Dietmar Jannach, Lukas Lerche

Perspektiven in der Offline-Evaluation von Empfehlungsalgorithmen

Empfehlungssysteme sind heutzutage ein zentraler Bestandteil vieler Online-Shops und stellen für die Betreiber ein wertvolles Mittel dar, Kunden bei der Produkt- oder Informationssuche zu helfen, sowie auf weitere interessante Produkte hinzuweisen. Die meisten Forschungsarbeiten zu Empfehlungssystemen verwenden explizite Produktbewertungen von Kunden als Eingabe für die Algorithmen und als Grundlage für die Empfehlungsgenerierung. In der Realität sind solche Bewertungen jedoch oft nicht in ausreichender Menge vorhanden, sodass für die Produktvorschläge auf andere Datenquellen – wie zum Beispiel Logdaten der Kundenaktionen – zurückgegriffen werden muss. In diesem Beitrag werden praktische Herausforderungen bei der Nutzung und Interpretation solcher weiteren Datenquellen für die Empfehlungsgenerierung besprochen sowie auf methodische Fragen der vergleichenden Bewertung von Empfehlungsalgorithmen eingegangen.

Inhaltsübersicht

1 Empfehlungssysteme

Empfehlungssysteme (engl.: Recommender Systems) sind intelligente Softwareanwendungen, die einen Online-Benutzer in personalisierter Form auf relevante Informationen und Produkte hinweisen. Solche Systeme finden sich heute in fast jedem Online-Shop, werden aber auch in anderen Domänen wie sozialen Netzen oder Nachrichtenportalen eingesetzt. Empfehlungssysteme helfen Kunden, sich in der Angebotsvielfalt zurechtzufinden bzw. auf Produkte hinzuweisen, die bislang nicht bekannt waren, sogenanntes Cross-Selling.

In den letzten beiden Jahrzehnten hat sich das Forschungsgebiet der Empfehlungssysteme als eigenständige wissenschaftliche Disziplin entwickelt. Getrieben unter anderem durch die zunehmende Nachfrage nach solchen Systemen in der Praxis, wurden in dieser Zeit eine Vielzahl von Empfehlungsalgorithmen entwickelt, welche zumeist auf Verfahren des Maschinellen Lernens aufbauen (vgl. [Jannach et al. 2010]).

Der Vergleich der tatsächlichen Wirksamkeit oder der Treffsicherheit unterschiedlicher Empfehlungsalgorithmen gestaltet sich jedoch aus verschiedenen Gründen gerade im Forschungsumfeld als schwierig. Während bei der Bewertung von realen web-basierten Lösungen und Lösungsalternativen auf A/B-Tests mit tatsächlichen Nutzern zurückgegriffen wird, bedient sich die Forschung oftmals experimentellen Designs, welche auf historischen Daten operieren und

Messgrößen aus dem Maschinellen Lernen oder dem Information Retrieval verwenden. Zur Beantwortung der Frage, welches Empfehlungsverfahren voraussichtlich am besten im Realeinsatz funktionieren wird, lässt sich ein Vergleich der Vorhersagegenauigkeit der einzelnen Verfahren in einem Vorabexperiment durchführen.

Oft nutzen heutige Forschungsarbeiten als Eingabe der Algorithmen ausschließlich vom Endbenutzer abgegebene, explizite Produktbewertungen, wobei häufig eine Bewertungsskala von eins bis fünf Sternen zugrunde gelegt wird. In realen E-Commerce-Anwendungen liegen solche Kundenbewertungen jedoch meist gar nicht oder nicht in ausreichender Menge vor. Dies ist vor allem bei kleineren Webshops mit wenigen Kunden der Fall, aber auch bei Produkten, welche Kunden nur wenige Male im Leben kaufen. In diesen Fällen kann versucht werden, implizites Feedback des Kunden in den Empfehlungsalgorithmus miteinzubeziehen. Als implizites Feedback gilt beispielsweise das Betrachten eines Produkts, das Hinzufügen des Produkts zum Warenkorb oder der tatsächliche Kauf. Sofern nicht schon in der E-Commerce-Plattform erfasst, kann dieser Kontext unter anderem aus dem üblichen Webserver-Log rekonstruiert werden. Wie die zusätzlichen Informationen in den Empfehlungsprozess optimal eingebaut werden können, ist leider bislang weitgehend offen und ist derzeit nur Gegenstand einiger weniger Forschungsarbeiten auf dem Gebiet.

Angesichts der geschilderten Einschränkungen erscheinen die bestehenden Methoden aus der Forschung zur Offline-Bewertung verschiedener Empfehlungsalgorithmen an vielen Stellen insgesamt nur bedingt geeignet, um mit den komplexen Gegebenheiten realer Empfehlungssysteme umzugehen. In diesem Beitrag gehen wir auf bestehende Einschränkungen aktueller Evaluierungsmethoden aus der Forschung ein und umreißen offene Herausforderungen bei der Entwicklung von umfassenderen und realitätsnäheren Verfahren.

2 Offline-Evaluation in der Forschung heute

In der Literatur zu Empfehlungssystemen der letzten Jahre werden verschiedene Empfehlungsalgorithmen vorwiegend in Offline-Analysen auf Basis von historischen Datensätzen verglichen. Dabei dominieren Metriken, die die Vorhersagegenauigkeit einzelner Verfahren messen oder Kennzahlen, die die Anzahl der Treffer in Top-N-Empfehlungslisten ermitteln [Jannach et al. 2012]. Zur Bestimmung der Kennzahlen wird ein Teil der vorhandenen Bewertungsdaten zum Lernen eines Modells verwendet, mithilfe dessen die verbleibenden und vom Empfehlungsverfahren versteckten Daten vorhergesagt werden. Diese Vorhersagen bzw. die Reihung der Produkte in den durch die Algorithmen erstellten Empfehlungslisten werden anschließend mit den tatsächlichen Bewertungen verglichen. Am genauesten sind folglich diejenigen Verfahren, deren Bewertungsvorhersagen im Mittel am nächsten an den verdeckten Werten sind, bzw. solche, die die meisten der tatsächlich vom Kunden gemochten Produkte (am besten in der richtigen Reihenfolge) in ihre Empfehlungslisten aufnehmen[1].

Der Vergleich von Empfehlungsverfahren nach diesen Kennzahlen liefert zweifelsohne wertvolle Einsichten über die Fähigkeit verschiedener Algorithmen, Produkte zu filtern oder den Präferenzen des Benutzers nach entsprechend zu sortieren. Die Verwendung dieser Metriken alleine zur Abschätzung des wahren Wertes eines Empfehlungssystems wird in der Forschungsgemeinschaft aber zunehmend hinterfragt. Neben methodischen Fragen, wie sie zum Beispiel bereits in [Herlocker et al. 2004] hinsichtlich der direkten Übertragbarkeit von Kennzahlen des Information Retrieval für Empfehlungssysteme gestellt wurden, haben die üblicherweise

[1] Ein detaillierte Darstellung der verwendeten Messprotokolle findet sich in [Herlocker et al. 2004].

verwendeten Genauigkeitsmetriken noch weitere Einschränkungen bezüglich ihrer tatsächlichen Aussagekraft über die praktische Wirksamkeit unterschiedlicher Verfahren.

2.1 Popularitätseffekte

Hinsichtlich der Trefferraten in den Empfehlungslisten – üblicherweise gemessen in Precision[2] und Recall[3] – besteht zum Beispiel das Problem, dass Verfahren, welche sich auf populäre Produkte konzentrieren und diese bevorzugt für jeden empfehlen, bereits sehr gute Ergebnisse erzielen. Diese können oft von komplizierteren personalisierten Verfahren nur in geringem Ausmaß übertroffen werden [Cremonesi et al. 2010]. Vergleichende Studien in realen Anwendungssituationen zeigen jedoch, dass die Empfehlung ausschließlich populärer Produkte meist nicht die verkaufsfördernden Effekte personalisierter Empfehlungsverfahren erreicht. Letztlich kann die Fokussierung auf populäre Produkte auch zu ungewünschten Verstärkungseffekten führen, sodass eine Konzentration auf eine kleine Menge von Produkten erfolgt und Nischenprodukte immer weniger angeboten und verkauft werden [Fleder und Hosanagar 2009]. Sich rein auf Treffermetriken in der Offline-Analyse zu verlassen, kann demzufolge zu irreführenden Schlussfolgerungen führen.

2.2 Statistische und reale Signifikanz

Die in der aktuellen Literatur berichteten Unterschiede hinsichtlich der Empfehlungsgenauigkeit einzelner Verfahren bewegen sich manchmal bis in den dritten Nachkommastellenbereich. Dies und die oft nicht vorhandene Analyse einer statistischen Signifikanz der beobachteten Differenz sind weitere Faktoren, die die Auswahl eines geeigneten Empfehlungsverfahrens für einen bestimmten Anwendungsbereich erschweren. Zudem ist in der Literatur manchmal zu beobachten, dass die Reihung der zu vergleichenden Algorithmen nicht nur von der Anwendungsdomäne und den Charakteristika des Bewertungsdatensatzes, sondern auch von der Wahl der Messmethode sowie spezifisch optimierten Algorithmusparametern abhängt. Damit ist es schwer, eine eindeutige Aussage treffen zu können. Letztlich bleibt unklar, ob eine geringe Verbesserung hinsichtlich einer Genauigkeitsmetrik tatsächlich einen messbaren Einfluss auf die Kunden hat.

2.3 Verfügbarkeit von Kundenfeedback und Domänenspezifika

Die Forschung der letzten Jahre konzentrierte sich letztlich auch aufgrund der öffentlichen Verfügbarkeit von umfangreichen Bewertungsdatensätzen auf Kinofilme. Der Vorteil dieser Anwendungsdomäne für Empfehlungssysteme liegt darin, dass es viele Benutzer gibt, die eine größere Menge von Filmen bewertet haben. Insofern können Verfahren des Maschinellen Lernens auf vergleichsweise große Datenbestände aufsetzen und folglich die Empfehlungen optimiert werden. In der Realität, zum Beispiel bei kleineren Web-Shops, liegen solche detaillierten Präferenzprofile meist nicht vor. Zusätzlich gibt es Anwendungsdomänen, in denen Kunden insgesamt nur wenige Transaktionen über größere Zeiträume hinweg durchführen oder das passende Produkt vielmehr von externen Faktoren, Rahmenbedingungen und aktuellen Bedürfnissen abhängt, wie es vielleicht bei Filmen der Fall ist. Die Auswahl eines geeigneten Empfehlungsverfahrens hängt oft von diesen Faktoren ab; eine Übertragung der Erkenntnisse hinsichtlich der Genauigkeit der Verfahren von einer Domäne auf eine andere birgt dementsprechend Gefahren. Anhand der Verfügbarkeit von Präferenzdaten lassen sich in letzter Zeit vermehrt Ansätze beobachten, die auch implizites Feedback des Kunden – etwa

[2] Anteil der relevanten Produkte in einer Empfehlung.

[3] Anteil der empfohlen relevanten Produkte bezogen auf alle relevanten Produkte.

Klickverhalten oder Kaufverhalten – besser berücksichtigen. Wie in diesen Fällen mit der enormen Fülle solcher Daten umzugehen ist, welche Informationen genutzt werden, wie die Daten interpretiert werden und ob aktuelle Algorithmen überhaupt ausreichend skalierbar sind, bleibt jedoch noch oft ungeklärt.

2.4 Vernachlässigung des Benutzerkontexts

Die Frage der Einbeziehung des Benutzerkontexts in den Empfehlungsprozess ist erst in den letzten Jahren weiter in den Mittelpunkt des Forschungsinteresses gerückt. Als Beispiel für den Einfluss des Kontexts auf die Adäquatheit einer Empfehlung lässt sich die Filmdomäne heranziehen. Die Auswahl eines Filmes kann davon abhängen, ob der Nutzer alleine, zweit oder in einer Gruppe ins Kino geht. Übertragen auf die Domäne von Web-Shops kann die Empfehlung davon abhängen, ob etwas für sich selbst oder für jemand anderen gesucht wird. Datensätze sowie Evaluierungsmethoden, die solche verfügbaren Informationen über den Benutzerkontext in wirklich angemessener Form berücksichtigen, sind in der heutigen Forschung noch nicht ausreichend vorhanden. Des Weiteren ist der Vergleich von Ergebnissen auf Basis von kontextualisierten mit unkontextualisierten Empfehlungen schwierig.

2.5 Weitere Qualitätsfaktoren für Empfehlungslisten

Bereits seit einigen Jahren wird auch im Forschungsumfeld darauf hingewiesen, dass ein Vergleich von Algorithmen, der allein auf Genauigkeitsanalysen basiert, zu kurz greifen kann. So können die Empfehlungen zwar treffgenau sein, die Elemente der Empfehlungsliste sind aber vielleicht zu eintönig oder zu offensichtlich. Ein typisches Beispiel für diesen Fall ist eine Empfehlungsliste, die nur aus Büchern eines einzigen Autors besteht, den der Kunde üblicherweise mag. In den letzten Jahren wurden daher verschiedene Metriken vorgeschlagen, mit denen die Diversität oder potentielle Überraschungseffekte bestimmt werden können. Ein einheitlicher Standard hat sich bislang jedoch noch nicht herausgebildet. Auch ist die Anzahl der Forschungsarbeiten, die sich mit der Balancierung der verschiedenen Faktoren – zum Beispiel Vorhersagegenauigkeit gegenüber Diversität – befasst, noch überschaubar.

2.6 Empfehlungsalgorithmen vs. Empfehlungssysteme

Insgesamt liegt der Fokus vieler aktueller Forschungsarbeiten mehr auf der Analyse einzelner *Algorithmen* und weniger auf dem Empfehlungs*system* als Ganzes. Die beabsichtigte und tatsächliche Wirkung des Informationssystems auf die Kunden und deren Konsumverhalten wird oft außer Acht gelassen. Einzelne Trends zu einer stärkeren Fokussierung auf benutzerzentrierte Evaluationsmethoden für Empfehlungssysteme sind jedoch bemerkbar, leiden aber naturgemäß an den Einschränkungen des Laborcharakters der zugrundeliegenden Studien.

3 Neue Ansätze in der Offline-Evaluation

In den letzten Jahren hat die Forschung ein wachsendes Bewusstsein für die beschränkte Aussagekraft von vergleichenden Offline-Analysen, welche nur auf Genauigkeitsmessungen und Bewertungsvorhersagen basieren, entwickelt. Dementsprechend können vermehrt experimentelle Analysen, welche mehrere bekannte Metriken gegeneinander abwägen, innovative Messgrößen vorschlagen oder weitere Datenquellen neben den Bewertungsmatrizen verwenden, in der Literatur beobachtet werden.

3.1 Nutzerzentrierte Evaluation und Offline-Analysen abgleichen

Naturgemäß werden vergleichende Analysen mit realen Systemen und Benutzern, sowie die Durchführung von dazugehörigen A/B-Tests, im Forschungsumfeld immer schwierig sein und eine Ausnahme bilden. Experimentelle Studien, die in Laborsituationen einzelne Aspekte eines Empfehlungssystems untersuchen, können an dieser Stelle sicherlich dazu beitragen, den Faktor Mensch stärker in den Evaluationsprozess miteinzubeziehen. Dementsprechend wurden in letzter Zeit verschiedene Vorschläge zur systematischeren und strukturierteren nutzer-zentrierten Evaluation von Empfehlungssystemen gemacht und die Anzahl der Benutzerstudien im Informatik-Umfeld scheint zuzunehmen, siehe zum Beispiel [Pu et al. 2011]. Solche Methoden und Studien sind im Bereich der Wirtschaftsinformatik (bzw. im Feld "Information Systems") klar etabliert. Nichtsdestotrotz muss die in manchen Fällen eingeschränkte und nicht verallgemeinerbare Aussagekraft solcher Experimente bewusst sein. Untersucht beispielsweise ein Experiment die Wirksamkeit von Empfehlungssystemen oder auch nur die wahrgenommene Qualität der Empfehlungslisten, ist die Laborsituation in den meisten Fällen nicht "real" in dem Sinne, dass die Experimentteilnehmer sich wirklich für ein Produkt entscheiden und es kaufen. Auch sind die Teilnehmer vielfach Studenten – oft aus einer einzigen Disziplin – und dementsprechend nicht repräsentativ für einen größeren Kundenkreis. Schließlich besteht wie in vielen anderen Experimentsituationen das Problem darin, eine unter Umständen nicht unwesentliche Menge potentieller Einflussfaktoren im Experiment konstant zu halten, um etwaige Effekte der Variation von einzelnen Variablen messen zu können. Letztlich ist die Verallgemeinerbarkeit von Beobachtungen – gegebenenfalls auch deren tatsächliche Relevanz und Gültigkeit – aus einem einzelnen durchgeführten Experiment schwer zu beurteilen. Die Wiederholung einer einmal durchgeführten Studie durch andere Forschungsgruppen zur Validierung ist wie in manch anderen Wissenschaftsdisziplinen leider kaum zu beobachten.

Trotz verschiedener Einschränkungen sollte jedoch die nutzerzentrierte Evaluation von Empfehlungssystemen weiterhin ein wichtiger Baustein für eine umfassende Bewertung von Empfehlungsalgorithmen sein. Zudem ist es wünschenswert, wenn in Zukunft vermehrt Evaluationsansätze zu beobachten wären, in denen Messungen aus Offline-Experimenten mit Beobachtungen aus Studien mit echten Benutzern abgeglichen und gegenübergestellt werden.

3.2 Zielorientierung in der Evaluation

Wie erwähnt beschränkt sich das Optimierungsziel vieler Forschungsarbeiten, die auf Offline-Experimenten beruhen, auf die Verwendung der einen oder anderen Genauigkeitsmessgröße, wobei in den einzelnen Arbeiten manchmal nicht klar begründet wird, warum eine bestimmte Größe gewählt wurde. Die Einsatzziele eines Empfehlungssystems können aus Betreibersicht aber durchaus unterschiedlich sein. Schon in frühen Arbeiten wurde zwischen den möglichen Zielsituationen "finde ein passendes Produkt" oder "finde alle relevanten Produkte" unterschieden [Herlocker et al. 2004], wobei in einem Fall ein hoher "Precision"-Wert wichtig ist, im anderen Fall die Metrik "Recall" die größere Bedeutung hat. Aus kaufmännischer Sicht könnte es aber auch Ziel sein, unbekannte Produkte aus dem sogenannten "Long Tail" des Katalogs für den Konsumenten durch Platzierung in Empfehlungslisten sichtbar zu machen. Andere Systeme sollen womöglich den Kunden auf eine ganz andere Kategorie von Produkten im Sinne von Cross-Selling-Bestrebungen hinweisen. Letztlich können in manchen Anwendungsszenarien falsche oder wenig passende Empfehlungen auch ein Risiko in dem Sinne darstellen, dass der Kunde die ihm präsentierten Empfehlungen für wenig wertvoll erachtet, den Nutzen des Ganzen insgesamt anzweifelt und daher das System meidet.

Die Tatsache, dass verschiedene Empfehlungsalgorithmen recht unterschiedliche Empfehlungslisten für einzelne Benutzer generieren können, ist seit langer Zeit bekannt. Inhaltsbasierte Empfehlungssysteme haben zum Beispiel wesensbedingt die Tendenz, den Kunden mehr vom Gleichen vorzuschlagen, da sie gerade diejenigen Produkte auswählen, die eine hohe Übereinstimmung zu bisher präferierten Dingen haben. Ähnliche Unterschiede zwischen Verfahren lassen sich auch hinsichtlich anderer Kriterien feststellen. So beschränken sich manche Algorithmen in ihren Empfehlungen auf einen kleinen Teil des Produktspektrums, etwa populäre Produkte. Andere Verfahren hingegen weisen keine solche Tendenz auf, unter anderem inhaltsbasierte Verfahren.

In der Praxis sollte daher bei der Evaluation von Empfehlungsalgorithmen jeweils das Ziel des Empfehlungssystems mitberücksichtigt und dementsprechend Metriken ausgewählt bzw. optimiert werden, welche für die Erreichung des Ziels entscheidend sind.

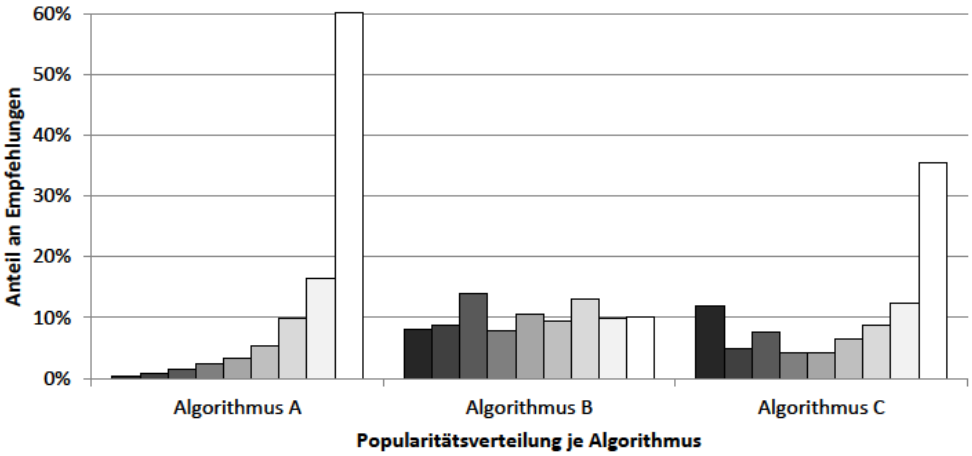

Abb. 1: Verteilung von Empfehlungen verschiedener Algorithmen

Abbildung 1 zeigt beispielhaft das Verhalten verschiedener Algorithmen hinsichtlich ihrer Tendenz, populäre Produkte zu empfehlen (vgl. [Jannach et al. 2013]). Für die Messung wurden für jeden Kunden die ersten zehn empfohlenen Produkte gesammelt, nach ihrer allgemeinen Beliebtheit aufsteigend sortiert und in neun gleich große Gruppen zusammengefasst, welche durch die Säulen in der Abbildung visualisiert werden. Algorithmus A tendiert offensichtlich dazu, vorwiegend bereits populäre Produkte zu empfehlen, wohingegen Algorithmus B Vorschläge aus dem ganzen Spektrum populärer und weniger populärer macht. Algorithmus C hingegen weist keine klare Tendenz auf, empfiehlt jedoch auch häufig populäre Produkte.

3.3 Verwendung mehrerer Metriken

Im Zusammenhang mit der Zielorientierung bei der Bewertung von Empfehlungsverfahren kann es zu einem Zielkonflikt kommen. Hohe Treffsicherheit und das damit verbundene geringe Risiko schlechter Empfehlungen stehen beispielsweise oft im Konflikt mit dem Wunsch eines erhöhten Überraschungseffekts oder der Diversität der Empfehlungsliste insgesamt. Gleichzeitig könnte es aus kaufmännischer Sicht auch sinnvoll sein, zusätzliche Produkte in den Listen zu platzieren, die gerade im besonderen Maße beworben werden sollen. Auch solche Maßnahmen können das Risiko einer letztlich unpassenden Empfehlung mit sich bringen. Werden in einem anderen Szenario im Sinne einer risikoarmen Strategie beispielsweise nur Produkte empfohlen, die im

Schnitt hoch bewertet werden, hat dies zudem eine Auswirkung auf die sogenannte Katalogabdeckung, d.h. die Menge der Produkte, die überhaupt empfohlen werden.

Entgegen der in der Forschung oft üblichen Praxis, nur eine einzige Kennzahl oder eine Menge verwandter Kennzahlen für den Vergleich von Algorithmen heranzuziehen, empfiehlt es sich daher in der Praxis ganz besonders, Algorithmen immer hinsichtlich unterschiedlicher Eigenschaften gleichzeitig zu untersuchen und die Charakteristika mit den eigentlichen Zielen des Empfehlungssystems abzugleichen. Bei vorhandenen Zielkonflikten kann versucht werden, die Parameter der verwendeten Algorithmen zu variieren oder hybride Methoden einzusetzen, welche einen Kompromiss bei den verschiedenen Kriterien gewährleisten. Ein Beispiel für eine solche Methode, die Genauigkeit und Katalogabdeckung gegeneinander abwägt, wird in [Adomavicius und Kwon 2012] beschrieben.

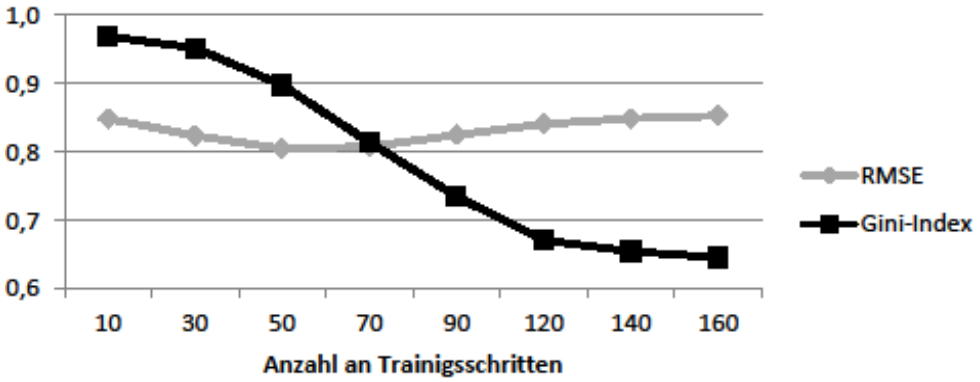

Abb. 2: Entwicklung verschiedener Metriken bei einem Matrixfaktorisierungsverfahren

Abbildung 2 zeigt beispielhaft einen Zielkonflikt für ein heute in der Forschung übliches Matrixfaktorisierungsverfahren, welches vor seiner Anwendung in mehreren Iterationen trainiert wird. Je nach Anzahl der Trainingsschritte, welche auch die Rechenzeit für die Modellbildung stark beeinflussen, variieren die Ergebnisse für zwei auf einem üblichen Filmdatensatz gemessene Metriken. Die Genauigkeit (RMSE[4]) ist hier bei ca. 60 Schritten optimal, wohingegen sich die Diversität der Empfehlungen (gemessen am Gini-Index[5]) erst nach ca. 120 Schritten dem Optimum annähert. Optimiert man nur nach der Vorhersagegenauigkeit, würde man – um Rechenzeit zu sparen – unter Umständen bereits nach 60 Iterationen abbrechen. Betrachtet man aber gleichzeitig, wie gut das Verfahren ein breiteres Produktspektrum abzudecken in der Lage ist, würde man weitere Trainingsrunden machen um den Wert des Gini-Koeffizienten weiter zu senken (vgl. wiederum [Jannach et al. 2013]).

3.4 Kontextualisierung und implizite Ratings

Erst in den letzten Jahren hat das Thema "Kontextualisierung von Empfehlungen" im Forschungsumfeld an Bedeutung gewonnen und erste algorithmische Ansätze hervorgebracht. Viele dieser Ansätze gehen weiterhin vom Standardevaluationsmodell der Forschung und somit Vorhandensein von expliziten Bewertungen aus, wobei es nun erforderlich sein kann, dass ein Benutzer ein Produkt in verschiedenen Situationen bewertet. Angesichts der ohnehin

[4] Root Mean Square Error: Eine Metrik, mit der die mittlere Abweichung der durch eine Vorhersagefunktion geschätzten Werte von den realen Werten berechnet werden kann.

[5] Ein statistisches Maß zur Darstellung von Ungleichverteilungen. Eine Gleichverteilung liegt bei einem Gini-Index von 0 vor, eine maximale Ungleichverteilung bei einem Wert von 1.

üblicherweise schon gegebenen Spärlichkeit der Daten, gestaltet sich die Analyse schwierig. Gleichzeitig ist bei vielen Anwendungsszenarien nicht immer klar, wie der aktuellen Kontext des Benutzers zum Empfehlungszeitpunkt zumindest teilweise automatisiert bestimmt oder abgeschätzt werden kann.

Derzeit in der Forschung zu Empfehlungssystemen kaum beachtet, aber in praktischen Fällen von eminenter Bedeutung, ist der Navigations- und Interessenskontext eines Benutzers während eines Webseitenbesuchs. Kehrt beispielsweise ein Stammkunde zu seinem favorisierten Online-Shop zurück, kann das Empfehlungssystem eine Vorschlagsliste auf Basis aller bisherigen Einkäufe erzeugen. Eine solche Empfehlungsliste kann durchaus auch berücksichtigen, für welche spezifische Produktkategorie der Kunde sich beim letzten Besuch der Seite interessiert hat und einen stärkeren Fokus auf diese Produkte setzen, sofern es vorher zu keiner erfolgreichen Transaktion gekommen ist. Ein weiterer Fall ist, dass der Besucher gleich beim Einstieg zur Detailseite eines Produkts einer bestimmten Kategorie navigiert. In solchen Situationen kann ein Empfehlungssystem erkennen, dass sich die aktuelle Interessenslage auf diese Kategorie bezieht und demzufolge bei der Empfehlungsgenerierung vermehrt ähnliche Produkte aus dieser Kategorie anzeigen.

Auf realen E-Commerce Plattformen sind die genannten Konzepte "Fortsetzen des letzten Besuchs" bzw. "Anzeigen von ähnlichen Produkten bei Besuch einer Detailseite" vielfach zu finden, wobei manchmal recht einfache Mechanismen dahinter vermutet werden können. Die Adaption von Webseitenelementen – zum Beispiel hinsichtlich der Menge der angezeigten Links – vom bisherigen Navigationsverlauf abhängen zu lassen, wurde bereits vor einiger Zeit vorgeschlagen, vgl. [Mobasher et al. 2001]. Gleichzeitig zeigen zumindest einzelne Studien aus der Forschung, dass die Wahl des geeigneten Empfehlungsverfahrens vom Navigationskontext abhängt. In der Realwelt-Studie von [Jannach und Hegelich, 2009] für eine Downloadplattform von Handyspielen wurde beispielsweise festgestellt, dass inhaltsbasierte Verfahren, die bei Offline-Experimenten hinsichtlich ihrer Genauigkeit eher schlecht abschneiden, hier zu den besten Verkaufseffekten geführt haben. Gleichzeitig zeigte sich, dass in bestimmten Situationen – zum Beispiel direkt nach einer Kauftransaktion – andere Algorithmen, deren Empfehlungen auf andere Teile des Produktkatalogs fokussieren, effektiver sein können.

Um den aktuellen Navigations- und Interessenskontext automatisch bestimmen zu können, ist es notwendig, das Navigationsverhalten des Besuchers – und auch das damit verbundene implizite Feedback zu Produkten und Produktkategorien – korrekt zu interpretieren. Auf Basis dieser Informationen können entsprechende Empfehlungsstrategien entwickelt werden, die beispielsweise situationsabhängig zwischen verschiedenen Verfahren umschalten. Für die tatsächliche Umsetzung solcher Strategien bedarf es aus heutiger Sicht noch einiger Forschung bzw. der Berücksichtigung verschiedener Herausforderungen.

3.5 Konkrete Herausforderungen bei der Verwendung von Log-Daten

In realen Online-Shop-Lösungen liegt üblicherweise eine Unmenge an Daten vor, die prinzipiell für die Personalisierung und Adaptierung genutzt werden können. Diese umfassen Logs über Seitenbesuche, Warenkorbaktionen, demographische Informationen über die Benutzer, Merkmale der gesehenen und gekauften Produkte inklusive deren Kategorisierung, Statistiken über die allgemeine Beliebtheit von Produkten sowie manchmal sogar einzelne Klicks oder Suchanfragen. Unter anderem können daher folgende praktische Herausforderungen identifiziert werden, die heute noch unzureichend in der Forschung betrachtet werden.

1. Welche dieser Daten soll überhaupt genutzt bzw. wie soll die Fülle an Informationen in eine Empfehlungsstrategie oder in einzelne Empfehlungsalgorithmen eingebunden werden?
2. Wie wird sichergestellt, dass die verwendeten Daten auch effizient verarbeitet werden? Moderne Algorithmen können mittlerweile viele Millionen explizite Kundenbewertungen verarbeiten. Bei Nutzung von weiteren Datenquellen und besonders bei Berücksichtigung von Klickdaten aus Weblogs sind die zu verarbeitenden Datenvolumina jedoch um ein Vielfaches größer, sodass möglicherweise neue Algorithmen und Ansätze notwendig sind, um mit solch großen Datenmengen umgehen zu können. In diesem Umfeld sind zum Beispiel der Einsatz moderner (NoSQL-)Datenbankkonzepte sowie die Verwendung verteilt arbeitender Software in Computerclustern nützlich und werden in der Praxis bereits erfolgreich eingesetzt.

Für die Forschung im Bereich von Empfehlungssystemen hat dies vor allem in Bezug auf den ersten der beiden Punkte weitergehende Auswirkungen und es müssen noch entsprechende Voraussetzungen für eine systematische Erforschung geschaffen werden.

1. **Datensätze:** Die heute in der Forschung zu Empfehlungssystemen verwendeten Datensätze beschränken sich wie beschrieben auf explizite Ratings und beinhalten zumeist keine detaillierten Informationen über die Produkte selbst, was die Vergleichbarkeit von Verfahren, welche weitere Informationen nutzen, sehr schwierig macht. Ziel der Forschung kann es sein, eine kleine Menge von Referenzdatensätzen aus verschiedenen Domänen zu erstellen. Jeder Referenzdatensatz sollte einen Satz von Zusatzinformationen beinhalten, welcher typischerweise bei realen Webshops zu finden ist, beispielsweise Informationen über einzelne Sitzungen der Nutzer oder Kategorisierungsinformation zu den Produkten.
2. **Evaluationsprotokolle:** Gleichzeitig sind neue Protokolle für die Evaluation von Empfehlungsverfahren notwendig, welche zum Beispiel den aktuellen Sitzungskontext berücksichtigen können. Aktuelle, unkontextualisierte Evaluationsverfahren, die unabhängig von der Kundensituation Empfehlungen aussprechen, können ein falsches Bild bei der Bewertung der Nützlichkeit von Empfehlungsalgorithmen für den realen Einsatz liefern. Der verstärkte Einsatz von simulationsbasierten Experimenten kann hier eine interessante Möglichkeit darstellen, die derzeit kaum im Umfeld der Empfehlungssystemforschung beachtet wird.

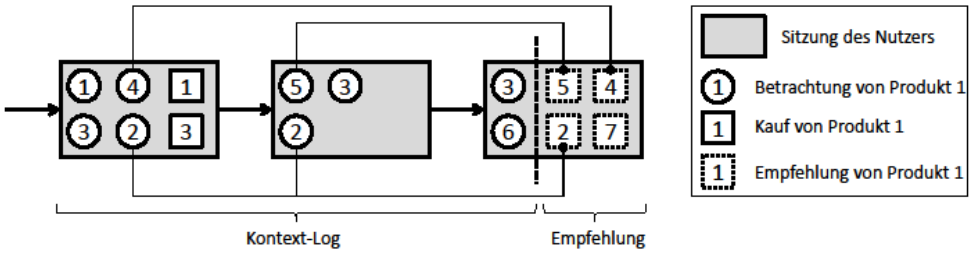

Abb. 3: Kontextbasierter Empfehlungsprozess

Abbildung 3 skizziert ein Beispiel für eine Empfehlungsstrategie, die Produkte anhand des Sitzungskontexts des Nutzers vorschlägt. Ausgehend vom Nutzerverhalten in zwei vorherigen und der aktuellen Sitzung werden Empfehlungen ermittelt. Produkte, die in der Vergangenheit

betrachtet, aber noch nicht gekauft wurden, werden empfohlen. Die Reihenfolge der Vorschläge kann beispielsweise anhand der Anzahl und des Zeitpunktes der Betrachtungen bestimmt werden. In diesem Beispiel wurden Produkte 4 und 5 je einmal, Produkt 2 zweimal betrachtet. Da die Betrachtung von Produkt 4 weiter in der Vergangenheit liegt, werden dem Nutzer die Produkte 2, 5 und 4 in dieser Reihenfolge vorgeschlagen. Obwohl Produkt 3 dreimal angesehen wurde, wird es genau wie Produkt 1 nicht empfohlen, da beide schon in der Vergangenheit erworben wurden. Produkt 6 wird ebenfalls nicht vorgeschlagen, da der Nutzer es gerade erst in der aktuellen Sitzung betrachtet hat. Schließlich wird ein bisher nicht bekanntes Produkt 7 empfohlen, da es der Bestseller aus einer Kategorie ist, in welcher sich viele der bisher betrachteten und gekauften Produkte des Nutzers befinden.

4 Künftige Entwicklungen

Eine große Anzahl von Beiträgen aus der aktuellen Forschung zu Empfehlungssystemen beschränkt sich bei den Eingabedaten auf zum Teil öffentlich verfügbare Bewertungsdatensätze aus verschiedenen Domänen. Dies vereinfacht nicht zuletzt die Vergleichbarkeit von Forschungsergebnissen, zumal sich die Offline-Analysen auch auf eine kleine Menge von Genauigkeitskennzahlen fokussieren. In der Realität sind solche expliziten Bewertungsdaten jedoch nicht in ausreichender Menge vorhanden. Gleichzeitig liegen jedoch fast immer weitere Informationen über die Nutzer, die Produkte oder über die aktuelle Kontextsituation der Benutzer vor.

In diesem Artikel wurden mögliche Einschränkungen hinsichtlich des praktischen Nutzens von heutigen Offline-Datenanalysen diskutiert und auf die Notwendigkeit einer zielorientierten und kontextabhängigen Evaluierung von Empfehlungsalgorithmen eingegangen. Einerseits wird vorgeschlagen, verstärkt Evaluierungsansätze zu verwenden, welche mehrere Metriken gleichzeitig betrachten. Gleichzeitig wird auf die Notwendigkeit der Entwicklung von neuen Methoden und Evaluierungsprotokollen, welche den aktuellen Interessens- und Navigationskontext besser berücksichtigen, hingewiesen. Kontextualisierung von Empfehlungen und der mehrdimensionale Einsatz von Metriken eröffnen neue Perspektiven zur Informationsgewinnung und werden wahrscheinlich weiter in den Fokus der Forschung rücken. Mithilfe in dieser Weise optimierter Analysen kann eine bessere Voraussetzung geschaffen werden, um die tatsächliche Wirksamkeit verschiedener Empfehlungsalgorithmen im realen Einsatz bereits im Offline-Experiment abzuschätzen.

Abschließend soll angemerkt werden, dass diese verbesserten Methoden zur Offline-Evaluation den Einsatz von Laborstudien oder Live-Tests nicht ersetzen sollen. Vielmehr soll in Zukunft in der Forschung zu Empfehlungssystemen verstärkt der Versuch gemacht werden, die verschiedenen Instrumente zu kombinieren und die Beobachtungen miteinander abzugleichen.

Literatur

[Adomavicius und Kwon 2012] *Adomavicius, G.; Kwon, Y.:* Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. IEEE Transactions on Knowledge and Data Engineering, 24(5), 2012, S. 896-911.

[Cremonesi et al. 2010] *Cremonesi, P.; Koren, Y.; Turrin, R.:* Performance of Recommender Algorithms on Top-N Recommendation Tasks. Proc. ACM Conference on Recommender Systems - RecSys '10, 2010, Barcelona, Spain S. 39-46.

[Fleder und Hosanagar 2009] *Fleder, D.; Hosanagar, K.:* Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. Management Science 55(5), 2009, S. 697-712.

[Herlocker et al. 2004] *Herlocker, J.; Konstan J.; Terveen L.; Riedl, J.:* Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems, 22(1), 2004, S. 5-53.

[Jannach und Hegelich 2009] *Jannach, D.; Hegelich, K.:* A Case Study on the Effectiveness of Recommendations in the Mobile Internet. Proc. ACM Conference on Recommender Systems – RecSys '09, New York, USA, 2009, S. 205-208.

[Jannach et al. 2010] *Jannach, D.; Zanker; M., Felfernig, A; Friedrich, G.:* Recommender Systems – An Introduction. Cambridge University Press, 2010.

[Jannach et al. 2012] *Jannach, D.; Zanker, M.; Ge, M.; Gröning, M.:* Recommender Systems in Computer Science and Information Systems - A Landscape of Research. Proc. 13th International Conference on Electronic Commerce and Web Technologies - EC-Web 2012, Vienna, Austria, 2012, S. 76-87.

[Jannach et al. 2013] *Jannach, D.; Lerche, L.; Gedikli, F.; Bonnin, G.:* What Recommenders Recommend - An Analysis of Accuracy, Popularity, and Sales Diversity Effects. 21st International Conference on User Modeling, Adaptation and Personalization - UMAP 2013, Rome, Italy, 2013.

[Mobasher et al. 2001] *Mobasher, B.; Dai, H.; Luo, T.; Nakagawa, M.:* Effective Personalization Based on Association Rule Discovery from Web Usage Data. Proc. 3rd International Workshop on Web Information and Data Management - WIDM '01, Atlanta, Georgia, USA, 2001, S. 9-15.

[Pu et al. 2011] *Pu, P.; Chen, L. und Hu, R.:* A User-centric Evaluation Framework for Recommender Systems. Proc. ACM Conference on Recommender Systems - RecSys '11, Chicago, Illinois, USA, 2011, S. 157-164.

Stichworte

Empfehlungssysteme, Datenanalyse, E-Commerce, Maschinelles Lernen, Information Retrieval, Evaluierung

Kolumnentitel

Offline-Evaluation von Empfehlungsalgorithmen