Ein Online-Spiel zur Benennung latenter Faktoren in Empfehlungssystemen

Johannes Kunkel, Benedikt Loepp, Jürgen Ziegler[1]

Universität Duisburg-Essen[1]

johannes.kunkel@uni-due.de, benedikt.loepp@uni-due.de,
juergen.ziegler@uni-due.de

Zusammenfassung

Empfehlungssysteme, die auf latenten Faktormodellen basieren, sind dafür bekannt sehr genaue Vorschläge zu generieren. Häufig werden diese Systeme jedoch von Nutzern als intransparent wahrgenommen. Semantische Beschreibungen der latenten Faktoren könnten helfen, dieses Problem zu lindern. Solche Beschreibungen automatisch zu ermitteln gestaltet sich allerdings aufgrund der statistischen Herleitung der Faktoren aus numerischen Bewertungsdaten als schwierig. In diesem Beitrag stellen wir ein Output-Agreement-Spiel vor, das Spieler dazu motiviert, anhand repräsentativer Produkte Beschreibungen zu den Faktoren zu erstellen. Eine durchgeführte Nutzerstudie zeigt, dass das Spiel viel Spaß bereitet und die erhobenen Beschreibungen realweltliche Eigenschaften der Faktoren widerspiegeln.

1 Einleitung

Aktuelle *Empfehlungssysteme* (ES) setzen häufig modellbasiertes *Collaborative Filtering* (CF) ein (Koren und Bell, 2015). *Matrix Factorization* (MF) ist ein besonders populäres statistisches Verfahren, bei dem automatisch latente Faktoren aus zuvor erhobenen Interaktionsdaten der Nutzerschaft, z. B. numerischen Bewertungen, hergeleitet werden (Koren und Bell, 2015). Mit ihrer Hilfe können dann für jeden Nutzer Vorhersagen über zukünftige Produktbewertungen berechnet bzw. diese für personalisierte Empfehlungen herangezogen werden. Während anhand von latenten Faktoren erzeugte Empfehlungen für ihre hohe Genauigkeit bekannt sind, werden sie häufig als intransparent wahrgenommen (Rossetti et al., 2013), was zu Ablehnung seitens der Nutzer führen kann (Pu et al., 2012). Einige Forschungsarbeiten weisen allerdings auf Verbindungen zwischen automatisch gelernten latenten Faktoren und realweltlichen Eigenschaften hin (Koren und Bell, 2015; Loepp, Donkers et al., 2018). So kann im Falle von Filmen beispielsweise ein Faktor den Grad widerspiegeln, zu dem eine Liebesgeschichte enthalten ist, ein anderer, wie lustig die Filme sind. Gleichwohl mangelt es bislang an wissenschaftlichen Ansätzen, solche semantischen Beziehungen aufzudecken. Sind entsprechende Zuordnungen jedoch

bekannt, kann dem oben genannten Problem etwa durch textuelle Beschreibungen der statistisch gelernten Faktoren begegnet werden (Rossetti et al., 2013; Loepp, Donkers et al., 2018). Die mittels MF generierten Empfehlungen könnten anhand von Begriffen erklärt werden, welche jene Faktoren beschreiben, die sowohl für den aktiven Nutzer als auch für die empfohlenen Items besonders stark ausgeprägt sind. Darüber hinaus könnten Nutzer in Kaltstartsituationen (d. h. wenn ein ES erstmalig genutzt wird) aus den Begriffen, welche die einzelnen Faktoren beschreiben, relevante auswählen und somit ad-hoc innerhalb des Faktormodells eingeordnet werden, sodass in Folge passende Empfehlungen generiert werden können. Letztendlich ließen sich über die Faktoren und deren zugeordnete Begriffe inhaltliche Produktbeschreibungen gewinnen, sollten entsprechende Metadaten nicht oder nur in unzureichender Form vorliegen.

In diesem Beitrag stellen wir ein *Game-with-a-Purpose* (GWAP) (von Ahn und Dabbish, 2008) vor, das Nutzer motiviert, zuvor automatisch generierte latente Faktoren semantisch zu beschreiben. Dazu werden – der *Output-Agreement*-Methode (von Ahn und Dabbish, 2008) folgend – zufällig Spielerpaare gebildet, denen für jeden Faktor einzeln repräsentative Produkte gezeigt werden. Die Aufgabe der Spieler besteht darin, solange Begriffe einzugeben welche die Gemeinsamkeiten der gezeigten Produkte beschreiben, bis sie unabhängig voneinander denselben Begriff eingeben, also eine Übereinstimmung erreichen. Wir präsentieren eine Nutzerstudie und eine Analyse der erhobenen Daten, mit der wir zeigen, dass die gesammelten Begriffe und Übereinstimmungen eine sinnvolle Benennung der latenten Faktoren ermöglichen.

2 Verwandte Arbeiten

Lange Zeit wurde in der Forschung primär auf eine Verbesserung der Empfehlungspräzision hingearbeitet. Zuletzt rückten jedoch auch nutzerorientierte Aspekte wie z. B. die wahrgenommene Transparenz der Empfehlungen in den Fokus (Pu et al., 2012; Jugovac und Jannach, 2017). Dennoch ist es in automatischen ES noch immer schwer zu verstehen, *warum* und *wie* Empfehlungen generiert werden (Tintarev und Masthoff, 2015). Entsprechend wurden diverse Ansätze vorgeschlagen, um Erklärungen zu empfohlenen Produkten zu präsentieren – über frühe (Herlocker et al., 2000) oder sehr einfache Varianten (etwa den typischen „Andere Kunden kauften auch…"-Erklärungen von *Amazon*) bis hin zu anspruchsvollen textuellen Erklärungen und Visualisierungen (Tintarev und Masthoff, 2015). Besonders herausfordernd ist das Generieren von Erklärungen bei modellbasierten CF-Ansätzen, da die zugrundeliegenden Modelle einzig auf der statistischen Auswertung von Interaktionsdaten wie z. B. numerischen Produktbewertungen basieren. Dennoch konnten erste Fortschritte dahingehend erzielt werden, die latenten Dimensionen der Modelle zu erklären. So wurden etwa aus unstrukturierten Daten extrahierte Themen mit den Faktoren assoziiert (Rossetti et al., 2013) oder die Faktoren beim Lernen der Modelle mit Tags verrechnet (Loepp, Donkers et al., 2018). Andere Ansätze stellen den Zusammenhang von vordefinierten Tags und Faktoren visuell dar (Németh et al., 2013) oder bestimmen eine Kartendarstellung der Produkte durch Reduktion der hochdimensionalen Faktorwerte, was die Hinzunahme zusätzlicher Produktdaten erübrigt (Kunkel et al., 2017).

Während all diese Ansätze auf dem Vorhandensein einer inhärenten Semantik der zugrundeliegenden Faktormodelle beruhen, erscheint es aus Systemsicht nach wie vor schwierig, die

tatsächliche Bedeutung der Faktoren aufzudecken – insbesondere ohne Zuhilfenahme vordefinierter Daten oder komplexer Visualisierungstechniken. Folglich scheint es vielversprechend, auf die Mitarbeit der Nutzer zu setzen um eine Ausgangsbasis für die Erklärung der Faktoren zu schaffen. Entsprechende Ansätze werden oft mit *Human Computation* überschrieben, wobei der Frage, wie Nutzer zur freiwilligen Leistung des erforderlichen Aufwands motiviert werden können, eine zentrale Bedeutung zukommt. GWAP haben in dieser Hinsicht besonderes Potenzial gezeigt (von Ahn und Dabbish, 2008; Cooper et al., 2010; Eickhoff et al., 2012; Banks et al., 2015). Eine häufig anzutreffende Variante sind *Output-Agreement*-Spiele (von Ahn und Dabbish, 2008), in denen zufällig gepaarte Spieler einen gemeinsamen Inhalt präsentiert bekommen, über den Informationen gewünscht werden. Ohne weitere Möglichkeiten der Kommunikation müssen die Spieler dann Begriffe zu dem gezeigten Inhalt eingeben und versuchen eine Übereinstimmung zu erzielen. Um zu gewinnen gilt es demzufolge, die einzige geteilte Information (also den dargestellten Inhalt) so genau wie möglich zu beschreiben, sodass eine gemeinsame Ausgabe entsteht. Derartige Spiele wurden in der Vergangenheit häufig zur Beschreibung von Bildern genutzt (Ho et al., 2009), aber auch auf andere Aufgaben übertragen, beispielsweise die Auflösung textueller Mehrdeutigkeiten (Seemakurty et al., 2010) oder – im Bereich von ES – die Erhebung von Präferenzen (Hacker und von Ahn, 2009) und Bestimmung von Ähnlichkeiten zwischen Produkten (Walsh und Golbeck, 2010).

3 Das Spiel

Das von uns entwickelte *Output-Agreement*-Spiel (Abbildung 1) stellt einen Ansatz dar, um unter Mithilfe der Spieler die automatisch ermittelten Faktoren eines MF-Modells zu benennen: In einer Offline-Phase werden zunächst auf Basis eines Datensatzes mit Produktbewertungen latente Faktoren statistisch berechnet, und anschließend zu jedem Faktor repräsentative Produkte bestimmt. In einer darauf folgenden Online-Phase werden dann zufällig Spielerpaare gebildet, die immer drei der Repräsentanten für jeweils einen Faktor angezeigt bekommen. Die Spieler haben nun das Ziel zu erraten, welche Gemeinsamkeiten die dargestellten Produkte aufweisen, und entsprechend passende Begriffe einzugeben. Sobald eine Übereinstimmung gefunden, also derselbe Begriff von beiden Spielern eingegeben wurde, endet die aktuelle Runde und es wird mit dem nächsten Faktor fortgefahren, d. h. es werden erneut drei Produkte – nun repräsentativ für diesen Faktor – angezeigt. Die eingegebenen Begriffe, und insbesondere jene die zu einer Übereinstimmung geführt haben, lassen sich im Anschluss als Beschreibung für den jeweiligen latenten Faktor, und auch darüber hinaus sehr vielseitig in ES verwenden.

3.1 Faktormodell und -repräsentanten

Bevor das Spiel eingesetzt werden kann, muss offline ein MF-Modell mit latenten Faktoren gelernt werden. Als Datengrundlage nutzen wir den in der Forschung zu ES weithin eingesetzten *MovieLens-20M*-Datensatz[1], bestehend aus 20 Mio. numerischen Bewertungen von 137 000 Nutzern für 27 000 Filme. Zur Faktorisierung verwenden wir den *ParallelSGDFactorizer* aus

[1] http://grouplens.org/datasets/movielens/20m/

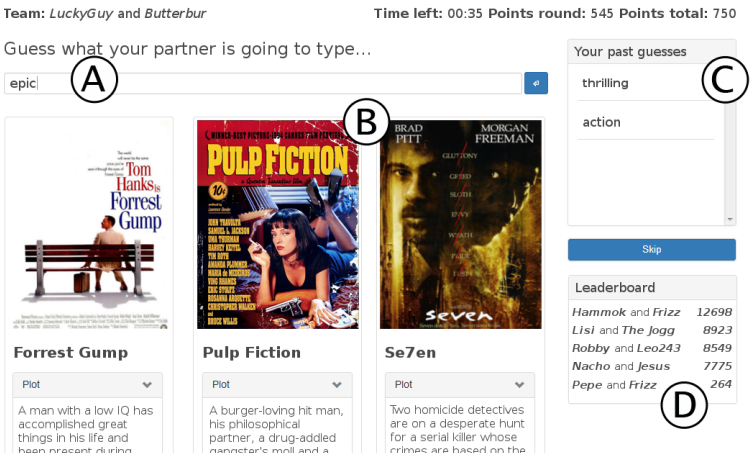

Abbildung 1: Einer der Spieler aus der Paarung „LuckyGuy“ und „Butterbur“ gibt gerade einen Begriff ein (A), den er für die drei gezeigten Filme (B) als passend erachtet. Bei den Filmen handelt es sich um Repräsentanten für einen Faktor des zugrundeliegenden latenten Modells. Zwei Begriffe hat der Spieler bereits zuvor eingegeben (C). Eine Rangliste zeigt die bisher erfolgreichsten Spielerpaare (D).

der weit verbreiteten *Mahout*-Bibliothek[2]. Als Parameter wählten wir $\lambda = 0.001$, 16 Iterationen und 10 Faktoren, was in einer eigens vorgenommenen Evaluation zu angemessenen Werten bezüglich der Empfehlungsgenauigkeit führte ($RMSE = 0.86$, $NDCG@10 = 0.82$). Während unser Ansatz prinzipiell unabhängig von der Anzahl der Faktoren ist, wählten wir bewusst einen tendenziell niedrigen Wert. Dies befindet sich im Einklang mit früheren Vorschlägen (Loepp, Hussein et al., 2014) und ist zudem für das vorgestellte Spiel von Vorteil: 1. können somit später mehr Begriffe pro Faktor erhoben werden, was zu einem besseren Verhältnis von geleistetem Aufwand und erhaltenen Beschreibungen führt, 2. werden Redundanzen zwischen den Faktoren unwahrscheinlicher, was ein abwechslungsreicheres Spielerlebnis ermöglicht.

Anschließend werden die Repräsentanten für jeden latenten Faktor des Modells ermittelt. Um dabei neben einer guten Unterscheidbarkeit sicherzustellen, dass die individuellen semantischen Eigenschaften der Faktoren bei der Auswahl der Repräsentanten zur Geltung kommen, folgen wir im Prinzip der Methodik von Loepp, Hussein et al. (2014). Zu diesem Zweck berechnen wir für jeden Faktor f und jeden Film i, dessen Faktorwert innerhalb der oberen 25 % der Werte für f liegt, einen Score s_{if} gemäß Gleichung (1). Anschließend wählen wir die 25 Filme[3] mit den höchsten Scores als Repräsentanten.

$$s_{if} = 0.4 \cdot pop(i) + 0.3 \cdot rel(i, f) + 0.3 \cdot spec(i, f) \tag{1}$$

Loepp, Hussein et al. (2014) folgend, berücksichtigt die Gleichung folgende Eigenschaften:

Popularität: Um die Wahrscheinlichkeit zu erhöhen, dass Spieler die Filme kennen, definieren wir $pop(i)$ derart, dass Filmen ein höherer Popularitätswert zugewiesen wird, je mehr Bewertungen sie in der Vergangenheit von der Nutzerschaft erhalten haben.

[2] http://mahout.apache.org/

[3] Gewichtungen für die Gleichung und Anzahl der Repräsentanten ermittelten wir in einer qualitativen Nutzerstudie.

Relevanz: Um sehr charakteristische Filme für f zu erhalten, weist die Funktion $rel(i, f)$ Filmen mit hohem Faktorwert für f einen hohen Relevanzwert zu.

Spezifität: Um darüber hinaus Filme zu erhalten, die möglichst spezifisch für f sind, definieren wir $spec(i, f)$ derart, dass Filme einen hohen Spezifizitätswert erhalten, wenn sie für f hohe, aber für die anderen Faktoren eher neutrale Faktorwerte besitzen.

3.2 Spielmechanik

Wie für *Output-Agreement*-Spiele empfohlen (von Ahn und Dabbish, 2008), setzten wir unser Spiel als Web-Applikation um. Übergeordnetes Spielziel ist es, so viele *Punkte* wie möglich zu sammeln, um auf der zentralen Rangliste (Abbildung 1, D) möglichst weit oben platziert zu werden. Punkte werden im Laufe von einzelnen *Spielen* gesammelt, wobei für jedes Spiel zwei Spieler zufällig einander zugeordnet werden. Jedes Spiel besteht aus mehreren *Runden*. Für jede Runde wird zufällig ein latenter Faktor des MF-Modells gewählt und anhand von 3 aus den 25 möglichen Repräsentanten zufällig ausgewählten Filmen dargestellt (B). Die Darstellung eines Films erfolgt durch Filmplakat, inhaltliche Beschreibung, Trailer und eine Liste von Regisseuren und Schauspielern[4]. Beide Spieler versuchen daraufhin Gemeinsamkeiten der Filme zu beschreiben, indem sie möglichst schnell viele treffende Begriffe eingeben (A), welche jeweils in einer Liste gesammelt werden (C). Dies wird so lange fortgeführt, bis ein Begriff von beiden Spielern eingegeben wurde. Hierdurch endet die aktuelle Runde und es werden entsprechend der Zeit, die benötigt wurde um zu der Übereinstimmung zu gelangen, Punkte gutgeschrieben. Sollten die Filme unbekannt sein oder die Spieler ihre Beschreibung als zu schwer empfinden, kann unter beidseitiger Zustimmung eine Runde übersprungen werden, was jedoch mit Negativpunkten bestraft wird. In jedem Fall startet nach Beendigung einer Runde automatisch die nächste, d. h. es werden erneut 3 Filme – nun für einen anderen Faktor – dargestellt. Insgesamt verfügen die Spieler über 3 Minuten Zeit, in denen sie so viele Runden wie möglich spielen können. Im Anschluss kann ein neues Spiel gestartet werden, was zu erneuter Zuweisung eines Spielpartners und Zurücksetzen von Spielzeit und Punktestand führt.

4 Evaluation

Bei der Evaluation eines GWAP, und speziell eines *Output-Agreement*-Spiels, kommt zwei Aspekten eine zentrale Bedeutung zu: Der generellen *Game Experience* im Umgang mit dem Spiel und der Qualität der von den Spielern produzierten Ausgaben. Im Folgenden stellen wir eine Nutzerstudie vor, bei der die Probanden gebeten wurden, das hier präsentierte Spiel zu spielen und anschließend einen Fragebogen auszufüllen. Letzterer wurde verwendet um die subjektive Wahrnehmung der Probanden zu erfassen. Zudem zeichneten wir sämtliche Interaktionsdaten auf, insbesondere die eingegebenen Begriffe und gefundenen Übereinstimmungen.

[4] Die entsprechenden Daten stammen von https://www.themoviedb.org.

4.1 Methodik

Durchführung: Die Studie wurde in einem einwöchigen Zeitraum durchgeführt und fand an unterschiedlichen Orten statt (Usability-Labor, bei Probanden zu Hause, in Kursen an der Universität). Dabei war immer ein Versuchsleiter anwesend, der kontrollierte, dass keinerlei Kommunikation außerhalb des Spiels stattfand. Mitunter wurde das Spiel auch online gespielt. In diesen Fällen konnte keine Kontrolle erfolgen. Die Gesamtspieldauer war nicht vorgegeben, einzige Bedingung war, mindestens ein Spiel mit so vielen Runden wie möglich zu spielen.

Fragebogen: Wir verwendeten eigens generierte Fragen um die subjektive Wahrnehmung von Aspekten des Spiels wie z. B. Schwierigkeitsgrad und Bekanntheit der Repräsentanten zu erfassen. Zudem wurden demographische Daten und Kenntnis der Filmdomäne abgefragt. Mittels existierender Fragebögen maßen wir *Game Experience* und Spielspaß. Hierzu kamen SUS (Brooke, 1996), UEQ (Laugwitz et al., 2008) und IMI (Ryan, 1982) zum Einsatz. Alle Fragen wurden auf einer positiven 5-stufigen Likert-Skala erhoben. Einzige Ausnahmen bildeten Fragen des UEQ (7-stufige bipolare Skala) und IMI (positive 7-stufige Likert-Skala).

Stichprobe: Insgesamt wurden 84 (42 weiblich) Probanden akquiriert, die den Fragebogen ausfüllten[5]. Das Alter lag zwischen 17 und 36 Jahren ($M = 21.23$, $\sigma = 4.53$). Die meisten Probanden waren Studierende ($82.1\,\%$) oder Angestellte ($15.5\,\%$). Der Großteil besaß Abitur ($84.5\,\%$) oder einen Hochschulabschluss ($13.1\,\%$). Teilweise gaben die Probanden an, ihren Spielpartner sehr gut zu kennen ($21.4\,\%$), was in manchen Situationen (zu Hause oder im Labor) gut zu erklären ist. In vielen anderen Fällen waren sich die Spielpartner jedoch völlig fremd ($41.7\,\%$). Insgesamt gaben die Probanden tendenziell an, Filme zu lieben ($M = 3.31$, $\sigma = 1.03$), und über eine mittlere Filmkenntnis zu verfügen ($M = 2.60$, $\sigma = 0.98$).

4.2 Ergebnisse

Während des Studienzeitraums wurden insgesamt 173 Spiele gespielt. Obwohl die Probanden nur mindestens ein Spiel absolvieren mussten, spielte jeder Spieler im Schnitt 4.12 Spiele. Dies legt nahe, dass die Probanden das Spiel gerne spielten, was mit einem Wert von $M = 4.79$ ($\sigma = 1.26$) auch durch Ergebnisse des IMI auf der 7-stufigen Skala zum Spielspaß belegt wird. Die Probanden gaben zudem an, dass sie nur manchmal Unsicherheit bei der Eingabe von Begriffen empfanden ($M = 2.75$, $\sigma = 1.05$). Als Resultat wurde der Schwierigkeitsgrad als eher einfach bewertet ($M = 3.21$, $\sigma = 1.24$).

Die ausgewählten Filme waren den Probanden zu einem gewissen Grad bekannt ($M = 2.77$, $\sigma = 0.99$). Mitunter wurde angemerkt, dass manche Filme sehr alt gewesen seien und diese besser nicht verwendet werden sollten. Dennoch gaben die Probanden insgesamt an, eher zu verstehen weshalb die jeweiligen Filme gemeinsam als Repräsentanten dargestellt wurden ($M = 2.74$, $\sigma = 0.96$). Unter Berücksichtigung aller gespielten Runden attestierten sie den Filmen zudem eine hohe Diversität ($M = 3.58$, $\sigma = 1.01$). Die präsentierten Informationen wurden als ausreichend wahrgenommen ($M = 3.06$, $\sigma = 0.95$), wobei dem Filmplakat der höchste Informationswert beigemessen wurde ($M = 4.32$, $\sigma = 0.91$), gefolgt vom Filmtitel ($M = 3.74$,

[5] Da nicht auszuschließen ist, dass das Spiel auch (insbesondere online) ohne Ausfüllen des Fragebogens gespielt wurde, liegt die tatsächliche Spielerzahl vermutlich höher.

Faktor	1	2	3	4	5	6	7	8	9	10
Beispiele für Repräsentanten	*The Lion King*, *Bad Boys*, *Home Alone*	*Forrest Gump*, *Fight Club*, *Se7en*	*Jurassic Park*, *Rocky V*, *Star Wars*	*Cast Away*, *Meet the Parents*, *Waterworld*	*The Doors*, *The Beach*, *Casino*	*Indiana Jones*, *Speed*, *Aliens*	*Nude Girls*, *American Pie*, *Harold & Kumar*	*The Net*, *Dave*, *Groundhog Day*	*Batman Forever*, *Deep Impact*, *Twister*	*Pretty Woman*, *Big*, *E.T.*
Begriffe	620	612	589	565	562	580	423	438	754	598
Übereinstimm.	54	57	51	49	55	65	42	50	66	56
Verhältnis	11.48	10.74	11.55	11.53	10.22	8.92	10.07	8.76	11.42	10.68
Übereinstimm. (#)	comedy (10) lustig (8) disney (4) action, liebe (3) kampf, sex (2)	action (13) ernst, kampf, mann (4) spannend (3) comedy, drama, hund, krieg, thriller (2)	action (12) krieg (5) kampf (4) comedy (3) drama (2)	action (13) comedy (8) drama, gruselig, lustig, thriller (2)	action (7) comedy, horror, liebe (5) gruselig (3) alt, erotic, mystery (2)	action (24) comedy (3) abenteuer, alien, alt, kampf, liebe, waffen (2)	comedy (10) sex (7) action, college, drama (2)	liebe (12) comedy (5) familie (3) action, amerika, boring, frau, romance, sex (2)	action (18) horror (6) gruselig, sci-fi (3) alien, aliens, batman, comedy, zukunft (2)	liebe (11) comedy (6) familie (4) action, romance, romantic, tiere (3) dramatisch (2)
Ähnl. Faktoren (Kosinusähnl.)	Faktor 4 (.47) Faktor 7 (.13) Faktor 8 (.11)	Faktor 3 (.28) Faktor 7 (.15) Faktor 6 (.05)	Faktor 2 (.28) Faktor 6 (.09) Faktor 4 (.07)	Faktor 1 (.47) Faktor 2 (.15) Faktor 5 (.13)	Faktor 9 (.50) Faktor 8 (.18) Faktor 6 (.15)	Faktor 5 (.15) Faktor 8 (.11) Faktor 3 (.09)	Faktor 8 (.16) Faktor 1 (.13) Faktor 4 (.06)	Faktor 10 (.54) Faktor 5 (.18) Faktor 7 (.16)	Faktor 5 (.50) Faktor 4 (.11) Faktor 6 (.08)	Faktor 8 (.54) Faktor 5 (.14) Faktor 6 (.08)

Tabelle 1: Faktoren mit Beispielrepräsentanten und statistischer Auswertung der Spieldaten: Anzahl eingegebener Begriffe und Übereinstimmungen pro Faktor, gefundene Übereinstimmungen (für Begriffe die min. zweimal genannt wurden), und die gemäß Kosinusähnlichkeit der TF-IDF-Termvektoren ähnlichsten drei Faktoren.

$\sigma = 1.01$) und der inhaltlichen Beschreibung ($M = 3.12$, $\sigma = 1.25$). Trailer und Listen der Regisseure bzw. Schauspieler wurden hingegen als wenig nützlich eingeschätzt.

Mit einem SUS-Score von 79 wurde die Usability als *gut* bewertet. Auch die unterschiedlichen Skalen des UEQ zeigten (sehr) positive Resultate, von 0.90 (*Stimulation*) über 1.14 (*Originalität*) und 1.27 (*Attraktivität*) bis hin zu 1.98 (*Durchschaubarkeit*)[6].

Insgesamt wurden $5\,741$ Begriffe eingegeben, durchschnittlich 574.10 pro Faktor ($\sigma = 93.29$) und 33.18 pro Spiel ($max = 102$, $\sigma = 20.60$). Hieraus resultierten insgesamt 545 gefundene Übereinstimmungen mit einem Durchschnitt von 54.50 pro Faktor ($\sigma = 7.23$) und 3.15 pro Spiel ($max = 39$, $\sigma = 5.05$). Unter Berücksichtigung der Spielerzahl hatte jeder Spieler folglich eine *Expected Contribution* (von Ahn und Dabbish, 2008) von 68.35 Begriffen und 6.49 Übereinstimmungen. Tabelle 1 zeigt die gesammelten Spieldaten für jeden Faktor.

Für die weitere Auswertung der Übereinstimmungen bereinigten wir den Datensatz und fassten Begriffe sinngemäß zusammen. Des Weiteren wählten wir $X = 2$ als *Good Label Threshold* (von Ahn und Dabbish, 2008), d. h. als minimale Anzahl von Übereinstimmungen damit ein Begriff als sinnvolle Beschreibung für einen Faktor angesehen wird. Unsere Entscheidung für $X > 1$ fiel bewusst, u. a. um Begriffe herauszufiltern, die einzig aufgrund eines besonders herausstechenden Repräsentanten eingegeben wurden. Von den 545 Übereinstimmungen blieben somit schließlich noch 325 Übereinstimmungen, bestehend aus 35 distinkten Begriffen, übrig. Tabelle 1 zeigt, welche Übereinstimmungen für jeden Faktor auftraten.

Um neben der qualitativen Betrachtung die gefundenen Übereinstimmungen auch quantitativ analysieren zu können, bestimmten wir aus den verbliebenen 35 Begriffen ein Wörterbuch. Auf dieser Grundlage definierten wir Inhaltsvektoren für jeden Faktor, deren Einträge wir mittels *TF-IDF* berechneten. Jeder Eintrag gab dabei an, wie oft der jeweilige Begriff für den Faktor zu Übereinstimmungen führte, im Verhältnis zu der Häufigkeit, mit der dies insgesamt der Fall war. Dieses Vorgehen erlaubte es uns die Ähnlichkeit dieser Inhaltsvektoren – und damit

[6] Der Wert für *Steuerbarkeit* war neutral (0.58). Allerdings erscheint diese Dimension nicht konsistent erfasst worden zu sein (Cronbach's $\alpha = 0.37$). Dies wirkt angesichts der berücksichtigten, im Kontext eines Spiels aber wenig aussagekräftigen Konzepte wie z. B. Vorhersagbarkeit und Unterstützung, jedoch nachvollziehbar.

der für die Faktoren generierten Beschreibungen – mithilfe des Kosinusmaßes zu berechnen und folglich die Faktoren quantitativ zu vergleichen. In Tabelle 1 sind jeweils die demgemäß ähnlichsten drei Faktoren aufgeführt[7].

4.3 Diskussion

Wie eingangs erwähnt, konnten bereits semantische Verbindungen der Dimensionen latenter Faktormodelle zu realweltlichen Eigenschaften aufgezeigt werden (Koren und Bell, 2015; Loepp, Donkers et al., 2018). Vor diesem Hintergrund stellten Loepp, Hussein et al. (2014) eine Methode vor, um Nutzern zu erlauben, anhand exemplarisch für die Faktoren ausgewählter Repräsentanten ihre Präferenzen auszudrücken. Wir folgten diesem Ansatz um eine inhaltliche Interpretation der Faktoren zu ermöglichen, indem wir Repräsentanten für jeden Faktor gemäß Gleichung (1) bestimmten. Die berichteten Fragebogenergebnisse deuten auf eine hohe Diversität zwischen den Faktoren hin, was die Annahme unterstützt, dass diese Repräsentanten tatsächlich unterschiedliche Bedeutungen der Faktoren beschreiben. Dies kann durch eine Inspektion der beispielhaft aufgeführten Repräsentanten in Tabelle 1 gut nachvollzogen werden. Konsequenterweise lassen sich die semantischen Unterschiede der Faktoren auch in den mithilfe des Spiels gesammelten Begriffen beobachten: Die gefundenen Übereinstimmungen und die Häufigkeit ihres Auftretens wirken innerhalb der Faktoren konsistent, zwischen den Faktoren scheinen sie jedoch unterschiedliche Konzepte zu beschreiben.

Auch die weitere quantitative Untersuchung belegt, dass die eingegebenen Begriffe überwiegend unterschiedliche semantische Beziehungen der Faktoren beschreiben, d. h. die mittels *TF-IDF* berechneten Inhaltsvektoren wirken relativ distinkt. Nur vereinzelt sind sich Faktoren ähnlicher, z. B. Faktor 8 und 10 (siehe Tabelle 1). Dies könnte einerseits darin begründet sein, dass nicht genug Daten vorliegen um eine höhere Trennschärfe zwischen den Begriffsmengen zu erreichen, andererseits könnte sich die Semantik mancher latenter Faktoren tatsächlich ähneln. Diese Ähnlichkeiten innerhalb des Faktormodells, und ihre Verstärkung bei steigender Faktoranzahl, sind vermutlich auch Grund dafür, dass bei einer Erhöhung der Faktoranzahl die Qualität des resultierenden Modells nur noch marginal ansteigt (Koren und Bell, 2015). Wir nehmen folglich an, dass eine höhere Faktoranzahl zu einem Anstieg der semantischen Überschneidungen der mit dem Spiel gesammelten Begriffe führen würde.

Bei genauerer Betrachtung der eingegebenen Begriffe die zu Übereinstimmungen führten zeigt sich eine Tendenz zu eher oberflächlichen Begriffen. So wurden beispielsweise Genres wie „action“ und „comedy“ vergleichsweise häufig als Beschreibung genutzt. Hierbei handelt es sich um einen generellen Effekt bei der Verwendung von GWAP (von Ahn und Dabbish, 2008), dem zwar mit verschiedenen Mechanismen entgegengewirkt werden kann (etwa durch Tabulisten mit unerlaubten Wörtern), wofür allerdings eine Datenbasis erforderlich ist, wie sie zu Beginn unserer Studie naturgemäß nicht gegeben war. Im Gegensatz dazu lassen sich trotz des gewählten *Good Label Threshold* mitunter sehr spezifische Begriffe in den Beschreibungen finden (etwa „hund“ für Faktor 2). Dies ist jedoch wahrscheinlich eher den Plakaten bestimmter, für den jeweiligen Faktor dargestellter Repräsentanten zuzuschreiben, als dass es sich hierbei um

[7] Eine Ähnlichkeit von 0 bedeutet, dass zwei Inhaltsvektoren keinerlei Gemeinsamkeiten aufweisen, während ein Wert von 1 bedeutet, dass zwei identische TF-IDF-Termvektoren vorliegen.

von den Spielern als Beschreibung für den Faktor gedachte Begriffe handelt. Insgesamt scheint dieser Effekt die erhobenen Daten allerdings nur in geringem Maße zu betreffen, und es ist zu erwarten, dass er mit größerer Datenmenge und damit einhergehend vielseitigerer Kombination von Repräsentanten bei gleichzeitiger Anpassung von X ganz verschwinden würde.

Des Weiteren lässt sich beobachten, dass es für manche Faktoren offenbar einfacher war passende Begriffe zu finden als für andere. Dies kann etwa durch einen Vergleich von Faktor 3 und Faktor 6 in Tabelle 1 nachvollzogen werden: Während für beide Faktoren ungefähr die gleiche Anzahl an Begriffen eingegeben wurde, kamen im Verhältnis mehr Übereinstimmungen für Faktor 6 zustande. Dass Faktor 6 scheinbar eher einfach zu beschreiben war, wird auch von den Häufigkeiten, wie oft einzelne Begriffe für diesen Faktor zu Übereinstimmungen führten, unterstrichen: Beispielsweise kam es mit dem Begriff „action“ um ein Vielfaches häufiger zu einer Übereinstimmung als mit jedem anderen Begriff. Dies legt nahe, dass dieser Begriff den latenten Faktor besonders treffend beschreibt.

Zusammengefasst erlauben die qualitativen und quantitativen Ergebnisse festzuhalten, dass unser Ansatz eine Möglichkeit darstellt, die unterschiedlichen realweltlichen Bedeutungen der Faktoren eines MF-Modells mit verhältnismäßig geringem Aufwand sinnvoll zu benennen.

5 Fazit und Ausblick

Das in diesem Beitrag vorgestellte Online-Spiel soll Nutzer motivieren, die mittels der in Empfehlungssystemen häufig eingesetzten Matrix Factorization gelernten latenten Faktoren zu beschreiben. Die durchgeführte Evaluation zeigt, dass das Spiel nicht nur Spaß bereitet, sondern auch sinnvolle Ausgaben erzeugt, die geeignet sind, die bislang verborgene Semantik der Dimensionen latenter Faktormodelle zu verstehen. Ferner belegt unsere Auswertung die bisherige Annahme, dass die Faktoren Muster in den zugrundeliegenden Bewertungsdaten beschreiben, und dabei tatsächlich in Verbindung zu unterschiedlichen realweltlichen Eigenschaften der Produkte stehen. Unser Ansatz stellt dementsprechend eine wertvolle Grundlage dar, um Daten zu sammeln damit etwa Empfehlungen textuell erklärt oder Produkte anhand ihrer Beziehung zu den Faktoren mit inhaltlichen Beschreibungen versehen werden können.

Ungeachtet der vielversprechenden Ergebnisse betrachten wir die Erweiterung der Spielmechanik als Teil künftiger Arbeit. Um zusätzliche Anreize zu schaffen, sollen Spielelemente wie z. B. vom Spielgeschehen abhängige Achievements eingeführt werden. Darüber hinaus ermöglicht die nun absolvierte erste Datenerhebung und die damit entstandene Grundmenge von Begriffen den Einsatz zusätzlicher GWAP-Elemente (von Ahn und Dabbish, 2008), etwa von Tabulisten, mit denen Spieler dazu bewegt werden könnten, ein breiteres aber auch spezifischeres Begriffsspektrum zu nutzen. Zudem könnte für Fälle, in denen sich online kein Partner finden lässt, eine Einzelspielerversion implementiert werden. Schließlich möchten wir das Spiel einer größeren Nutzerschaft zugänglich machen, um eine größere Menge an Beschreibungen – auch für anders konfigurierte Faktormodelle – zu erlangen. Die Qualität der Ausgabe planen wir gemeinsam mit den erwähnten Anwendungsmöglichkeiten der Begriffe in Empfehlungssystemen, z. B. zur Erklärung von Empfehlungen, in künftigen Nutzerstudien eingehender zu evaluieren.

Literaturverzeichnis

Banks, S., Rafter, R. & Smyth, B. (2015). The recommendation game: Using a game-with-a-purpose to generate recommendation data. In *RecSys '15* (S. 305–308). ACM.

Brooke, J. (1996). SUS – A quick and dirty usability scale. In *Usability Evaluation in Industry* (S. 189–194). Taylor & Francis.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., … Foldit players. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*, 756–760.

Eickhoff, C., Harris, C. G., de Vries, A. P. & Srinivasan, P. (2012). Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *SIGIR '12* (S. 871–880). ACM.

Hacker, S. & von Ahn, L. (2009). Matchin: Eliciting user preferences with an online game. In *CHI '09* (S. 1207–1216). ACM.

Herlocker, J. L., Konstan, J. A. & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *CSCW '00* (S. 241–250). ACM.

Ho, C.-J., Chang, T.-H., Lee, J.-C., Hsu, J. Y.-j. & Chen, K.-T. (2009). KissKissBan: A competitive human computation game for image annotation. In *HCOMP '09* (S. 11–14). ACM.

Jugovac, M. & Jannach, D. (2017). Interacting with recommenders – Overview and research directions. *ACM TiiS*, *7*(3), 10:1–10:46.

Koren, Y. & Bell, R. M. (2015). Recommender Systems Handbook. (Kap. Advances in collaborative filtering, S. 77–118). Springer US.

Kunkel, J., Loepp, B. & Ziegler, J. (2017). A 3D Item space visualization for presenting and manipulating user preferences in collaborative filtering. In *IUI '17* (S. 3–15). ACM.

Laugwitz, B., Held, T. & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work* (S. 63–76). Springer.

Loepp, B., Donkers, T., Kleemann, T. & Ziegler, J. (2018). Interactive Recommending with Tag-Enhanced Matrix Factorization (TagMF). *Int. J. Hum. Comput. Stud.*

Loepp, B., Hussein, T. & Ziegler, J. (2014). Choice-based preference elicitation for collaborative filtering recommender systems. In *CHI '14* (S. 3085–3094). ACM.

Németh, B., Takács, G., Pilászy, I. & Tikk, D. (2013). Visualization of movie features in collaborative filtering. In *SoMeT '13* (S. 229–233).

Pu, P., Chen, L. & Hu, R. (2012). Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Model. User-Adap. 22*(4-5), 317–355.

Rossetti, M., Stella, F. & Zanker, M. (2013). Towards explaining latent factors with topic models in collaborative recommender systems. In *DEXA '13* (S. 162–167).

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *J. Pers. Soc. Psychol. 43*(3), 450–461.

Seemakurty, N., Chu, J., von Ahn, L. & Tomasic, A. (2010). Word sense disambiguation via human computation. In *HCOMP '10* (S. 60–63). ACM.

Tintarev, N. & Masthoff, J. (2015). Recommender Systems Handbook. (Kap. Explaining recommendations: Design and evaluation, S. 353–382). Springer US.

von Ahn, L. & Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM*, *51*(8), 58–67.

Walsh, G. & Golbeck, J. (2010). Curator: A game with a purpose for collection recommendation. In *CHI '10* (S. 2079–2082). ACM.