Ermittlung nutzungsbasierter Ähnlichkeiten zwischen Objekten zur Unterstützung von Empfehlungssystemen im Umgang mit selten genutzten Objekten[1]

Katja Niemann[2]

Abstract: Empfehlungssysteme weisen eine stetig wachsende Bedeutung in vielen Anwendungsdomänen auf. Damit wachsen auch die Erwartungen der Nutzer an die Empfehlungen, z.B. in Bezug auf Präzision, Diversität und Neuartigkeit. Unter anderem hindert die dabei oft bestehende geringe Datendichte die Systeme jedoch daran, ihr volles Potential zu entfalten, und inbesondere selten genutzte Objekte werden kaum empfohlen. In dieser Arbeit werden Lösungen konzipiert und empirisch belegt, um Ähnlichkeiten zwischen Objekten basierend auf ihrer Nutzung zu entdecken. Auf diese Art wird eine neue Informationsquelle geschaffen, welche genutzt werden kann, um bestehende Empfehlungssysteme zu erweitern und selten genutzte Objekte zuverlässiger zu empfehlen.

1 Einführung

Für die meisten Internetnutzer sind Empfehlungssysteme allgegenwärtig geworden. Sie bekommen Empfehlungen, wenn sie online einkaufen, Musik hören, einen Urlaub planen oder eine E-Learning Anwendungen nutzen. Empfehlungssysteme wählen dabei aus einer Menge von Objekten diejenigen aus, welche für die Nutzer von besonderer Relevanz sind.

Ein Problem für Empfehlungssysteme stellt die geringe Datendichte dar, welche in den meisten Domänen vorliegt [Zh14]. In Anwendungen, in welchen nur sehr spärliche Informationen vorhanden sind, können häufig keine personalisierten Empfehlungen generiert werden. Dies ist oft bei online verfügbaren Lernportalen der Fall und führt dazu, dass die Lernenden durch das System nicht optimal in ihrem Lernprozess unterstützt werden [Ve11]. In anderen Anwendungen können zwar personalisierte Empfehlungen erstellt werden, die häufig genutzten Objekte werden jedoch überproportional oft empfohlen und die selten genutzten Objekte, welche für die meisten Nutzer aufgrund ihres in der Regel niedrigeren Bekannheitsgrades schwerer zu finden sind, werden vernachlässigt [AK12]. Dieses Problem betrifft z.B. Portale, in welchen Filme angesehen werden können. Nutzer dieser Portale schätzen jedoch insbesondere Empfehlungen für unbekannte Filme, welche sie selber nicht gefunden hätten, wohingegen Blockbuster weniger interessant sind [Go10].

Das Ziel dieser Arbeit liegt darin, Möglichkeiten aufzuzeigen, mit welchen auch spärliche Informationen über die Nutzung von Objekten so ausgewertet werden können, dass geeignete Empfehlungen für diese Objekte erstellt werden können. Dabei sollen keine Meta-

[1] Original title: Discovery of Usage-based Item Similarities to Support Recommender Systems in Dealing with Rarely Used Items

[2] XING AG, katja.niemann@xing.com

daten, wie beispielsweise Angaben zum Inhalt oder Genre eines Films, benötigt werden. Hieraus ergibt sich die erste Forschungsfrage.

FF 1: Lassen sich basierend auf der Analyse der Kontexte, in welchen Objekte genutzt werden, Ähnlichkeitsrelationen zwischen den Objekten aufdecken?

Die nutzungsbasierten Ähnlichkeiten sollen daraufhin genutzt werden, um auch selten genutzte Objekte in die personalisierten Empfehlungen für Nutzer einbinden zu können. Aus diesem Anliegen ergibt sich die zweite Forschungsfrage.

FF 2: Kann der Einsatz von nutzungsbasierten Objektähnlichkeiten Empfehlungssysteme im Umgang mit selten genutzten Objekten unterstützen?

Die Zusammenfassung ist wie folgt gegliedert. Kapitel 2 gibt einen kurzen Überblick über den Stand der Wissenschaft, während Kapitel 3 die in der Arbeit analysierten Datenmengen beschreibt. Kapitel 4 und 5 befassen sich mit jeweils einer der beiden Forschungsfragen. Abschließend diskutiert Kapitel 6 die Ergebnisse und Implikationen der Arbeit.

2 Stand der Wissenschaft

Inhaltsbasierte Empfehlungssysteme nutzen die Attribute der Objekte und die Präferenzen der Nutzer, um Empfehlungen zu generieren. Dabei können die Informationen über die Objekte entweder manuell oder automatisch generiert werden, z.B. durch die Extraktion von Schlüsselwörtern aus Texten. Die Interessen der Nutzer können explizit erfragt oder durch Analyse der genutzten Objekte abgeleitet werden. Zur Empfehlungserstellung werden die Objektprofile mit den Nutzerprofilen abgeglichen und die am besten passenden Objekte empfohlen [LdGS11]. Inhaltsbasierte Systeme können erfolgreich eingesetzt werden, sobald Informationen über die Nutzer und die Objekte vorhanden sind, was jedoch häufig nicht der Fall ist. Zudem tendieren sie zur Überspezialisierung und berücksichtigen keine subjektiven Faktoren, wie die subjektiv empfundene Qualität eines Films.

Im Gegensatz dazu beruhen Verfahren, welche kollaboratives Filtern nutzen, ausschließlich auf Matrizen, welche die expliziten oder impliziten Bewertungen der Nutzer für die Objekte enthalten. Hierbei können entweder die Nutzer, basierend auf den von ihnen bewerteten Objekten, oder die Objekte, basierend auf den Nutzern, welche sie bewertet haben, verglichen werden. Empfohlen werden daraufhin entweder die gut bewerteten Objekte ähnlicher Nutzer oder die Objekte, welche den gut bewerteten Objekten eines Nutzers ähnlich sind. Ein weiteres Verfahren besteht darin, die vorhandene Matrix in kleinere, vollbesetzte Matrizen zu zerlegen und somit die Objekte und Nutzer auf dieselben latenten Faktoren abzubilden. Durch Multiplikation der Matrizen lässt sich jedes Feld der ursprünglichen Matrix und somit jede bisher unbekannte Bewertung schätzen [KB11]. Die Vorteile des kollaborativen Filterns liegen darin, dass keine semantischen Informationen über die Objekte vorhanden sein müssen und auch Ähnlichkeiten zwischen Objekten gefunden werden können, welche nicht auf den ersten Blick ersichtlich sind. Allerdings müssen die Objekt- und Nutzerprofile sich erst durch Nutzeraktivitäten entwickeln, bevor zufriedenstellende Empfehlungen erstellt werden können.

Hybride Techniken kombinieren verschiedene Empfehlungssysteme, um von ihren Vorteilen zu profitieren und ihre Nachteile zu kompensieren [Bu07]. Empfehlungen verschiedener Systeme können z.B. basierend auf einem Gewichtungsschema kombiniert werden. Ein weiteres Beispiel stellt die Merkmalserweiterung dar, bei welcher ein Empfehlungssystem Informationen erstellt, welche vom nächsten Empfehlungssystem genutzt werden.

Es ist jedoch nicht immer ausreichend, Objekte zu empfehlen, welche die Nutzer mögen. Nutzer möchten positiv überrascht werden und Empfehlungen für Objekte bekommen, welche sie ohne Hilfe nicht gefunden hätten. Zudem möchten z.B. Online-Shops, dass eine möglichst große Anzahl ihrer angebotenen Produkte empfohlen wird. Daher werden immer mehr Ansätze entwickelt, welche die Diversität der empfohlenen Objekte und die Anzahl der Empfehlungen für selten genutzte Objekte erhöhen sollen. Solche Ansätze kombinieren in der Regel entweder Nutzungsdaten mit inhaltlichen Daten, welche jedoch häufig nicht verfügbar sind, oder verschlechtern die Präzision der Empfehlungen, wenn z.B. bewusst selten genutzte Objekte bei der Empfehlungserstellung bevorzugt werden, für welche jedoch aufgrund der niedrigen Nutzung keine zuverlässige Bewertungen geschätzt werden können [Ta13, AK12].

3 Anwendungsdomänen und verwendete Datenmengen

Die Domänen, in welchen die hier vorgestellten Ansätze angewendet werden können, sind nicht begrenzt. Diese Arbeit konzentriert sich jedoch auf die Analyse von Datenmengen, welche in den Lernportalen MACE[3] und Travel well[4], bzw. in den Filmportalen Movielens[5] und Netflix[6] gesammelt wurden.

Das Webportal des MACE-Projektes verknüpft Lernmaterialen aus dem Bereich der Architektur über die Grenzen einzelner Repositorien hinweg und ermöglicht seinen Nutzern somit ein einfacheres Auffinden von Objekten wie Zeichnungen und Videos. Für die Analyse konnten Interaktionen von 620 registrierten Nutzern mit 12.176 verschiedenen Objekten ausgewertet werden. Die Nutzeraktionen wurden über einen Zeitraum von 3 Jahren gesammelt und beinhalten das Aufrufen, Bewerten sowie Taggen der Objekte. Das Travel-well-Webportal bietet Lernmaterialien für den Sprachunterricht in Schulen an. Die Datenmenge enthält die Interaktionen von 98 registrierten Nutzern mit 1.924 Objekten, welche innerhalb von sechs Monaten gesammelt wurden. Die gespeicherten Nutzeraktionen umfassen hier jedoch nur das Bewerten und Taggen der Objekte. Für ca. 80% der Objekte beider Datenmengen sind entweder Tags, welche von Nutzern hinzugefügt wurden, oder Klassifikationen, welche von Experten hinzugefügt wurden, vorhanden.

Für MovieLens und Netflix stehen ausschließlich die expliziten Bewertungen von Nutzern für Filme zur Verfügung. Für MovieLens liegen 1.000.000 Bewertungen von 6.040 Nutzern für 3.952 Filme vor. Die hier verwendete Netflix-Datenmenge beinhaltet 1.863.197 Bewertungen von 9.006 Nutzern für 17.208 Filme.

[3] http://www.fit.fraunhofer.de/de/fb/cscw/projects/mace.html
[4] http://lreforschools.eun.org/web/guest/travelwell-all
[5] http://www.grouplens.org/node/73
[6] http://www.netflixprize.com/

4 Nutzungsbasierte Ähnlichkeiten zwischen Objekten

Dieses Kapitel behandelt die erste Forschungsfrage: *Lassen sich basierend auf der Analyse der Kontexte, in welchen Objekte genutzt werden, Ähnlichkeitsrelationen zwischen den Objekten aufdecken?* Zunächst wird die Motivation für dieses Vorgehen erläutert. Daraufhin wird der Begriff des Nutzungskontextes in Bezug auf die jeweiligen Domänen definiert, die Berechung der Ähnlichkeit von Objekten basierend auf ihren Nutzungskontexten erklärt und schließlich ein ausgewähltes Experiment vorgestellt.

4.1 Motivation

Die Arbeit folgt der Annahme aus dem *Context Aware Computing*, dass die Aktivitäten eines Nutzers durch sein vorhandenes Wissen und seinen aktuellen Kontext beeinflusst werden und diese somit implizit in den Nutzungsinformationen der genutzten Objekte enthalten sind [AM07]. Es mag zum Beispiel Nutzer geben, welche an sonnigen Sommertagen andere Musik hören als an regnerischen Wintertagen und ein mit Software-Engineering vertrauter Nutzer wird vermutlich andere Bücher zu diesem Thema lesen als ein Studienanfänger. Auch wenn basierend auf den genutzten Objekten nicht alle Kontextinformationen ermittelt werden können, so bestehen sie doch als kontextuelle Verbindung zwischen den Objekten. Diese Arbeit stellt nun die Hypothese auf, dass zwei Objekte, welche in ähnlichen, aber nicht notwendigerweise in denselben Kontexten genutzt wurden, in einer Ähnlichkeitsbeziehung zueinander stehen.

Diese Idee kann als Analogie zu Ansätzen aus der Korpuslinguistik verstanden werden, in welchen linguistische Entitäten (z.B. Wörter) durch ihre Nutzungskontexte, d.h. durch die Entitäten, mit welchen sie gemeinsam genutzt wurden, beschrieben werden. Basierend auf den Arbeiten von Saussure Anfang des 20ten Jahrhunderts, prägte Harris [Ha54] in den 1950er Jahren den Begriff der distributionellen Hypothese, welche aussagt, dass Wörter, die in ähnlichen Kontexten genutzt werden, häufig eine ähnliche Bedeutung aufweisen. Ein Beispiel: in vielen Sätzen kann der Ausdruck *Auto* durch den Ausdruck *Wagen* ersetzt werden. Das bedeutet, dass die beiden Ausdrücke in ähnlichen Kontexten genutzt werden, welche zum Beispiel die Ausdrücke *Werkstatt* und *Autobahn* enthalten. Daher kann angenommen werden, dass die Ausdrücke *Auto* und *Wagen* eine semantische Relation aufweisen [Ho09]. Die vorliegende Arbeit untersucht nun die Annahme, dass Objekte, welche in Nutzungskontexten (z.B. *Web Sessions*) genutzt wurden, in Analogie zu Wörtern, welche in Sätzen genutzt wurden, analysiert werden können.

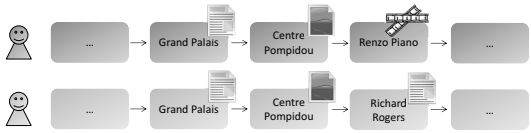

Abb. 1: Beispielhafte Nutzungskontexte

Abbildung 1 zeigt Objekte, auf welche von zwei unterschiedlichen Nutzern in einem Online-Portal zugegriffen wurde. Es kann festgestellt werden, dass der Film über Renzo

Piana und das Textdokument über Richard Rodgers getrennt voneinander und von unterschiedlichen Nutzern ausgewählt wurden. Dennoch wurden sie in ähnlichen Nutzungskontexten, d.h. gemeinsam mit denselben Objekten (mit einem Textdokument über das Grand Palais und einem Bild, welches das Centre Pompidou zeigt), genutzt. Daher kann angenommen werden, dass die beiden Objekte in einer Ähnlichkeitsrelation zueinander stehen. Tatsächlich referenzieren sie jeweils einen der beiden Architekten, welche das Centre Pompidou gemeinsam entworfen haben.

Ein weiterer Ansatz aus der Korpulinguistik berechnet Kookkurrenzen höherer Ordnung, um Gruppen zu bilden, welche semantisch ähnliche linguistische Entitäten enthalten. Ebenso wie der erste Ansatz, lässt sich dieses Vorgehen adaptieren und auf Objekte anwenden (siehe [NW14, Ni12]), kann hier jedoch nicht weiter ausgeführt werden.

4.2 Definition des Nutzungskontextes für die verwendeten Datenmengen

Die Datenmengen aus dem E-Learning wurden in Webportalen gesammelt, in welchen Nutzer nach Lernmaterialien suchen und sie direkt verwenden können. Hier kann ein Nutzungskontext als äquivalent zu einer *Web Session* verstanden werden, also einer Abfolge von Aktivitäten, welche von einem einzelnen Nutzer bei einem Besuch auf einer Webseite ausgeführt wurden. Diese Definition ist übertragbar auf Domänen mit ähnlichen Rahmenbedingungen, z.B. Webportale, in welchen die Nutzer kurze Videoclips ansehen können.

Die Definition des Nutzungskontextes für Filme ist nicht so naheliegend wie für Lernmaterialien. Gewöhnlicherweise werden nicht mehrere Filme nacheinander in einer kurzer Zeit, sondern an verschiedenen Tagen, konsumiert. Eine Möglichkeit zur Bildung sinnvoller Kontexte besteht daher in der Ausnutzung detaillierter Kontextinformationen, wie dem Zeitpunkt, zu dem ein Film angesehen wurde (z.B. am Wochende oder an einem Wochentag), oder die Begleitung (z.B. mit Freunden oder mit Kindern). Häufig liegen jedoch nur explizite Bewertungen der Filme vor und keine detaillierten Kontextinformationen. In dieser Arbeit werden daher die Nutzerprofile, welche alle Bewertungen eines Nutzers enthalten, verwendet, um Nutzungskontexte zu erstellen. Ein Nutzerprofil kann z.B. in zwei Nutzungskontexte aufgeteilt werden, wobei einer alle über- und der andere alle unterdurchschnittlich bewerteten Filme enthält. Hierfür wird jeweils der Bewertungsdurchschnitt des entsprechenden Nutzers berechnet.

4.3 Repräsention von Objekten und Objektähnlichkeit

In Analogie zur Korpuslinguistik werden zwei Objekte als Kookkurrenzen bezeichnet, sobald sie wenigstens einmal gemeinsam in einem Nutzungkontext aufgetreten sind. Zur Beschreibung eines Objektes eignen sich jedoch nur diejenigen Objekte, mit welchen es signifikant häufig und nicht nur zufällig genutzt wurde. Daher wird zunächst für alle Kookkurrenzen ein Signifikanzwert errechnet, welcher die Häufigkeit des gemeinsamen Auftretens, aber auch die Auftretenshäufigkeiten der einzelnen Objekte sowie die Anzahl aller Nutzungskontexte berücksichtigt. Einfache Assoziationsmaße wie *Mutual Information* (MI) vergleichen für jedes Objektpaar die Häufigkeit des erwarteten und des

tatsächlichen gemeinsamen Auftretens. Andere Assoziationsmaße wie der χ^2-Test beruhen auf einer Kontingenztabelle und vergleichen somit auch die erwarteten Werte für das alleinige Vorkommen der Objekte. Da die Vorkommenshäufigkeit der einzelnen Objekte stark variieren kann und viele Objekte kaum genutzt werden, bietet sich hier für den χ^2-Test die Verwendung der Kontinuitätskorrektur nach Yates an (cor-χ^2). Weitere getestete Assoziationsmaße sind *Log-Likelikhood (LL)* und ein *Poisson*-basiertes Assoziationsmaß (PAM), siehe [NW13b].

Nach Berechnung der Signifikanzwerte werden für jedes Objekt die signifikanten Kookkurrenzen ausgewählt. Hierfür gibt es jedoch keinen Standardgrenzwert [Ev08]. In der Arbeit werden daher zwei Herangehensweisen untersucht. Erstens, für jedes Objekt werden die *n* signifikantesten Kookkurrenzen ausgewählt. Zweitens, für jedes Objekt wird die mittlere Signifikanz aller seiner Kookkurrenzen ermittelt und als Grenzwert für die Bestimmung der signifikanten Kookkurrenzen genutzt. Jedes Objekt wird nun durch einen Vektor beschrieben, welcher die signifikanten Kookkurrenzen des Objektes inklusive der errechnten Signifikanzwerte beinhaltet. Zum Vergleich zweier Objekte wird das Kosinus-Ähnlichkeitsmaß genutzt, welches die Richtung zweier Vektoren vergleicht.

4.4 Experiment: Nutzungsbasierte Ähnlichkeit von Lernmaterialien

Für alle Objektpaare aus den Datenmengen, welche in den Lernportalen MACE und Travel well gesammelt wurden, werden nutzungsbasierte Ähnlichkeiten erstellt und mit den dazugehörigen inhaltsbasierten Ähnlichkeiten verglichen. Zur Berechnung der nutzungsbasierten Ähnlichkeiten werden zunächst die in Kapitel 4.3 vorgestellten Assoziationsmaße und Ansätze zur Auswahl der signifikanten Kookkurrenzen eingesetzt. Hierbei wird die Anzahl *n* der signifikanten Kookurrenzen zwischen 10 und 1.500 (MACE), bzw. zwischen 10 und 150 (Travel well) variiert. Die nutzungsbasierten Ähnlichkeiten ergeben sich durch die Berechnung der Kosinus-Ähnlichkeit der Kookkurrenzvektoren. Zudem wird jedes Objekt durch eine Menge, welche seine Tags und Klassifikationswerte enthält, beschrieben. Zur Ermittlung der inhaltsbasierten Ähnlichkeit zweier Objekte wird der Jaccard-Koeffizient dieser Mengen berechnet. Abschließend wird der Pearson-Korrelationskoeffizient zwischen den nutzungsbasierten und den inhaltsbasierten Ähnlichkeiten berechnet sowie die Anzahl der Objekte, für welche mindestens eine Ähnlichkeitsbeziehung zu einem anderen Objekt gefunden werden konnte, ermittelt.

Es zeigt sich eindeutig, dass je mehr Kookkurrenzen als signifikant eingestuft werden, also je größer *n*, desto mehr Ähnlichskeitsbeziehungen können zwischen den Objekten gefunden werden. Dies kann intuitiv dadurch erklärt werden, dass je mehr Kookkurrenzen genutzt werden, um die Objekte zu beschreiben, desto höher ist die Wahrscheinlichkeit, dass zwei Objekte mindestens eine Kookkurrenz teilen. Zudem zeigt sich der Trend, dass mit einem höheren Wert für *n* der Korrelationskoeffizient zunächst stark steigt und ab einem gewissen Punkt wieder leicht zu sinken beginnt. Werden objektspezifische Grenzwerte zur Auswahl der signifikanten Kookkurrenzen berechnet, können ähnlich gute Ergebnisse wie mit den besten Werten für *n* erzielt werden und es muss kein zusätzlicher Parameter bestimmt werden. Weiterhin zeigt sich, dass das Assoziationsmaß *Mutual Information* für

beide Datenmengen zu den besten Ergebnissen führt. Die Empfehlung liegt daher auf der Nutzung von *Mutual Information* in Kombination mit objektspezifischen Grenzwerten.

Die mit diesem Verfahren erzielten Korrelationskoeffizienten von 0.47 (MACE) und 0.33 (Travel well) deuten auf eine mittlere Korrelation hin. Die genutzten semantischen Informationen sind jedoch spärlich und stellen nur eine oberflächliche Inhaltsrepräsentation der Objekte dar. Eine zusätzliche manuelle Untersuchung der 100 Objektpaare mit den höchsten nutzungsbasierten Ähnlichkeiten zeigt, dass über 95% eine inhaltliche Ähnlichkeit aufweisen, welche in 33% der Fälle nicht aus den semantischen Metadaten ersichtlich ist. Aus diesen Ergebnissen kann gefolgert werden, dass die nutzungsbasierten Ähnlichkeiten der Objekte aus den untersuchten Lernportalen einen Hinweis auf ihre inhaltliche Nähe geben. Für weitere Experimente und ausführliche Ergebnisse, siehe [Ni11, Ni10].

5 Verbesserung von Empfehlungen

Dieses Kapitel behandelt die zweite Forschungsfrage: *Kann der Einsatz von nutzungsbasierten Objektähnlichkeiten Empfehlungssysteme im Umgang mit selten genutzten Objekten unterstützen?* Zunächst wird erläutert, wie die Objektähnlichkeiten für die Vorhersage von Bewertungen genutzt werden können. Daraufhin wird ihr Einsatz zur Empfehlungserstellung in Lern- und Filmportalen diskutiert.

5.1 Vorhersagen von Bewertungen

Die nutzungsbasierten Objektähnlichkeiten und die Objektbewertungen, welche Nutzer bereits abgegeben haben, können genutzt werden, um für jeden Nutzer u vorherzusagen, welche Bewertung $\hat{r}_{ui}$ er für ein von ihm bisher nicht bewertetes Objekt i abgeben würde, siehe Gleichung 1. $P(u)$ bezeichnet dabei das Nutzerprofil von Nutzer u, welches jedes bereits bewertete Objekt j inklusive Bewertung r_{uj} enthält. Die Objektbewertungen werden daraufhin kombiniert und basierend auf der Objektähnlichkeit $sim(i, j)$ gewichtet.

$$\hat{r}_{ui} = \frac{\sum_{j \in P(u), i \neq j} (sim(i, j) * r_{uj})}{\sum_{j \in P(u), i \neq j} sim(i, j)} \tag{1}$$

5.2 Experiment: Empfehlung von Lernmaterialien

Die Evaluationen auf den Datenmengen, welche in den Lernportalen MACE und Travel well gesammelt wurden, zeigen, dass die Kombination von nutzungsbasierten Ansätzen mit kollaborativen Filtern in Domänen mit geringer Datendichte sehr vielversprechend ist. Die Anzahl der Bewertungen, für welche eine Vorhersage getroffen werden kann, steigt dabei im Vergleich zu den kollaborativen Verfahren von 14,8% auf 67,9% (MACE), bzw. von 31,5% auf 94,7% (Travel well), ohne dass die Qualität der Vorhersagen sich verschlechtern. Die ausführliche Beschreibung dieser Experimente ist in [NW13a] zu finden und wurde bei der *EC-TEL '13* mit dem *Best Student Paper Award* ausgezeichnet.

5.3 Experiment: Empfehlung von Filmen

Die Bewertungen der MovieLens- und Netflix-Datenmengen werden zufällig in je fünf Teilmengen aufgeteilt, um eine 5-fache Kreuzvalidierung zu ermöglichen. Dem Ansatz in [AK12] folgend, werden alle Nutzer, welche weniger als 20 hohe Bewertungen (mind. 4 von 5 Sternen) in der jeweiligen Testmenge aufweisen, aus der Testmenge entfernt und zur Trainingsmenge hinzugefügt. Die so erstellten Testmengen enthalten im Schnitt 147.494 Bewertungen (85.309 hohe and 62.185 niedrige Bewertungen) von 2.152 Nutzern (MovieLens), bzw. 305.132 Bewertungen (177.551 hohe and 127.581 niedrige Bewertungen) von 3.439 Nutzern (Netflix). Für jeden Nutzer aus der jeweiligen Testmenge werden nun die Bewertungen für die in der Testmenge vorhandenen Filme geschätzt und die zehn Filme mit der jeweils höchsten geschätzten Wertung ausgewählt. Der hier vorgestellte Ansatz (*Usage Context-based Collaborative Filtering, UC-BCF*) wird dabei in Kombination mit verschiedenen Assoziationsmaßen (siehe Kapitel 4.3) getestet. Dabei wird die Anzahl der signifikanten Kookurrenzen n von 10-100 variiert. Zudem werden folgende Ansätze des kollaborativen Filterns genutzt, um Vergleichswerte zu erstellen: *Item-based Collaborative Filtering* (IBCF), *User-based Collaborative Filtering* (UBCF), *Single Value Decomposition* (SVD) aus der PREA[7] Bibliothek und *Biased Matrix Factorization* (BMF) aus der MyMediaLite[8] Bibliothek.

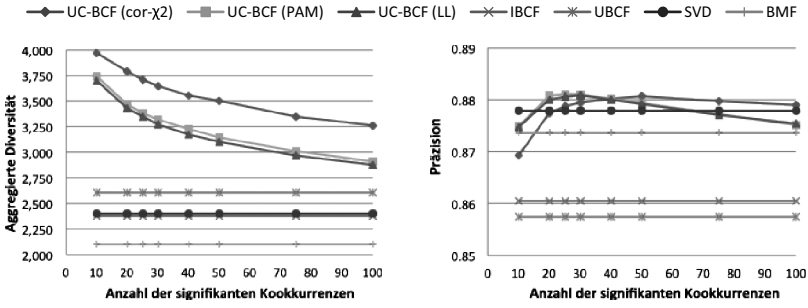

Abb. 2: Ergebnisse für Netflix (Top-10)

Abbildung 2 zeigt eine Auswahl an Ergebnissen für Netflix. Die aggregierte Diversität gibt an, wieviele unterschiedliche Filme insgesamt empfohlen wurden. Die Präzision gibt die relative Anzahl der Filme an, welche für einen Nutzer ausgewählt und von diesem auch tatsächlich mit mindestens 4 von 5 Sternen bewertet wurden. Die Evaluation zeigt für beide Datenmengen, dass der χ^2-Test mit Kontinuitätskorrektur in Kombination mit n=25 die besten Ergebnisse liefert. Somit kann die Anzahl der empfohlenen Filmen um 42,37% (UBCF) bis zu 76,52% (BMF) gesteigert werden. Im Gegensatz zu anderen Verfahren, welche die aggregierte Diversität erhöhen, wird hier jedoch die Präzision nicht verringert, sondern sogar leicht angehoben. Weitere Untersuchungen zeigen, dass nicht nur insgesamt mehr, sondern tatsächlich mehr selten genutzte Filme empfohlen werden und dieser Effekt durch Kombinationen mit anderen Verfahren noch verstärkt wird (siehe [NW13b]).

[7] http://mloss.org/software/view/420/
[8] http://www.mymedialite.net/

6 Schlussfolgerung

Die Erstellung von Empfehlungen für selten genutzte Objekte ist anspruchsvoll, aber in vielen Anwendungen wünschenswert, z.B. um Nutzer positiv zu überraschen und zufriedenstellen zu können oder um sie in ihren Lernprozessen optimal unterstützen zu können. Das Anliegen dieser Arbeit ist es daher, zum Stand der Wissenschaft für Empfehlungssysteme beizutragen, indem Methoden entwickelt und evaluiert werden, mit welchen trotz einer geringen Datendichte, nützliche Empfehlungen erstellt werden können. Dafür wurden zunächst Lösungen konzipiert, mit welchen Ähnlichkeiten zwischen Objektpaaren allein durch die Analyse ihrer Nutzung gefunden werden können. Hierbei hat sich gezeigt, dass es zwischen den Domänen Unterschiede bei der Auswahl der besten Werkzeuge (wie z.B. Assoziationsmaße und Anzahl signifikanter Kookkurrenzen) gibt. Innerhalb einer Domäne sind die Werkzeuge jedoch zwischen den Datenmengen übertragbar. Danach wurde die Nützlichkeit dieser nutzungsbasierten Ähnlichkeiten bei der Erstellung von Empfehlungen evaluiert. Hierbei hat sich gezeigt, dass die nutzungsbasierten Ansätze, welche in dieser Arbeit entwickelt wurden, sehr viel besser geeignet sind, um Empfehlungen für selten genutzte Objekte zu erstellen, als aktuelle Empfehlungssysteme. Die nutzungsbasierten Ansätze können dabei als eigenständige Empfehlungssysteme eingesetzt oder auch mit anderen Systemen kombiniert werden, um noch mehr, bzw. noch präzisere Empfehlungen erstellen zu können als jedes System für sich alleine.

Literaturverzeichnis

[AK12] Adomavicius, Gediminas; Kwon, Youngok: Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. IEEE Transactions on Knowledge and Data Engineering, 24(5):896–911, 2012.

[AM07] Anand, Sarabjot Singh; Mobasher, Bamshad: Contextual Recommendation. In: Proc. of the PKDD Workshop on Web Mining (WebMine '06). Springer, S. 142–160, 2007.

[Bu07] Burke, Robin: Hybrid Web Recommender Systems. In: The Adaptive Web: Methods and Strategies of Web Personalization, Kapitel 12, S. 377 – 408. Springer, 2007.

[Ev08] Evert, Stefan: Corpora and collocations. In: Corpus Linguistics. An International Handbook, Kapitel 57, S. 1197–1211. Mouton de Gruyter, Berlin, 2008.

[Go10] Goel, Sharad; Broder, Andrei; Gabrilovich, Evgeniy; Pang, Bo: Anatomy of the long tail. In: Proc. of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10). ACM Press, New York, NY, USA, S. 201–210, 2010.

[Ha54] Harris, Zellig S.: Distributional Structure. Word, 10(23):146–162, 1954.

[Ho09] Hoey, Michael: Corpus linguistics and word meaning. In: Corpus Linguistics. An International Handbook, Kapitel 45, S. 972–987. de Gruyter, Berlin, 2nd. Auflage, 2009.

[KB11] Koren, Yehuda; Bell, Robert: Advances in Collaborative Filtering. In: Recommender Systems Handbook, Kapitel 5, S. 145–186. Springer, 2011.

[LdGS11] Lops, Pasquale; de Gemmis, Marco; Semeraro, Giovanni: Content-based Recommender Systems: State of the Art and Trends. In: Recommender Systems Handbook, S. 73–105. Springer, 2011.

[Ni10] Niemann, Katja; Scheffel, Maren; Friedrich, Martin; Kirschenmann, Uwe; Schmitz, Hans-Christian; Wolpers, Martin: Usage-based Object Similarity. Journal of Universal Computer Science, 16(16):2272–2290, 2010.

[Ni11] Niemann, Katja; Schmitz, Hans-Christian; Scheffel, Maren; Wolpers, Martin: Usage Contexts for Object Similarity: Exploratory Investigations. In: Proc. of the 1st International Conference on Learning Analytics and Knowledge (LAK '11). ACM Press, New York, NY, USA, S. 81–85, 2011.

[Ni12] Niemann, Katja; Schmitz, Hans-Christian; Kirschenmann, Uwe; Wolpers, Martin; Schmidt, Anna; Krones, Tim: Clustering by Usage: Higher Order Co-occurrences of Learning Objects. In: Proc. of the 2nd International Conference on Learning Analytics & Knowledge (LAK '12). ACM Press, New York, NY, USA, S. 238–247, 2012.

[NW13a] Niemann, Katja; Wolpers, Martin: A New Collaborative Filtering Approach for Increasing the Aggregate Diversity of Recommender Systems. In: Proc. of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13). ACM Press, New York, NY, USA, S. 955–963, 2013.

[NW13b] Niemann, Katja; Wolpers, Martin: Usage Context-Boosted Filtering for Recommender Systems in TEL. In: Proc. of the 8th European Conference on Technology Enhanced Learning (EC-TEL '13). Springer, Berlin Heidelberg, S. 246–259, 2013.

[NW14] Niemann, Katja; Wolpers, Martin: Usage-Based Clustering of Learning Resources to Improve Recommendations. In: Proc. of the 9th European Conference on Technology Enhanced Learning (EC-TEL '14). Springer, Berlin Heidelberg, S. 317–330, 2014.

[Ta13] Taramigkou, Maria; Bothos, Efthimios; Christidis, Konstantinos; Apostolou, Dimitris; Mentzas, Gregoris: Escape the bubble. In: Proc. of the 7th ACM Conference on Recommender Systems (RecSys '13). ACM Press, New York, NY, USA, S. 335–338, 2013.

[Ve11] Verbert, Katrien; Drachsler, Hendrik; Manouselis, Nikos; Wolpers, Martin; Vuorikar, Riina; Duval, Erik: Dataset-driven Research for Improving Recommender Systems for Learning. In: Proc. of the 1st International Conference on Learning Analytics and Knowledge (LAK '11). ACM Press, New York, NY, USA, S. 44–53, 2011.

[Zh14] Zhang, Mi; Tang, Jie; Zhang, Xuchen; Xue, Xiangyuang: Addressing Cold Start in Recommender Systems: A Semi-supervised Co-training Algorithm. In: Proc. of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14). ACM Press, New York, NY, USA, 2014.

Dr. Katja Niemann wurde 2015 an der RWTH Aachen in der Informatik promoviert und hat zuvor Computerlinguistik und Informatik in Heidelberg studiert. Von 2008 bis 2016 arbeitete sie als wissenschaftliche Mitarbeiterin beim Fraunhofer-Institut für Angewandte Informationstechnik (FIT) in Sankt Augustin und engagierte sich überwiegend in EU-geförderten Forschungsprojekten wie MACE, OpenScout, OpenDiscoverySpace und CloudTeams. Seit 2016 arbeitet sie als Data Scientist bei der XING AG und befasst sich u.a. mit der Erstellung geeigneter Job-Empfehlungen. In ihrer Dissertation hat sie die Nutzung von Datenobjekten in Webportalen analysiert, um Empfehlungssysteme bei der Empfehlung selten genutzter Objekte zu unterstützen. Ihre Forschungsinteressen umfassen u.a. Empfehlungssysteme, Data Mining, Text Mining, Information Extraction und Learning Analytics.