhalten einer Webseite individuell anzupassen. Zur Protokollierung und Auswertung der Besucherdaten wird außerdem das Open-Source Webanalytik-Werkzeug *Piwik*[2] eingesetzt, das ebenfalls über eine API zum Exportieren der Besucherdaten verfügt und es erlaubt, direkt Zugriff auf die dahinter angeschlossene Datenbank zu nehmen.

Auf der Startseite von China-Gadgets.de werden stets die aktuellen Beiträge in absteigender chronologischer Reihenfolge präsentiert. Unterhalb des ersten dargestellten Beitrags erfolgt die Einblendung der Empfehlungen. Diese beziehen sich dabei, je nach verwendetem Verfahren, entweder direkt auf den vorangegangenen Beitrag oder sind für den jeweiligen Besucher personalisiert.

Dazu wird, wie in Abbildung 3.2 verdeutlicht, nach dem Laden der Seite eine AJAX-Anfrage an den Server zurückgeschickt (1), die die aktuelle Besucher-ID von Piwik, die Beitrag-ID des obigen Beitrags und ebenfalls die aktuelle Bildschirmbreite (zur Erkennung von mobilen Webseitennutzern) enthält. Diese Anfrage wird anschließend von einer WordPress-Erweiterung (2) verarbeitet und an den Empfehlungsserver weitergeleitet (3), der wiederum eine Liste von Beitrag-IDs zurückgibt (4), die den empfohlenen Artikeln ent-

[1] `https://wordpress.org/`
[2] `http://piwik.org/`

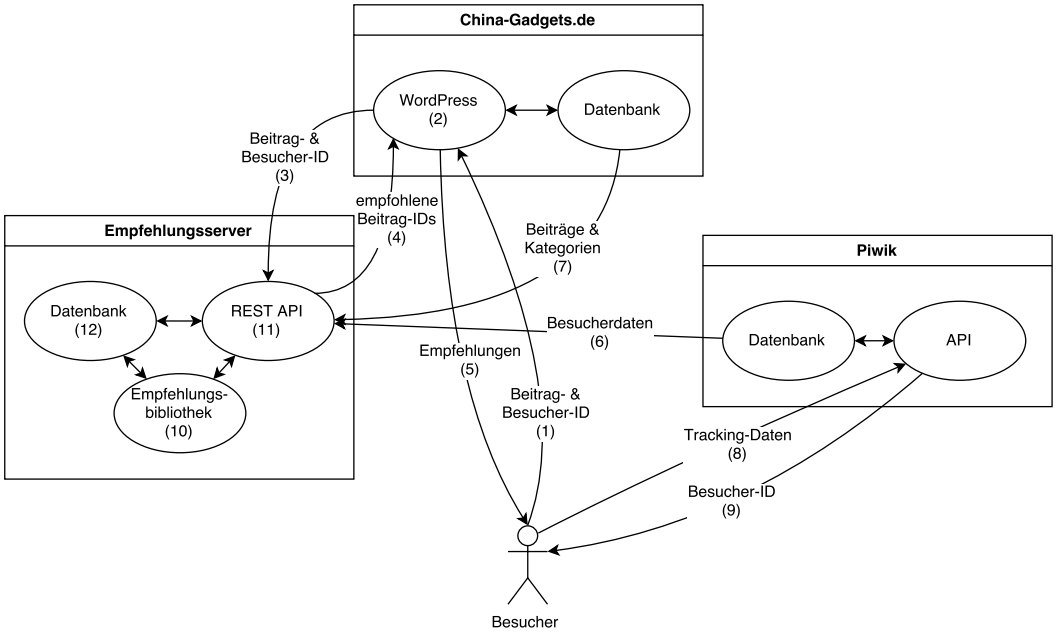

Abbildung 3.2: Server-Kommunikation

sprechen. Diese Artikel werden schließlich von WordPress mit den zugehörigen Titeln und Beitragsbildern aufgefüllt, formatiert und an den Browser zur Ausgabe gegeben (5). Außerdem wird eine eindeutige Referenz-ID übertragen, die über eine Datenbank eine spätere Zuordnung der empfohlenen Beiträge zu den jeweils verwendeten Verfahren ermöglicht.

Zusätzlich wird zur anschließenden Datenauswertung das Verhalten des Benutzers überwacht. Sobald die Empfehlungen, die anfangs in der Regel unterhalb des sichtbaren Bereichs (*below the fold*) im Browser liegen, durch Scrollen auf dem Bildschirm erscheinen, wird dies dem Empfehlungsserver mitgeteilt und die ausgegebene Empfehlung in der Datenbank als „gesehen“ markiert. Klickt dieser anschließend eine Empfehlung an, so wird dies, genauso wie das finale Besuchen der Webseite des Anbieters eines Produktes, ebenfalls zur Protokollierung zurückgemeldet.

Um den bereits bestehenden Datenbestand (d.h. die Beiträge und verfügbaren Besucherdaten) komfortabel nutzen zu können, wurden verschiedene kleine Werkzeuge entwickelt, die es ermöglichen, den gesamten nötigen Datenbestand aus WordPress und Piwik im JSON-Format zu exportieren. Diese Daten werden schließlich auf einem weiteren dritten Server, der ausschließlich für die Generierung der Empfehlungen vorgesehen ist, wieder eingelesen (6,7). Die Wahl dieses unabhängigen Servers wurde getroffen, um eine möglicherweise negative Beeinflussung der Hauptseite (z.B. durch Programmfehler) so gering wie möglich zu halten und außerdem genügend Ressourcen für das Empfehlungssystem selbst zur Verfügung zu haben. Piwik protokolliert dabei völlig unabhängig die aufgerufe-

nen Beiträge mit (8) und setzt entsprechende Cookies um eine eindeutige zurückverfolgbare Besucher-ID zu generieren (9).

Gelangt ein Besucher das erste Mal auf die Seite, so wird ihm eines von sechs zufälligen Verfahren zugelost und seiner einmaligen Benutzer-Kennung zugeordnet. Ist dieser bereits ein wiederkehrender Besucher, so wird das vorherige Verfahren erneut verwendet. Wurde ein personalisiertes Verfahren ausgewählt, welches eine bestimmte Mindestanzahl an Interaktionen voraussetzt, um zuverlässige Empfehlungen zu erzeugen, wird temporär bis zum Erreichen dieser Mindestanzahl auf ein sekundäres, nicht personalisiertes Verfahren zurückgegriffen.

3.3 Empfehlungssystem

Das Empfehlungssystem selbst besteht aus drei wesentlichen Komponenten:

- Einer Anbindung der **Empfehlungsbibliothek** (Abbildung 3.2 (10)) *Recommender101* [14], die mit den exportierten Besucher- und Beitragsdaten gefüllt wird,
- einer **REST-Schnittstelle** (3.2 (11)), die mittels der Bibliothek *RESTX*[3] realisiert wird und insbesondere vom Hauptserver eingehende Empfehlungsanfragen, die die Besucher-ID und Beitrag-ID enthalten, mit den jeweiligen empfohlenen Beiträgen beantwortet
- und einer *MySQL*-**Datenbankanbindung** (Abbildung 3.2 (12)), die alle ausgegebenen Empfehlungen protokolliert und ebenfalls nach Rückmeldung vom Hauptserver festhält, welche Beiträge vom Besucher aufgerufen wurden.

Die REST-Schnittstelle ist der zentrale Kommunikationspunkt und wird sowohl zum Importieren der Benutzer- und Beitragsdaten, zum Ausgeben der Empfehlungen, als auch zur Rückmeldung der angeklickten Artikel verwendet. Die importierten Daten werden dabei in das Datenmodell von Recommender101 und in eigene Datenstrukturen eingefügt, wobei letztere hauptsächlich dazu verwendet werden, triviale Algorithmen wie die Ausgabe der zuletzt angeklickten Artikel zu realisieren, die von Recommender101 nicht zur Verfügung gestellt werden.

3.3.1 Datenimport

In regelmäßigen Abständen (d.h. durchschnittlich alle 15 Minuten) wird ein neuer Datenexport gestartet. Das bedeutet, dass die neuen Besucherdaten aus Piwik und alle Beiträge aus WordPress mit ihren zugehörigen Kategorien abgeholt werden. Dadurch bleibt das Empfehlungssystem immer synchronisiert und kann auf einen möglichst aktuellen Datenbestand zugreifen, der dem Benutzer schnellstmöglich zugute kommt. Die Beiträge

[3] `http://restx.io/`

werden anschließend nach der Bereinigung von URLs, HTML-Tags und anderen Formatierungen vom *tfidfVectorizer* aus der Werkzeug-Sammlung von Recommender101 in TF-IDF-Vektoren umgerechnet. Diese geben dabei an, welche Wörter aus einem Beitrag wie charakteristisch für ihn sind. Dabei werden die Wörter ebenfalls, soweit es möglich ist, auf ihre Wortstämme reduziert (*stemming*), um Abweichungen z.B. aufgrund verschiedener Numeri und Konjugationen zu minimieren. Ist dies geschehen, wird der Import-Vorgang angestoßen und die Aktualisierung über die REST-Schnittstelle in die laufende Instanz des Empfehlungsservers geladen. Dieser füllt damit die Datenstrukturen neu auf, berechnet die Kosinus-Ähnlichkeiten der TF-IDF-Vektoren und initialisiert die Algorithmen. Dies geschieht parallel zum laufenden Betrieb und dauert ca. 10 Minuten. Im Anschluss werden die alten Daten verworfen und die neuen für den Live-Betrieb verwendet.

3.3.2 Verfahren

Das Empfehlungssystem implementiert sechs verschiedene Verfahren, die zufällig auf die Benutzer verteilt und ausgewählt werden. Von diesen sechs Verfahren sind die ersten drei nachfolgend beschriebenen personalisiert und die darauffolgenden drei unpersonalisiert.

Letzte Artikel (RecItems) Bei diesem Verfahren werden dem Benutzer seine zuletzt angeklickten Artikel in umgekehrter Reihenfolge erneut als Erinnerung präsentiert, wie dies ebenfalls beim Online-Versandhändler *Amazon*[4] der Fall ist. Die Aktualisierung erfolgt dabei jedoch nicht sofort, sondern erst nach erneutem Import der Besucherdaten aus Piwik.

Bayesian Personalized Ranking (BPR) Für BPR [20] wird Recommender101 [14] verwendet und der Algorithmus mit den Parametern $numFeatures = 20$, $initialSteps = 150$, $regU = 0,0025$, $regI = 0,0025$ und $regJ = 0,00025$ initialisiert. Während der zweiten Testphase (s. Kapitel 4) wurden *numFeatures* und *initialSteps* auf 30 und 1000 angepasst, die in vorab getätigten Untersuchungen für den gegebenen Datensatz die besten Ergebnisse erzielten. Anschließend werden die empfohlenen Artikel für einen gegebenen Benutzer zurückgegeben.

Content-based Personalized Ranking (CBP) Beim CBP wird ebenfalls Recommender101 [14] verwendet. Dazu werden die zuvor generierten TF-IDF-Vektoren der besuchten Beiträge eines Benutzers zu einem persönlichen Durchschnittsvektor verrechnet und mit denen der restlichen Beiträge verglichen. Die Beträge, die die größte Kosinus-Ähnlichkeit zum Benutzerprofil aufweisen, werden anschließend zurückgegeben.

Populäre Artikel (PopItems) Zur Ausgabe der populärsten Artikel werden die Beträge zuerst nach ihren zugeordneten Oberkategorien gruppiert und anschließend innerhalb

[4] `http://www.amazon.de/`

dieser Kategorisierung nach ihrer Aufrufzahl absteigend sortiert. Bei einer eingehenden Empfehlungsanfrage werden zuerst die Kategorien des Beitrags, unter dem die Empfehlungen dargestellt werden sollen, bestimmt und danach die in diesen Kategorien populärsten Artikel zurückgegeben.

Ähnliche Artikel (SimItems) Um weitere ähnliche Artikel zu einem gegebenen Beitrag zu finden, werden die Beiträge zurückgegeben, die aufgrund ihrer Kosinus-Ähnlichkeit der TF-IDF-Vektoren die größte Übereinstimmung zu ihm haben.

Co-occurrence patterns (CoOcc) Zur Erzeugung von Empfehlungen nach dem Verfahren „Kunden, die diesen Artikel kauften, kauften auch...“ von Amazon, werden zusammengehörige Aufrufe von Artikeln eines einzelnen Benutzers ausgewertet. Dazu wird eine Artikel-Artikel-Tabelle erzeugt, dessen Felder den Wert „0“ beinhalten. Für alle möglichen Beitragspaare, die jeder einzelne Benutzer nun während des gesamten Aufnahmezeitraums besucht hat, wird der entsprechende Wert in der Tabelle um eins erhöht. Alternativ wäre es hier ebenfalls möglich, statt des gesamten Aufnahmezeitraums nur die aufgerufenen Beiträge innerhalb eines einzelnen Besuches zu verwenden. Da die Besuche in den exportierten Besucherdaten jedoch nicht mehr getrennt voneinander vorliegen und ebenfalls keine Zeitinformationen mehr zur Verfügung stehen, wurde sich hier dazu entschieden, den gesamten Aufnahmezeitraum eines Besuchers auszuwerten. Sollen nun zu einem Beitrag Empfehlungen nach diesem Verfahren ausgegeben werden, so werden die anhand der erzeugten Tabelle am meisten zusammen mit diesem Artikel aufgerufenen Beiträge zurückgegeben.

3.3.3 Datenbank

Die Informationen darüber, welche Verfahren für die einzelnen Besucher verwendet wurden, werden zusammen mit den protokollierten Tracking-Daten in einer MySQL-Datenbank in den Tabellen *user*, *recommendation* und *post* abgelegt.

Wie in Abbildung 3.3 zu erkennen ist, wird dabei für jeden Benutzer ein primäres und ein sekundäres Verfahren bestimmt und zusammen mit einem Zeitstempel abgelegt. Das primäre wird zufällig aus allen sechs verfügbaren ausgewählt, das sekundäre nur aus den unpersonalisierten Verfahren. Handelt es sich bei dem primären Verfahren um ein personalisiertes Verfahren und sollten zu dem Zeitpunkt nicht ausreichend Daten des Besuchers für eine zuverlässige Empfehlung zur Verfügung stehen, so wird auf das sekundäre Verfahren zurückgegriffen.

Für jede getätigte Empfehlung wird protokolliert, wann diese ausgegeben wurde, auf welchen Beitrag sie sich ggf. bezieht und ob und wann diese auf dem Bildschirm sichtbar war. Um anschließend ebenfalls eventuelle Unterschiede zwischen verschiedenen verwen-

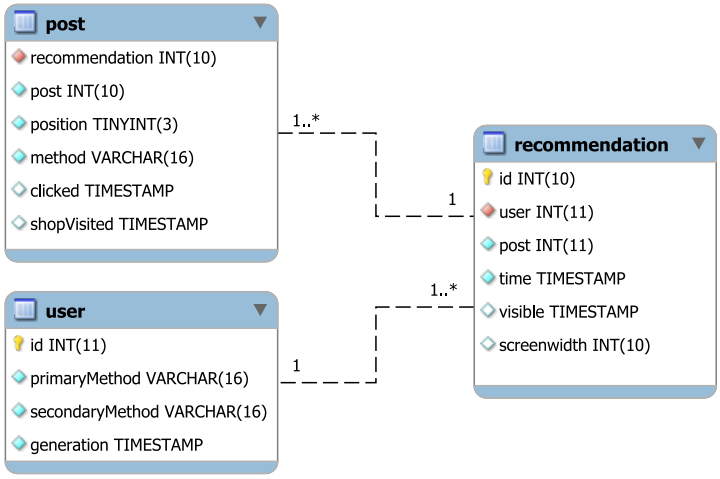

Abbildung 3.3: Datenbankmodell

Tabelle 3.1: Erläuterung zu Abbildung 3.3

Tabellenspalte	Bedeutung
Tabelle: **user**	
id	ID des Besuchers
primaryMethod	primäres Verfahren
secondaryMethod	sekundäres Verfahren
generation	Zeitpunkt, zu dem der Besucher erstmalig eine Empfehlung angezeigt bekommen hat
Tabelle: **recommendation**	
id	ID des angezeigten Empfehlungsblocks
user	ID des Besuchers
post	ID Beitrags oberhalb der eingeblendeten Empfehlungen
time	Generierungszeitpunkt
visible	ggf. Zeitpunkt, zudem die Empfehlung durch Scrollen sichtbar wurde
screenwidth	Breite des Browsers
Tabelle: **post**	
recommendation	ID des Empfehlungsblocks
post	angezeigte Beitrag-ID
position	Position des Beitrags innerhalb des Empfehlungsblocks (1-4)
method	verwendetes Verfahren
clicked	ggf. Zeitpunkt, zu dem der Beitrag angeklickt wurde
shopVisited	ggf. Zeitpunkt, zu dem der Shop besucht wurde

deten Geräten, d.h. Smartphones, Tablets und Desktop-PCs erkennen zu können, wird außerdem die verwendete Bildschirmbreite mit abgelegt.

Eine einzelne Empfehlung besteht dabei aus mehreren Beiträgen, für die die Positionen innerhalb des Blocks, die verwendeten Verfahren und die Tracking-Daten über Klicks und ausgehende Anbieter-Besuche mit abgelegt werden. An dieser Stelle ist anschließend ebenfalls erkennbar, ob die Empfehlungen mit weiteren Beiträgen aus anderen Verfahren aufgefüllt werden mussten, wenn nicht genügend zur Verfügung standen.

Kapitel 4

Evaluation

Um den Erfolg eines live eingesetzten Empfehlungssystems zu messen, stehen verschiedene Möglichkeiten zur Verfügung. Die aussagekräftigste Metrik wäre es, die Umsatzsteigerung zu messen, die durch das Empfehlungssystem bzw. die verschiedenen eingesetzten Verfahren erzielt wurde. Diese Zahl ist jedoch insbesondere bei einem Schnäppchenblog wie China-Gadgets.de schwierig zu erfassen, da sich diese über Provisionen finanzieren, die von den jeweiligen Anbietern der Produkte gezahlt werden und nicht immer präzise genug ermittelt werden können. Das liegt daran, dass dabei in der Regel mit diversen Werbenetzwerken zusammengearbeitet wird, die die geschäftliche Abwicklung der verschiedenen Shops übernehmen. Allerdings stellen diese Netzwerke oftmals keine einheitliche zuverlässige Schnittstelle zur Verfügung, die einen eindeutigen Rückschluss der Verkäufe auf die einzelnen Besucher zulässt.

Alternativ bleibt nur die Möglichkeit, die Klicks auf die empfohlenen Beiträge zu zählen und ebenfalls zu protokollieren, wenn einem ausgehenden Verweis auf einen Shop gefolgt wurde und dies als Kaufinteresse aufzufassen. Als Metrik kann dazu die *click-through rate* (s. Kapitel 2.4) verwendet werden. Diese wurde dabei sowohl für die Empfehlungen auf der Startseite als auch für die Shop-Verlinkungen in einem Artikel bestimmt.

4.1 Messung und Basisdaten

Der Testzeitraum wurde in zwei Hälften unterteilt, wobei nach der ersten noch einige kleine Änderungen vorgenommen wurden. Der erste Testzeitraum erstreckte sich dabei vom 11.06.2015 bis 10.07.2015 und der zweite vom 10.07.2015 bis 06.08.2015. Innerhalb eines jeden Zeitraums wurde für jeden Benutzer protokolliert, welche Beiträge ihm mittels welcher Verfahren angezeigt wurden. Außerdem wurde ggf. aufgezeichnet, wenn dieser eine Empfehlung gesehen, diese angeklickt oder den Shop besucht hat. Als **gesehen** galt eine Empfehlung dann, wenn sie durch Scrollen des Benutzers in das Sichtfeld seines Webbrowsers gebracht wurde. Über diese Aufzeichnungen können dann im Anschluss die

Tabelle 4.1: Anzahl sichtbar empfohlener Beiträge in den jeweiligen Testzeiträumen

Verfahren	Beiträge in Zeitraum 1	Beiträge in Zeitraum 2
BPR	74.588	74.056
CBP	71.236	70.780
RecItems	74.977	70.750
PopItems	182.671	153.892
CoOcc	185.383	170.579
SimItems	205.321	216.051

click-through rates bestimmt werden. Dabei kann sowohl eine Beitrag-CTR, die die Klicks zu den Impressionen eines empfohlenen Beitrags in Beziehung setzt, als auch eine Shop-CTR bestimmt werden, die den Anteil der Besucher angibt, die nach einer angeklickten Empfehlung schließlich den Shop besucht haben. Durch Multiplikation dieser beiden Werte ergibt sich die Shoprate, die den Anteil der Shopbesuche gegenüber den ursprünglichen Impressionen eines empfohlenen Beitrags aufzeigt.

Jedem Besucher wurde eines der 6 Verfahren zugelost, wobei ggf. auf ein alternatives unpersonalisiertes Verfahren zurückgegriffen wurde, wenn ein Besucher zu dem Zeitpunkt bislang weniger als 5 Beiträge angeklickt hatte.

Während beider Testzeiträume kamen dabei die Verfahren BPR, CBP, RecItems, PopItems, CoOcc und SimItems (s. Kapitel 3.3.2) zum Einsatz, wobei CoOcc im ersten Testzeitraum nicht gewertet werden konnte, da ein Fehler in der Implementierung vorlag, der erst für den zweiten Teil behoben wurde. Außerdem wurden im Zuge dessen die Parameter von BPR von $numFeatures = 20$ und $initialSteps = 150$ auf $numFeatures = 30$ und $initialSteps = 1000$ angepasst, da sich diese nach einer Offline-Analyse als geeigneter herausstellten.

Die Verfahren wurden dabei zu Beginn des Experiments mit den Beiträgen der letzten 12 Monate und mit den Besucherdaten der letzten 6 Monate angelernt und ungefähr alle 15 Minuten mit den ggf. neu hinzugekommenen Daten ergänzt. Am Ende der Evaluierung umfasste der Datensatz somit insgesamt 1.523 Beiträge und gut 2 Millionen Aufrufe.

China-Gadgets.de verzeichnete während des gesamten Testzeitraums pro Tag durchschnittlich 10.874 Besuche von denen 71% wiederkehrend waren. Ein Besuch ist **wiederkehrend**, wenn es mindestens der zweite Besuch eines Besuchers ist. In diesen Zahlen sind die Besuche, die über die zugehörige App erfolgt sind, bereits ausgenommen, da dort keine Empfehlungen angezeigt wurden und daher nicht weiter betrachtet werden. Dabei haben 53% der wiederkehrenden Besucher die Seite bereits mindestens 9 Mal besucht. Auf der Startseite wurden pro Tag durchschnittlich 9.525 Empfehlungen mit jeweils 4 Beiträgen generiert, die zu 69% vom Benutzer durch Scrollen sichtbar wurden. Insgesamt lagen am Ende des Experiments 4.128 veröffentliche Beiträge in der Datenbank.

Die Anzahl der jeweils empfohlenen Beiträge ist in Tabelle 4.1 abhängig von den verwendeten Verfahren aufgezeigt und ergibt geteilt durch 4 die Anzahl der ausgegebenen Empfehlungsblöcke. Da ein Besucher erst nach einer Mindestanzahl von 5 Interaktionen zu einem personalisierten Verfahren zugewiesen wurde, ist die Anzahl der empfohlenen Beiträge bei den unpersonalisierten Verfahren insgesamt deutlich höher. Außerdem wurde zusätzlich SimItems vereinzelt als letztes Rückfall-Verfahren verwendet, falls insgesamt weniger als 4 Beiträge empfohlen werden konnten. Dies trat vor allem bei neu veröffentlichten Beiträgen innerhalb der ersten ungefähr 15 Minuten auf, bei denen nicht-personalisierte kollaborative Verfahren aufgrund fehlender Besucherdaten noch nicht anwendbar waren.

4.2 Signifikanztest

Um sicherstellen zu können, dass die beobachteten Unterschiede zwischen den einzelnen Verfahren nicht durch Zufall entstanden sind, bietet es sich an, Signifikanztests durchzuführen. Damit kann die Wahrscheinlichkeit, dass allein der Zufall zu den vorliegenden Ergebnissen geführt hat, abgeschätzt werden. Da pro Verfahren jeweils nur zwei verschiedene Merkmalsausprägungen möglich sind (entweder der Benutzer hat die Empfehlung bzw. den Shop angeklickt oder nicht), können diese paarweise mittels des Vierfeldertests ausgewertet werden. Der **Vierfeldertest** [23] ist eine spezielle Form des **Chi-Quadrat-Tests** und berechnet für zwei Merkmale (in diesem Fall für die Klickverhalten zwei verschiedener Verfahren) eine Prüfgröße χ^2, die schließlich in den p-Wert umgerechnet werden kann. Der p-Wert [10] gibt dabei an, wie wahrscheinlich es ist, dass allein durch Zufall das vorliegende Ergebnis oder ein extremeres entstanden ist.

Zunächst muss eine Vierfeldertafel erstellt werden, in der die gemessenen absoluten Häufigkeiten abhängig von den jeweiligen Verfahren gegeneinander aufgetragen werden:

Tabelle 4.2: Vierfeldertafel

	Verfahren 1	Verfahren 2	Σ
angeklickt	a	b	$a+b$
nicht angeklickt	c	d	$c+d$
Σ	$a+c$	$b+d$	$n=a+b+c+d$

Aus dieser Tafel lässt sich anschließend die **Prüfgröße** χ^2 bestimmen:

$$\chi^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot n}{(a+c) \cdot (b+d) \cdot (a+b) \cdot (c+d)} \tag{4.1}$$

Der **p-Wert** berechnet sich im Anschluss anhand einer Näherungsformel wie folgt [6]:

$$p = \frac{1}{2} \cdot 10^{-\frac{\chi^2}{3,84}} \tag{4.2}$$

Über den bestimmten p-Wert kann nun nach Festlegung eines Signifikanzniveaus entschieden werden, ob das vorliegende Ergebnis statistisch signifikant ist oder nicht. Das Signifikanzniveau α ist eine Wahrscheinlichkeit und dient als Schwellenwert zur Ablehnung der Nullhypothese. Die Nullhypothese gilt es zu widerlegen und besagt, dass zwischen beiden Verfahren kein Unterschied besteht. Das übliche Signifikanzniveau, das in der Wissenschaft häufig verwendet wird, liegt bei $\alpha = 0,05$. Liegt der bestimmte p-Wert nun unterhalb dieses Niveaus, so wird davon ausgegangen, dass die Nullhypothese abgelehnt werden kann, da sie wahrscheinlich nicht als Erklärung für das vorliegende Ergebnis in Frage kommt.

Hier ist jedoch zu beachten, dass das Signifikanzniveau durch mehrfaches Testen auf demselben Datensatz nach unten angepasst werden sollte, da sich das Gesamtrisiko, bei einem der vielen Tests die Nullhypothese fälschlicherweise verworfen zu haben, erhöht. Die Anpassung hängt von der Anzahl der durchgeführten Tests i ab und liegt bei 6 Verfahren und paarweisem Testen bei insgesamt 15 Tests. Damit ergibt sich für das angepasste Signifikanzniveau anhand der **Bonferroni-Korrektur** [1] der folgende Wert:

$$\alpha_i = 1 - (1 - 0,05)^{1/i} \implies \alpha_{15} \approx 0,00341 \tag{4.3}$$

Im nachfolgenden Kapitel werden einige Berechnungen damit durchgeführt.

4.3 Ergebnisse

Abbildung 4.1 zeigt die *click-through rates* der durch die verschiedenen Verfahren empfohlenen Beiträge. Dabei wird deutlich, dass insbesondere im zweiten Testzeitraum PopItems zusammen mit CoOcc am besten abgeschnitten haben. Hingegen hat BPR die schlechtesten Ergebnisse erzielt, wobei durch Anpassung der Parameter eine leichte Steigerung im Vergleich zum ersten Testzeitraum erkennbar ist. Ebenfalls ist nach Implementierung des korrekten CoOcc-Verfahrens in der zweiten Testphase eine deutliche Verbesserung zu erkennen.

Beim Vergleich der inhaltsbasierten Verfahren SimItems und CBP fällt auf, dass über die Personalisierung in CBP eine höhere CTR erreicht werden kann. Es scheint sich demnach also zu rentieren, das Profil eines Besuchers mit in das Empfehlungsverfahren einfließen zu lassen. Um darüber jedoch eine zuverlässige Aussage machen zu können, sollte an dieser Stelle ein Signifikanztest durchgeführt werden. Dazu werden zunächst die absoluten Häufigkeiten der beiden Verfahren innerhalb des zweiten Testzeitraums in einer Vierfeldertafel dargestellt (s. Tabelle 4.3).

Anhand der Darstellungen aus Kapitel 4.2 ergibt sich damit $\chi^2 = 7,274$ und ein p-Wert von 0,00638, der jedoch aufgrund des gewählten Signifikanzniveaus von $a_{15} = 0,00341$ zu hoch ist. Die Ergebnisse sind demnach also nicht extrem genug, um zuverlässige Aussagen machen zu können.

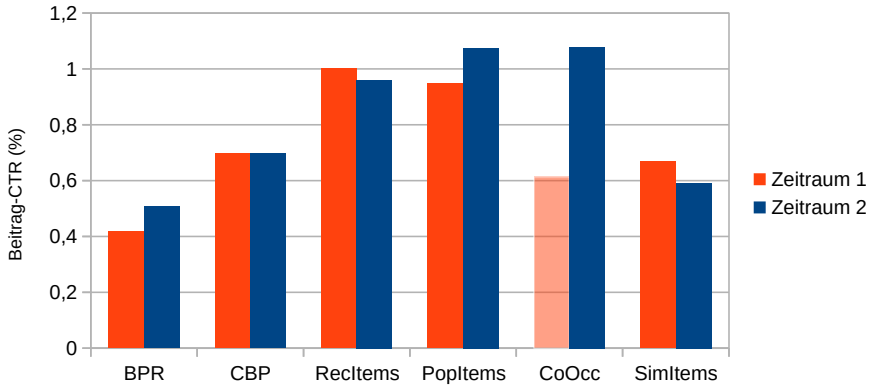

Abbildung 4.1: CTR pro empfohlenem Beitrag bei verschiedenen Verfahren in den Testzeiträumen 1 und 2

Tabelle 4.3: Angeklickte und nicht angeklickte Beiträge von CBP und SimItems innerhalb des zweiten Testzeitraums

	CBP	SimItems	Σ
angeklickt	494	1.285	1.779
nicht angeklickt	100.234	300.810	401.044
Σ	100.728	320.589	402.823

Nach Durchführung der 14 weiteren Tests innerhalb des zweiten Testzeitraums ergeben sich die signifikant unterschiedlichen Gruppen PopItems, CoOcc, RecItems - CBP, SimItems - BPR. Für den ersten Testzeitraum ergeben sich analog die Gruppen RecItems, PopItems - CBP, SimItems, CoOcc - BPR. Gut erkennbar ist die (abgesehen von CoOcc) gleiche Aufteilung der Gruppen. Trotz des im zweiten Zeitraum leichten Vorsprunges von PopItems gegenüber RecItems werden beide dennoch in eine gemeinsame Gruppe eingeteilt, was für eine sinnvolle Bestimmung des Signifikanzniveaus spricht.

Abbildung 4.2 zeigt die *click-through rates* der ausgehenden Shop-Verlinkungen eines Beitrags, nachdem der Benutzer über eine Empfehlung auf ihn gelangt ist. Dort ist gut zu erkennen, dass PopItems im Verhältnis zu RecItems und CoOcc weniger Erfolg zu erzielen scheint. Eine mögliche Ursache dafür ist, dass insbesondere populäre Artikel zwar durch ihre oftmals auffallenden Produktbilder die Aufmerksamkeit erregen, aber schließlich dennoch kein ernsthaftes Kaufinteresse erzeugen können. Es handelt sich dort häufig um Produkte, die besonders ungewöhnlich aussehen, aber dann nicht als relevant erachtet werden oder um Produkte, die zwar interessant erscheinen, jedoch schlussendlich zu teuer für einen spontanen Kauf sind. Weiterhin ist festzustellen, dass RecItems besonders gut abschneidet. Das Verfahren kommt damit offensichtlich erfolgreich seiner Erinnerungsfunktion nach. So waren nach detaillierterer Betrachtung der Benutzerdaten

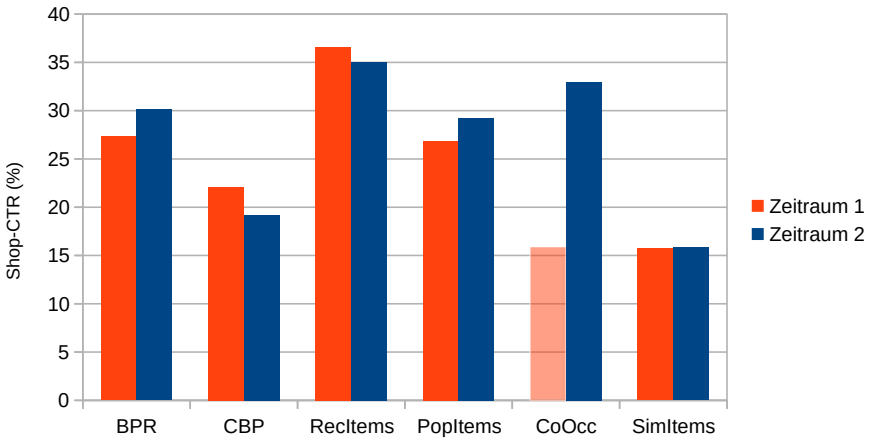

Abbildung 4.2: CTR der Shopverlinkungen bei verschiedenen Verfahren nach dem Klick auf eine Empfehlung in den Testzeiträumen 1 und 2

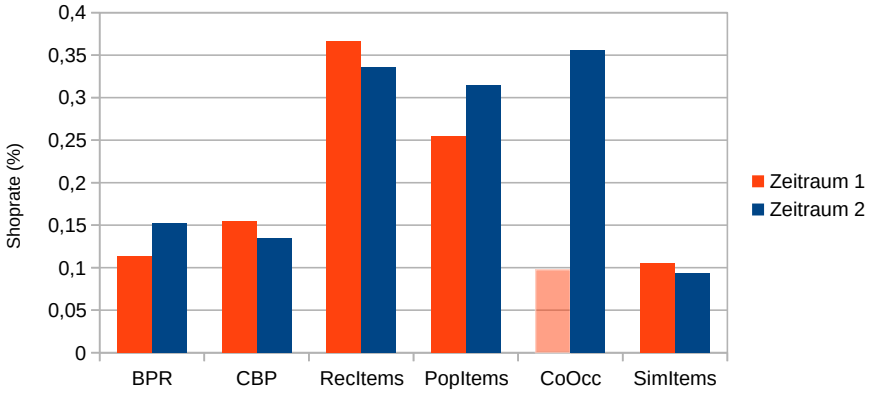

Abbildung 4.3: Shoprate pro empfohlenem Beitrag bei verschiedenen Verfahren in den Testzeiträumen 1 und 2

ebenfalls mehrere Besucher nachverfolgbar, die durch RecItems auf Produkte hingewiesen wurden, die sie teils vor mehreren Wochen zuletzt betrachtet haben und schließlich noch einmal beim jeweiligen Shop aufgerufen wurden. Auffällig ist außerdem, dass BPR besser abschneidet als in Abbildung 4.1. Das legt die Vermutung nahe, dass durch BPR Produkte empfohlen werden, die zwar im ersten Moment weniger Aufmerksamkeit erzielen, jedoch schlussendlich ein größeres Kaufinteresse hervorrufen.

Nach Kombination der Ergebnisse von Abbildungen 4.1 und 4.2 ergibt sich die dargestellte Shoprate in Abbildung 4.3, die angibt, wie viele Shops schlussendlich pro empfohlenem Beitrag besucht wurden. Hier ist deutlich zu erkennen, dass insbesondere im zweiten Testzeitraum BPR aufgrund der höheren Shop-CTR insgesamt sogar minimal besser als CBP und SimItems abgeschnitten hat. Insgesamt führten CoOcc und RecItems jedoch zu den meisten Shopbesuchen. Eine mögliche Ursache für das verhältnismäßig schlechte Ergebnis von SimItems ist, dass zusammen mit dem Ursprungsbeitrag und den 4 darunter

Tabelle 4.4: Anzahl unterschiedlicher Beiträge, auf die sich 50% bzw. 95% der Empfehlungen im zweiten Testzeitraum verteilen

	BPR	CBP	RecItems	PopItems	CoOcc	SimItems
50%	4	8	55	16	6	67
95%	125	116	490	67	100	276

eingeblendeten Empfehlungen insgesamt 5 Beiträge vorliegen, die sehr ähnlich zueinander sind. Der Benutzer wird sich deswegen wahrscheinlich nur für einen der 5 entscheiden, sodass die Chance recht gering ist, dass mehrere der dort beworbenen Produkte gekauft werden. Außerdem erwartet der Benutzer möglicherweise im Fluss der Startseite nicht, dass sich die empfohlenen Produkte auf den darüber liegenden Beitrag beziehen und ist von den scheinbar gleichen Produkten eher irritiert.

Hier ergeben sich nach einem Signifikanztest für den zweiten Zeitraum die signifikant unterschiedlichen Gruppen CoOcc, RecItems, PopItems - BPR,CBP, SimItems, wobei in diesem Fall bereits BPR signifikant besser als SimItems abgeschnitten hat, jedoch zu dicht an CBP liegt.

Ein weiteres relevantes Kriterium ist die Vielfalt der empfohlenen Beiträge. Diese ist in Tabelle 4.4 dargestellt. Dort wird deutlich, dass BPR, CBP, CoOcc und PopItems insgesamt relativ wenige unterschiedliche Produkte empfehlen. Ein Grund dafür ist, dass ein Großteil der Benutzer regelmäßig die Startseite nach neuen Beiträgen absucht und ältere Beiträge eher wenig beachtet. Das führt dazu, dass die Benutzerprofile der meisten Benutzer recht ähnlich zueinander sind. Zusätzlich kommt hinzu, dass es Produkte gibt, die unabhängig von ihrem Erscheinungsdatum überdurchschnittlich beliebt sind. Das lässt sich daran festmachen, dass z. B. in der zweiten Testphase 50% aller Artikelaufrufe auf nur 42 Beiträge gefallen sind. Dadurch tendieren insbesondere die eingesetzten kollaborativen Verfahren dazu, nur ein recht kleines Produktspektrum abzudecken. Dieses Verhalten kann im schlechtesten Fall dazu führen, dass der Benutzer wenig Abwechslung feststellen kann und beginnt, diese zu ignorieren, was zu niedrigeren Klickraten führt. Dieser Effekt müsste im Rahmen einer weiteren Untersuchung ggf. genauer analysiert werden und kann in den nachfolgenden Erläuterungen zu Abbildung 4.4 nur abgeschätzt werden.

RecItems und SimItems streuen zudem relativ stark, was bei RecItems der Erinnerungsfunktion zu Gute kommt, jedoch bei SimItems möglicherweise auch dazu führt, dass zu unspannende oder alte Produkte erscheinen (s. Tabelle 4.5).

Bei Betrachtung des Alters der empfohlenen Beiträge in Tabelle 4.5 fällt auf, dass BPR die aktuellsten Beiträge empfiehlt, währenddessen SimItems die ältesten Beiträge zurückgibt. Der Grund liegt darin, dass SimItems auf die Artikel des vergangenen Jahres zurückgreifen kann, währenddessen z. B. BPR zunächst mittels des kleineren Besucherdatensatzes von ca. 6 Monaten trainiert werden muss. Jedoch sind BPR und CoOcc dabei

Tabelle 4.5: Durchschnittliche Anzahl Tage zwischen Veröffentlichung (bzw. Überarbeitung) und ausgegebener Empfehlung eines Beitrags im Testzeitraum 2

	BPR	CBP	RecItems	PopItems	CoOcc	SimItems
Tage	34,03	112,37	63,43	87,94	40,81	175,04
Standardabweichung	51,37	89,72	171,29	66,74	88,53	122,80

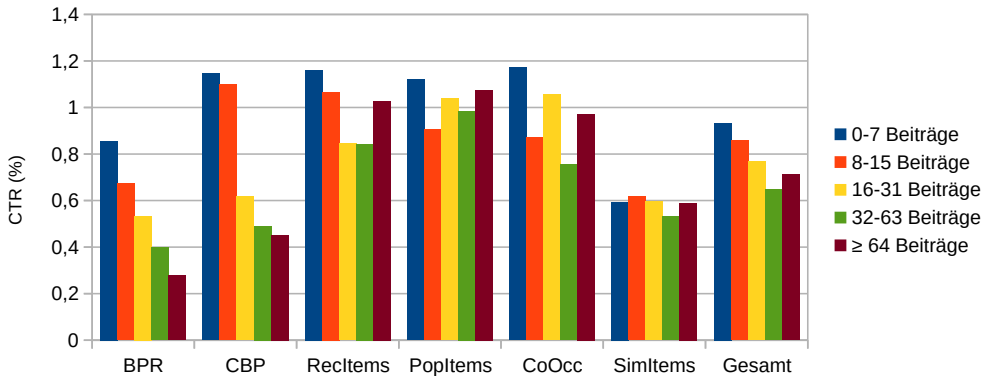

Abbildung 4.4: CTR pro empfohlenem Beitrag bei verschiedenen Verfahren im Testzeitraum 2 nach bereits aufgerufener Anzahl Beiträge eines Benutzers

möglicherweise ebenfalls von den relativ ähnlichen Benutzerprofilen derjenigen beeinflusst, die regelmäßig die Startseite absuchen, was dazu zu führen scheint, dass die dort erscheinenden aktuellen Produkte somit in enger Beziehung zueinander stehen und daher als Empfehlung in Frage kommen. Dennoch wäre es im Rahmen ein weiteren Untersuchung sinnvoll, den Datenzeitraum für die Besucher- und Produktdaten gleich zu wählen.

Dahinter gliedert sich das Verfahren RecItems an, bei dem vor allem anhand der hohen Standardabweichung deutlich wird, dass ebenfalls relativ alte Produkte besucht werden und durchaus Interesse wecken. Da die Redaktion die beliebtesten Artikel regelmäßig aktualisiert und pflegt, sind die durch PopItems empfohlenen Beiträge ebenfalls verhältnismäßig aktuell. Die ältesten Artikel werden durch die beiden inhaltsbasierten Verfahren CBP und SimItems empfohlen, die auf den Produktbestand der letzten 12 Monate zurückgreifen können und aufgrund des geringen zeitlichen Einflusses dieses Ergebnis liefen. Wohingegen sich CBP aufgrund der Personalisierung und der sich über die Zeit vermutlich minimal veränderten Trends noch vor SimItems absetzen kann, liegt SimItems an letzter Stelle.

Grundsätzlich kann das eher schlechte Ergebnis von BPR ebenfalls dadurch zustande kommen, dass zu wenige Benutzerdaten vorlagen, um zuverlässige Empfehlungen zu generieren. Dies lässt sich hier jedoch nicht abschließend beurteilen, da die Klickrate bei zunehmender Anzahl bereits aufgerufener Beiträge für BPR und CBP sogar abnimmt (s. 4.4), was gegen die These sprechen würde, dies jedoch dem Verhalten dieser Algorithmen

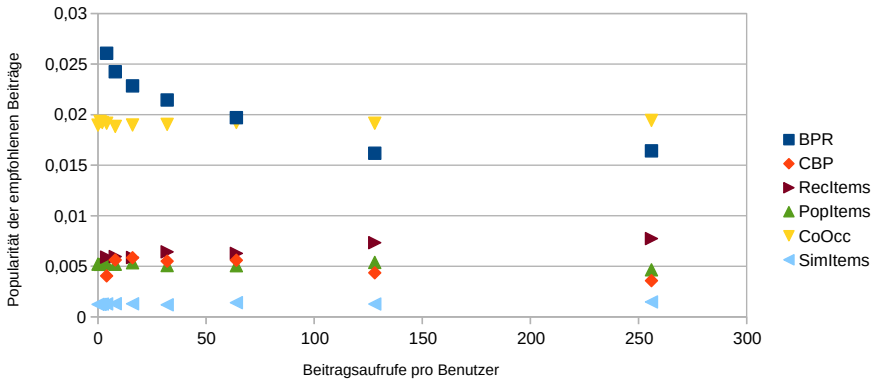

Abbildung 4.5: Popularität der empfohlenen Beiträge abhängig von der Anzahl der bis zu diesem Zeitpunkt aufgerufenen Beiträge eines Benutzers

geschuldet ist, keine bereits besuchten Beiträge zu empfehlen. Somit wurden die populärsten Artikel, die hohe Klickraten versprechen, von den meisten regelmäßigen Besuchern bereits angeklickt und tauchen bei ihnen im Gegensatz zu Besuchern, die Empfehlungen mittels RecItems und PopItems erhalten, nicht mehr auf. Insbesondere BPR tendiert bei steigender Menge an Beitragsaufrufen pro Benutzer deshalb dazu, immer weniger populäre Beiträge zu empfehlen (s. Abbildung 4.5). Der Abfall der Klickrate bei zunehmender Anzahl von besuchten Beiträgen ist daher eher auf eine künstliche Reduzierung der Popularität der Empfehlungen von BPR und CBP zurückzuführen als das Ergebnis der bereits angesprochenen Ermüdung der Benutzer.

Abbildungen 4.5 und 4.4 verdeutlichen jedoch die Wichtigkeit von populären Beiträgen. Diese scheinen vor allem aufgrund des Zusammenhangs beider Abbildungen zu einer Erhöhung der Klickrate beitragen zu können. Jedoch ist zu beachten, dass die letztendliche Kaufbereitschaft dadurch nicht zwangsläufig erhöht wird, da trotz des zunehmenden Wegfalls der bereits bekannten populären Produkte bei BPR in Abbildung 4.2 daraus dennoch eine verhältnismäßig hohe Shop-CTR resultiert. Dies lässt vermuten, dass populäre Artikel den Benutzer zwar auf den Beitrag klicken lassen, das eigentliche Kaufinteresse jedoch nicht zu steigern scheinen.

4.4 Zusammenfassung

Insgesamt wird deutlich, dass die eingesetzten Verfahren recht unterschiedlich gut abgeschnitten haben. Je nach verwendeter Metrik lassen sich zudem ebenfalls leicht unterschiedliche Ergebnisse erzielen. Es lässt sich jedoch festhalten, dass RecItems, PopItems und CoOcc die besten Ergebnisse erzielt haben, währenddessen BPR, CBP und SimItems in der Regel weniger gut abgeschnitten haben. Außerdem ist erkennbar, dass die Popularität der empfohlenen Produkte eine große Rolle im Bezug auf die Klickrate spielt,

Tabelle 4.6: Anzahl sichtbar empfohlener Beiträge in der Anschlussuntersuchung

Verfahren	Beiträge
BPR	22.932
CBP	23.296
RecItems	24.125
PopItems	50.586
CoOcc	55.727
SimItems	67.818

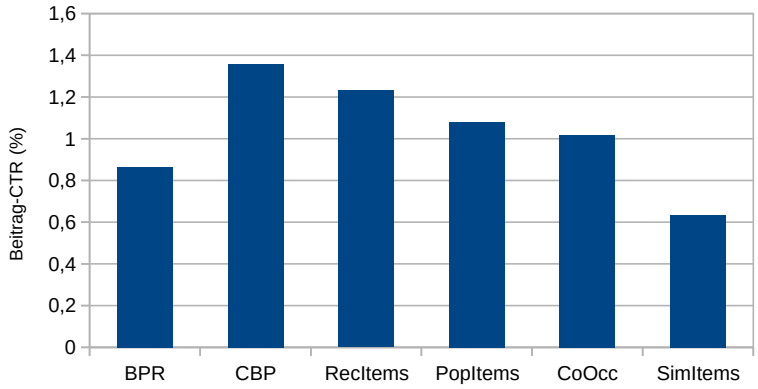

Abbildung 4.6: CTR pro empfohlenem Beitrag bei verschiedenen Verfahren nach Anpassung von BPR und CBP

allerdings nicht zwangsläufig auch zu vielen Shop-Besuchen führt. Hier wäre es zudem hilfreich, präzisere Informationen zu den konkreten Verkaufszahlen zu erhalten.

Ein weiteres Problem, das aufgetaucht ist, ist die bei einigen Verfahren relativ geringe Vielfalt der empfohlenen Beiträge und die allgemein große Popularität einiger weniger Beiträge, dessen Auswirkungen in einer längerfristigen Untersuchung gezeigt werden müssten.

Leider wurde zu Beginn des Experiments nicht bedacht, dass bei BPR und CBP das Ausblenden bereits aufgerufener Artikel ein Nachteil sein kann. So führt dies z. B. dazu, dass ebenfalls die populärsten Artikel nicht mehr angezeigt werden, da diese von den meisten Benutzer bereits betrachtet wurden, jedoch dennoch von Interesse sein können. Deshalb wurde im Anschluss an die beiden ersten Testzeiträume noch eine weitere kleine Untersuchung durchgeführt. Diese erstreckte sich vom 29.08.2015 bis zum 08.09.2015 und führte zu der in Tabelle 4.6 dargestellten Anzahl sichtbar empfohlener Beiträge. Dabei wurde eine leichte Anpassung der beiden Verfahren vorgenommen, sodass diese auch bereits betrachtete Beiträge noch einmal empfehlen können. Dadurch konnte dem ersten Anschein nach zu urteilen zumindest für CBP eine erhebliche Verbesserung erzielt werden, die in Abbildung 4.6 dargestellt ist.

Ein Signifikanztest führt hier zu dem Ergebnis, dass CBP signifikant besser als alle anderen Verfahren mit Ausnahme von RecItems abgeschnitten hat und SimItems das signifikant schlechteste Ergebnis erzielte. Die Anpassung scheint somit ebenfalls BPR verbessert zu haben, wenn auch in etwas geringerem Maße. Außerdem ist ein leichter Anstieg von RecItems erkennbar, der sich jedoch von CoOcc und PopItems nicht signifikant unterscheidet und möglicherweise auf den Zufall zurückzuführen ist.

Dadurch wird zusammenfassend deutlich, dass neben der Popularität auch die Erinnerungsfunktion eine sehr wichtige Rolle spielt. So führen selbst Produkte, die der Besucher bereits angeklickt hat und die dem Besucher aufgrund ihrer hohen Popularität bereits bekannt sein müssten, dennoch zu einer hohen Klickrate. Selbst die daran anschließende Shop-CTR ist bei CBP mit 34,8% an einer Spitzenposition, sodass dadurch ein ernsthaftes Kaufinteresse geweckt werden kann. Zudem spielt CBP hier vermutlich seine Qualitäten in der Personalisierung aus, wodurch selbst ein im ersten Test sehr gut funktionierendes Verfahren wie PopItems nachweisbar überboten werden konnte.

Kapitel 5

Zusammenfassung, Fazit und Ausblick

Ziel dieser Arbeit war es, die Wirkung und den Erfolg einiger Empfehlungsverfahren am Beispiel des Schnäppchenblogs China-Gadgets.de zu untersuchen.

Dazu wurden zunächst die wichtigsten Grundprinzipien der verschiedenen Empfehlungsverfahren vorgestellt. Dabei musste zwischen personalisierten und unpersonalisierten und zwischen kollaborativen und inhaltsbasierten Verfahren unterschieden werden. Außerdem wurde erklärt, dass die Verfahren anhand von Benutzerdaten trainiert werden müssen. Diese Daten können auf der Basis von explizit getätigten Bewertungen oder der aufgerufenen Artikel indirekt erzeugt werden. Als Beispielverfahren wurden hier das Nächste-Nachbarn-Verfahren und ein Algorithmus, der auf *Co-occurrence Patterns* basiert, vorgestellt, sowie erläutert, wie inhaltsbasierte Verfahren auf der Basis von TF-IDF-Vektoren arbeiten.

Anschließend ist auf die Evaluierung von Empfehlungssystemen eingegangen worden, wobei grundsätzlich zwischen online und offline durchgeführten Untersuchungen unterschieden werden musste. Bei einer Offline-Evaluation wird mit historischen Datensätzen gearbeitet und versucht, die durch einige Verfahren empfohlenen Artikel mittels verschiedener Metriken zu bewerten. Dazu wurden einige Metriken wie *Precision*, *Recall* und MAE vorgestellt, aber auch der Nutzen von Alternativen wie Artikelähnlichkeit und Neuheit erläutert. Diese Metriken werden z. B. im Rahmen einer Kreuzvalidierung und eines ggf. daran anschließenden Signifikanztests untersucht. Wird eine Evaluation hingegen online durchgeführt, so wird normalerweise die *click-through rate* oder die Umsatzsteigerung gemessen. Dazu kann ein A/B-Test durchgeführt werden, der verschiedenen Benutzern verschiedene Verfahren präsentiert und diese auswertet. Hier wurden einige bestehende Online-Evaluationen näher betrachtet und dessen Ergebnisse gegeneinander dargelegt. Dabei konnte zusammenfassend festgestellt werden, dass die offline bestimmten Ergebnisse nicht immer die einer Online-Evaluierung widerspiegeln, allerdings bei richtiger Anwen-

dung einen guten Anhaltspunkt bieten können [14, 3]. Außerdem wurden einige Problempunkte bei Verwendung der klassischen Metriken aufgezeigt, die zum Teil durch neuere Metriken reduziert werden können [19, 26]. Des Weiteren wurde deutlich, dass ein psychologischer Effekt wie die Überzeugungskraft einer Empfehlung und das Vertrauen eines Benutzers in das System eine wichtige Rolle spielen kann [19, 7, 2] und sich der Aufwand für personalisierte Empfehlungen im Gegensatz zu unpersonalisierten Empfehlungen durchaus zu lohnen scheint [9, 13].

Für die durchgeführte Fallstudie musste ein Empfehlungssystem aufgesetzt werden, das in der Lage war, für einen Besucher möglichst interessante Beiträge auszugeben. Diese mussten in Echtzeit für jeden Aufruf der Startseite anhand eines von sechs individuell zugewiesenen Verfahren erzeugt werden. Dafür wurde ein eigener Server installiert, der dieser Aufgabe nachzukommen hatte und über eine Schnittstelle mit China-Gadgets.de kommunizierte.

Die anschließende Auswertung brachte zu Tage, dass es vor allem wichtig ist, den Besucher an Produkte zu erinnern, die er vor längerer Zeit bereits einmal besucht hatte. Das wurde am guten Ergebnis von RecItems deutlich und spiegelte sich auch nach der Anpassung von BPR und CBP in einem starken Anstieg der Beitrag-CTR wider. Das geht zudem mit der Tatsache einher, dass ebenfalls PopItems gut funktioniert, da die dadurch empfohlenen Beiträge vielen Besuchern bereits bekannt sein dürften und, falls nicht, möglicherweise wegen ihrer häufig besonders auffälligen Produktbilder hervorstechen. Außerdem wurde deutlich, dass SimItems vermutlich zu ähnliche Produkte empfiehlt, die auf der Startseite im Lesefluss möglicherweise nicht derart erwartet werden und z. B. auf den jeweiligen Beitragsseiten besser aufgehoben wären. Insgesamt lieferte im zweiten Testzeitraum vor allem auch CoOcc sehr gute Ergebnisse, sodass sich die Personalisierung hier auszuzahlen scheint. Eine auffällige Beobachtung, die noch für BPR festgestellt werden konnte, war die verhältnismäßig hohe Shop-CTR. Diese erweckte den Anschein, dass das Verfahren im Vergleich zu den anderen zwar grundsätzlich weniger Aufmerksamkeit erzeugte, die empfohlenen Produkte jedoch schlussendlich besser passten und zum Besuch des Shops führten.

Insgesamt bleiben noch einige Ansatzpunkte offen, die in Zukunft genauer analysiert werden könnten. So konnten aufgrund zu weniger und unzuverlässiger Daten keine genauen Angaben zur Umsatzsteigerung gemacht werden, die aus Unternehmenssicht sicherlich noch interessanter wären. Hier könnte die dafür notwendige Anbindung an die Werbenetzwerke zur Rückmeldung über die einzelnen Käufe verbessert und ausgebaut werden. Außerdem ließe sich offline eine Analyse der eingesetzten Verfahren durchführen und prüfen, inwiefern sich die erhaltenen Ergebnisse auch dort wiederfinden lassen. Nicht zuletzt bleibt auch die Möglichkeit, weitere Verfahren auszuprobieren oder die derzeitigen weiter anzupassen und Detaileinstellungen zu verbessern.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Bland, J. und D. Altman: *Multiple significance tests: the Bonferroni method.* BMJ, 310(6973):170, 1995.

[2] Cremonesi, P., F. Garzotto und R. Turrin: *Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study.* ACM Transactions on Interactive Intelligent Systems (TiiS), 2(2):11, 2012.

[3] Cremonesi, P., F. Garzotto und R. Turrin: *User-centric vs. system-centric evaluation of recommender systems.* In: *Human-Computer Interaction - INTERACT 2013*, Seiten 334–351. 2013.

[4] Demšar, J.: *Statistical comparisons of classifiers over multiple data sets.* The journal of machine learning research, 7:1–30, 2006.

[5] Dias, M., D. Locher und M. Li: *The value of personalised recommender systems to e-business: A case study.* In: *Proc. RecSys '08*, Seiten 291–294, 2008.

[6] Dubben, H.: *Statistische Signifikanz.* `http://secunda.uke.uni-hamburg.de/institute/biometrie/downloads/institut-medizinische-biometrie-epidemiologie/WS05T2_W08.pdf`, 2006. [Abruf am 08.09.2015].

[7] Ekstrand, M., F. Harper, M. Willemsen und J. Konstan: *User perception of differences in recommender algorithms.* In: *Proc. RecSys '14*, Seiten 161–167, 2014.

[8] Garcin, F., C. Dimitrakakis und B. Faltings: *Personalized news recommendation with context trees.* In: *Proc. RecSys '13*, Seiten 105–112, 2013.

[9] Garcin, F., B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin und A. Huber: *Offline and online evaluation of news recommender systems at swissinfo.ch.* In: *Proc. RecSys '14*, Seiten 169–176, 2014.

[10] Goodman, Steven: *A dirty dozen: twelve p-value misconceptions.* In: *Seminars in hematology*, Seiten 135–140, 2008.

[11] Herlocker, J., J. Konstan, L. Terveen und J. Riedl: *Evaluating collaborative filtering recommender systems.* TOIS, 22(1):5–53, 2004.

[12] Hofmann, T.: *Probabilistic latent semantic indexing.* In: *Proc. SIGIR '99*, Seiten 50–57, 1999.

[13] Jannach, D. und K. Hegelich: *A case study on the effectiveness of recommendations in the mobile internet.* In: *Proc. RecSys '09*, Seiten 205–208, 2009.

[14] Jannach, D., L. Lerche, G. Gedikli und G. Bonnin: *What recommenders recommend - An analysis of accuracy, popularity, and sales diversity effects.* In: *Proc. UMAP '13*, Seiten 25–37, 2013.

[15] Jannach, D., M. Zanker, A. Felfernig und G. Friedrich: *Recommender systems - An introduction.* 2010.

[16] Kirshenbaum, E., G. Forman und M. Dugan: *A live comparison of methods for personalized article recommendation at Forbes.com.* In: *Proc. ECML/PKDD '12*, Seiten 51–66, 2012.

[17] Klahold, A.: *Empfehlungssysteme.* 2009.

[18] Koren, Y., R.t Bell und C. Volinsky: *Matrix factorization techniques for recommender systems.* Computer, (8):30–37, 2009.

[19] McNee, S., J. Riedl und J. Konstan: *Being accurate is not enough: How accuracy metrics have hurt recommender systems.* In: *Proc. ACM CHI '06*, Seiten 1097–1101, 2006.

[20] Rendle, S., C. Freudenthaler, Z. Gantner und L. Schmidt-Thieme: *BPR: Bayesian personalized ranking from implicit feedback.* In: *Proc. UAI '09*, Seiten 452–461, 2009.

[21] Said, A., A. Bellogin und A. de Vries: *A top-n recommender system evaluation protocol inspired by deployed systems.* In: *Proc. ACM LSRS '13*, 2013.

[22] Schilit, B., N. Adams und R. Want: *Context-aware computing applications.* In: *Proc. WMCSA '94*, Seiten 85–90, 1994.

[23] Swinscow, T., M. Campbell et al.: *Statistics at square one*, Kapitel 8. 2002.

[24] Voorhees, E. et al.: *The TREC-8 question answering track report.* In: *Proc. TREC '99*, Seiten 77–82, 1999.

[25] Zheng, H., D. Wang, Q. Zhang, H. Li und T. Yang: *Do clicks measure recommendation relevancy? An empirical user study.* In: *Proc. RecSys '10*, Seiten 249–252, 2010.

[26] Ziegler, C.N., S.M. McNee, J.A. Konstan und G. Lausen: *Improving Recommendation Lists through Topic Diversification*. In: *Proc. of WWW 2005*, Seiten 22–32, 2005.